Contemporary Engineering Sciences, Vol. 7, 2014, no. 7, 299 - 314
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.413

Continuous Iterative Guided Spectral Class Rejection Classification Algorithm for Hyperspectral Data

Rhonda D. Phillips[1], Layne T. Watson[1,2], Randolph H. Wynne[3], and Naren Ramakrishnan[1]

Departments of Computer Science[1], Mathematics[2], and
Forest Resources and Environmental Conservation[3]
Virginia Polytechnic Institute & State Univ., Blacksburg, VA 24061 USA

Abstract

This paper discusses a hyperspectral application of the continuous iterative guided spectral class rejection (CIGSCR) classification method based on the iterative guided spectral class rejection (IGSCR) classification method for remotely sensed data. Both CIGSCR and IGSCR use semisupervised clustering to locate clusters that are associated with classes in a classification scheme. In CIGSCR and IGSCR, training data are used to evaluate the strength of the association between a particular cluster and a class, and a statistical hypothesis test is used to determine which clusters should be associated with a class and used for classification and which clusters should be rejected and possibly refined. Experimental results indicate that the soft classification output by CIGSCR is reasonably accurate (when compared to IGSCR), and the fundamental algorithmic changes in CIGSCR (from IGSCR) result in CIGSCR being less sensitive to input parameters that influence iterations, as well as being highly parallelizable. Furthermore, evidence is presented that the semisupervised clustering in CIGSCR produces more accurate classifications than classification based on clustering without supervision.

Keywords: fuzzy clustering, semisupervised learning, remote sensing

Introduction

The conversion of the iterative guided spectral class rejection (IGSCR) classification method ([6], [2], [4]) from a hard classification to a soft classification method called continuous iterative guided spectral class rejection (CIGSCR) based on soft clustering will require the soft cluster evaluation and refinement methods developed in [3]. IGSCR evaluates hard clusters using a statistical hypothesis test based on a binomial random variable. This discrete random variable can be used to model hard cluster and class memberships. CIGSCR uses the hypothesis test developed in [3] that is based on both discrete random variables modeling class memberships of training data and continuous random variables modeling soft cluster memberships. The iterative cluster refinement in IGSCR assumes samples are attributed to only one cluster (hard clustering), but the soft cluster refinement proposed in [3] uses soft memberships to seed new clusters, taking advantage of soft clustering and providing an alternative mechanism for cluster refinement.

This paper describes how the soft cluster evaluation and refinement are incorporated into the IGSCR framework to form CIGSCR, and provides experimental results demonstrating that CIGSCR can produce superior classifications to IGSCR, especially in circumstances that are ideally suited for soft clustering and classification, such as hyperspectral data classification. The following sections summarize the CIGSCR algorithm, discuss how alternative clustering distance (dissimilarity) functions may be used within CIGSCR, present experimental results and detailed discussion, and conclude.

CIGSCR Algorithm

Let $\{x^{(i)}\}_{i=1}^{n}$ be the multivariate data points to be clustered into K clusters and classified into C classes. $U^{(j)}$ denotes the jth cluster prototype and w_{ij} is the weight (or probability or fuzzy membership) of point $x^{(i)}$ in cluster j. CIGSCR, like IGSCR, begins by clustering an image, but unlike IGSCR, CIGSCR uses soft clusters where each sample has partial membership in each cluster. Each soft cluster is then evaluated using the association significance test based on one of two standard normal random variables derived in [3],

$$\hat{z} = \frac{\sqrt{n_c}(\overline{w}_{c,j} - \overline{w}_j)}{S_{\overline{w}_j}}, \tag{1}$$

where n_c is the number of samples labeled with the cth class ($1 \leq c \leq C$), $\overline{w}_{c,j}$ is the average weight of samples labeled with the cth class for the jth cluster ($1 \leq j \leq K$), $\overline{w}_j$ is the sample mean of all weights in the jth cluster, and $S_{\overline{w}_j}$

is the sample standard deviation of the weights for the jth cluster, and

$$\hat{z} = \frac{y_{c,j} - n_c \overline{w}_j}{\sqrt{p_c \sum_{d=1}^{C} n_d (S^2_{\overline{w}_{d,j}} + (1 - p_c)\overline{w}^2_{d,j})}}, \tag{2}$$

where $y_{c,j}$ is the sum of weights of samples labeled with the cth class for the jth cluster, p_c is the estimated probability of the cth class, and $S_{\overline{w}_{d,j}}$ is the sample standard deviation of weights of samples labeled with the dth class for the jth cluster. Clusters that fail the test are refined in subsequent iterations. Unless termination criteria are met, a new cluster is introduced using information in an existing cluster, effectively splitting that cluster into two clusters. The full CIGSCR algorithm is precisely defined in pseudocode in [3] using the soft k-means clustering algorithm and various classification rules. For brevity that pseudocode, association significance test, and classification rules DR (maximum likelihood), IS (iterative stacking), and IS+ (IS followed by DR) are not repeated here.

