# Land Cover Classification Using Fuzzy Rules and Aggregation of Contextual Information Through Evidence Theory

Arijit Laha, *Member, IEEE*, Nikhil R. Pal, *Fellow, IEEE*, and Jyotirmoy Das

***Abstract*—Land cover classification using multispectral satellite images is a very challenging task with numerous practical applications. We propose a multistage classifier that involves fuzzy rule extraction from the training data and then the generation of a possibilistic label vector for each pixel using the fuzzy rule base. To exploit the spatial correlation of land cover types, we propose four different information aggregation methods which use the possibilistic class label of a pixel and those of its eight spatial neighbors for making the final classification decision. Three of the aggregation methods use the Dempster–Shafer theory of evidence, while the remaining one is modeled after the fuzzy k-NN rule. The proposed methods are tested with two benchmark seven-channel satellite images, and the results are found to be quite satisfactory. They are also compared with a Markov random field model-based contextual classification method and found to perform consistently better.**

***Index Terms*—Classifier, evidence theory, fuzzy k-NN, fuzzy rules, rule extraction.**

## I. Introduction

LAND cover classification in remotely sensed images is considered to be a cost effective and reliable method for generating up-to-date land cover information [1]. Usually, such images are captured by multispectral scanners (such as Landsat TM) that acquire data at several distinct spectral bands producing multispectral images. Automated analysis of such data calls for sophisticated techniques for data fusion and pattern recognition. The most widely used techniques include statistical modeling involving discriminant analysis and maximum likelihood classification and artificial neural networks-based approaches [2]–[5]. However, these classifiers usually classify a sample to the class for which maximum support is obtained, no matter how small this amount of support may be or if there may be another class for which the support is very close to the maximum. This particular feature is often criticized [1] in context of their applicability in land cover classification. "Soft" classifiers can be useful for such problems as they can produce a measure of confidence in support of the decision as well as indicate measures of confidence in support of alternative decisions, which can be used for further processing using auxiliary information. This can result in a more robust and accurate system.

Manuscript received June 4, 2004; revised September 15, 2005.

A. Laha is with the Institute for Development and Research in Banking Technology, Castle Hills, Hyderabad 500 057, India (e-mail: alaha@idrbt.ac.in).

N. R. Pal and J. Das are with the Electronics and Communication Science Unit, Indian Statistical Institute, Calcutta 700 108, India (e-mail: nikhil@isical.ac.in; jdas@isical.ac.in).

Digital Object Identifier 10.1109/TGRS.2006.864391

In developing soft classifiers for land cover analysis, two approaches have gained popularity. These are based on 1) the fuzzy set theory [7] and 2) Dempster and Shafer's evidence theory [6]. There is, of course, the probabilistic approach that we do not pursue further.

Different fuzzy methodologies for land cover classification in multispectral satellite images have been investigated by various researchers. For example, [1] uses the fuzzy c-means algorithm, while Kumar *et al.* [9] applied a fuzzy integral method. Fuzzy rule-based systems have been used for classification by many researchers [7], [8] for diverse fields of application. Fuzzy rules are attractive because they are interpretable and can provide an analyst a deeper insight into the problem. Use of fuzzy rule-based systems for land cover analysis is a relatively new approach. Recently, Bárdossy and Samaniego [10] have proposed a scheme for developing a fuzzy rule-based classifier for analysis of multispectral images, where the randomly generated initial rules are fine tuned by simulated annealing. In [11], Kulkarni and McCaslin used fuzzy neural networks for rule extraction.

The other approach to design soft classifiers use the evidence theory developed by Dempster and Shafer [6], [12], [13]. As observed by Lee *et al.* in [14], for multispectral image analysis there may be a great incentive for applying Dempster–Shafer theory of evidence. Since the theory of evidence allows one to combine evidences obtained from diverse sources of information in support of a hypothesis, it seems a natural candidate for analyzing multispectral images for land cover classification [15]–[17]. In all of these works, the approach is to treat each channel image as a separate source of information. Each image is analyzed to associate each pixel with some degree of belief pertaining to its belonging to each member of a set of hypotheses known as the *frame of discernment*. Usually, some probabilistic techniques are employed to assign the degree of belief. In the next stage, these belief values from all images for a pixel are combined using Dempster's rule [6] to calculate the total support for each hypothesis. In a recent paper [18], Jouan and Allard used evidence theory for combining information from multiple sources for land use mapping.

In a satellite image, usually the landcover classes form spatial clusters, i.e., a pixel belonging to a particular class is more likely to have neighboring pixels from the same class rather than from other classes. Thus, the inclusion of contextual information from the neighboring pixels is likely to increase classification accuracy. An overview of common contextual pattern recognition methods can be found in [19]. In this paper, we propose several schemes for classifier design that uses both fuzzy

sets theory and Dempster–Shafer evidence theory. These are two stage schemes. In the first stage, a fuzzy rule-based classifier is developed using a small training set. This classifier is noncontextual and for each pixel it generates a possibilistic label vector. In the next stage, we aggregate the responses of the fuzzy rules over a $3 \times 3$ *spatial* neighborhood of a pixel to make the classification decision about that pixel. Thus, the decision making process takes into consideration the information available from the spatial neighborhood of the pixel. Here, we propose four methods for contextual decision making. In the experimental results, we compare the performance of the contextual classifiers with the fuzzy rule-based noncontextual classifiers designed in the first stage. We also provide a comparison of the performance of the proposed classification schemes with a recent Markov random field (MRF) model-based contextual classification scheme proposed by Sarkar *et al.* [20].

## II. Designing the Fuzzy Rule Base

We use a multistage scheme for designing fuzzy rule-based classifiers. The set of gray values corresponding to a pixel in the channel images is used as the feature vector for that pixel. In the first stage, a set of labeled prototypes representing the distribution of the training data is generated using a Self-organizing Feature Map (SOFM) [23] based algorithm developed in [24]. The algorithm dynamically decides the number of prototypes based on the training data. Then, each of these prototypes is converted to a fuzzy rule.

Next, the fuzzy rules are tuned by modifying the position of the peaks as well as the spreads of the fuzzy sets associated with the rules. The tuned rules can readily be used to classify unknown samples based on the firing strengths of the rules. Typically, a test sample is classified to the class of the rule generating the highest firing strength. For the sake of completeness, we first give a brief description of SOFM.

