

Fuzzy Rules and Evidence Theory for Satellite Image Analysis

Arijit Laha

National Institute of Management, Calcutta
arijitl@yahoo.com

J. Das

Indian Statistical Institute, Calcutta
jdas@isical.ac.in

Abstract

Design of a fuzzy rule based classifier is proposed. The performance of the classifier for multispectral satellite image classification is improved using Dempster-Shafer theory of evidence that exploits information of the neighboring pixels. The classifiers are tested rigorously with two known images and their performance are found to be better than the results available in the literature. We also demonstrate the improvement of performance while using D-S theory along with fuzzy rule based classifiers over the basic fuzzy rule based classifiers for all the test cases.

1. Introduction

Analysis of satellite images has many important applications such as prediction of storm and rainfall, estimation of natural resources, estimation of crop yields, assessment of damage caused by natural disasters, and land cover classification. In this paper we focus on land cover classification from multi-spectral satellite images.

The most widely used techniques for this problem employ discriminant analysis, maximum likelihood classification, and neural networks [6], [3]. Such classifiers cannot handle the fact that for land cover a pixel may correspond to more than one types of objects. For example, the area covered by a pixel may correspond to 30% land and 70% water. Note that, the uncertainty involved in classifying such a pixel is not probabilistic, but fuzzy in nature and thereby it demands “soft” classifiers. In developing soft classifiers for land cover analysis two approaches have gained popularity. These are based on (1) fuzzy set theory and (2) Dempster and Shafer’s (DS) evidence theory [7].

Numerous fuzzy classification techniques have been developed by many researchers to solve problems in diverse fields. A comprehensive account of such works can be found in [2]. Fuzzy rules are attractive because they are interpretable and provides an analyst a deeper insight into the problem. Use of fuzzy rule based systems for land cover analysis is relatively new. In a recent paper Bárdossy and Samaniego [1] have proposed

a scheme for developing a fuzzy rulebased classifier for analysis of multispectral images.

The other approach for designing soft classifiers is to use the evidence theory developed by Dempster and Shafer [7]. Since the theory of evidence allows one to combine evidences obtained from diverse sources of information in support a hypothesis, it seems a natural candidate for analyzing multispectral images for land cover classification.

Here we propose a scheme for designing fuzzy rule-based classifiers for land cover types that uses evidence theory for decision making. This is a two stage process. First we find a good set of fuzzy rules using information from all channels. In the next stage, the responses of the fuzzy rules over a 3×3 neighborhood are used to define 8 Basic Probability Assignment which are then combined by DS rule to exploit contextual information to make a better decision. The problem of high variation in the variances of different features, which often degrades the performance of a distance based classifier substantially, is handled in a natural manner by fuzzy rules due to the atomic nature of the antecedent clauses.

2. Designing the Fuzzy Rule base

The proposed scheme has several stages. First a set of labeled prototypes is generated. Then the prototypes are converted into fuzzy rules. The fuzzy rules are further tuned for improving their performance. Labeled prototypes can be generated using any clustering algorithm followed by labeling the cluster centers. However, for most of such algorithms the number of clusters is a predefined parameter. Here we use the prototype generation scheme described in [5]. It is a two stage algorithm involving unsupervised and supervised learning that dynamically decides the number of prototypes and extract them using the training data. For details the readers are referred to [5].

2.1. Designing the fuzzy rulebase

A prototype (representing a cluster of points) \mathbf{v}_i for class k can be translated into a fuzzy rule of the form :

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

The fuzzy set CLOSE TO v_{ij} is modeled by a Gaussian membership function :

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

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 .

Each fuzzy set is characterized by two parameters v_j and σ_{ij} . The v_{ij} s of the rules can be initialized with the components of the final set of prototypes, V^{final} , generated by our SOFM based algorithm, $V^0 = V^{final} = \{\mathbf{v}_1^{final}, \dots, \mathbf{v}_{\hat{c}}^{final}\} = \{\mathbf{v}_1^0, \dots, \mathbf{v}_{\hat{c}}^0\}$ where $v_{ij}^0 = v_{ij}^{final}$. The notation V^0 is used to indicate that it corresponds to the initial centers of the membership functions. The initial estimates of the σ_{ij} s are computed as follows.

