Continuous Iterative Guided Spectral Class Rejection Classification Algorithm

Rhonda D. Phillips, *Member, IEEE*, Layne T. Watson, *Fellow, IEEE*,
Randolph H. Wynne, *Member, IEEE*, and Naren Ramakrishnan, *Member, IEEE*

***Abstract*—This paper presents a new semiautomated soft classification method that is a hybrid between supervised and unsupervised classification algorithms for the classification of remote sensing data. Continuous iterative guided spectral class rejection (IGSCR) (CIGSCR) is based on the IGSCR classification method, a crisp classification method that automatically locates spectral classes within information class training data using clustering. This paper outlines the model and algorithm changes necessary to convert IGSCR to use soft clustering to produce soft classification in CIGSCR. This new algorithm addresses specific challenges presented by remote sensing data including large data sets (millions of samples), relatively small training data sets, and difficulty in identifying spectral classes. CIGSCR has many advantages over IGSCR, such as the ability to produce soft classification, less sensitivity to certain input parameters, potential to correctly classify regions that are not amply represented in training data, and a better ability to locate clusters associated with all classes. Furthermore, evidence is presented that the semisupervised clustering in CIGSCR produces more accurate classifications than classification based on clustering without supervision.**

***Index Terms*—Fuzzy clustering, land cover classification, partially supervised learning, remote sensing, soft clustering, statistical learning.**

List of Symbols

a_{ij}	$E(W_{ij})$.
b_{ij}^2	$\mathrm{Var}(W_{ij})$.
c_i	Class i.
k	Cluster.
m	Number of samples in cluster.
n	Number of points.
n_c	Number of points in class/distribution c.
p	Homogeneity threshold.
$p()$	Probability.
p_0	User-supplied threshold.
w_{ij}	Weight.
$\hat{w}$	Weight in next iteration.
$\overline{w}_{c,j}$	Sample mean of weights in class c and cluster j.
$\overline{w}_j$	Sample mean of weights in cluster j.
x	Multivariate sample.
z	Test statistic.
B	Number of bands.
C	Number of classes.
E	Expected value.
F	Cumulative distribution function.
I	Index set of class.
J	Index set of cluster.
$J(\rho)$	Objective function.
K	Number of clusters.
Q	Number of distributions.
S_w	Sample standard deviation.
T	Transpose operation.
U	Cluster mean.
V	Binomial random variable.
$V_{c,j}$	Count of samples in class c and cluster j.
V_{ij}	Bernoulli random variable.
X	Data point random variable.
Y	Sum of random variables V and W.
Z	Normal random variable.
ρ	Distance.
α	Type-I error.
α_{qj}	Expected value of weights drawn from qth distribution.
Σ	Covariance.
δ_{ij}	Kronecker delta.
$\Phi(i)$	Label.
$\Psi(i)$	Distribution.
β_{qj}^2	Variance of weights from qth distribution.

Manuscript received January 18, 2011; revised July 8, 2011; accepted September 11, 2011.

R. D. Phillips is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02420 USA (e-mail: rhonda.phillips@ll.mit.edu).

L. T. Watson and N. Ramakrishnan are with the Department of Computer Science and the Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA (e-mail: ltw@cs.vt.edu; naren@cs.vt.edu).

R. H. Wynne is with the Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA (e-mail: wynne@vt.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2011.2173802

I. Introduction

In REMOTE sensing, the identification of spectral classes required for supervised classification is a tedious and laborious process. While information/land cover classes (such as forest or row crop) have physical meaning, spectral classes are defined mathematically and often have statistical requirements for good classification. Finding a comprehensive set of spectral classes that fully represent the image's spectrum with good statistical properties (for supervised classification) is nontrivial. A common strategy for addressing this issue is to use an unsupervised technique such as clustering to learn the inherent spectral classes within the image of interest. Clustering has the advantage of producing mathematically defined spectral classes that are guaranteed to sufficiently cover the image.

There are multiple examples of using unsupervised learning of spectral classes within a partially supervised framework

within the remote sensing community. Perhaps, one of the first such algorithms is described in [1]. Clustering is used to define spectral classes, and clusters are manually analyzed using a variety of statistical methods to determine the spectral classes to be used for classification. Additionally, unsupervised spectral class formation can locate missing spectral classes [2], and unsupervised approaches can even be used to locate all classes when training data are only present for one class of interest [3]–[5]. These techniques proved to be suitable for the identification of one class of interest using supervised classification, with the advantage of allowing an analyst to focus training resources on only the class of interest [5]. In some cases, spectral class training data are unavailable for some images but exist for similar images/areas of interest such as multiple images taken from the same scene at different times. Several partially supervised spectral class identification approaches are proposed to leverage an existing training data set [6]–[9]. In situations where information class/land cover class training data exist, unsupervised techniques such as clustering may be used to locate spectral classes that correspond to information classes, allowing statistical classification methods to produce a final land cover classification where the land cover classes do not have desired statistical properties. Fernandez-Prieto proposed an iterative method that could learn land cover maps using information class training samples [10], and Bauer *et al.* introduced an algorithm called "guided clustering" used to learn spectral classes from information class training data directly [11]. The training data are clustered to reveal spectral classes that correspond to an information class in the original data set. A fully automated spectral class detection algorithm is the iterative guided spectral class rejection (IGSCR) classification method [12]–[14]. IGSCR applies clustering to the entire input image and uses information class training data to determine the most likely information class assignment for each spectral class. Furthermore, training data are used to statistically test spectral classes for information class purity, providing automatic spectral class rejection required to form a robust set of spectral classes. Due to its high accuracy and automation, IGSCR is a frequently used hybrid classification method in the remote sensing community [15]–[18].

One limitation present in the existing partially supervised classification methods (including IGSCR) is the dependence on hard clustering algorithms to produce hard (crisp) classification results. Crisp classifications assign each pixel or sample to one class in the particular classification scheme, which can be interpreted as picking the class that has the highest probability of containing the sample. Alternatively, soft classifications contain information on possible memberships in multiple classes, not just the most likely class. Soft or subpixel classifications are of considerable interest to the remote sensing community as this type of classification can effectively model geographic data whose natural boundaries rarely coincide with pixel boundaries. Pixels can also contain multiple species that are commingled, leading to classification difficulty (the resolution of the image is not sufficiently high to ensure that each pixel contains only one class). Furthermore, individual classes within the classification scheme can have overlapping electromagnetic reflectance spectra, making it difficult to discriminate between these classes and locate spectral classes. Scientists have successfully used soft classification for applications such as land cover mapping [19], vegetation mapping [20], and the classification of snow [21], to name a few. Popular methods for obtaining soft classifications of remote sensing images include fuzzy c-means [22] and spectral unmixing [23].

The purpose of this paper is to develop a partially supervised classification algorithm that uses soft clustering to locate difficult-to-detect spectral classes. Note that, when good spectral class training data exist to perform a supervised classification, those data should be used in a supervised classification. This paper specifically addresses the issue of learning spectral classes when only information class training data are available. The approach taken in this work is to adapt the clustering rejection and refinement framework in IGSCR to use soft clustering and produce soft classifications. This framework will potentially affect other classification algorithms that have labeled data and involve clustering. Soft clustering retains all information regarding the proximity of data points to clusters and will therefore directly produce a soft classification and will potentially provide better training spectral classes for a supervised decision rule (DR). The major challenges in converting the discrete IGSCR to a fully continuous algorithm producing soft classification are in converting the underlying inherently discrete models and algorithms to suitable continuous models and algorithms while preserving the automated spectral class identification properties of IGSCR. More specifically, a hypothesis test that is fundamental to IGSCR is based on the discrete binomial probability distribution. A hypothesis test based on a new continuous probability distribution is necessary in continuous IGSCR (CIGSCR). IGSCR uses an iterative cluster refinement framework that breaks down under soft clustering, and therefore, a new iterative cluster refinement method is developed for CIGSCR.

The remainder of this paper is organized as follows. Section II describes IGSCR in detail, and Section III introduces CIGSCR. Section IV rigorously derives the association significance test, a hypothesis test based on a new distribution that will be suitable for evaluating the class associations to soft clusters. Section V discusses changes necessary for the iterative refinement of soft clusters and precisely states the complete CIGSCR algorithm. Section VI concerns distance functions, with experimental results following in Section VII. Section VIII concludes this paper.

II. IGSCR

IGSCR is a classification method that uses clustering to generate a classification model $p(c_i|x)$ where x is a multivariate sample to be classified and $c_i,\ i = 1, \ldots, C$, is the ith class where there are C classes in the classification scheme. IGSCR uses clustering to estimate $p(k_j|x)$ in the expression

$$p(c_i|x) = \sum_{j=1}^{K} p(c_i, k_j|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x) \quad (1)$$

where $k_j,\ j = 1, \ldots, K$, is the jth cluster out of K total clusters. IGSCR also uses the clusters to train a DR using

Bayes' theorem [24]

$$p(k_j|x) = \frac{p(x|k_j)p(k_j)}{\sum_{i=1}^{K} p(x|k_i)p(k_i)}. \quad (2)$$

The prior probabilities of the clusters $p(k_j)$ are assumed to be equal since no *a priori* knowledge of the clusters is available.