### A. Kohonen's SOFM Algorithm

SOFM is formed of neurons placed on a regular (usually) one-dimensional (1-D) or two-dimensional grid. Thus, each neuron is identified with a index corresponding to its position in the grid (the viewing plane). Each neuron $i$ is represented by a weight vector $\mathbf{w}_i \in \Re^p$ where $p$ is the dimensionality of the input space. In the $t$-th training step, a data point $\mathbf{x} \subset \Re^p$ is presented to the network. The winner node with index $r$ is selected as $r = \underbrace{\arg\min}_{i}\{\|\mathbf{x} - \mathbf{w}_{i,t-1}\|\}$. $\mathbf{w}_{r,t-1}$ and the other weight vectors associated with cells in the spatial neighborhood $N_r(r)$ of $r$ are updated using the rule

$$\mathbf{w}_{i,t} = \mathbf{w}_{i,t-1} - \alpha(t) h_{ri}(t)(\mathbf{x} - \mathbf{w}_{i,t-1})$$

where $\alpha(t)$ is the learning rate and $h_{ri}(t)$ is the neighborhood kernel (usually Gaussian). The learning rate and the radius of the neighborhood kernel decrease monotonically with time. During the iterative training the SOFM behaves like a flexible net that folds onto the "cloud" formed by the input data. A trained SOFM exhibits remarkable and useful properties of *topology preservation* and *density matching* and often used for visualization of metric-topological relationships and distributional density properties of training data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\Re^p$ through their mapping onto the viewing plane. From a clustering viewpoint, SOFM has the advantage of avoiding underutilization of the prototypes.

### B. Generation of Prototypes

First, we train a 1-D SOFM with $c$ nodes, where $c$ is the number of classes. We do so because the smallest number of rules that may be required is equal to the number of classes. At the end of the training, the weight vector distribution of the SOFM reflects the distribution of the input data. Then the training data is divided into a $c$ partition according to their closeness to $c$ weight vectors $V^0 = \{\mathbf{v}_1^0, \mathbf{v}_2^0, \ldots, \mathbf{v}_c^0\} \subset \Re^p$. Each of the prototypes is labeled based on the class information of corresponding partition using *majority voting*. However, such a set of prototypes may not classify the data satisfactorily because the class information is not used in the training resulting in the possibility that a prototype may represent data from more than one class significantly.

We use the prototype refinement scheme described in [24]. The basic idea behind this refinement algorithm is that a useful prototype should satisfy two criteria: 1) it should represent an adequate number of points and 2) only one of the classes should be strongly represented by it.

The prototype refinement scheme applies four operations, *deletion*, *modification*, *merging*, and *splitting* on the set of prototypes while trying to fulfill the above conditions. In the training data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, let there be $N_j$ points from class $j$. The refinement stage uses just *two* parameters $K_1$ and $K_2$ to dynamically generate $c + 1$ retention thresholds known as a global retention threshold $\alpha$ and a set of class-wise retention threshold $\beta_k$ (one for each class) to evaluate the performance of each prototype. $\alpha$ and $\beta_k$ are computed dynamically (not fixed) for the $t$-th iteration using the following formulae:

$$\alpha^t = \left(K_1 \left| V^{t-1} \right|\right)^{-1} \text{ and } \beta_k^t = \left(K_2 \left| V_k^{t-1} \right|\right)^{-1}$$

where $V_k^{t-1} = \{\mathbf{v}_i \mid \mathbf{v}_i \in V^{t-1}, C_i = k\}$.

Here, $V^{t-1}$ is the set of prototypes obtained after $(t-1)$ iterations of the algorithm. To consider a prototype useful, it must represent more than $\alpha^t N$ training points. Further, a prototype must represent more than $\beta_k^t N_k$ points from class $k$ to be considered a (potential) prototype for class $k$.

Finally, the set of prototypes is again refined by SOFM algorithm with winner-only update strategy. After a few iterations, this algorithm produces a set of adequate number of prototypes that represents the training data much better than the initial one. For details, the readers are referred to [24]. The final set of prototypes $V^{\text{final}} = \{\mathbf{v}_1^{\text{final}}, \ldots, \mathbf{v}_{\hat{c}}^{\text{final}}\}$, where $\hat{c} \geq c$, is used to generate the set of initial fuzzy rules.

### C. Designing Fuzzy Rule Base

A prototype $\mathbf{v}_i$ represents a cluster of points for class $k$. This cluster can be described by a fuzzy rule of the form: $R_i$: If $\mathbf{x}$ is CLOSE TO $\mathbf{v}_i$ then class is $k$. This rule can be further translated into

$R_i$: If $x_1$ is CLOSE TO $v_{i1}$ AND
$\cdots$ AND $x_p$ is CLOSE TO $v_{ip}$ then class is $k$.

Note that this is just one possible interpretation of "$\mathbf{x}$ is CLOSE TO $\mathbf{v}_i$." The first form requires a multidimensional membership function, while the second form requires several 1-D membership functions and a conjunction operator. In general, the two forms will not produce the same output. Depending on the choice of the membership function and the conjunction operator, the forms may lead to the same output, but since the representation by the atomic clauses is a plausible realization of the multidimensional form, the performance of the systems built using either form is not going to be much different.

The fuzzy set CLOSE TO $v_{ij}$ can be modeled by triangular, trapezoidal, or Gaussian membership function. In this investigation, we use the Gaussian membership function

$$\mu_{ij}(x_j; v_{ij}, \sigma_{ij}) = \exp\left(-\frac{(x_j - v_{ij})^2}{\sigma_{ij}^2}\right).$$

Given a data point $\mathbf{x}$ with unknown class, we first find the firing strength of each rule. Let $\alpha_i(\mathbf{x})$ denote the firing strength of the $i$-th rule on a data point $\mathbf{x}$. We assign the point $\mathbf{x}$ to class $k$ if $\alpha_r = \max_i(\alpha_i(\mathbf{x}))$ and the $r$-th rule represents class $k$.

The performance of the classifier depends crucially on the adequacy of the number of rules used and proper choice of the membership functions. In our case, each fuzzy set is characterized by two parameters $v_{ij}$ and $\sigma_{ij}$. Let the initial set of fuzzy rules be $R^0 = \{R_i^0 \mid i = 1, 2, \ldots, c\}$. The parameters $v_{ij}^0$ and $\sigma_{ij}^0$ for fuzzy sets in the antecedent part of a rule $R_i^0 \in R^0$ are obtained from the prototype $\mathbf{v}_i^{\text{final}} \in V^{\text{final}}$ as follows:

$$v_{ij}^0 = v_{ij}^{\text{final}} \text{ and } \sigma_{ij}^0 = \frac{k_w\left(\sqrt{\sum_{\mathbf{x}_k \in X_i}(x_{kj} - v_{ij}^{\text{final}})^2}\right)}{|X_i|}$$

where $X_i$ is the set of training data closest to $\mathbf{v}_i^{\text{final}}$ and $k_w > 0$ is a constant parameter that controls the initial width of the membership function. If $k_w$ is small, then the specificity of the fuzzy sets defining the rules will be high, and, hence, each rule will model a small area in the input space. On the other hand, a high $k_w$ will make each rule cover a bigger area. Since the spreads are tuned, in principle, $k_w$ should not have much impact on the final performance, but, in practice, the value of $k_w$ may have a significant impact on the classification performance for complicated data sets because of the local minimum problem of gradient descent techniques. In the current work, the values of $k_w$s are found experimentally. One can use a validation set for this.