For each prototype \mathbf{v}_i^0 in the set $V^0 = \{\mathbf{v}_i^0 \mid i = 1, \dots, \hat{c}, \mathbf{v}_i^0 \in \mathbb{R}^p\}$ let X_i be the set of training data closest to \mathbf{v}_i^0 . For each \mathbf{v}_i^0 the set

$$S_i = k_w \{\sigma_{ij} \mid \sigma_{ij} = (\sqrt{(\sum_{\mathbf{x}_k \in X_i} (x_{kj} - v_{ij}^0)^2)}) / |X_i|\}$$

is computed and is associated with the prototype. We use the $k_w \sigma_{ij}$ as the spread of the membership function whose center is at v_{ij} ; $k_w > 0$ is a constant parameter and its value can have a significant impact on the classification performance for complex data sets.

2.2. Tuning the rulebase

The initial rulebase R^0 thus obtained is further refined to achieve better performance. The exact tuning algorithm depends on the conjunction operator used for computation of the firing strengths. The firing strength can be calculated using any T-norm [2]. Use of different T-norms results in different classifiers. The **minimum** and the **product** are among the most popular T-norms used as conjunction operators. It is much easier to formulate a calculus based tuning algorithm if product is used. However, if there are many clauses in the antecedent, the firing strength of a rule tends to have low numerical values even when the membership value of each individual clause is quite high. Though computationally this does not pose any problem (we are interested in relative firing strengths of the rules), it is conceptually somewhat unattractive - especially from the interpretability viewpoint.

Thus to avoid the use of the product and at the same time to be able to derive update rules easily we use a soft-min operator.

The **soft-match** of n positive number x_1, x_2, \dots, x_n is defined by

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

where q is any real number. SM is known as an aggregation operator with upper bound of value 1 when $x_i \in [0, 1] \forall i$. It is easy to see that $\lim_{q \rightarrow \infty} SM(x_1, x_2, \dots, x_n, q) = \max(x_1, x_2, \dots, x_n)$ and $\lim_{q \rightarrow -\infty} SM(x_1, x_2, \dots, x_n, q) = \min(x_1, x_2, \dots, x_n)$. 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

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

In the present study we use $q = -10.0$.

Let $\mathbf{x} \in X$ be from class c and R_c be the rule from class c giving the maximum firing strength α_c for \mathbf{x} . Also let R_{-c} be the rule from the incorrect classes having the highest firing strength α_{-c} for \mathbf{x} .

We use the error function $E = \sum_{\mathbf{x} \in X} (1 - \alpha_c + \alpha_{-c})^2$.

We minimize E with respect to v_{cj} , v_{-cj} and σ_{cj} , σ_{-cj} of the two rules R_c and R_{-c} using gradient descent. Here the index j corresponds to clause number in the corresponding rule. Minimizing E will refine the rules with respect to their contexts in the *feature space*. Note that, the context referred here is different from the context of a pixel defined in terms of its spatial neighborhood. The tuning process is repeated until the rate of decrement in E becomes negligible resulting in final rule base R_{final} .

3. Using the theory of evidence for Rule aggregation

For the sake of completeness, we briefly introduce the Dempster-Shafer theory of evidence. Let Θ be the universal set and $P(\Theta)$ be its power set. A Belief measure is a function $Bel : P(\Theta) \rightarrow [0, 1]$ that satisfies the axioms [7]. $b1 : Bel(\emptyset) = 0$ and $Bel(\Theta) = 1$.

$b2 : \text{For every } A, B \in P(\Theta), \text{ if } A \subset B \text{ then } Bel(A) \leq Bel(B).$

$b3 : Bel(A_1 \cup A_2 \cup \dots \cup A_n) \geq \sum_i Bel(A_i) - \sum_{i < j} Bel(A_i \cap A_j) + \dots + (-1)^n Bel(A_1 \cap \dots \cap A_n),$ for every n and for every collection of subsets of Θ .