Clustering is performed using a discrete clustering method such as k-means that minimizes the objective function

$$J(\rho) = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}\rho_{ij} \quad (3)$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1$$

where $w_{ij} \in \{0,1\}$ is the value in the ith row and jth column of the partition matrix $W \in \Re^{n\times K}$, $U^{(j)} \in \Re^B$ is the prototype for the jth cluster k_j, $x^{(i)} \in \Re^B$ is the ith data point, and $\rho_{ij} = \|x^{(i)} - U^{(j)}\|_2^2$. The clusters $k_1, \ldots, k_K$ form a partition of $\{x^{(i)}\}_{i=1}^n$. The algorithm for k-means requires K initial cluster prototypes and iteratively assigns each sample to the closest cluster using

$$w_{ij} = \begin{cases} 1, & \text{if } j = \arg\min\limits_{1\le j\le K} \rho_{ij} \\ 0, & \text{otherwise} \end{cases}$$

followed by the cluster prototype (mean) recalculation

$$U^{(j)} = \sum_{i=1}^{n} \left(w_{ij}x^{(i)}\right) \Big/ \sum_{i=1}^{n} w_{ij}$$

once W has been calculated [25]. This process, guaranteed to terminate in a finite number of iterations, continues until no further improvement is possible, terminating at a local minimum point of (3).

IGSCR uses labeled data in a semisupervised clustering framework to locate clusters that correspond to classes in a given classification scheme. IGSCR requires a labeled set of training data composed of individual samples within the image to be classified and corresponding class labels. Rather than using the labeled data to train a DR directly, the entire image is clustered, thereby capturing the inherent structure of all the data and not just the labeled samples. The clusters represent spectral classes, and each spectral class ideally corresponds to exactly one class in the final classification scheme. Once clusters are generated, each cluster must be mapped to one class or rejected as impure. While, theoretically, each cluster should contain samples belonging to only one information class, in practice, clusters (spectral classes) that contain predominantly samples of one class can contain a few samples from other classes because of inherent errors. However, if a cluster contains too many samples from different classes, the cluster itself is considered confused and should not be labeled with one class. Impure clusters are rejected and can be further refined in the iterative part of the algorithm.

Once clusters are generated, the test for cluster purity is performed using the labeled training set. Let $V_{c,j}$ be the binomial random variable denoting the number of labeled samples assigned to the jth cluster that are labeled with a particular cth class. Let p be the user-supplied cluster homogeneity threshold ($p = 0.9$ would indicate that a cluster is 90% pure with respect to the majority class), and let α be the user-supplied acceptable one-sided type-I error for a statistical hypothesis test. Then, if c is the majority class represented in the jth cluster, the jth cluster is rejected if $P(Z < \hat{z}) < 1 - \alpha$ where Z is a standard normal random variable, m is the number of labeled samples in the jth cluster, and

$$\hat{z} = \frac{v_{c,j} - mp}{\sqrt{mp(1-p)}}. \quad (4)$$

Typically, a continuity correction of 0.5 is added in the numerator of (4) to closer approximate a binomial distribution when m is small.

If a cluster is rejected, the samples making up that cluster can be reclustered in subsequent iterations. All samples belonging to pure clusters are removed from the image being clustered, resulting in only samples belonging to impure clusters being reclustered. Once more clusters are generated, those clusters are evaluated for purity and removed from the image, and clustering is performed again until termination criteria are met. The termination criteria include all samples belonging to pure clusters, leaving no remaining samples to be clustered. Also, no pure clusters could be found in the previous iteration, meaning that the clustering would continue to be performed on the same data resulting in the same impure clusters (assuming deterministic cluster seeding). Finally, a set number of iterations can be reached resulting in termination of the iteration. Note that deterministic seeding ensures that the iteration will terminate, even without specifying a maximum number of iterations.

Once the iterative clustering is complete, one or more classifications are performed. The first classification is called the iterative stacked (IS) classification because it is the result of combining or "stacking" all cluster assignments over all iterations (each sample will be assigned to at most one accepted cluster). Assume that all samples not assigned to an accepted cluster are combined to form one cluster k_{K+1}, and the class assignment for that cluster is "unclassified" or c_{C+1}. Then, the IS assignment for a pixel using (1) is

$$\text{IS}(x) = \arg\max_{1\le i\le C+1} p(c_i|x) = \arg\max_{1\le i\le C+1} \sum_{j=1}^{K+1} p(c_i|k_j, x)p(k_j|x)$$

where

$$p(c_i|k_j, x) = \begin{cases} 1, & \text{if } k_j \text{ is labeled } c_i \\ 0, & \text{otherwise} \end{cases}$$

$$p(k_j|x) = \begin{cases} 1, & \text{if } x \in k_j \\ 0, & \text{otherwise} \end{cases}$$

since cluster assignments are discrete.

The second possible classification, the DR classification, uses the pure clusters to form a DR. Recall in (2) that

$$p(k_j|x) = \frac{p(x|k_j)}{\sum_{i=1}^{K} p(x|k_i)}$$

when all the $p(k_j)$ values are equal. Traditionally, the maximum likelihood DR, assuming a multivariate normal distribution

$$p(x|k_j) = 2\pi^{-B/2}|\Sigma_j|^{-1/2} e^{-\frac{1}{2}(x-U^{(j)})^T \Sigma_j^{-1}(x-U^{(j)})}$$

is used where Σ_j is the covariance matrix of the jth cluster [26]. Since IGSCR produces hard classifications, the full probability need not be calculated as determining only the cluster associated with the maximum probability is necessary. The DR classification function is

$$\mathrm{DR}(x) = \arg\max_{1\le i\le C} p(c_i|x) = \arg\max_{1\le i\le C} \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x) \tag{5}$$

where

$$p(k_j|x) = \begin{cases} 1, & \text{if } j = \arg\max\limits_{1\le j\le K} \\ & \quad \times \left(-\ln|\Sigma_j| - \left(x-U^{(j)}\right)^T \Sigma_j^{-1}\left(x-U^{(j)}\right)\right) \\ 0, & \text{otherwise.} \end{cases}$$

A final classification, the IS plus (IS+) classification, combines the DR and IS classifications. If a sample is labeled as unclassified in the IS classification, the DR class value is used for the IS+ classification; otherwise, the IS class value is used for that particular sample. The IS+ classification function is

$$\mathrm{IS}+(x) = \begin{cases} \mathrm{IS}(x), & \text{if } x \notin k_{K+1} \\ \mathrm{DR}(x), & \text{otherwise.} \end{cases}$$

III. CIGSCR

CIGSCR uses a similar semisupervised clustering framework to the one established in IGSCR to produce a soft or probabilistic classification instead of a hard classification and uses continuous algorithms and models instead of discrete algorithms and models. Recall in (1) that $p(c_i|k_j, x)$ and $p(k_j|x)$ are either zero or one (discrete) in practice in IGSCR. $p(c_i|k_j, x)$ is necessarily discrete because, while several clusters can comprise one class, only one class (theoretically) can label the members of a particular cluster, but there are no similar restrictions on $p(k_j|x)$. In fact, the clustering algorithm and the maximum likelihood DR indicate positive probabilities that a sample is associated with each cluster, but IGSCR makes an assignment only to the cluster with the highest probability.

Consider a soft clustering algorithm that minimizes the objective function [27]

$$J(\rho) = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}^p \rho_{ij} \quad \text{subject to: } \sum_{j=1}^{K} w_{ij} = 1 \quad \text{for each } i \tag{6}$$

where $w_{ij} \in (0, 1)$ is the value in the ith row and jth column of the weight matrix $W \in \Re^{n\times K}$ [analogous to the partition matrix W in (3)], $U^{(j)} \in \Re^B$ is the jth cluster prototype, $p > 1$, and $\rho_{ij} = \rho(x^{(i)}, U^{(j)}) = \|x^{(i)} - U^{(j)}\|_2^2$ is the Euclidean distance squared. The algorithm that minimizes this objective function is similar to that of k-means in that it first calculates

$$w_{ij} = \frac{(1/\rho_{ij})^{1/(p-1)}}{\sum\limits_{k=1}^{K} (1/\rho_{ik})^{1/(p-1)}}$$

for all i and j followed by calculating updated cluster prototypes

$$U^{(j)} = \sum_{i=1}^{n} w_{ij}^p x^{(i)} \Big/ \sum_{i=1}^{n} w_{ij}^p.$$

This iteration (recalculation of the weights followed by recalculation of cluster prototypes, following by recalculation of the weights, etc.) is guaranteed to converge (with these definitions of ρ_{ij}, $U^{(j)}$, and w_{ij}) for $p > 1$ [28].

With a continuous alternative to the discrete hypothesis test and a continuous alternative to the IGSCR iterative cluster refinement that follows in Sections V and VI, the classification function for IS classification is

$$IS(x) = p(c_i|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x) \tag{7}$$

where $p(k_j|x)$ is estimated using w_{ij} and $p(c_i|k_j, x)$ does not change from IGSCR. The classification function for the DR classification is

$$\begin{aligned} DR(x) &= p(c_i|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x) \\ &= \frac{\sum\limits_{j=1}^{K} p(c_i|k_j, x)\left[\frac{2e^{-\frac{1}{2}\left(x-U^{(j)}\right)^T \Sigma_j^{-1}\left(x-U^{(j)}\right)}}{\pi^{B/2}|\Sigma_j|^{1/2}}\right]}{\sum\limits_{l=1}^{K}\left[\frac{2e^{-\frac{1}{2}\left(x-U^{(l)}\right)^T \Sigma_l^{-1}\left(x-U^{(l)}\right)}}{\pi^{B/2}|\Sigma_l|^{1/2}}\right]}. \end{aligned} \tag{8}$$

An analog for the IS+ classification is unnecessary in CIGSCR as all samples will be part of pure clusters and will be classified.

IV. Association Significance Test

A key component in the IGSCR semisupervised clustering framework is the homogeneity test used to determine if a cluster contains a statistically significant proportion of one class. This test provides a basis for rejecting a cluster for further refinement, the second phase of the semisupervised clustering.

A cluster might be composed of more than one class because the cluster should be split into multiple clusters. A cluster might also contain more than one class because the initial clusters were determined in such a way as to prevent a cluster from moving toward a particular class. It would be useful to determine which clusters are not spectrally pure (contain more than one class with high probability) so that the cluster can be further refined, and if no refinement is possible (any number of iteration ending criteria are met), the cluster should not be used

in the classification model. Statistical hypothesis tests provide a mechanism for determining class purity once an appropriate statistical model is selected for the data.