The initial rule base $R^0$ thus obtained can be further fine tuned to achieve better performance, but the exact tuning algorithm depends on the conjunction operator (implementing AND operation for the antecedent part) used for computation of the firing strengths. The firing strength can be calculated using any T-norm [7]. The use of different T-norms results in different classifiers. The **product** and the **minimum** are the most popular choices of T-norms.

Though it is much easier to formulate a calculus-based tuning algorithm if the product is used, its use is conceptually somewhat unattractive. To illustrate the point, let us consider a rule having two atomic clauses in its antecedent. If the two clauses have truth values $a$ and $b$, then, intuitively, the antecedent is satisfied at least to the extent of $\min(a, b)$. However, if the product is used as the conjunction operator, we always have $ab \leq \min(a, b)$. Thus, we always underdetermine the importance of the rule. This does not cause any problem for nonclassifier fuzzy systems because the defuzzification operator usually performs some kind of normalization with respect to the firing strength, but in classifier-type applications, a decision may appear to be made with a very low confidence, when it is actually not the case. In certain cases, such a situation may lead to overemphasis on total ignorance under evidence theory framework. Thus, to avoid the use of the product and, at the same time, to be able to apply calculus to derive update rules, we use a soft-min operator.

The **soft-match** of $n$ positive numbers $x_1, x_2, \ldots, x_n$ is defined by

$$SM(x_1, x_2, \ldots, x_n, q) = \left\{\frac{(x_1^q + x_2^q + \cdots + x_n^q)}{n}\right\}^{1/q}$$

where $q$ is any real number. $SM$ is known as an aggregation operator with upper bound of value 1 when $x_i \in [0, 1] \forall i$. This operator is used by different authors [25], [26] for different purposes. It is easy to see that as $q \to -\infty$, the soft-match operator behaves like the "min" operator. Thus, we define the softmin operator as the soft match operator with a sufficiently negative value of the parameter $q$. The firing strength of the $r$th rule computed using softmin is

$$\alpha_r(\mathbf{x}) = \left\{\frac{\sum_{j=1}^{j=p}(\mu_{rj}(x_j; v_{rj}, \sigma_{rj}))^q}{p}\right\}^{1/q}.$$

At the value $q = -10$, its behavior resembles very closely the min operator, so we use $q = -10.0$. A very low value of $q$ may lead to numerical instability in the computation.

### D. Tuning of the Rule Base

Let $\mathbf{x} \in X$ be from class $c$ and $R_c$ be the rule from class $c$ giving the maximum firing strength $\alpha_c$ for $\mathbf{x}$. Also, let $R_{\neg c}$ be the rule from the incorrect classes having the highest firing strength $\alpha_{\neg c}$ for $\mathbf{x}$. We use the error function $E$

$$E = \sum_{\mathbf{x} \in X}(1 - \alpha_c + \alpha_{\neg c})^2. \quad (1)$$

This kind of error function has been used in [29]. We minimize $E$ with respect to $v_{cj}$, $v_{\neg cj}$ and $\sigma_{cj}$, $\sigma_{\neg cj}$ of the two rules $R_c$ and $R_{\neg c}$ using gradient decent. Here, the index $j$ corresponds to clause number in the corresponding rule. The tuning process is repeated until the rate of decrement in $E$ becomes negligible resulting in the rule base $R^{\text{final}}$.

Since a Gaussian membership function is extended to infinity, for any data point, all rules will be fired to some extent. In our implementation, if the firing strength is less than a threshold $\epsilon$ ($\simeq 0.01$), then the rule is not assumed to be fired. The threshold is set considering approximate $2\sigma$ limit of the Gaussian membership functions. Thus, under this situation, the rule base extracted by the system may not be complete with respect to the training data. This can happen even when we use membership functions with triangular or trapezoidal shapes. This is not a limitation but a distinct advantage, although, for the data sets we

used, we did not encounter such a situation. If no rule is fired by a data point, then that point can be thought of as an outlier. If such a situation occurs for some test data, then that will indicate an observation not close enough to the training data, and, consequently, no conclusion should be made about such test points.

## III. Decision Making With Aggregation of Spatial Contextual Information

In this section, we describe four decision making schemes. The first one involves a simple aggregation of the fuzzy label vectors of the neighboring pixels. The scheme is modeled after the well known fuzzy k-nearest neighbor algorithm proposed by Keller *et al.* [27]. The other three methods use the Dempster–Shafer theory of evidence. For the sake of completeness, we give a brief introduction of Dempster–Shafer theory of evidence before we describe the decision making methods in detail.

### A. Dempster–Shafer Theory of Evidence

Let $\Theta$ be the universal set and $P(\Theta)$ be its power set. A function $m : P(\Theta) \rightarrow [0, 1]$ is called a basic probability assignment (BPA) whenever $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. Here, $m(A)$ is interpreted as the degree of evidence supporting the claim that the "truth" is in $A$ and in absence of further evidence no more specific statement can be made. Every set $A \subset P(\Theta)$ for which $m(A) > 0$ is called a *focal element* of $m$. Evidence obtained in the same context from two distinct sources and expressed by two BPAs $m^1$ and $m^2$ on the same power set $P(\Theta)$ can be combined by Dempster's rule of combination to obtain a joint BPA $m^{1,2}$ as

$$m^{1,2}(A) = \begin{cases} \dfrac{\sum_{B \cap C = A} m^1(B) m^2(C)}{1 - K}, & \text{if } A \neq \emptyset \\ 0, & \text{if } A = \emptyset. \end{cases} \tag{2}$$

Here, $K = \sum_{B \cap C = \emptyset} m^1(B) m^2(C)$.

Equation (2) is often expressed with the notation $m^{1,2} = m^1 \oplus m^2$. The rule is commutative and associative. Evidence from any number (say $k$) of distinct sources can be combined by repetitive application of the rule as $m = m^1 \oplus m^2 \oplus \cdots \oplus m^k = \bigoplus_{i=1}^{k} m^i$.

To select the optimal decision from the evidence embodied in a BPA, typically, we construct a probability function $P^\Theta$ on $\Theta$. This is done through a transformation known as *pignistic transformation* [28]. The pignistic probability for $\theta \in \Theta$ can be expressed in terms of BPAs as follows:

$$P^\Theta(\theta) = \sum_{A \subseteq \Theta, \theta \in A} \frac{m(A)}{|A|}. \tag{3}$$

Optimal decision can now be chosen in favor of $\theta_0$ with the highest pignistic probability.

### B. Aggregation of Context Information for Decision Making

Let there be $c$ classes, $\mathcal{C} = \{C_1, C_2, \ldots C_c\}$. For each pixel $P$, the fuzzy rule base generates a fuzzy label vector $\boldsymbol{\alpha} \in \Re^c$ such that the value of the $k$th component of $\boldsymbol{\alpha}$, $\alpha_k$ ($0 \leq \alpha_k \leq 1$) represents the confidence of the rule base supporting the fact that the pixel $P$ belongs to class $C_k$. Strictly speaking, $\boldsymbol{\alpha} \in \Re^c$ is a possibilistic label vector [7]. The value of $\alpha_k$ is computed as follows.