Distance Functions

For positive real numbers ρ_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, K+1$, and weights w_{ij} computed in a particular way (such as in soft k-means), the addition of a cluster will result in a smaller value of the objective function $J(\rho) = \sum_{i=1}^{n} \sum_j w_{ij}^2 \rho_{ij}$ (refer to [3]). Although the soft clustering iteration for the objective function $J(\rho) = \sum_{i=1}^{n} \sum_j w_{ij}^2 \rho_{ij}$ is only guaranteed to converge when ρ_{ij} is Euclidean distance squared (between the ith sample $X^{(i)}$ and the jth cluster prototype $U^{(j)}$), [1] suggests that other functions may be used. The Euclidean distance squared is a special case of a radial function: $f : \Re^B \to \Re$ is *radial* if $f(x) = f(y)$ for $\|x\|_2 = \|y\|_2$. Thus $\rho_{ij} = f(x^{(i)} - U^{(j)}) = \|x^{(i)} - U^{(j)}\|_2^2$ is radial. Some alternative radial functions include

$$f(x) = \exp(\|x\|_2^q)$$

and

$$f(x) = \|x\|_2^q$$

where $q \geq 1$ and $\rho_{ij} = f(x^{(i)} - U^{(j)})$. The advantage of using a radial function is that distances can be magnified so the difference between large and small cluster weights will be more extreme, approaching hard clustering.

None of the aforementioned metrics or radial functions influence the assignment of cluster weights based on the prelabeled points. Semisupervised clustering uses prior information to influence a clustering method. Although the association significance test and iteration are indirectly doing this, a modified objective function could directly use prior information to influence clusters.

Consider the modified objective function component

$$J_i = \sum_{j=1}^{K} w_{ij}^2 \rho_{ij}(1 + \beta L_{ij}), \quad i = 1, \ldots, n,$$

where the term βL_{ij} is the penalty associated with assigning a labeled pixel to a cluster with a different associated label [3]. $\phi(i) = c$ is the class label of the ith labeled pixel, and let $\phi(i) = \Omega \notin \{c_1, \ldots, c_C\}$ if the ith pixel is unlabeled,

$$C(j) = \begin{cases} c, & \text{if the } j\text{th cluster is associated with the} \\ & c\text{th class,} \\ \Omega, & \text{otherwise,} \end{cases}$$

$$L_{ij} = \begin{cases} 1, & \text{if } \phi(i) \neq \Omega,\ \phi(i) \neq C(j),\ C(j) \neq \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

The distance function $f(x^{(i)} - U^{(j)}) = d_{ij} = \rho_{ij}(1 + \beta L_{ij})$ can be substituted for ρ_{ij} in the CIGSCR algorithm to magnify the weights of pixels labeled with the cth class to clusters associated with the cth class. Note that in place of using the Euclidean distance squared for ρ_{ij} in the soft clustering algorithm, one of the radial functions of Euclidean distance or a distance function with a penalty (β) could be used.

Experimental Results and Discussion

The first dataset used to obtain experimental results for IGSCR and CIGSCR is a mosaicked Landsat Enhanced Thematic Mapper Plus (ETM+) satellite image taken from Landsat Worldwide Reference System (WRS) path 17, row 34, located in Virginia, USA, shown in Fig. 1. This image, hereafter referred to as VA1734, was acquired on November 2, 2003 and consists largely of forested, mountainous regions, and a few developed regions that are predominantly light blue and light pink in Fig. 1. Fig. 1 contains a three color representation of VA1734 where the red color band in Fig. 1 corresponds to the near infrared wavelength in VA1734, the green color band to the red wavelength in VA1734, and the blue color band to the green wavelength in VA1734.

The training data for this image was created by the interpretation of point locations from a systematic, hexagonal grid over Virginia Base Mapping Program (VBMP) true color digital orthophotographs. A two class classification was performed (forest/nonforest), and classification parameters and results are given in Table 1 (DR classification) and Table 2 (IS/IS+ classification). Classification images for this dataset are given in Figs. 2 though 5.

Validation data in the form of point locations at the center of USDA Forest Service Forest Inventory and Analysis (FIA) ground plots were used to assess the accuracy of this classification. Since these validation data are typically used to evaluate crisp classifications, only homogeneous FIA plots were