In hard IGSCR with hard clustering, the notion of a pure cluster is clear. Each sample will belong to one and only one cluster. A cluster can be 100% homogeneous when all labeled samples contained within that cluster belong to only one class. Although this is possible, it is unlikely that one cluster contains only one class because of inherent error in the labeling process, two different information class categories can contain spectrally similar samples, and one pixel can be spectrally mixed. Once a homogeneity level is determined, a rigorous hypothesis test can be applied to select clusters that contain a certain percentage of one class, with that percentage unlikely to be observed in a particular cluster randomly.

Using soft clusters introduces complications to assessing and determining cluster purity. The first question might be whether a soft cluster can be spectrally pure, because being soft might indicate that clusters are naturally composed of multiple classes. However, just as the goal in IGSCR is to determine clusters that are representative of just one predominant class, that goal holds in CIGSCR with soft clusters. Soft clusters are composed of different portions of each sample or pixel within an image, meaning that each sample has a positive probability of being in different individual classes or clusters. When samples labeled with different classes have a positive probability of belonging to the same cluster, that does not indicate that the cluster really contains two different classes, but rather perhaps that, while the pixels have strong associations with different classes, there is also a positive (although possibly small) probability that each pixel actually belongs to or partially belongs to the majority class within the cluster. Both cases (the cluster is confused or the cluster is not confused but the pixels labeled with different classes still have small associations with the same class) are possible in soft clustering. The appropriate test for soft clusters is not which pixels "belong" to a particular cluster (they all "belong" to some degree) rather how strongly pixels from different classes belong to a particular cluster. If pixels from only one class have strong associations with a cluster when compared to pixels labeled with other classes, then the cluster should be labeled with that most strongly associated class. In this manner, each pixel/sample is associated by varying degrees with multiple spectrally pure clusters that are mapped to individual classes, ultimately producing a soft classification output when each sample is then mapped to different individual classes with varying probabilities.

A. Distribution

Developing a hypothesis test to assess the purity of clusters requires a random variable and knowledge of the distribution of that random variable. In IGSCR, a cluster can be considered pure and labeled with a class if the number of labeled samples belonging to the class is high compared to the number of labeled samples not belonging to the class. The random variable of interest, $V_{c,j} = \sum_{i \in I_j} V_{ic}$, is the count of the number of labeled samples belonging to the cth class for a particular jth cluster where i is the pixel index, I_j is the index set of labeled pixels in the jth cluster, and V_{ic} is the Bernoulli random variable corresponding to the ith pixel being associated with the cth class. A hypothesis test can be developed using the binomial distribution or the less computationally intensive normal distribution, which approximates the binomial distribution well when the number of labeled samples is large.

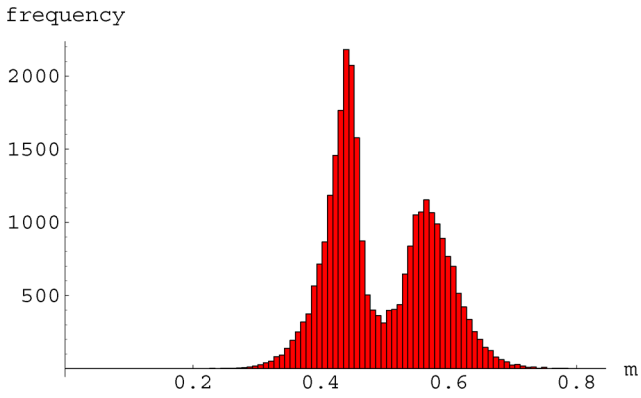


Fig. 1. Histogram of cluster weights m for all data in one cluster $K = 2$.

In CIGSCR, the random variable and distribution are more complicated as there are class memberships (either zero or one) and cluster memberships (between zero and one). Building a test on only the class memberships is not useful as each labeled sample will have some positive probability of belonging to a particular cluster, making the results of the test the same for each cluster unless memberships are also considered. In this case, the association of a sample to a particular class (the majority class, for example) is still a Bernoulli trial. Each pixel also has a weight vector $w_{i\cdot}$ indicating the probability of membership to each cluster. The random variable of interest is the sum of the memberships for the cth class and weights to the jth cluster

$$Y_{c,j} = V_{1c}W_{1j} + V_{2c}W_{2j} + \cdots + V_{nc}W_{nj}$$

where n is the total number of labeled samples. The labels of the classified pixels are independent of cluster assignment, making an assumption that V_{ic} and W_{ij} are independent reasonable. Furthermore, the training samples are labeled prior to clustering, making the random variable of interest

$$Y_{c,j}|(V_{1c}, V_{2c}, \ldots, V_{nc}) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c}$$

where $\phi(i)$ is the label of the ith pixel and

$$\delta_{\phi(i),c} = \begin{cases} 0 & \text{if } \phi(i) \neq c \\ 1 & \text{if } \phi(i) = c \end{cases}$$

is the Kronecker delta. The probability density function (pdf) of $Y_{c,j}|(V_{ic}, i = 1, \ldots, n) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c}$ is the pdf of a sum of individual cluster weights.

Fig. 1 shows the experimental frequency histograms of weights w_{ij} for two clusters $(K = 2)$ of the satellite image used to generate experimental results in this paper. The distribution of the cluster weights appears to be multimodal, which is consistent with the data having multiple inherent classes, indicating

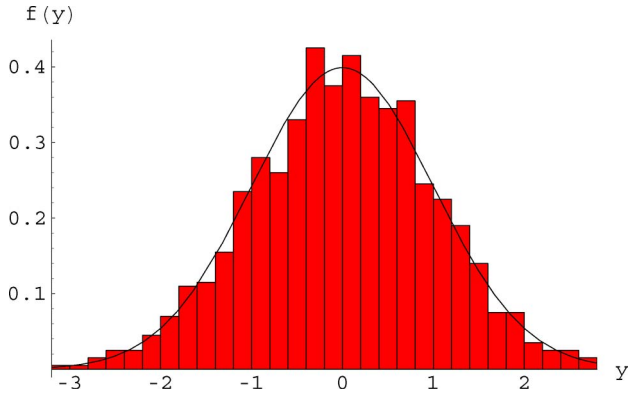


Fig. 2. Pdf of Y (with sample mean subtracted and divided by the standard deviation) compared to a standard normal distribution.

that W_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, K$, would not be identically distributed. A closed-form distribution is not readily available for W_{ij}, but a closed-form distribution, or at least a reasonable approximate closed-form distribution, for $W_{+j} = \sum_{i=1}^{n} W_{ij}$ exists, as will be shown in the next section.

B. Normal Approximation to $Y_{c,j}$

Suppose that an image x contains n pixels $x^{(i)} \in \Re^B$, $i = 1, \ldots, n$. For K fixed cluster centers, $U^{(k)} \in \Re^B$, $k = 1, \ldots, K$, the assigned weight of the ith pixel to the jth cluster is

$$w_{ij} = \frac{1 \Big/ \left\| x^{(i)} - U^{(j)} \right\|_2^2}{1 \Big/ \sum_{k=1}^{K} \left\| x^{(i)} - U^{(k)} \right\|_2^2}$$

which is the inverse of the distance squared over the sum of the inverse squared distances (such inverse distance weights are widely used, e.g., by Shepard's algorithm for sparse data interpolation) [29]. Note that this is the specific case in the soft clustering algorithm described earlier when $p = 2$. In this case, where a remote sensing image is to be clustered, it is reasonable to assume that $x^{(i)}$, $i = 1, \ldots, n$, values are generated from a finite number of multivariate normal distributions. The act of clustering assumes that the data are generated from a finite number of distributions, and remote sensing Earth data are assumed to be generated from normal distributions. The proof in Appendix A demonstrates that, under these assumptions (pixels are generated from a finite number of normal distributions), the Lindeberg condition is satisfied, and therefore, the central limit theorem applies to the sum of a sequence of cluster weight random variables $\sum_{i=1}^{n} W_{ij}$.

Experimental results match this theoretical result, as illustrated by one experiment in Fig. 2.

C. Association Significance Test

The hypothesis test used in IGSCR to assess the significance of a cluster association to a class is based on the normal approximation to the binomial distribution (4). The null hypothesis is that the true probability of a pixel belonging to the majority class (for the cluster of interest) is less than p_0, a user-supplied value. If $P(Z > \hat{z}) < \alpha$, where α is the user-provided type-I error, then the null hypothesis is rejected. The null hypothesis corresponds to the case when the cluster is impure, and rejecting the null hypothesis equates with labeling the cluster pure; if the null hypothesis is *not* rejected, the cluster is impure and the cluster is "rejected."

The hypothesis test for pure clusters in CIGSCR is different as the Bernoulli trials are fixed and testing the probability p of a success is no longer relevant. A pure soft cluster should have large weights for the majority class and comparatively small weights for other classes. One possible hypothesis test compares the average weight for one particular cth class with the overall average weight for all classes in the jth cluster. Starting with the normal approximation for the sum of the cluster weights, the standard normal test statistic would be

$$\hat{z} = \frac{\sum_{i \in J_c} (w_{ij} - \mathrm{E}[W_{ij}])}{\sqrt{\sum_{i \in J_c} \mathrm{Var}[W_{ij}]}}$$

where J_c is the index set of pixels prelabeled with the cth class. $\mathrm{E}[W_{ij}]$ and $\mathrm{Var}[W_{ij}]$ are unknown but can be reasonably approximated using the sample mean

$$\overline{w}_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$$

and sample unbiased standard deviation

$$S_{\overline{w}_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (w_{ij} - \overline{w}_j)^2}.$$

The Wald statistic is then

$$\hat{z} = \frac{\sqrt{n_c}(\overline{w}_{c,j} - \overline{w}_j)}{S_{\overline{w}_j}} \tag{9}$$

where $n_c = |J_c|$ and

$$\overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}.$$

Since $\hat{z}$ is generated (approximately) by the standard normal distribution, a hypothesis test can be formed where the null hypothesis is that the average cluster weights corresponding to the cth class *are not* significantly different from the average cluster weights corresponding to all classes, and the alternate hypothesis is that the average cluster weights corresponding to the cth class *are* significantly different from the average cluster weights corresponding to all classes. Again, since class memberships are known *a priori* and all pixels have some positive membership with all clusters, testing for class memberships is not meaningful, but testing for significantly different cluster weights is meaningful. If $P(Z > \hat{z}) < \alpha$, the probability of observing the difference in the average cluster weights associated with c and the average cluster weights associated with all classes in the jth cluster is significant, and the null hypothesis is rejected. If the null hypothesis is *not* rejected, the cluster itself is rejected as impure, and further refinement is necessary.