Let $r$ be the number of rules in the fuzzy rule base. Since $c < r$, there could be multiple rules corresponding to a class. Then, $\alpha_k$ is the highest firing strength produced by the rules corresponding to the class $C_k$ for pixel $P$. We treat this value as the confidence measure of the rule base pertaining to the membership of pixel $P$ to the class $C_k$. However, if $\alpha_k$ is less than a threshold, say 0.01, it is set to 0.

In our decision-making schemes, we consider a pixel together with the pixels within its $3 \times 3$ spatial neighborhood. We identify the pixel $p$ under consideration as $p^0$ and its eight neighbors as $p^1, p^2, \ldots p^8$. The corresponding possibilistic label vectors assigned to these nine pixels are denoted as $\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^1, \ldots \boldsymbol{\alpha}^8$.

*1) Method 1: Aggregation of Possibilistic Labels:* This is a simple aggregation scheme modeled after the fuzzy k-NN method of Keller *et al.* [27]. Given a set of nine label vectors $\{\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^1, \ldots \boldsymbol{\alpha}^8\}$, an aggregated possibilistic label vector $\boldsymbol{\alpha}^A$ is computed as follows:

$$\boldsymbol{\alpha}^A = \frac{\sum_{i=0}^{8} \boldsymbol{\alpha}^i}{9}. \tag{4}$$

The pixel $p^0$ is assigned to class $C_k$, such that $\alpha_k^A \geq \alpha_i^A \forall i = 1, 2, \ldots c$. Note that, though the label vector $\boldsymbol{\alpha}^0$ corresponds to the pixel to be classified, no special emphasis (importance) is given to it in this aggregation scheme.

*2) Method 2: Aggregation With Bayesian Belief Modeling:* This aggregation method as well as the next two use the evidence theoretic framework of Dempster–Shafer. Within this framework, the set of classes $\mathcal{C}$ is identified as the *frame of discernment*. A body of evidence is represented by a BPA $m$ over subsets of $\mathcal{C}$. The value $m(A)$ denotes the belief in support of the proposition that *the true class label of the pixel of interest is in* $A \subset \mathcal{C}$. In the context of our problem, we shall: 1) identify distinct bodies of evidence (BOE); 2) formulate a realistic scheme of assigning the BPAs to the relevant subsets of $\mathcal{C}$; 3) combine evidences provided by all BOEs; 4) compute the pignistic probability for each class from the combined evidence; and 5) make a decision using the maximum pignistic probability principle.

In this scheme, we identify eight BOEs for eight neighbor pixels with corresponding BPAs denoted as $m^1, m^2, \ldots, m^8$. We consider the Bayesian belief structure, i.e., each focal element has only one element. Assigning BPA to a subset essentially involves committing some portion of belief in favor of the proposition represented by the subset. So, the scheme followed for assigning BPAs must reflect some realistic assessment of the information available in favor of the proposition. We define $m^i$ as follows:

$$m^i(\{C_k\}) = \frac{(\alpha_k^i + \alpha_k^0)}{S}, k = 1, 2, \ldots, c \tag{5}$$

where $S = \sum_{k=1}^{k=c} (\alpha_k^i + \alpha_k^0)$ is a normalizing factor. Thus, each BPA contains $c$ focal elements, one corresponding to each class, and the assigned value $m^i(\{C_k\})$ is influenced by the magnitudes of firing strengths produced by the rule base in support of class $C_k$ for the pixel of interest $p^0$ and its $i$-th neighbor $p^i$. Clearly, the label vector $\boldsymbol{\alpha}^0$ influences all eight BPAs. Hence, it is expected that, in the final decision making, the influence of

$\boldsymbol{\alpha}^0$ will be higher than any neighbor. Such an assignment is motivated by the fact the spatial neighbors usually are highly correlated, i.e., pixel $p^0$ and its immediate neighbors are expected from the same class.

Thus, for eight neighboring pixels, we obtain eight separate BPAs. These BPAs are combined to get the global BPA applying the Dempster's rule repeatedly. It can be easily seen that the global BPA $m^G$ is also Bayesian and can be computed as

$$m^G(\{C_k\}) = \frac{\prod_{i=1}^{8} m^i(\{C_k\})}{\sum_{l=1}^{c} \prod_{i=1}^{8} m^i(\{C_l\})}. \quad (6)$$

In this case, the pignistic probability $P^G(C_k)$ is the same as $m^G(\{C_k\})$. So, the pixel $p^0$ is assigned to class $C_k$ such that $m^G(\{C_k\}) > m^G(\{C_l\}) \forall C_l \in \mathcal{C}$.

*3) Method 3: Aggregation With Non-Bayesian Belief:* Here, the BOEs are also identified in the same way as in the previous method. However, we allow the assignment of belief to subsets of $\mathcal{C}$ having two elements, i.e., the subsets $\{C_l, C_m : l < m$ and $l, m = 1, 2, \ldots, c\}$. We define $m^i$ as follows:

$$m^i(\{C_k\}) = \frac{(\alpha_k^i + \alpha_k^0)}{S}, k = 1, 2, \ldots, c \quad (7)$$

$$m^i(\{C_l, C_m : l < m\}) = \frac{(\alpha_l^i + \alpha_m^0) - (\alpha_m^i + \alpha_l^0)}{2S} \quad (8)$$

$l, m = 1, 2, \ldots, c$, where

$$S = \sum_{k=1}^{k=c} (\alpha_k^i + \alpha_k^0) + \sum_{l=1}^{l=c-1} \sum_{m=l+1}^{m=c} [(\alpha_l^i + \alpha_m^0) - (\alpha_m^i + \alpha_l^0)].$$

In Method 2, we exploited only the correlation of spatial neighbors. However, for satellite images, when a pixel falls at the boundary of some land cover type, it may correspond to more than one land cover type. Since the chance of a pixel to have representation from more than two land cover types is not high, we restrict the cardinality of the focal sets to two. Using (8) for eight neighboring pixels, we obtain eight separate bodies of evidence. The global BPA for the focal elements is then computed applying Dempster's rule repeatedly. In our previous method, we could use (6) to compute the global BPA since we dealt with singleton focal elements only. In this case, we have to compute the global BPA in steps, combining two bodies of evidence at a time and preparing an intermediate BPA which will be combined with another BPA, and so on. Once the global BPA is computed, the class label is assigned according to highest pignistic probability.