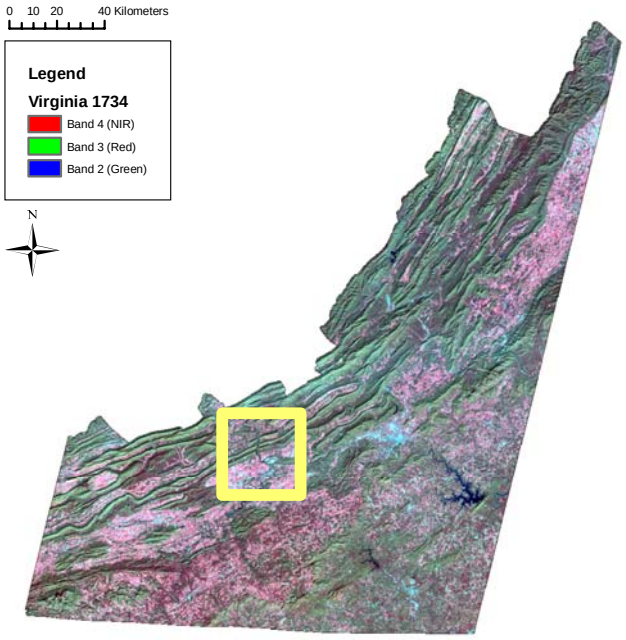

Figure 1: Landsat ETM+ path 17/row 34 over Virginia, USA with area of interest highlighted.

used (either 100 percent 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 1–4 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 1–4 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 .5 and .9, with $\alpha = .01$ for all IGSCR classifications. A threshold of .9 would indicate a homogeneous cluster, but a threshold of .5 is perhaps more analogous to the new association significance test used in CIGSCR. Experimental runs of CIGSCR used traditional Euclidean distance squared in addition to two proposed radial functions $f(X^{(i,j)} - U^{(k)}) = \|X^{(i,j)} - U^{(k)}\|_2^4$ and $f(X^{(i,j)} - U^{(k)}) = \exp(\|X^{(i,j)} - U^{(k)}\|_2)$. For all reported CIGSCR runs, $\alpha = .0001$ (values of $\hat{z}$ tend to be high for the association significance test). All reported CIGSCR classifications used hypothesis test (2). Only three out of 24 total CIGSCR classifications reported in this paper were different using (1) and (2), and the difference in resulting classification accuracies was not significant and did not show that one test consistently resulted in higher classification accuracies than the other test. Values of $\hat{z}$ are slightly smaller using (2) than (1), resulting in more potential for

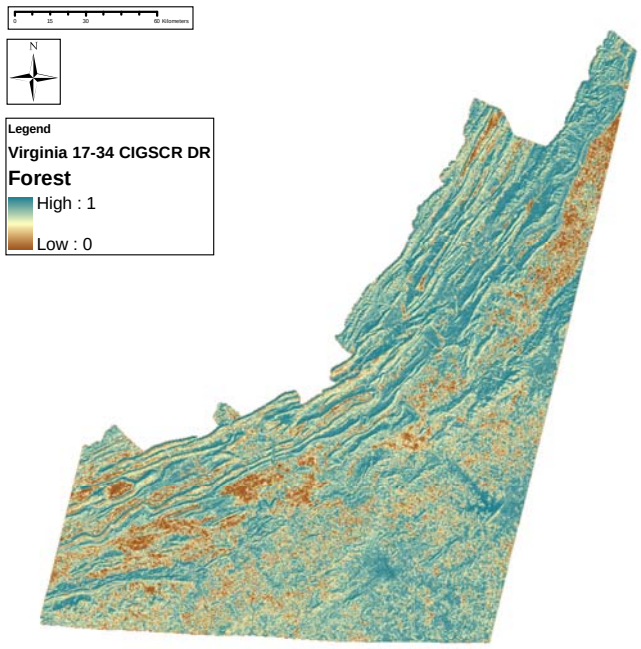

Figure 2: CIGSCR DR classification using ten initial clusters and Euclidean distance squared.

cluster refinement. Additionally, the distance function with penalty was used in classification, although results are not reported in Tables 1 and 2 because incorporating the penalty into the distance function did not increase classification accuracies in any experimental runs. Large values of β produced less accurate classification results. 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.

The second dataset used to obtain experimental results for IGSCR and CIGSCR is a hyperspectral image of the Appomattox Buckingham State Forest in Virginia, USA. The AVIRIS 224-band, low-altitude flight lines were acquired in the winter of 1999 and ranged from approximately 400-2500nm (10nm spectral resolution) with 3.4m spatial resolution [5]. The AVIRIS data were geometrically and radiometrically corrected (to level 1B at-sensor radiance, units of microwatts per square centimeter per nanometer per steradian) by the Jet Propulsion Laboratory (JPL; Pasadena, California, USA). The three flight lines used for this study were registered (8–12 control points per flight line) to an existing 0.5m orthophoto of the area. Resampling resulted in root mean square errors (RMSE) ranging between 0.23 and 0.24 pixels [5].