V. Iteration

Together with the cluster association significance test, the iteration forms the semisupervised clustering framework in CIGSCR. The application of a hypothesis test determines which clusters should be used for classification, and an iteration works to produce a set of associated clusters with each class being represented by at least one associated cluster. This is accomplished by introducing new clusters that are likely to be associated and, when necessary, are associated with a class not already represented by a cluster.

In IGSCR, pure hard clusters are removed from the image that is clustered in subsequent iterations, focusing further refinement on clusters that failed to pass the purity test. K clusters are used for each iteration, presumably producing smaller clusters as less data are divided into the same number of clusters. The underlying assumption is that clusters that fail to pass the purity test could actually be composed of multiple clusters that would pass the purity test individually, and clustering the remaining data into K more clusters will reveal these smaller clusters. This method will not directly work on soft clusters as soft clusters cannot be removed simply by removing any sample associated with a pure cluster—all samples have a positive probability of belonging to any particular cluster.

In CIGSCR, unassociated clusters are targeted for refinement by using their information to create new clusters that will likely be associated. IGSCR is effectively locating smaller clusters that, when combined to form a larger cluster, would have been rejected. IGSCR accomplishes this by finding the same number of clusters (K) in the original data set and then in successively smaller subsets of that original data set. A similar approach that would locate smaller pure clusters in rejected clusters is "splitting" a cluster, employed by Ball and Hall [30] in Iterative Self-Organizing Data Analysis Techniques (ISODATA). Clusters are split by partitioning a cluster into two new clusters and recalculating new means. Soft clusters are represented by cluster means, and splitting a soft cluster would equate with replacing one cluster mean with two cluster means (calculated based on data associated with a cluster).

A cleaner algorithmic solution is to add one new cluster using the information contained in the target cluster (the cluster that would be split), which effectively splits the cluster into two clusters. When using a clustering algorithm based on objective function (6), adding a new cluster guarantees a smaller function value (shown hereinafter) when $p = 2$. Using only the labeled samples belonging to the majority class (as determined in the cluster association significance test) to seed a new cluster would have the effect of pulling the new cluster toward those samples. Once another clustering iteration is completed, the targeted cluster would produce one cluster that is likely to be associated with the majority class and another cluster that retains relatively strong associations with all other classes. In CIGSCR, once the association significance test is performed, if at least one cluster is unassociated (and there are no unassociated classes), the cluster with the lowest value of $\hat{z}$ is used to generate a new cluster. The new cluster mean is determined using

$$U^{(K+1)} = \frac{\sum_{i \in \phi^{-1}(c_k)} w_{ik} X^{(i)}}{\sum_{i \in \phi^{-1}(c_k)} w_{ik}} \tag{10}$$

where k is the cluster with the lowest value of $\hat{z}$ and c_k is the majority class in cluster k, and recall that $\phi^{-1}(c)$ is the index set of labeled samples whose label is c. This formula also works when a class other than the majority class is used to seed a new cluster mean.

A shortcoming in IGSCR is that there is no guarantee that any clusters will be created and labeled with any particular class, and if a particular class is not represented by a cluster, the desired classification cannot be performed. In CIGSCR, this issue is addressed by adding a new cluster using the information from a particular class if that class is not represented in the associated clusters. If a class c is not represented in the associated clusters, the cluster that is closest to being associated with c is used to generate a new cluster using (10) with $c_k = c$. The "closest" cluster is determined to be the cluster with the highest ratio of the average membership of class c to the average membership of the majority class.

When there are classes not represented by associated clusters and there are unassociated clusters, only one method can be used to determine the creation of a new cluster. If a cluster is unassociated, it is simply not used in classification. It is more important to have each class represented by the associated clusters than to refine an unassociated cluster, because the desired classification cannot be applied unless all classes are represented by associated clusters. Therefore, adding a new cluster so that all classes will be represented takes precedence over adding a new cluster because an existing cluster is unassociated.

Finally, the theorem proving that adding one cluster mean will result in a smaller value of (6) is presented in Appendix A.

Assuming that the clustering algorithm locates a local minimum point of the objective function, the combination of the clustering algorithm and this cluster prototype addition are guaranteed to move toward a smaller objective function value. If left unchecked, infinitely many clusters could be added, and the algorithm would continue to find smaller objective function values. The association significance test plays a crucial role in the termination of this iterative process. Once all clusters pass the association significance test and each class has at least one associated cluster, the iteration stops because the higher level objective has been met: Clusters that significantly correspond to all classes have been located. The iteration also terminates when a maximum number of clusters is reached, and only those clusters that pass the association significance test are used for classification. Pseudocode for the full CIGSCR algorithm is presented in Appendix B.

VI. Distance Functions

In [28], Bezdek suggests that other functions may be used to calculate ρ_{ij} in place of squared Euclidean distance. To more accurately model the assumed normal probability distribution of the data, $\rho_{ij} = \exp(\|x^{(i)} - U^{(j)}\|_2^q)$ will be used. Note that, in the previous proof, showing that adding a soft cluster will result in a smaller objective function value, the proof holds for positive values of ρ_{ij} (ρ_{ij} is not required to be squared Euclidean distance). When using a crisp clustering function such as k-means, calculating the exponential function is not necessary as each sample is simply assigned to the cluster that

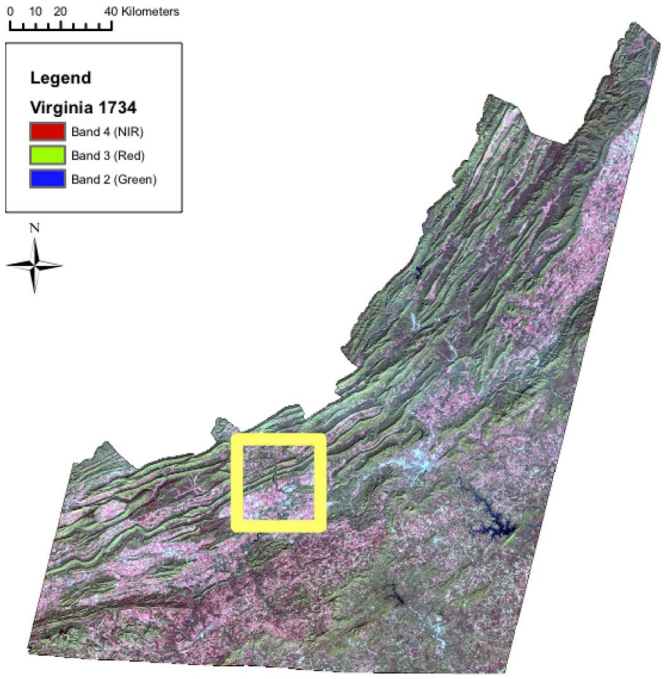


Fig. 3. Landsat ETM+ path 17/row 34 over Virginia, U.S., with the area of interest highlighted.

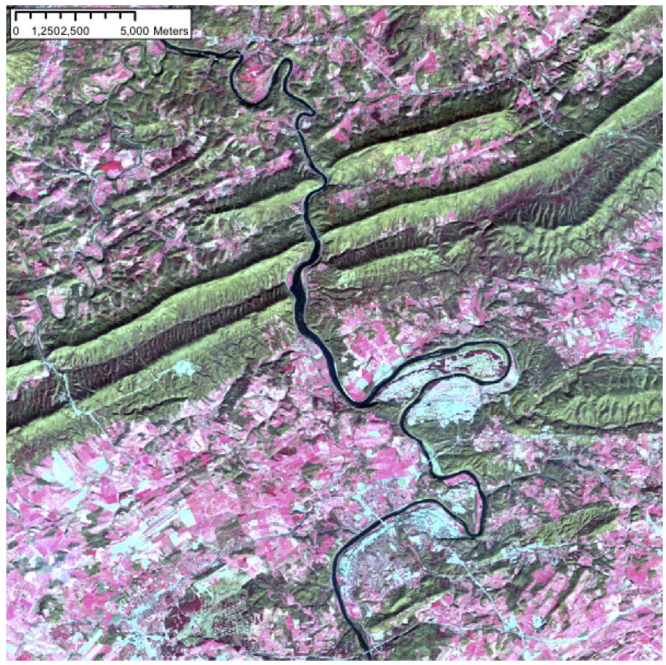


Fig. 4. Landsat ETM+ path 17/row 34 area of interest over Virginia, U.S.

maximizes the posterior probability. Since the soft memberships are being retained as probability estimates, calculating the exponential function is necessary (and results in more accurate clustering, as shown in the result section).