*4) Method 4: Aggregation Using Evidence Theoretic k-NN Rule:* This method is fashioned after the evidence theoretic k-NN method [12]. Here, nine BPAs $m^0, m^1, \ldots, m^8$ are identified corresponding to the possibilistic label vectors $\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^8$. The BPA $m^i$ is assigned as follows: Let $q = \underbrace{\arg\max}_{k}(\alpha_k^i)$, then

$$m^i(\{C_q\}) = \alpha_q^i \quad (9)$$

$$m^i(\mathcal{C}) = 1 - \alpha_q^i \quad (10)$$

$$m^i(A) = 0 \forall A \in P(\mathcal{C}) \backslash \{\mathcal{C}, \{C_q\}\}. \quad (11)$$

Thus, there is only one focal element containing one element in each BOE. The rest of the belief is assigned to the frame of discernment $\mathcal{C}$, which can be interpreted as the amount of ignorance. The evidences are then combined using Dempster's rule and the class label is assigned according to highest pignistic probability.

Like Method 1, in this method, no special emphasis is given to the pixel under consideration $p^0$ over the neighboring pixels, whereas in Methods 2 and 3, $\boldsymbol{\alpha}^0$ influences all the BPAs, i.e., $\boldsymbol{\alpha}^0$ plays a special role. It can be seen in Section IV that these two approaches provide different classification efficiencies depending on the nature of the spatial distribution of the classes. Intuitively, if pixels corresponding to different land cover types are scattered all over the image, neighboring pixels may not be given the same importance as that of the central pixel for optimal performance of the classifier. On the other hand, if pixels of a particular land cover type form spatial clusters, then giving equal importance to the neighboring pixels may be desirable. However, the optimal weight to be given to the neighboring pixels to get the best performance should depend on the distribution of different land cover types on the image being analyzed. To realize this, we modify the BPA assignment scheme of current method as follows. Let $q = \underbrace{\arg\max}_{k}(\alpha_k^i)$, then

$$m^0(\{C_q\}) = \alpha_q^0, \text{ for } i = 0 \quad (12)$$

$$m^i(\{C_q\}) = w\alpha_q^i, \text{ otherwise} \quad (13)$$

$$m^0(\mathcal{C}) = 1 - \alpha_q^0, \text{ for } i = 0 \quad (14)$$

$$m^i(\mathcal{C}) = 1 - w\alpha_q^i, \text{ otherwise} \quad (15)$$

$$m^i(A) = 0 \forall A \in P(\mathcal{C}) \backslash \{\mathcal{C}, \{C_q\}\} \quad (16)$$

where $0 \leq w \leq 1$ is a weight factor that controls the contribution of the neighboring pixels in the decision making process. The optimal value of $w$ for best classification performance depends on the image under investigation and can be learnt during training using grid search.

## IV. Experimental Results and Discussions

We report the performances of the proposed classifiers for two multispectral satellite images. We call them *Satimage1* and *Satimage2*. Satimage1 is a Landsat-TM image available along with full ground truth in the catalog of sample images of the ERDAS software and used for testing various algorithms [9]. The image covers the city of Atlanta, GA, and its surroundings. Satimage2 is also a Landsat-TM image depicting outskirts of the city of Vienna, Austria [3].

The Satimage1 is of size $512 \times 512$ pixels captured by seven detectors operating in different spectral bands from Landsat-TM3. Each of the detectors generates an image with pixel values varying from 0 to 255. The $512 \times 512$ ground truth data provide the actual distribution of classes of objects captured in the image. From this data, we produce the labeled data set with each pixel represented by a seven-dimensional feature vector and a class label. Each dimension of a feature vector comes from one channel, and the class label comes from the ground truth data. Fig. 1(a) shows the ground truth for Satimage1 where different classes are indicated with different shades of gray.

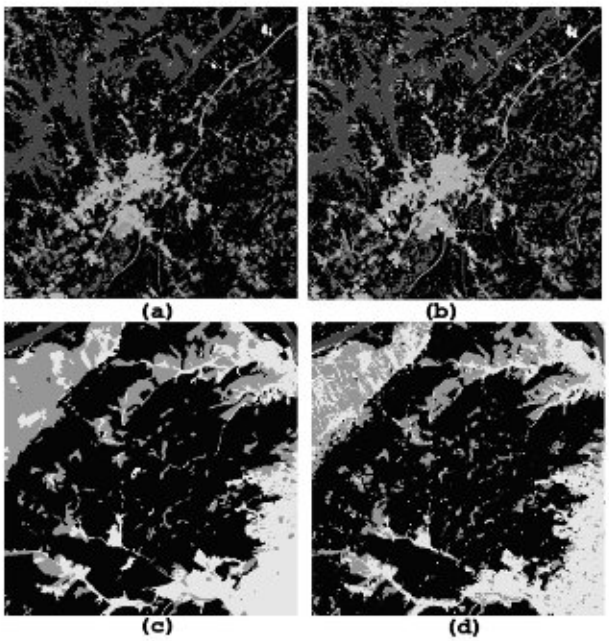


Fig. 1. (a) Ground truth for Satimage1. (b) Classified image for Satimage1 (Method 3, Training set 1). (c) Ground truth for Satimage2. (d) Classified image for Satimage2 (Method 4, Training set 1).

Satimage2 also is a seven channel Landsat-TM image of size $512 \times 512$. However, due to some characteristic of the hardware used in capturing the images, the first row and the last column of the images contain gray value 0. So, we did not include those pixels in our study and effectively worked with $511 \times 511$ images. The ground truth containing four classes is used for labeling the data. Fig. 1(c) shows the ground truth for Satimage2 where different classes are indicated with different shades of grey.

In our study, we generated four sets of training samples for each of the images. For Satimage1, each training set contains 200 data points randomly chosen from each class. For Satimage2, we include in each training set 800 randomly chosen data points from each of the four classes. The classifiers designed with the training data are tested on the whole images.

The classification results of the noncontextual fuzzy rule-based systems are summarized in Table I(a) and (b) for Satimage1 and Satimage2, respectively. Table I(a) and (b) summarizes the performances of the four classifiers using different methods of aggregation of spatial information for Satimage1 and Satimage2, respectively. We used the same fuzzy rule bases for respective training sets as used previously, but the decision-making step is modified to use the proposed four methods.