Training data were acquired by collecting 142 field locations [5] surrounded by homogeneous areas of single pine species (64 loblolly (*Pinus taeda*), 30 shortleaf (*Pinus echinata*), and 48 Virginia pine (*Pinus virginiana*)) with dif-

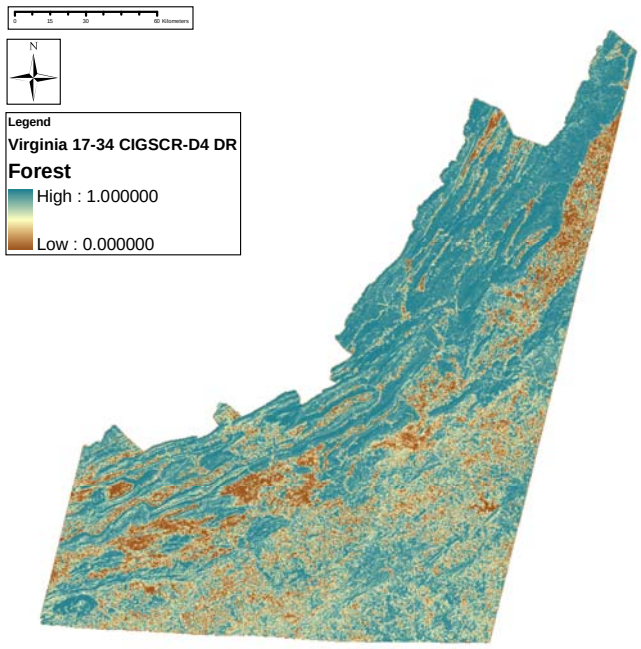

Figure 3: CIGSCR DR classification using ten initial clusters and Euclidean distance to the fourth power.

ferentially corrected global positioning system (GPS) coordinates. These locations were used in a region growing algorithm to obtain a sufficient number of points for training and validation, and nonpine training data were acquired using knowledge of the area and maps of known stands in the region. The image (shown in Fig. 6 and hereafter referred to as ABSF) contains various tree stands that include the three species of pines listed above, hardwoods, and mixed (evergreens and hardwoods).

400 points were randomly selected to serve as validation data for these four classes (loblolly, shortleaf, and Virginia pines, and nonpine). Classification results for these data are reported in Tables 3 and 4, Fig. 7 contains the IGSCR IS classification image using 25 initial clusters and a homogeneity threshold of .5, and Figs. 8a–d contain the CIGSCR IS classification images using 10 initial clusters and Euclidean distance to the fourth power. Classifications were run using the same parameters as classifications reported in Tables 1 and 2. An asterisk (*) indicates that the classification failed because at least one class had no associated clusters. Tables 5 and 6 report the number of pure clusters (IGSCR), and the number of clusters produced and number of associated clusters (CIGSCR).

Discussion.

The soft clustering and soft classification in CIGSCR can result in qualitatively different classifications than IGSCR. Even when the final classifications

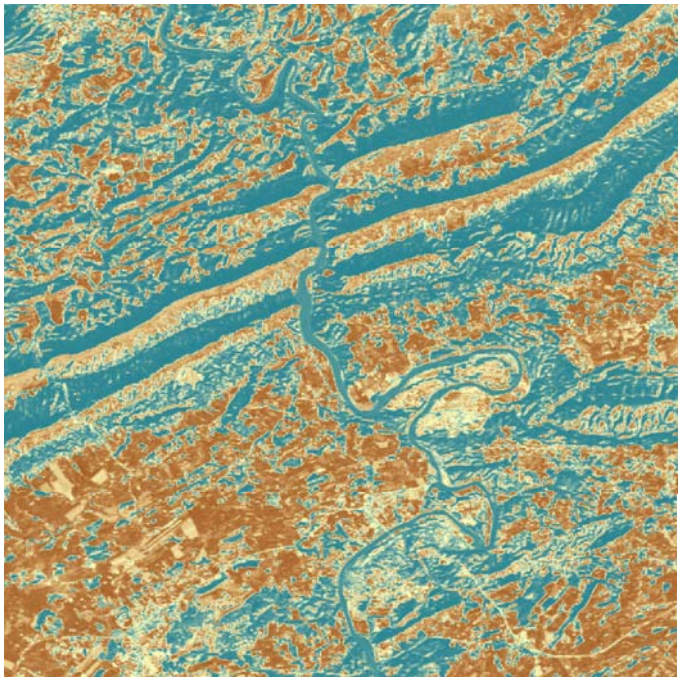

Figure 4: CIGSCR IS classification using ten initial clusters and Euclidean distance squared.

are similar, CIGSCR provides more information through soft classification. The soft classifications in Figs. 2 and 3 compared to the hard classification (shown in [3]) show that even when the hard clustering/classification in IGSCR and the soft clustering/classification in CIGSCR identify the same general regions as likely to be forest or likely to be nonforest, the soft classifications in Figs. 2 and 3 provide extra information relating to how strongly a particular sample is forest or nonforest. The dark green and dark brown colors indicate a high probability of forest and nonforest, respectively. Lighter shades of both colors indicate lower probabilities of membership in respective classes, and the beige regions indicate that the probabilities of that region being forest or nonforest are almost equal. The classifications in Figs. 7 and 8a–d show that in addition to providing more information, CIGSCR can produce qualitatively different classifications than IGSCR. The classifications present in Figs. 7 and 8a–d are the IS classifications that result from clustering, showing that soft clustering in CIGSCR produces different clustering and classification than the hard clustering in IGSCR. The regions identified by CIGSCR as being likely to contain individual pine species are different from the regions identified by IGSCR, although both algorithms identified similar nonpine regions.