VII. Experimental Results and Discussion

Experimental results for IGSCR and CIGSCR were obtained using a mosaicked Landsat Enhanced Thematic Mapper Plus (ETM+) satellite image taken from Landsat Worldwide Reference System path 17, row 34, located in Virginia, U.S., shown in Fig. 3 [31]. This image, hereafter referred to as VA1734, was acquired on November 2, 2003 and consists largely of forested mountainous regions and a few urban areas that are predominantly light blue and light pink in Fig. 3. Fig. 3 contains a three-color representation of VA1734 where the red color band corresponds to the near-infrared wavelength in VA1734, the green color band corresponds to the red wavelength in VA1734, and the blue color band corresponds to the green wavelength in VA1734. Fig. 4 shows a zoomed area of interest. VA1734 has a resolution of 30 m, 8720 rows, 8575 columns, and 6 bands.

The training data for this image were created by the interpretation of point locations from a systematic hexagonal grid over Virginia base mapping program true color digital orthophotographs [32], [33]. Twenty-nine thousand points were included in the training data set (approximately 8000 nonforest and 21 000 forest). A two-class classification was performed (forest/nonforest), and the classification parameters and results are given in Table I (DR classification) and Table II (IS/IS+ classification). Classification images for this data set are given in Figs. 5–9.

Test data in the form of point locations at the center of U.S. Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) ground plots were used to assess the accuracy of this classification. Nine hundred fifty-nine uniformly scattered (throughout the image) points were included in the test set. Since these test data are typically used to evaluate crisp classifications, only homogeneous FIA plots were used (either 100% forest or nonforest), and these plots were obtained between 1997 and 2001. Accuracy was assessed based on an error matrix where classification results for specific points (not included in the training data set) are compared against known class values. The accuracies reported in Tables I and II were obtained by first converting all soft classifications to hard classifications for the purpose of comparing hard classification values to hard ground truth values. The classification results reported in Tables I and II used 10, 15, 20, and 25 initial clusters for IGSCR and CIGSCR. Experimental runs of IGSCR used homogeneity thresholds (test probabilities of observing the majority class in a particular cluster) of 0.5 and 0.9, with $\alpha = 0.01$ for all IGSCR classifications. A threshold of 0.9 would indicate a homogeneous cluster, but a threshold of 0.5 is perhaps more analogous to the new association significance test used in CIGSCR. Experimental runs of CIGSCR used $\rho_{ij} = \|X^{(i,j)} - U^{(k)}\|_2$ (squared Euclidean distance) and $\rho_{ij} = \exp(\|X^{(i,j)} - U^{(k)}\|_2)$. For all reported CIGSCR runs, $\alpha = 0.0001$ (values of $\hat{z}$ tend to be high for the association significance test). Finally, classification was performed using just clustering without the semisupervised framework to evaluate the effect of the combination of the association significance test and iteration in CIGSCR on classification accuracies. Results for this classification are found in the last column of Tables I and II. Comparisons between IGSCR and supervised and unsupervised classifications are available in [13].

While fuzzy k-means is described earlier as an analog to k-means in IGSCR and was used to acquire the experimental results, other soft clustering methods could be used instead. Fuzzy k-means has the advantage of being efficient and is straightforward to use for the derivations in this paper. A more

TABLE I
IGSCR AND CIGSCR DR CLASSIFICATION ACCURACIES FOR VA1734

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$)		clustering
clusters	$p = .5$	$p = .9$	$\rho = \|\|x - U\|\|_2^2$	$\rho = e^{\|\|x-U\|\|_2}$	(no iteration)
10	85.81	75.49	88.74	87.70	72.26
15	88.22	74.56	80.50	86.97	73.72
20	84.78	89.57	79.87	88.74	76.54
25	87.49	84.25	81.44	88.74	77.58
average	86.5	80.97	82.64	88.04	75.03
st. deviation	1.62	7.21	4.12	.86	2.46

TABLE II
IGSCR IS+ AND CIGSCR IS CLASSIFICATION ACCURACIES FOR VA1734

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$)		clustering
clusters	$p = .5$	$p = .9$	$\rho = \|\|x - U\|\|_2^2$	$\rho = e^{\|\|x-U\|\|_2}$	(no iteration)
10	68.30	75.39	83.63	85.09	72.26
15	86.34	74.56	76.96	85.19	72.99
20	84.46	88.95	75.60	86.86	76.85
25	66.63	83.94	78.52	87.28	76.75
average	76.43	80.71	78.68	86.11	74.71
st. deviation	10.41	6.94	3.51	1.13	2.43

Fig. 5. IGSCR DR classification using ten initial clusters and a homogeneity threshold of 90%. Tan is nonforest, and green is forest.

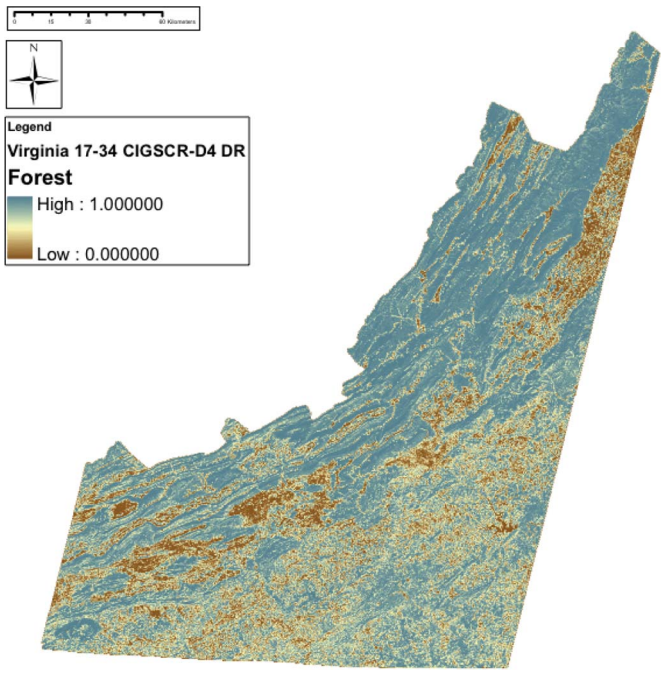


Fig. 6. CIGSCR DR classification using ten initial clusters and $\rho_{ij} = e^{\|X_i - U_j\|_2}$. Brown is nonforest, green is forest, and tan indicates both.

computationally expensive method such as Gaussian mixture model clustering may provide better clustering results. Furthermore, such a soft clustering approach could also be used in IGSCR if the soft memberships are converted to hard memberships. A further advantage of fuzzy k-means is that it is implemented on parallel computers quite efficiently. Results for this paper were generated in less than 10 min on a multicore computer.

A. Discussion

The classification results in Figs. 5 and 6 show that soft classification provides more information than hard classification, even if classification results are similar. While IGSCR (hard classification) separates the data in Fig. 5 into one of two classes, CIGSCR (soft classification method) retains information about how likely a sample is to belong to each class. Since only two classes are used in this classification, this information can be displayed using one figure where dark green indicates high probability of forest and dark brown indicates high probability of nonforest. The beige areas between green and brown indicate regions of high uncertainty, information that is not present in the hard classification in Fig. 5. The zoomed images in Figs. 7–9 contain IS classifications from IGSCR and CIGSCR, and these images further illustrate this point. The IGSCR IS classification contains a class for samples (shown in

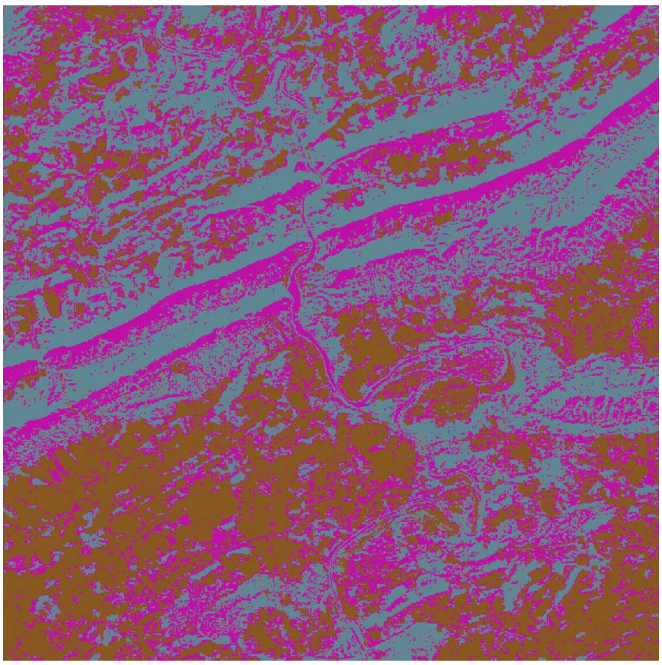


Fig. 7. IGSCR IS classification using ten initial clusters and a homogeneity threshold of 90%. Tan is nonforest, green is forest, and pink indicates no assigned class during the clustering phase.

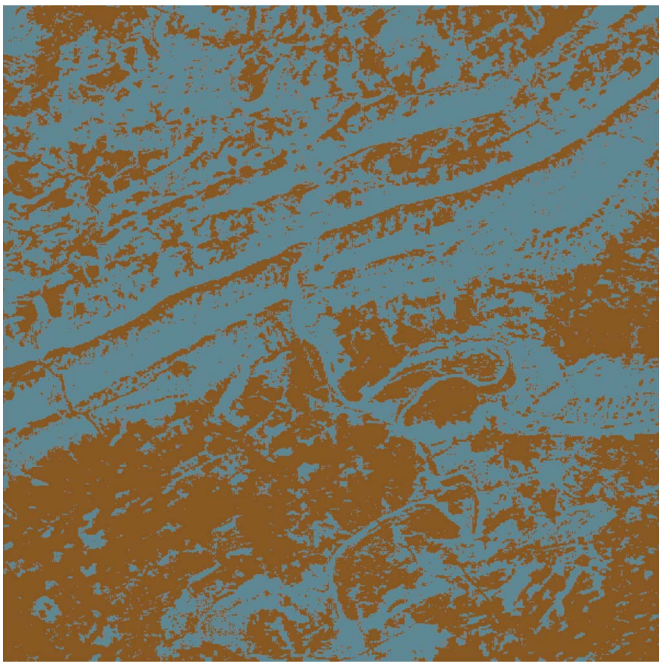


Fig. 8. IGSCR IS+ classification using ten initial clusters and a homogeneity threshold of 90%. Tan is nonforest, and green is forest.

pink) that are not clustered, meaning that they were originally part of confused or impure clusters. A logical correspondence seems likely between confused clusters in IGSCR and uncertain classes in CIGSCR, and in fact, most of the uncertain samples in the CIGSCR IS classification are also unclassified in the IGSCR IS classification, although the reverse is not true. Soft classification is necessary in order to have information on classification certainty (and uncertainty).