There is a plausibility measure with each belief measure defined by $Pl(A) = 1 - Bel(A^c) \forall A \in P(\Theta)$.

Every belief measure and its dual plausibility measure can be expressed in terms of a Basic Probability Assignment (BPA) function m . $m : P(\Theta) \rightarrow [0, 1]$ is called a BPA iff $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. A belief measure and a plausibility measure are uniquely determined by m through the formulas:

$$Bel(A) = \sum_{B \subseteq A} m(B). \quad (1)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A \subset \Theta. \quad (2)$$

Every set $A \in P(\Theta)$ 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 some power set $P(\Theta)$ can be combined by Dempster's rule of combination to obtain a joint BPA $m^{1,2}$ as:

$$m^{1,2}(A) = \begin{cases} \frac{\sum_{B \cap C = A} m^1(B)m^2(C)}{1-K} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \quad (3)$$

Here

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

Eq. (3) 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 $m = m^1 \oplus m^2 \oplus \dots \oplus m^k = \oplus_{i=1}^k m^i$.

3.1. Pignistic probability

Given a belief measure we are often required to make decisions based on the available evidence. In such case Θ becomes the set of decision alternatives and the function Bel denote our belief about the choice of the optimal decision $\theta_0 \in \Theta$. However, in general it is not possible to select the optimal decision directly from the evidence embodied in the function Bel . In such cases, we use the *pignistic transformation*, Γ_Θ , to construct a probability function for selecting the optimal decision [8]. Thus

$$P^\Theta = \Gamma_\Theta(Bel).$$

P^Θ is called a *pignistic probability*, which can be used for making decision. The pignistic probability for $\theta \in \Theta$ can be expressed in terms of BPAs as follows:

$$P^\Theta(\theta) = \sum_{A \subseteq \Theta, \theta \in A} \frac{m(A)}{|A|} \quad (4)$$

Optimal decision can now be chosen in favor of θ_0 , if θ_0 has the highest pignistic probability.

3.2. Scheme for decision making

In our problem the frame of discernment is the set of classes, $\mathcal{C} = \{C_1, C_2, \dots, C_c\}$, where c is the number of classes. The propositions take the form *the true class label of the pixel of interest is in* $A \subset \mathcal{C}$.

Let us denote the pixel of interest as p^0 and its eight spatial neighbors as p^1, p^2, \dots, p^8 . We use the firing strengths produced by the rulebase in support of different classes for p^0 and one of its neighbors, say p^i as the i -th source of evidence. Let r be the number of rules in the fuzzy rulebase. Since $c \leq r$, there could be multiple rules corresponding to a class. Let α_k^0 be the highest firing strength produced by the rules corresponding to the class C_k for p^0 . We treat this value as the confidence measure of the rulebase pertaining to the membership of p^0 to the class C_k . Thus, the set of values $CM^0 = \{\alpha_k^0 : k = 1, 2, \dots, c\}$ contain the confidence measures for all the classes for p^0 (if a confidence measure is less than a threshold, say 0.01, it is set to 0). A similar set of confidence measures CM^i can be constructed for every p^i ; $i = 1, \dots, 8$.

Now we use CM^0 and CM^i to define the i -th BPA m^i to the subsets of \mathcal{C} . There are 2^c possible subsets of \mathcal{C} , i.e., members of the power set of \mathcal{C} . Each subset corresponds to the proposition that the “true” class of p^0 is contained in that subset. We shall consider the subsets containing one and two elements only. The subsets containing one element correspond to propositions of the form “the class contained in the subset is the true class for p^0 ” and the subsets containing two elements corresponds to propositions of the form “the true class label of p^0 is any one of the two classes contained in the subset”. 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:

$$m^i(\{C_k\}) = \frac{\frac{(\alpha_k^i + \alpha_k^0)}{2} \exp^{-(\alpha_k^i - \alpha_k^0)^2}}{S}, k = 1, 2, \dots, c \quad (5)$$