Based on accuracies reported in Tables 1 and 2, CIGSCR is less sensitive to the number of initial clusters than IGSCR, especially when the alternative radial functions are used. As shown in Tables 1 and 2, IGSCR can be sen-

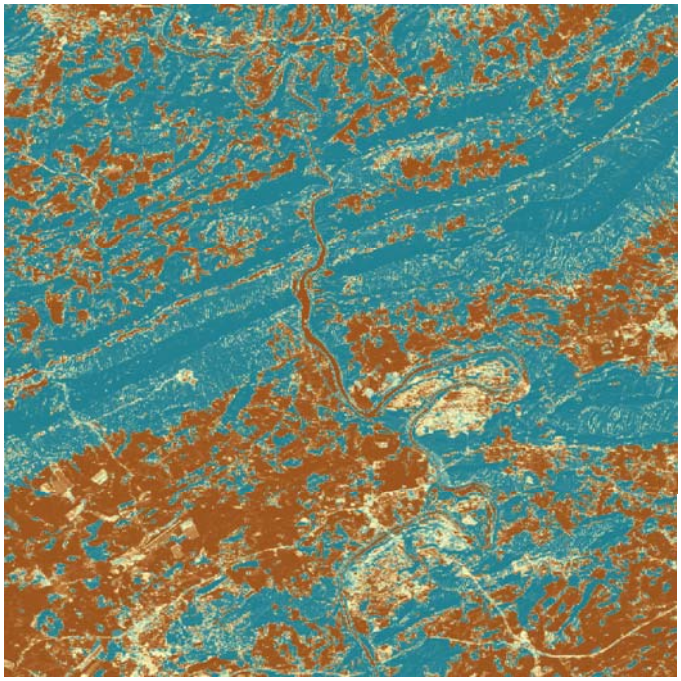

Figure 5: CIGSCR IS classification using ten initial clusters and Euclidean distance to the fourth power.

sitive to the number of initial clusters and the homogeneity threshold. 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 2 occur when a small number of iterations occurs, which can be greatly influenced by the number of initial clusters and the homogeneity test. The classification accuracies reported for CIGSCR in Tables 1 and 2 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 evaluates the observed values against a user supplied probability of observing a specific class (within a cluster), but the association significance test determines if the average cluster memberships per class are statistically significantly different (requiring no user specified probability). Secondly, 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

Table 1
IGSCR and CIGSCR decision rule (DR) classification accuracies for VA1734.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$), different ρ			clustering
clusters	$p = .5$	$p = .9$	$\|x-U\|_2^2$	$\|x-U\|_2^4$	$e^{\|x-U\|_2}$	(no iteration)
10	85.81	75.49	88.74	87.07	87.70	72.26
15	88.22	74.56	80.50	88.53	86.97	73.72
20	84.78	89.57	79.87	89.68	88.74	76.54
25	87.49	84.25	81.44	89.47	88.74	77.58

Table 2
IGSCR iterative stacked plus (IS+) and CIGSCR iterative stacked (IS) classification accuracies for VA1734.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$), different ρ			clustering
clusters	$p = .5$	$p = .9$	$\|x-U\|_2^2$	$\|x-U\|_2^4$	$e^{\|x-U\|_2}$	(no iteration)
10	68.30	75.39	83.63	84.67	85.09	72.26
15	86.34	74.56	76.96	86.03	85.19	72.99
20	84.46	88.95	75.60	85.40	86.86	76.85
25	66.63	83.94	78.52	88.32	87.28	76.75

(especially 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 is a desirable property.

The CIGSCR classifications shown in Figs. 2–5 experimentally validate the discussion earlier that radial functions magnify the difference between the largest and smallest cluster weights and will more closely approximate hard clustering. The classifications based on clustering with Euclidean distance to the fourth power have significantly fewer samples with almost equal probabilities of being in either class (corresponding to the beige color in the classification images). The classifications based on clustering with an exponential function of Euclidean distance (not pictured) are even closer to hard classification. Some beige areas remain in Figs. 3 and 5, indicating that although classifications based on these functions become more like hard classifications, in practice these classifications retain desirable properties of soft classification. Based on accuracies reported in Tables 1 and 2, these CIGSCR classifications with alternative radial functions are often the most accurate classifications for a given number of initial clusters. CIGSCR with alternative radial functions is accurate, can approximate hard classification when hard classification is desired, still provides more information than strict hard classification, and is less sensitive to input parameters than IGSCR.