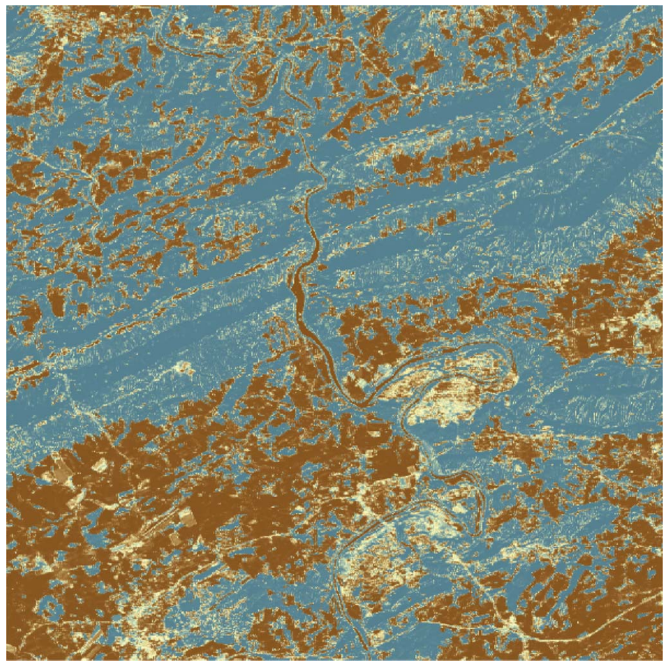


Fig. 9. CIGSCR IS classification using ten initial clusters and $\rho_{ij} = e^{\|X_i - U_j\|_2}$. Green is forest, brown is nonforest, and tan indicates a mixture.

One advantage of soft clustering is the potential ability to correctly identify spectral classes that are close to the boundary (or in the overlapping boundary) between information classes. In the experimental data (shown in Figs. 5 and 6), water appears similar to forest but should be labeled as nonforest. In the experimental results (Figs. 7–9), IGSCR incorrectly labels water as forest (see Fig. 8), highlighting a limitation in the use of hard clusters to locate spectral classes. IGSCR either was unable to generate a pure cluster that corresponded to water or lacked sufficient water training data to correctly label that spectral class. Pink regions in Fig. 7 show which pixels in the area of interest were not attributed to a pure spectral class, and water pixels are included. CIGSCR correctly labeled the water as nonforest, as shown in Fig. 9. Similarly, the shadowed forested regions were not part of a pure spectral class in IGSCR (Fig. 7), were ultimately incorrectly labeled as nonforest using IGSCR (Fig. 8), and were correctly labeled as forest using CIGSCR (Fig. 9). In this experiment, the correct identification of these spectral classes is the direct result of using soft clustering.

In addition to inherently providing more information through soft memberships, CIGSCR has other desirable properties over IGSCR, namely, it is frequently more accurate and less sensitive to input parameters. Based on accuracies reported in Tables I and II, CIGSCR is less sensitive to the number of initial clusters than IGSCR (particularly when $\rho = e^{\|x-U\|_2}$ is used). As shown in Tables I and II, IGSCR can be sensitive to the number of initial clusters and the homogeneity threshold and is less accurate on average than CIGSCR using $\rho = e^{\|x-U\|_2}$ (based on the data being normally distributed). Note that, particularly for the IS+ classifications obtained directly from clustering, the standard deviation of experimental IGSCR accuracies is several times larger than those from CIGSCR, indicating the sensitivity to parameters. The set of clusters ultimately used for

classification in IGSCR is directly affected by the number of initial clusters and the homogeneity test, and furthermore, when all clusters fail the homogeneity test, the iteration terminates, and no more clusters are found. The number of clusters used for classification can vary widely depending on the number of iterations completed as each iteration potentially produces several pure clusters. The low accuracies reported for the IGSCR IS+ classifications in Table II occur when a small number of iterations occur, which can be greatly influenced by the number of initial clusters and the homogeneity test. The classification accuracies reported for CIGSCR in Tables I and II are more consistent as CIGSCR does not have the same sensitivity issues. First, the association significance test no longer requires a user input threshold like the homogeneity test. The homogeneity test in IGSCR evaluates the observed values against a user-supplied probability of observing a specific class (within a cluster), but the association significance test in CIGSCR determines if the average cluster memberships per class are statistically significantly different (requiring no user-specified probability). Second, the iteration in CIGSCR is fundamentally different from the iteration in IGSCR. While each iteration in IGSCR locates multiple clusters, each iteration in CIGSCR adds one additional cluster, and terminating this iteration potentially excludes many fewer clusters from the final classification than terminating the iteration in IGSCR (particularly when few iterations occur). As classification methods are already sensitive to training data and clustering methods are sensitive to initial prototype locations, classifications being sensitive to fewer parameters are a desirable property.

The final column in Tables I and II provides compelling evidence that the semisupervised clustering frameworks in IGSCR and CIGSCR ultimately lead to more accurate classifications than merely clustering without supervision. The spectrally pure clusters located using IGSCR and CIGSCR lead to classification accuracies that are on the order of 10% higher. Furthermore, both classification methods iteratively locate clusters beyond the initial set of clusters, also potentially leading to higher classification accuracies.

VIII. Conclusion

This paper has presented a continuous analog to IGSCR that rejects and refines clusters to automatically classify a remote sensing image based on information class training data. This new algorithm addressed specific challenges presented by remote sensing data including large data sets (millions of samples), relatively small training data sets, and difficulty in identifying spectral classes. The resulting classifications are fundamentally different from IGSCR (the discrete predecessor to CIGSCR) classifications, even when converting the CIGSCR soft classifications to hard classifications. CIGSCR has many advantages over IGSCR, such as the ability to produce soft classification, less sensitivity to certain input parameters, potential to correctly classify regions that are not amply represented in training data, and a better ability to locate clusters associated with all classes. The semisupervised clustering framework within CIGSCR has been shown here to improve classification accuracies over clustering alone. This semisupervised clustering framework could be incorporated into many classification algorithms that use clustering.

The highly automated CIGSCR classification algorithm is a contribution to the remote sensing community that has few, if any, partially supervised soft classification algorithms analogous to the many partially supervised hard classification algorithms that exist. Future work includes using this soft classifier for many applications of classification in remote sensing.

Appendix A
Theorems

Theorem: Let $X^{(i)}$, $i = 1, 2, \ldots$, be B-dimensional random vectors having one of Q distinct multivariate normal distributions. Let $q = \psi(i)$ denote the distribution from which $X^{(i)}$ was sampled. For $i = 1, 2, \ldots$ and $j = 1, \ldots, K$, define the random variables

$$W_{ij} = W_j\left(X^{(i)}\right) = \frac{1/\left\|X^{(i)} - U^{(j)}\right\|_2^2}{\sum_{k=1}^{K} 1/\left\|X^{(i)} - U^{(k)}\right\|_2^2}$$

where K is the number of clusters and $U^{(k)} \in \Re^B$ is the kth cluster center (and is considered fixed for weight calculation). Then, for any $j = 1, \ldots, K$

$$P\left\{\frac{1}{B_{nj}}\sum_{i=1}^{n}(W_{ij} - a_{ij}) < x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz$$

as $n \to \infty$, where $a_{ij} = \mathrm{E}[W_{ij}]$, $b_{ij}^2 = \mathrm{Var}[W_{ij}]$, and $B_{nj}^2 = \sum_{i=1}^{n} b_{ij}^2$.

Proof: W_{ij} is a bounded $(0 \le W_{ij} \le 1)$ measurable function of a normal random variable and is, therefore, a random variable with finite mean and variance. Fix j for the remainder of the proof, and let $q = \psi(i)$ denote which of the Q distributions $X^{(i)}$ is from. In order to prove

$$P\left\{\frac{1}{B_{nj}}\sum_{i=1}^{n}(W_{ij} - a_{ij}) < x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz$$

it is sufficient to verify the Lindeberg condition [24]

$$\lim_{n\to\infty}\frac{1}{B_{nj}^2}\sum_{i=1}^{n}\int_{|x-a_{ij}|>\tau B_{nj}}(x - a_{ij})^2 dF_{\psi(i),j}(x) = 0$$

for any constant $\tau > 0$ where $F_{\psi(i),j}(x)$ is the cumulative distribution function for W_{ij}.

For each q, $1 \le q \le Q$, define $I_q = \psi^{-1}(q) = \{i | \psi(i) = q, 1 \le i \le n\}$ and $n_q = |I_q|$, and for $i \in I_q$, let $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ and $\mathrm{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$. Now, considering only the independent and identically distributed (i.i.d.) random variables W_{ij}, $i \in I_q$, the Lindeberg condition holds

$$\lim_{n_q\to\infty}\frac{1}{n_q\beta_{qj}^2}\sum_{i\in I_q}\int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}}(x - \alpha_{qj})^2\,dF_{qj}(x)$$

$$= \lim_{n_q\to\infty}\frac{1}{\beta_{qj}^2}\int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}}(x - \alpha_{qj})^2\,dF_{qj}(x) = 0.$$

Since β_{qj} is positive and finite and the integral is finite, the limit of the integral is zero as $\sqrt{n_q}\beta_{qj} \to \infty$.