For $l, m = 1, 2, \dots, c$, $m^i(\{C_l, C_m : l < m\}) =$

$$\frac{\frac{(\alpha_l^i + \alpha_m^0)}{2} \exp^{-(\alpha_l^i - \alpha_m^0)^2} + \frac{(\alpha_m^i + \alpha_l^0)}{2} \exp^{-(\alpha_m^i - \alpha_l^0)^2}}{2S} \quad (6)$$

where $S = \sum_{k=1}^{k=c} \frac{(\alpha_k^i + \alpha_k^0)}{2} \exp^{-(\alpha_k^i - \alpha_k^0)^2} + \sum_{l=1}^{l=c-1} \sum_{m=l+1}^{m=c} \frac{(\alpha_l^i + \alpha_m^0)}{2} \exp^{-(\alpha_l^i - \alpha_m^0)^2} + \frac{(\alpha_m^i + \alpha_l^0)}{2} \exp^{-(\alpha_m^i - \alpha_l^0)^2}$.

The numerators in the right hand side of the above formulae are measures of confidence in favor of the respective propositions. A closer look on (5) shows that

the numerator is a product of two terms. The first term is the average of the confidence measures of p^0 and p^i for the class C_k , while the second term is an exponential one that reflects the degree of closeness of the confidence measures. Thus as a whole a high value of the numerator reflects two facts: (1) both p^0 and p^i has high confidence value for class C_k and (2) the confidence values are close to each other. Eq. (6) is a straightforward extension of the same concept when we define the confidence in favor of a pair of classes.

Thus for the eight neighboring pixels we obtain eight combinable sources of evidence. The global BPA can be computed by applying the Dempster's rule repeatedly. The combined global BPA m^G is computed as follows:

$$m^G = \oplus_{i=1}^8 m^i = (\dots((m^1 \oplus m^2) \oplus m^3) \oplus \dots m^8). \quad (7)$$

It is easily seen that:

$$m^{(i,j)}(\{C_k\}) = m^i(\{C_k\}) \oplus m^j(\{C_k\})$$

$$= \frac{\left\{ \begin{array}{l} m^i(\{C_k\})m^j(\{C_k\}) \\ + m^i(\{C_k\}) \sum_{l \neq k} m^j(\{C_k, C_l\}) \\ + m^j(\{C_k\}) \sum_{l \neq k} m^i(\{C_k, C_l\}) \\ + \sum_{l \neq k} m^i(\{C_k, C_l\}) \sum_{m \neq k, l} m^j(\{C_k, C_m\}) \end{array} \right\}}{1-K},$$

$k = 1, 2, \dots, c$
and

$$m^{(i,j)}(\{C_l, C_m\}) = m^i(\{C_l, C_m\}) \oplus m^j(\{C_l, C_m\})$$

$$= \frac{m^i(\{C_l, C_m\})m^j(\{C_l, C_m\})}{1-K} \quad (8)$$

$l, m = 1, 2, \dots, c, l \neq m$; where K is given by

$$K = \sum_{k=1}^{c-1} m^i(\{C_k\}) \sum_{l=k+1}^c m^i(\{C_l\})$$

$$+ \sum_{k=1}^c m^i(\{C_k\}) \sum_{l, m \neq k}^c m^j(\{C_l, C_m\})$$

$$+ \sum_{k=1}^c m^j(\{C_k\}) \sum_{l, m \neq k}^c m^i(\{C_l, C_m\})$$

$$+ \sum_{l \neq r, s, \text{ and } m \neq r, s}^c m^i(\{C_l, C_m\})m^j(\{C_r, C_s\}).$$

Once m^G is obtained the pignistic probability for each class is computed. The following formula is used for computing the pignistic probability of class C_k :

$$P^C(C_k) = m^G(\{C_k\}) + \frac{\sum_{l=1, l \neq k}^c m^G(\{C_k, C_l\})}{2} \quad (9)$$

The pixel p^0 is assigned to the class C_k such that

$$P^C(C_k) \geq P^C(C_l) \quad \forall C_l \in \mathcal{C}.$$

4. Experimental results and discussions

We report the performances of the proposed classifiers for two multispectral satellite images. We call them **Satimage1** and **Satimage2**.