All classification methods can be expected to perform poorly when training data are insufficient (samples within the dataset are not represented in the training set). This is especially true in IGSCR where spectrally pure hard clusters containing multiple training samples must be located in order for samples

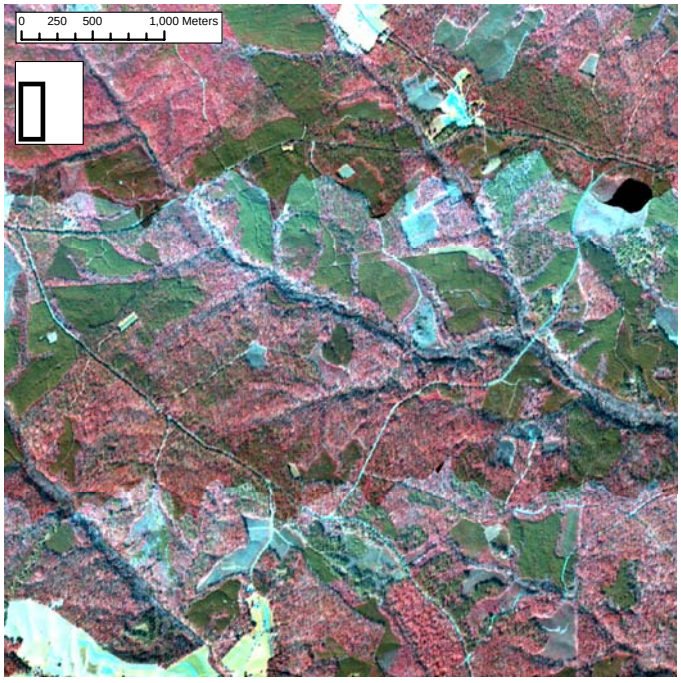

Figure 6: AVIRIS image (three flight lines) taken over Appomattox Buckingham State Forest in Virginia, USA.

to be labeled with that particular class. In the VA1734 dataset, an example of a spectral class with insufficient training data is water, and although water is technically nonforest, water is often classified as forest because water and forest are spectrally similar in certain wavelength regions. This is the case in Fig. 4 where the New River running vertically through the zoomed area of interest has been identified as forest. In the IGSCR IS classification image in [3], this region is "unclassified" meaning that these pixels are not part of a pure cluster as expected (few or no water training samples are identified for this image/training dataset). Another misclassification occurs as a result of shadows in the forested mountains running diagonally in the upper half on the zoomed image (Figs. 4 and 5). The IGSCR IS classification image in [3] indicates a likelihood that there is insufficient training data for these regions. Ultimately these water and shadow regions are misclassified using the decision rule in IGSCR (not pictured), and these regions are classified incorrectly using CIGSCR with Euclidean distance squared. However, notice in Fig. 5 that the CIGSCR IS using Euclidean distance to the fourth power correctly classified the river and the shadow regions. With soft clustering, different clusters were formed, allowing these features to potentially be correctly placed in similar clusters, even though these clusters likely contained small percentages of the training data. In this case, it is potentially useful to know that these features are unclassified (in IGSCR) allowing for modification of the training data, and

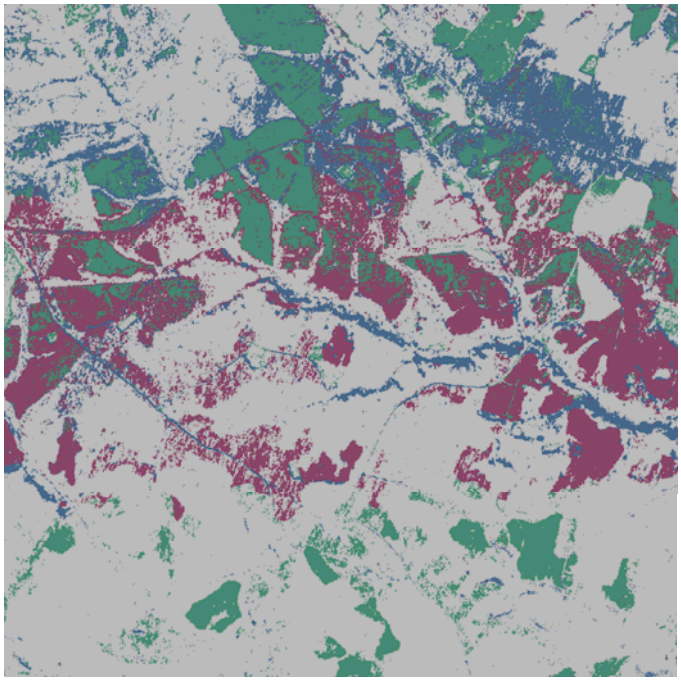

Figure 7: IGSCR IS classification of ABSF.

unfortunately CIGSCR does not have this capability. However, when more training samples are not available, CIGSCR can potentially provide a better estimate of the correct class for these data that are not well represented in the training data (although this is obviously not guaranteed as CIGSCR using two different distance functions produced different classification results). Also of interest is that the uncertainty in the soft classifications (regions in beige) does not necessarily match the unclassified regions in the IGSCR IS classification image in [3]. There does not appear to be a correlation between samples that are not part of pure clusters in IGSCR and samples that may belong to multiple classes in CIGSCR.