W_{ij}, $i = 1, 2, \ldots$, denotes the random variables from Q i.i.d. distributions F_{qj}, $q = 1, \ldots, Q$, where the mean of the qth distribution is α_{qj}, the variance is β_{qj}^2, and the number of random variables from that distribution is n_q, where $\sum_{q=1}^{Q} n_q = n$. As $n \to \infty$, there is at least one q for which $n_q \to \infty$. For this sequence of independent random variables from Q distributions, the Lindeberg condition is

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{B_{nj}^2} \sum_{i=1}^{n} \int_{|x-a_{ij}|>\tau B_{nj}} (x - a_{ij})^2 \, dF_{\psi(i),j}(x) \\
&= \lim_{n\to\infty} \frac{1}{\sum_{k=1}^{Q} n_k \beta_{kj}^2} \sum_{q=1}^{Q} n_q \\
&\quad \cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x - \alpha_{qj})^2 \, dF_{qj}(x) \\
&= \lim_{n\to\infty} \sum_{q=1}^{Q} \frac{n_q}{\sum_{k=1}^{Q} n_k \beta_{kj}^2} \\
&\quad \cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x - \alpha_{qj})^2 \, dF_{qj}(x) \\
&\leq \lim_{n\to\infty} \sum_{q=1}^{Q} \frac{1}{\beta_{qj}^2} \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x - \alpha_{qj})^2 \, dF_{qj}(x) = 0.
\end{aligned}$$

Since each variance β_{qj}^2 is positive and finite and $B_{nj} = \sqrt{n_1\beta_{1j}^2 + \cdots + n_Q\beta_{Qj}^2} \to \infty$ as at least one $n_q \to \infty$, each integral converges to zero as $n \to \infty$, and the Lindeberg condition is verified. ■

Remark: The assumption that the $X^{(i)}$, $i = 1, 2, \ldots$, is generated from a finite number of normal distributions is stronger than necessary. This proof holds if $X^{(i)}$, $i = 1, 2, \ldots$, is generated from a finite number of arbitrary distributions.

Theorem: Given an integer $K > 0$, positive real numbers ρ_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, K+1$, defining a point $\rho \in \Re^{n \times K+1}$, and the objective function

$$J^{(K)}(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^2 \rho_{ij}$$

for K clusters where

$$w_{ij} = \frac{1/\rho_{ij}}{\sum_{k=1}^{K} 1/\rho_{ik}}$$

the objective function

$$J^{(K+1)}(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K+1} \hat{w}_{ij}^2 \rho_{ij}$$

for $K+1$ clusters where

$$\hat{w}_{ij} = \frac{1/\rho_{ij}}{\sum_{k=1}^{K+1} 1/\rho_{ik}}$$

satisfies

$$J^{(K+1)}(\rho) < J^{(K)}(\rho).$$

Proof: Note that the ρ_{ij} does not change with the addition of the $(K+1)$st cluster prototype; however, $\hat{w}_{ij} < w_{ij}$ for $j < K+1$ because the denominator of $\hat{w}_{ij}$ has an additional term. Let $J_i^{(K)} = \sum_{j=1}^{K} w_{ij}^2 \rho_{ij}$ and $J_i^{(K+1)} = \sum_{j=1}^{K+1} \hat{w}_{ij}^2 \rho_{ij}$. It is sufficient to show that $J_i^{(K+1)} < J_i^{(K)}$ for each i to prove that $J^{(K+1)} < J^{(K)}$.

Let

$$S_1 = \sum_{k=1}^{K} 1/\rho_{ik} \qquad S_2 = \sum_{k=1}^{K+1} 1/\rho_{ik}.$$

Then

$$w_{ij}^2 = \frac{(1/\rho_{ij})^2}{S_1^2} \qquad \hat{w}_{ij}^2 = \frac{(1/\rho_{ij})^2}{S_2^2}.$$

$$\begin{aligned}
J_i^{(K)} - J_i^{(K+1)} &= \sum_{j=1}^{K} \frac{(1/\rho_{ij})}{S_1^2} - \sum_{j=1}^{K+1} \frac{(1/\rho_{ij})}{S_2^2} \\
&= \frac{S_2^2 \sum_{j=1}^{K} (1/\rho_{ij}) - S_1^2 \sum_{j=1}^{K+1} (1/\rho_{ij})}{S_1^2 S_2^2}.
\end{aligned}$$

Examining only the numerator in the previous term

$$\begin{aligned}
&(S_1 + (1/\rho_{i,K+1}))^2 \sum_{j=1}^{K} (1/\rho_{ij}) \\
&\qquad - S_1^2 \left(\sum_{j=1}^{K} (1/\rho_{ij}) + (1/\rho_{i,K+1}) \right) \\
&= (S_1 + (1/\rho_{i,K+1}))^2 S_1 - S_1^2 (S_1 + (1/\rho_{i,K+1})) \\
&= S_1^3 + 2S_1^2(1/\rho_{i,K+1}) + S_1(1/\rho_{i,K+1})^2 \\
&\qquad - S_1^3 - S_1^2(1/\rho_{i,K+1}) \\
&= S_1^2(1/\rho_{i,K+1}) + S_1(1/\rho_{i,K+1})^2 \\
&> 0
\end{aligned}$$

yielding

$$J_i^{K+1} < J_i^{(K)}.$$

■

APPENDIX B
PSEUDOCODE

Algorithm CIGSCR
Input: X % multispectral image
ϕ^{-1} % set of (row, col) indices for each class
$Kinit$ % number of initial clusters
$Kmax$ % maximum number of clusters
C % number of classes
ϵ % convergence threshold
α % Type-I error for one-sided hypothesis test
Output: DR % decision rule classification
IS % iterative stacked classification
begin
% Initialization
Initialize cluster means U along the axis defined by the mean plus or minus the standard deviation of the image X;
$K := Kinit$;
% Begin Iteration
for $iteration := Kinit$ **step** 1 **until** $Kmax$ **do**
begin
$w := 0$; $convergence := 1$;
while $convergence > \epsilon$ **do**
begin
% Cluster Data
$num := 0$; $denom := 0$;
for $i := 1$ **step** 1 **until** $rows$ **do**
for $j := 1$ **step** 1 **until** $cols$ **do**
for $k := 1$ **step** 1 **until** K **do**
begin

$$\hat{w}_{ij,k} := \frac{1/\|X^{(ij)} - U^{(k)}\|_2^2}{\sum_{l=1}^{K} 1/\|X^{(ij)} - U^{(l)}\|_2^2};$$

% update sums for mean calculations.
$num^{(k)} := num^{(k)} + \hat{w}_{ij,k}^2 X^{(ij)}$;
$denom_k := denom_k + \hat{w}_{ij,k}^2$;
end
% update cluster means
for $k := 1$ **step** 1 **until** K **do**

$$U^{(k)} := \frac{num^{(k)}}{denom_k};$$

$$convergence := \max_{i,j,k} |w_{ij,k} - \hat{w}_{ij,k}|;$$

$w := \hat{w}$;
end
% Determine Good Clusters
for $k := 1$ **step** 1 **until** K **do**
begin
Determine majority class c of cluster k;
$c_k := c$;

$$Z_k := \frac{\sqrt{n_c}(\overline{w}_{c,k} - \overline{w}_k)}{s_{\overline{w}_k}};$$

end
if any class is not associated with a cluster **then**
begin
$c :=$ first unassociated class
$k := \arg\max_k(\overline{w}_{c,k}/w_{c_k,k})$
$K := K + 1$

$$U^{(K)} = \frac{\sum_{ij \in \phi^{-1}(c)} w_{ij,k} X^{(ij)}}{\sum_{ij \in \phi^{-1}(c)} w_{ij,k}};$$

end
elseif (any$(Z_k < Z(\alpha), k = 1, \ldots, K)$ **then**
begin
$k := \arg\min_k Z_k$;
$K := K + 1$;

$$U^{(K)} = \frac{\sum_{ij \in \phi^{-1}(c_k)} w_{ij,k} X^{(ij)}}{\sum_{ij \in \phi^{-1}(c_k)} w_{ij,k}};$$

end
else
exit for loop;
end
end
for $k := 1$ **step** 1 **until** K **do**
begin
% initialize for covariance calcs.
$\Sigma_k := 0$;
$denom_k := 0$;
end
% IS classification
for $i := 1$ **step** 1 **until** $rows$ **do**
for $j := 1$ **step** 1 **until** $cols$ **do**
begin
$csum := 0$;
for $k := 1$ **step** 1 **until** K **do**
if $(Z_k > Z(\alpha))$ **then**
$csum_{c_k} := csum_{c_k} + w_{ij,k}$;
for $c := 1$ **step** 1 **until** C **do**

$$IS_{ij,c} := \frac{csum_c}{\sum_{k=1}^{C} csum_k};$$

% calculate covariance matrices
for $k := 1$ **step** 1 **until** K **do**
begin
$\Sigma_k := \Sigma_k + w_{ij,k}$
$\cdot (X^{(ij)} - U^{(k)})(X^{(ij)} - U^{(k)})^T$;
$denom_k := denom_k + w_{ij,k}$;
end
end
for $k := 1$ **step** 1 **until** K **do**
$\Sigma_k := 1/denom_k \cdot \Sigma_k$;
% DR classification
for $i := 1$ **step** 1 **until** $rows$ **do**
for $j := 1$ **step** 1 **until** $cols$ **do**
begin
$csum := 0$;
for $k := 1$ **step** 1 **until** K **do**
if $(Z_k > Z(\alpha))$ **then**

begin

$$p := \frac{2e^{-(1/2)(X^{(ij)}-U^{(k)})^T \Sigma_k^{-1}(X^{(ij)}-U^{(k)})}}{\pi^{B/2}|\Sigma_k|^{1/2}};$$

$csum_{c_k} := csum_{c_k} + p;$

else

$csum_{c_k} := 0;$

end

for $c := 1$ **step** 1 **until** C **do**

$$DR_{ij,c} := \frac{csum_c}{\sum_{k=1}^{C} csum_k};$$

end

end

REFERENCES

[1] M. D. Fleming, J. S. Berkebile, and R. M. Hoffer, "Computer-aided analysis of LANDSAT-1 MSS data: A comparison of three approaches, including a 'modified clustering' approach," LARS Technical Reports, Paper 96, 1975.