A comparison between Tables I(a) and II(a) shows that, for Satimage1, Method 3 performs the best with improvements varied between 1.12% and 2.05%, and the best performing classifier (for training set 4) achieves an error rate as low as 10.45%. This is closely followed by Method 2. Methods 1 and 4 show marginal improvement for training sets 1 and 2, while, for training sets 3 and 4, their performance degrades a little. A comparison of Tables I(b) and II(b) shows that, although the classifiers using Method 3 increase the classification accuracy by about 1.5%, Method 1 is clearly the best performer

TABLE I
PERFORMANCES OF NONCONTEXTUAL FUZZY RULE-BASED CLASSIFIERS

| Trng Set | No. of rules | $k_{cp}$ | Error Rate in Training Data | Error Rate in Whole Image |
|---|---|---|---|---|
| (a) Satimage1 | | | | |
| 1. | 30 | 5.0 | 12.0% | 13.6% |
| 2. | 25 | 6.0 | 14.5% | 14.47% |
| 3. | 25 | 5.0 | 12.0% | 13.03% |
| 4. | 27 | 4.0 | 12.6% | 12.5% |
| (b) Satimage2 | | | | |
| 1. | 14 | 2.0 | 16.3% | 14.14% |
| 2. | 14 | 2.0 | 16.3% | 14.04% |
| 3. | 12 | 2.0 | 17.09% | 14.01% |
| 4. | 11 | 2.0 | 17.34% | 14.23% |

TABLE II
PERFORMANCES OF FUZZY RULE-BASED CONTEXTUAL CLASSIFIERS

| Trng. Set | Error Rate in Whole Image | | | |
|---|---|---|---|---|
| | Method 1 | Method 2 | Method 3 | Method 4 |
| (a) Satimage1 | | | | |
| 1. | 13.32% | 11.93% | 11.48% | 13.43% |
| 2. | 14.00% | 13.01% | 12.62% | 14.38% |
| 3. | 13.66% | 11.54% | 11.23% | 13.90% |
| 4. | 12.98% | 11.01% | 10.45% | 13.18% |
| (b) Satimage2 | | | | |
| 1. | 11.55% | 12.96% | 12.64% | 11.75% |
| 2. | 11.55% | 12.75% | 12.48% | 11.65% |
| 3. | 11.24% | 12.45% | 12.23% | 11.35% |
| 4. | 11.24% | 12.53% | 12.28% | 11.44% |

for Satimage2 with improvements ranging between 2.49% and 2.99%. Similar performances are achieved by Method 4. Method 4 used $w = 1$ for both Table II(a) and (b). Fig. 1(b) shows the classified image corresponding to training set 1 and Method 3 for Satimage1. Fig. 1(d) displays classified image (for training set 1, Method 4) for Satimage2.

Our experimental results demonstrate that for Satimage1 Methods 3 and 2 work well while the other methods work better for Satimage2. These differences of performances can be explained if we look into the way the neighborhood information is aggregated in each method and the nature of the spatial distribution of the classes in the images. In Method 1, the fuzzy label vectors of the central pixel (the pixel of interest) and its eight neighboring pixels are treated in same way for aggregation. The same is true for Method 4, though the evidence theoretic approach is used for information aggregation. In Methods 2 and 3, eight BPAs are defined; each of them is assigned using the possibilistic label vector of the central pixel and that of one of the eight neighboring pixels. Thus, all the BPAs are heavily influenced by the central pixel and, consequently, so is the final decision. Hence, it is expected that for the images dominated by large stretches of homogeneous areas (i.e., area covered by single land cover type) can be classified better by Methods 1 and 4. Also, for Method 4, there may exist an optimal weight (different from 1) with improved performance. We shall see later that this is, indeed, the case. Comparing Fig. 1(c) (ground truth for Satimage2) with Fig. 1(a) (ground truth for Satimage1) reveals that Satimage1 mostly consists of small (often very small) patches of land cover types, while Satimage2 has large patches of land cover types. So, for Satimage1, neighborhood information needs to be used judiciously to get an improved classifier. This is what is achieved by Methods 2 and 3.

TABLE III
CLASS-WISE AVERAGE CLASSIFICATION PERFORMANCE FOR THE FOUR PROPOSED METHODS 1–4

| Landcover types (Frequency) | Average classification rate (%) | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| (a) Satimage1 | | | | |
| Forest (176987) | 93.25 | 90.80 | 91.33 | 94.47 |
| Water (23070) | 90.95 | 90.42 | 90.64 | 92.29 |
| Agriculture (26986) | 68.49 | 83.60 | 83.12 | 60.12 |
| Bare ground (740) | 58.15 | 65.46 | 65.92 | 55.10 |
| Grass (12518) | 53.19 | 73.75 | 75.65 | 55.59 |
| Urban area (11636) | 80.82 | 80.83 | 79.76 | 75.32 |
| Shadow (3197) | 48.97 | 96.09 | 95.68 | 40.42 |
| Clouds (358) | 99.78 | 99.78 | 99.78 | 99.79 |
| (b) Satimage2 | | | | |
| Built-up land (185812) | 85.83 | 84.09 | 84.38 | 85.55 |
| Forest (161522) | 95.16 | 94.00 | 94.22 | 95.68 |
| Water (2643) | 96.02 | 93.77 | 94.32 | 96.13 |
| Agriculture (48554) | 68.20 | 68.03 | 67.79 | 66.92 |

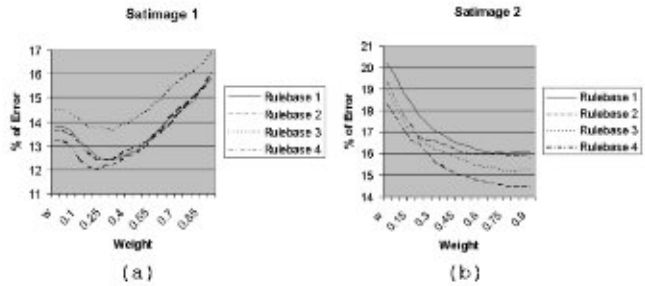


Fig. 2. Result of grid search for optimal value of $w$ using $100 \times 100$ subimages of (a) Satimage1 and (b) Satimage2. The numbers beside the plots correspond to the rule base number.

TABLE IV
CLASSIFICATION PERFORMANCES OF CLASSIFIERS USING MODIFIED METHOD 4 WITH $w = 0.35$ ON SATIMAGE1

| Training Set | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Error rate | 11.10% | 12.24% | 11.31% | 10.33% |

To have a closer look at class-wise performances of the proposed methods, we have presented the class-wise classification performances in Table III. It can be seen from the table that, for Satimage2, all four methods perform comparably for each of the four classes, with a slender edge in favor of Methods 1 and 4. However, for Satimage1, which contains more complicated spatial distribution of the classes, there is significant class-specific variation of performance among different methods. For the classes Forest, Water, Urban area, and Clouds, all methods perform nearly equally well (the variation is within 5% approximately); however, for other classes, the performances differ significantly. For Agriculture, the second largest class, Methods 2 and 3 have an accuracy of over 83%, while Methods 1 and 4 are 68.49% and 60.12% accurate, respectively. For the class Bare ground, Methods 2 and 3 outperform the others comfortably. For the class Shadow, there is a huge performance gap between the Methods 2–3 ($\approx$96%) and the Methods 1 and 4 (49% and 40%). However, since the frequency of the Shadow class is very small, this variation does not affect the overall accuracy significantly.

It can be observed from Tables I-III that, in a few cases, classifiers with fewer number of rules perform better than those with larger number of rules. This is due to the fact that each classifier is trained with different randomly generated training sets and there is some randomness involved in the SOFM-based prototype generation scheme. These result in different rule bases, where one with more number of rules may have a few rules in partial conflict.