The accuracies reported for the classification of ABSF tend to be lower than the classification accuracies reported for VA1734, which is reasonable considering the classification of ABSF is attempting to discriminate between spectrally similar pine species, ABSF is noisy, and ABSF contains several heterogeneous areas, making training difficult. Also note that the VA1734 DR classifications were almost always more accurate than corresponding IS classifications, but ABSF DR classifications are often less accurate than corresponding IS classifications. All ABSF classifications (IGSCR DR and IS and CIGSCR DR and IS) reasonably separated pines from nonpines, but IGSCR and CIGSCR differed in the identification of individual pines species. Both classification methods identified individual pines in mixed hardwood/pine stands in the top left corner of the image (Figs. 7 and 8a–d). A visual inspection of the classification

Table 3
IGSCR and CIGSCR decision rule (DR) classification accuracies for ABSF.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$), different ρ			clustering
clusters	$p = .5$	$p = .9$	$\|x-U\|_2^2$	$\|x-U\|_2^4$	$e^{\|x-U\|_2}$	(no iteration)
10	83.50	*	47.50	79.50	72.50	*
15	*	*	62.50	83.50	79.75	*
20	*	*	66.75	73.50	74.25	*
25	51.00	51.00	63.00	75.00	78.75	*

Table 4
IGSCR iterative stacked plus (IS+) and CIGSCR iterative stacked (IS) classification accuracies for ABSF.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$), different ρ			clustering
clusters	$p = .5$	$p = .9$	$\|x-U\|_2^2$	$\|x-U\|_2^4$	$e^{\|x-U\|_2}$	(no iteration)
10	83.75	*	51.75	84.50	72.75	*
15	*	*	51.00	84.50	83.25	*
20	*	*	51.00	84.00	81.50	*
25	91.00	75.25	51.00	76.75	83.00	*

images reveals that IGSCR and CIGSCR classifications disagree on loblolly (IGSCR has underestimated those stands) and shortleaf (both overestimated). IGSCR incorrectly picked out patches of shortleaf along the "veins" of the image, and both classifications overestimated Virginia pines.

Another potential advantage of CIGSCR with an alternative radial function is the ability to locate clusters associated with classes, even when there is overlap between classes or there is a small amount of training data for a class. IGSCR failed to locate enough pure clusters to perform classification, indicated by an asterisk in Tables 3 and 4, in most ABSF classification attempts. CIGSCR using Euclidean distance squared produced classifications, although the accuracies are low. CIGSCR using alternative radial functions performed reasonable classifications no matter the number of initial clusters. In highly heterogeneous sites like this where limited training data is available for multiple classes, IGSCR has difficulty locating pure clusters. Since multiple classes are spectrally similar, soft clustering allows for small differences between classes in a cluster to be detected. Hard clusters containing one species would be likely to contain a significant amount of the other species, and would therefore fail the hypothesis test (for reasonable p and α. With soft clustering, portions of both species would be attributed to a soft cluster, but if there is statistical significance of the difference in the memberships of the species, the cluster can be associated and used for training purposes. Furthermore, soft clustering allows for alternative functions to be used to determine cluster assignments. Recall that these radial functions magnify the difference between small and large probabilities, allowing clusters containing these less well represented classes to be formed and allowing samples to have high probabilities

Table 5
For VA1734 IGSCR, number of pure clusters. For VA1734 CIGSCR, the pairs (a,b) = (number of clusters produced, number of associated clusters).

no. init.	IGSCR		CIGSCR		
clusters	$p = .5$	$p = .9$	$\rho = \|x - U\|_2^2$	$\rho = \|x - U\|_2^4$	$\rho = e^{\|x-U\|_2}$
10	19	6	15,13	11,11	12,12
15	15	6	20,16	20,19	20,20
20	20	18	25,21	21,21	24,24
25	52	17	30,25	30,28	30,29

Table 6
For ABSF IGSCR, number of pure clusters. For ABSF CIGSCR, the pairs (a,b) = (number of clusters produced, number of associated clusters).

no. init.	IGSCR		CIGSCR		
clusters	$p = .5$	$p = .9$	$\rho = \|x - U\|_2^2$	$\rho = \|x - U\|_2^4$	$\rho = e^{\|x-U\|_2}$
10	16	8	15,15	10,10	11,11
15	14	11	20,19	15,15	15,15
20	19	9	25,24	20,20	20,20
25	23	15	30,29	25,25	26,26

of belonging to those clusters.