[2] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.

[3] L. Gomez-Chova, D. Fernandez-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, "Partially supervised hierarchical clustering of SAR and multispectral imagery for urban areas monitoring," in *Proc. SPIE, Conf. Image Signal Process. Remote Sens. X*, 2004, pp. 138–149.

[4] B. Jeon and D. A. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 1073–1079, Mar. 1999.

[5] C. Sanchez-Hernandez, D. S. Boyd, and G. M. Foody, "One-class classification for mapping a specific land-cover class: SVDD classification of fenland," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 1061–1073, Apr. 2007.

[6] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, Nov. 2006.

[7] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[8] R. Cossu, S. Chaudhuri, and L. Bruzzone, "A context-sensitive Bayesian technique for the partially supervised classification of multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 3, pp. 352–356, Jul. 2005.

[9] L. Bruzzone and D. F. Prieto, "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 456–460, Feb. 2001.

[10] D. Fernandez-Prieto, "An iterative approach to partially supervised classification problems," *Int. J. Remote Sens.*, vol. 23, no. 18, pp. 3887–3892, 2002.

[11] M. E. Bauer, T. E. Burke, A. R. Ek, P. R. Coppin, S. D. Lime, T. A. Walsh, D. K. Walters, W. Befort, and D. F. Heinzen, "Satellite inventory of Minnesota forest resources," *Photogramm. Eng. Remote Sens.*, vol. 60, no. 3, pp. 287–298, 1994.

[12] J. P. Wayman, R. H. Wynne, J. A. Scrivani, and G. A. Reams, "Landsat TM-based forest area estimation using iterative guided spectral class rejection," *Photogramm. Eng. Remote Sens.*, vol. 67, no. 10, pp. 1155–1166, Oct. 2001.

[13] R. F. Musy, R. H. Wynne, C. E. Blinn, J. A. Scrivani, and R. E. McRoberts, "Automated forest area estimation via iterative guided spectral class rejection," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 8, pp. 949–960, 2006.

[14] R. D. Phillips, L. T. Watson, and R. H. Wynne, "Hybrid image classification and parameter selection using a shared memory parallel algorithm," *Comput. Geosci.*, vol. 33, no. 7, pp. 875–897, Jul. 2007.

[15] H. Jiang, J. R. Strittholt, P. A. Frost, and N. C. Slosser, "The classification of late seral forests in the Pacific Northwest, USA using Landsat ERM+ imagery," *Remote Sens. Environ.*, vol. 91, no. 3/4, pp. 320–331, 2004.

[16] M. Kelly, D. Shaari, Q. H. Guo, and D. S. Liu, "A comparison of standard and hybrid classifier methods for mapping hardwood mortality in areas affected by 'sudden oak death'," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 11, pp. 1229–1239, 2004.

[17] R. Sivanpillai, C. T. Smith, R. Srinivasan, M. G. Messina, and X. Ben Wu, "Estimating regional forest cover in East Texas using enhanced thematic mapper (ETM plus) data," *Forest Ecol. Manage.*, vol. 218, no. 1–3, pp. 342–352, 2005.

[18] R. H. Wynne, K. A. Joseph, J. O. Browder, and P. M. Summers, "Comparing farmer-based and satellite-derived deforestation estimates in the Amazon basin using a hybrid classifier," *Int. J. Remote Sens.*, vol. 28, no. 6, pp. 1299–1315, Mar. 2007.

[19] Z. Sha, Y. Bai, Y. Xie, M. Yu, and L. Zhang, "Using a hybrid fuzzy classifier (HFC) to map typical grassland vegetation in Xilin River Basin, Inner Mongolia, China," *Int. J. Remote Sens.*, vol. 29, no. 8, pp. 2317–2337, Apr. 2008.

[20] A. Kumar, S. K. Ghosh, and V. K. Dadhwal, "Full fuzzy land cover mapping using remote sensing data based on fuzzy c-means and density estimation," *Can. J. Remote Sens.*, vol. 33, no. 2, pp. 81–87, 2007.

[21] M. Pepe, L. Boschetti, P. A. Brivio, and A. Rampini, "Accuracy benefits of a fuzzy classifier in remote sensing data classification of snow," in *Proc. IEEE Int. Conf. Fuzzy Syst. FUZZ-IEEE*, 2007, vol. 1–4, pp. 492–497.

[22] F. Okeke and A. Karnieli, "Methods for fuzzy classification and accuracy assessment of historical aerial photographs for vegetation change analyses. Part I: Algorithm development," *Int. J. Remote Sens.*, vol. 27, pp. 153–176, 2006.

[23] D. E. Sabol, J. B. Adams, and M. O. Smith, "Quantitative subpixel spectral detection of targets in multispectral images," *J. Geophys. Res.—Planets*, vol. 97, no. E2, pp. 2659–2672, Feb. 1992.

[24] B. V. Gnedenko, *Theory of Probability*, 6th ed. Amsterdam, The Netherlands: Gordan and Breach Sci. Publ., 1997, 497pp.

[25] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA: SIAM, 2007, pp. 161–163.

[26] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 3rd ed. Berlin, Germany: Springer-Verlag, 1999, 363 pp.

[27] J. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1974.

[28] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 1, pp. 1–8, Jan. 1980.

[29] W. I. Thacker, J. Zhang, L. T. Watson, J. B. Birch, M. I. Iyer, and M. W. Berry, "Algorithm XXX: SHEPPACK: Modified shepard algorithm for interpolation of scattered multivariate data," Computer Science, Virginia Tech., Blacksburg, VA, Tech. Rep.TR-09-13, 2009.

[30] G. H. Ball and D. J. Hall, "ISODATA, A novel method of sata analysis and pattern classification," Stanford Research Inst., Menlo Park, CA, NTIS Rep. AD699 616, 1965.

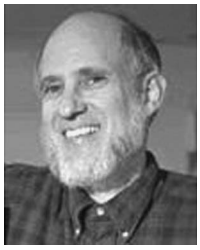
[31] A. M. Mika, "Three decades of Landsat instruments," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 7, pp. 839–852, Jul. 1997.

[32] W. A. Bechtold and C. T. Scott, "The forest inventory and analysis plot design," "The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures," W. A. Bechtold and P. L. Patterson, Eds., USDA Forest Service Southern Research Station, Asheville, NC, Gen. Tech. Rep. SRS-80, pp. 27–42, 2005.

[33] G. A. Reams, W. D. Smith, M. H. Hansen, W. A. Bechtold, R. A. Roesch, and G. G. Molsen, "The forest inventory and analysis sampling frame," "The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures," W. A. Bechtold and P. L. Patterson, Eds., USDA Forest Service Southern Research Station, Asheville, NC, Gen. Tech. Rep. SRS-80, pp. 11–26, 2005.

Rhonda D. Phillips (M'08) received the Ph.D. degree from the Virginia Polytechnic Institute and State University, Blacksburg, in 2009.

She is currently with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, as a Technical Staff Member. Her research interests include remote sensing optical image processing, synthetic aperture radar image processing, statistical inference, data reduction, and high-performance computing.



Layne T. Watson (F'93) received the B.A. degree (*magna cum laude*) in psychology and mathematics from the University of Evansville, Evansville, IN, in 1969, and the Ph.D. degree in mathematics from the University of Michigan, Ann Arbor, in 1974.

He was with Sandia National Laboratories, Albuquerque, NM, and General Motors Research Laboratories and served on the faculties of the University of Michigan, Michigan State University, East Lansing, and University of Notre Dame, Notre Dame, IN. He is currently a Professor of computer science and mathematics with the Virginia Polytechnic Institute and State University, Blacksburg. He serves as a Senior Editor of *Applied Mathematics and Computation* and an Associate Editor of *Computational Optimization and Applications*, *Evolutionary Optimization*, *Engineering Computations*, and the *International Journal of High Performance Computing Applications*. He is a Fellow of the National Institute of Aerospace. He has published well over 290 refereed journal articles and 200 refereed conference papers. His research interests include fluid dynamics, solid mechanics, numerical analysis, optimization, parallel computation, mathematical software, image processing, and bioinformatics.

Prof. Watson is a Fellow of the International Society of Intelligent Biological Medicine.

Randolph H. Wynne (M'01) received the Ph.D. degree from the University of Wisconsin, Madison, in 1995.

He is currently a Professor with the Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, where he also serves as an Associate Director of the Conservation Management Institute and a Codirector of the Center for Environmental Applications of Remote Sensing. He is a member of the Landsat Science Team. His current remote sensing research focuses on algorithm development, image time series analysis of forest disturbance and recovery, ecosystem service assessment and modeling, and terrestrial ecology.

Naren Ramakrishnan (M'98) received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1997.

He is currently a Professor and the Associate Head for Graduate Studies with the Department of Computer Science, Virginia Polytechnic Institute and State University (Virgina Tech), Blacksburg. His research interests include temporal data mining, designing algorithms that can flexibly incorporate domain knowledge, and new knowledge discovery abstractions for problems in intelligence analysis, sustainability, neuroscience, systems biology, and other areas. At Virginia Tech, he directs the Discovery Analytics Center, a university-wide effort that studies data mining problems in important areas of national interest. He is an ACM Distinguished Scientist.