The experiments are carried out using a desktop computer with P4 processor (3.0 GHz) and 256-MB memory. Apart from the image sizes, the computation time depends largely on the number of rules and number of classes. For an image, the computation time can be divided into two parts: computation of the fuzzy label vectors using rule base and decision making by applying the evidence theoretic methods. The first part is the same across all four methods and requires approximately 50 s for Satimage1 (25–30 rules and eight classes) and 25 s for Satimage2 (11–14 rules and four classes). The decision making using Methods 1–4 require 2, 10, 50, and 5 s, respectively, for Satimage1 and 1, 5, 10, and 2 s, respectively, for Satimage2. As expected, Method 1 is the fastest one because it involves simple averaging of the label vectors. On the other hand, Method 3 requires dealing with both singleton and doubleton sets of classes that increase the computation time almost quadratically with number of classes.

To investigate the possibility of optimization of Method 4, we now use the modified (12)–(16) that incorporate a weight factor $w$ for controlling the contribution of the information from the neighborhood in the decision making process. The value $w = 1.0$ makes the method same as the original Method 4. To find an optimal value of $w$, we use a $100 \times 100$ subimage from each image and find an optimal $w$ based on these blocks. We use grid search to find the optimal $w$. Note that the rule bases are the same as used earlier. For example, for Satimage1, for each of the four rule sets, we find the optimal $w$ using the classification error on the selected block of image. Fig. 2(a) depicts the variation of classification error as a function of $w$ for Satimage1. It is interesting to observe that for Satimage1 for all four rule bases the best performances are achieved around $w = 0.35$. So we should use the modified Method 4 with $w = 0.35$ for each of the four rule bases and Table IV displays the classification errors for the whole image. Comparing Table IV with column five of Table II(b), we find a consistent improvement with $w = 0.35$ in all four cases. The improvement varied between 2.14% and 2.75%. We also tried to find an optimal $w$ for each of the four rule bases for Satimage2. Fig. 2(b) displays the classification error as a function of $w$. In this case, for all four rule bases we find an optimal value of $w = 1.0$. This again confirms the fact that when different land cover types form spatially compact clusters, neighbors and central pixels play equally important roles in decision making.

### A. Comparison With an MRF Model-Based Classifier

Recently, contextual methods based on MRF models have become popular for classification of multispectral [20], as well as multisource [21], [22] satellite images. Typically, a Bayesian framework is used to model the posterior probability. To estimate the model parameters, an energy function is minimized using an optimization technique like simulated annealing, genetic algorithm, etc. These methods usually utilize

TABLE V
COMPARISON OF CLASSIFICATION PERFORMANCE (PERCENTAGE ERRORS) OF MRF-BASED METHOD WITH THE PROPOSED METHODS

| Image | MRF method | Fuzzy rulebase | Proposed contextual methods | | | |
|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 |
| Satimage 1 | 18.79 | 9.86 | 5.96 | 8.58 | 8.18 | 5.81 |
| Satimage 2 | 17.81 | 16.41 | 12.55 | 14.68 | 14.47 | 12.45 |

the contextual information in the training stage. Also, their accuracy depends on the correctness of assumption about the class-conditional density functions. Here, we compare the proposed methods with the MRF model-based approach introduced by Sarkar *et al.* [20]. This approach is based on capturing intrinsic characters of tonal and textural regions of a multispectral image. Using an initially oversegmented image and the multispectral image, it defines an MRF over the region adjacency graph (RAG) of the initial segmented image. Because of the energy minimization of the associated MRF, the oversegmented regions are likely to be merged. To compare the similarity of adjacent regions, a multivariate statistical test is incorporated while minimizing the energy function associated with the underlying MRF. A cluster validation technique is also used for finding optimal segments.

The implementation of the MRF-based scheme used for the experiments here is capable of handling only up to four channels. Therefore, we used channels 3, 4, 5, and 7 for Satimage1 and channels 2, 3, 5, and 7 for Satimage2. The choice is based on visual inspection of the channel images as well as some experiments with various combinations of channels with a view to finding a good combination of features suitable for the MRF-based method. Using the same training-test data sets as used by the previous experiments, for Satimage1, the MRF-based method fails to discriminate some of the classes, for example, a significant part of the water becomes forest. So, we used a more detailed ten-class ground truth data. Unfortunately, this detailed ground truth is incomplete, i.e., does not cover the entire image. Therefore, we used a training set with 1463 data points and a test set with 1480 data points. For Satimage2, the training set consisted of 3200 points as earlier. The classification performance of the MRF-based method, noncontextual fuzzy rule-based method, and proposed contextual decision making schemes on the test sets is summarized in Table V. All experiments are conducted with the same training and test data sets for respective images as described in this section.

Table V shows that for Satimage1 the error rate of the ordinary (noncontextual) fuzzy rule-based classifier is almost half of that by the MRF-based method. All of the four proposed contextual decision making schemes reduce the error rate further with Method 4 achieving the lowest error rate of 5.81%. In the case of Satimage2, the error rate for the noncontextual fuzzy rule-based classifier is also about 1.4% less than the MRF-based method. The proposed contextual schemes again produce further improved classification performance. Here, again, Method 4 exhibits the lowest error rate of 12.45% closely followed by Method 1 (12.55%). Thus, it can be observed that for both the images the ordinary fuzzy rule-based classifiers, as well as the four contextual classifiers perform consistently better than the MRF-model based method tested here.

## V. CONCLUSION

We proposed a comprehensive scheme for designing contextual classifiers for land cover classification in multispectral satellite images. This is a multistage scheme. First, an SOFM-based dynamic prototype generation algorithm is used to generate an adequate number of prototypes. Then, the prototypes are converted into fuzzy rules and they are further fine-tuned to design an efficient fuzzy rule-based classifier. However, this classifier is a noncontextual one. Since the landcover classes usually appear in spatial clusters, a classification scheme using contextual information can be more efficient than noncontextual one. Hence, we develop four decision making schemes which use information from spatial neighborhood of the pixels. For assigning a class label to a pixel, the schemes use the information generated by the fuzzy rule base in the form of possibilistic label vectors for the pixel under consideration as well as its eight spatial neighbors. Three of the proposed schemes use evidence theoretic framework and the remaining one is based on the fuzzy k-NN rule.

The suitability of a particular scheme depends to some extent on the nature of the image to be classified. The performance of the methods on the training/validation data can be used to decide on the best classifier for a given situation. One of the methods has a parameter that can be tuned based on the training/validation data to get a classifier with improved performance.

## ACKNOWLEDGMENT

The authors would like to thank A. Sarker for help provided to generate some results using the MRF-based method. They would also like to thank A. Pinz for providing them with the Satimage2 data for the experiments, as well as the referees for their valuable suggestions.

## REFERENCES

[1] G. M. Foody, "Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1317–1340, 1996.