Finally, perhaps the most important question about this semisupervised clustering scheme is whether using the combination of the association significance test and the iteration improves the clustering for the purposes of classification. Each cluster is labeled with the class that has the highest average membership in the cluster. Observe in experimental runs in Tables 1 and 2 that **all** classification accuracies using just clustering are lower than corresponding classification accuracies using CIGSCR with Euclidean distance. In Tables 3 and 4, iterative refinement was necessary to locate enough clusters (such that each class was represented by at least one cluster) for classification using Euclidean distance squared. Accuracies are much higher using alternative distance functions, but little or no iterative refinement was used. Based on the available results in Tables 1–4, the semisupervised clustering scheme in CIGSCR improves classification accuracies when training data are available to influence clustering.

Conclusions

This paper presented a continuous analog to IGSCR that rejects and refines clusters to automatically classify a remotely sensed image based on informational class training data. This new algorithm addressed specific challenges presented by remotely sensed data including large datasets (millions of samples), relatively small training datasets, 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, ability to use alternative distance functions that often produce more accurate classifications, potential to correctly classify regions that are not amply represented in training data, a better ability to locate clusters associated with all classes, and the ability to efficiently exploit parallel and GPU computer hardware. 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 radial functions used in CIGSCR resulted in consistently accurate classifications.

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

References

[1] J. C. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2 (1980), 1–8.

[2] 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, Photogrammetric Engineering & Remote Sensing, 72 (2006), 949–960.

[3] R. D. Phillips, A probabilistic classification algorithm with soft classification output, Ph.D. Thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2009.

[4] R. D. Phillips, L. T. Watson, and R. H. Wynne, Hybrid image classification and parameter selection using a shared memory parallel algorithm, Computers & Geosciences, 33 (2007), 875–897.

[5] J. A. N. Van Aardt and R. H. Wynne, Examining pine spectral separability using hyperspectral data from an airborne sensor: an extension of field-based results, International Journal of Remote Sensing, 28 (2007), 431–436.

[6] 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, Photogrammetric Engineering & Remote Sensing, 67 (2001), 1155–1166.

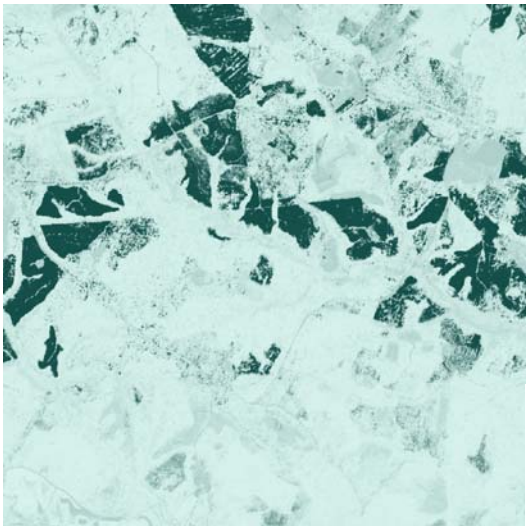

Fig. 8a. CIGSCR IS classification (loblolly pines).

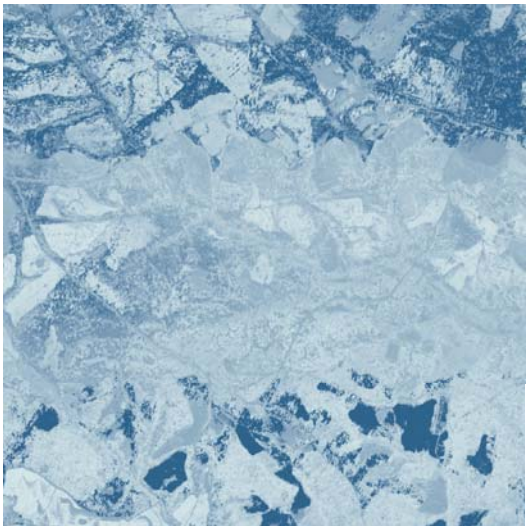

Fig. 8b. CIGSCR IS classification (shortleaf pines).

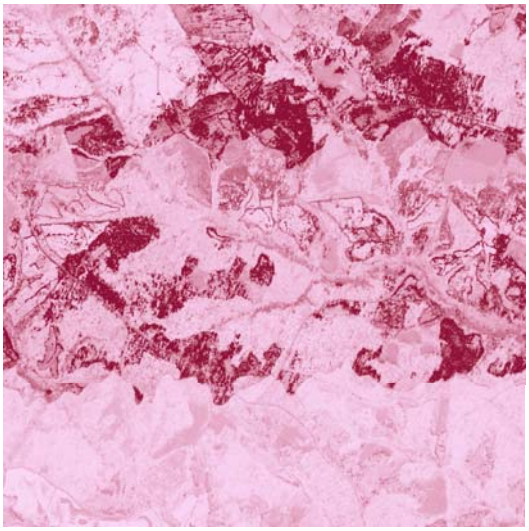

Fig. 8c. CIGSCR IS classification (Virginia pines).

Fig. 8d. CIGSCR IS classification (nonpine).

Received: January 19, 2014