[2] J. D. Paola and R. A. Schowengerdt, "A detailed comparison of backpropagation neural network and maximum likelihood classifiers for urban land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 4, pp. 981–996, Jul. 1995.

[3] H. Bischof, W. Schneider, and A. J. Pinz, "Multispectral classification of landsat-images using neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 3, pp. 482–490, May 1992.

[4] C. Y. Ji, "Land-use classification of remotely sensed datta using Kohonen self-organizing feature map neural networks," *Int. J. Remote Sens.*, vol. 18, no. 12, pp. 1451–1460, Dec. 2000.

[5] G. M. Foody, "Hard and soft classifications by a neural network with a nonnexhaustively defined set of classes," *Int. J. Remote Sens.*, vol. 23, no. 18, pp. 3853–3864, 2002.

[6] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press, 1976.

[7] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Norwell, MA: Kluwer, 1999.

[8] L. Kuncheva, *Fuzzy Classifier Design*. New York: Physica-Verlag, 2000.

[9] A. S. Kumar, S. Chowdhury, and K. L. Majumder, "Combination of neural and statistical approaches for classifying space-borne multispectral data," in *Proc. ICAPRDT*, 1999, pp. 87–91.

[10] A. Bárdossy and L. Samaniego, "Fuzzy rule-based classification of remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 2, pp. 362–374, Feb. 2002.
[11] A. Kulkarni and S. McCaslin, "Knowledge discovery from multispectral satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 246–250, Apr. 2004.
[12] T. Denœux, "A k-nearest neighbor classification rule based on Dempster–Shafer theory," *IEEE Trans. Syst. Man Cybern.*, vol. 25, no. 5, pp. 804–813, May 1995.
[13] L. M. Zouhal and T. Denœux, "An evidence theoretic k-NN rule with parameter optimization," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 28, no. 2, pp. 263–271, May 1998.
[14] T. Lee, J. A. Richards, and P. H. Swain, "Probabilistic and evidential approaches for multispectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, no. 2, pp. 283–293, Mar. 1987.
[15] A. Srinivasan and J. A. Rechards, "Knowledge-based techniques for multi-source classification," *Int. J. Remote Sens.*, vol. 11, pp. 501–525, 1990.
[16] D. R. Peddle, "An empirical comparison of evidential reasoning, linear discriminant analysis and maximum likelihood algorithms for land cover classification," *Can. J. Remote Sens.*, vol. 19, pp. 31–44, 1993.
[17] H. Kim and P. H. Swain, "Evidential reasoning approach to multisource-data classification in remote sensing," *IEEE Trans. Syst. Man Cybern.*, vol. 25, no. 8, pp. 1257–1265, Aug. 1995.
[18] A. Jouan and Y. Allard, "Land use mapping with evidential fusion of features extracted from polarimetric synthetic aperture radar an hyperspectral imagery," *Inf. Fusion*, vol. 5, pp. 251–267, 2004.
[19] R. M. Haralick, "Decision making in context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 4, pp. 417–428, Jul. 1983.
[20] A. Sarkar, M. K. Biswas, B. Kartikeyan, V. Kumar, K. L. Majumder, and D. K. Pal, "A MRF model-based segmentation approach to classification for multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1102–1113, May 2002.
[21] A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for clasification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.
[22] B. C. K. Tso and P. M. Mather, "Classification of multisource remote sensising imagery using a gentic algorithm and Markov random fields," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, no. 3, pp. 1255–1260, May 1999.
[23] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.
[24] A. Laha and N. R. Pal, "Some novel classifiers designed using prototypes extracted by a new scheme based on self-organizing feature map," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 31, no. 6, pp. 881–890, Dec. 2001.
[25] H. Dyckhoff and W. Pedrycz, "Generalized means as models of compensative connectives," *Fuzzy Sets Syst.*, vol. 14, pp. 143–154, 1984.
[26] N. R. Pal and C. Bose, "Context sensitive inferencing and "reinforcement type" tuning algorithms for fuzzy logic systems," *Int. J. Knowledge-Based Intell. Eng. Syst.*, vol. 3, no. 4, pp. 230–239, 1999.
[27] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[28] P. Smets and R. Kennes, "The transferable belief model," *Artif. Intell.*, vol. 66, pp. 191–234, 1994.
[29] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *J. Int. Fuzzy Syst.*, vol. 2, pp. 267–278, 1994.

**Arijit Laha** (M'05) received the B.Sc. (with honors) and M.Sc. degrees in physics from the University of Burdwan, Burdwan, India, in 1991 and 1993, respectively, and the M.Tech. degree in computer science from the Indian Statistical Institute (ISI), Calcutta, in 1997.

He was with Wipro Infotech Global R&D, Indian Statistical Institute and National Institute of Management, Kolkata. Currently, he is a faculty at the Institute for Development and Research in Banking Technology, Hyderabad, India. His research interests include computational intelligence techniques for pattern recognition, building expert systems, data compression, data warehousing, data mining, and parallel algorithms. He was involved in two projects for the development of mission-critical expert systems. He is a Consultant to a leading Indian bank in their data warehousing project.

**Nikhil R. Pal** (F'05) received the B.Sc. degree in physics and the Master of Business Management degree from the University of Calcutta, Calcutta, India, in 1978 and 1982, respectively, and the M.Tech. and Ph.D. degrees in computer science from the Indian Statistical Institute (ISI), Calcutta, in 1984 and 1991, respectively.

Currently, he is a Professor in the Electronics and Communication Sciences Unit, ISI. He has coauthored and edited/co-edited several books covering various facets of computational intelligence and data mining. His research interests include image processing, pattern recognition, fuzzy systems, neural networks, evolutionary computation, and bioinformatics.

Dr. Pal serves as the Editor-in-Chief of the IEEE TRANSACTIONS ON FUZZY SYSTEMS. He serves the editorial/advisory board of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, *Fuzzy Sets and Systems*, the *International Journal of Approximate Reasoning*, the *International Journal of Hybrid Intelligent Systems*, *Neural Information Processing—Letters and Reviews*, the *International Journal of Knowledge-Based Intelligent Engineering Systems*, the *Iranian Journal of Fuzzy Systems*, and the *International journal of Neural Systems*. He has served as the president of the Asia Pacific Neural Net Assembly. He was the Program Chair of the 4th International Conference on Advances in Pattern recognition and Digital Techniques, Calcutta, and the General Chair of 2002 AFSS International Conference on Fuzzy Systems, Calcutta. He was the General Chair of the 11th International Conference on Neural Information Processing, ICONIP 2004.

**Jyotirmoy Das** received the M.Tech. degree in radio physics and electronics and the Ph.D. degree in computer memory technology from the University of Calcutta, Calcutta, India.

He was a Senior Professor and the Head of Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta. His current research interest includes atmospheric science problems, computational intelligence, and millimeterwave propagation.

Dr. Das is a Fellow of Indian Science Academy (Allahabad).