Trng Set	No. of rules	k_w	Error Rate in Training Data	Error Rate in Whole Image
Satimage1				
1.	30	5.0	12.0%	13.6%
2.	25	6.0	14.3%	14.47%
3.	25	5.0	12.0%	13.03%
4.	27	4.0	12.6%	12.5%
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 1: Performances of fuzzy rulebased classifiers using firing strength for decision making for different training sets

The Satimage1 is a 256-level Landsat-TM image of size 512×512 pixels captured by seven sensors operating in different spectral bands. Each sensor generates an image with pixel values varying from 0 to 255. The 512×512 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 7-dimensional feature vector and a class label. Satimage2 also is a seven channel 256-level Landsat-TM image of size 512×512 . 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 511×511 images. The ground truth containing four classes is used for labeling the data.

In our study we generated 4 training sets of samples for each of the images. For Satimage1, each training set contains 200 data points randomly chosen from each of eight classes. This choice is made to conform to the protocol followed in [4]. For Satimage2 we include in each training set 800 randomly chosen data points from each of four classes. Bischof et al. [3] used more training points / class than that of ours.

First we report the performances of the fuzzy rule-based classifiers using firing strengths directly for decision making and compare the results with the published results. Then we report the performances of the fuzzy classifiers using evidence theoretic approach for decision making. The performances of fuzzy rulebased classifiers using firing strengths directly for decision making is summarized in the Table 1.

For Satimage1 the best result reported in [4] uses a fuzzy integral based method and gives the classification rate 78.15%. In our case, *even the worst result is about 5% better than that*.

For Satimage2 the reported result in [3] shows 84.7% accuracy with the maximum likelihood classifier (MLC) and 85.9% accuracy with neural network based classifier. In our case for all training-test partitions the fuzzy rulebased classifiers outperform the MLC and at

Training Set	No. of rules	Error Rate in Whole Image
Satimage1		
1.	30	12.3%
2.	25	13.37%
3.	25	11.6%
4.	27	11.03%
Satimage2		
1.	14	12.7%
2.	14	12.65%
3.	12	12.4%
4.	11	12.51%

Table 2: Performances of the evidence theoretic fuzzy classifiers for different training sets

par with the results reported for neural networks.

Tables 2 summarizes the performances of the fuzzy rulebased classifiers using evidence theoretic approach. We used the same set of fuzzy rules as used previously, but the rule outputs are aggregated using the evidence theory.

Comparison of Table 2 with Table 1 clearly shows that in every case there is a consistent improvement in the classification performance. In case of Satimage1 the improvements varied between 1.1% and 1.5% and the best performing classifier (for training set 4) achieves error rate as low as 11.03%. For Satimage2 also the improvement varied between 1.4% and 1.7%. So the overall improvement for Satimage1 over the existing methods is more than 7%. For Satimage2 also we achieved consistent improvements using training sets of smaller size. For applications like crop yield estimation even a small improvement will have a significant impact on the overall estimate.

5. Conclusion

We proposed two classifiers: one is fuzzy rule based and the other integrates outputs of fuzzy rules using theory of evidence. Fuzzy rules are extracted with the help SOFM. The system automatically decides on the number of rules.

The fuzzy rule based classifier is of general nature and can be applied in any classification problem, while the evidence theoretic classifier exploits the spatial information available for an image to make the classification decision.

In the evidence theoretic framework we use the pixel under consideration and one of its neighbors to provide a body of evidence in support of different propositions regarding the class membership (to a particular class as well as a pair of classes) of the pixel. The BPAs for the propositions are calculated from the mutual confidences of the pixels in support of respective propositions. Eight bodies of evidence is obtained for eight neighbors of the pixel. Now the evidences are combined to obtain a global body of evidence. Then pignistic

probability for each class is computed and the pixel is assigned to the class with highest pignistic probability. The proposed system demonstrates a consistent improvement in performance.

Acknowledgement: Authors thank Prof. N. R. Pal for his continuous support and advices. They also thank Dr. A. S. Kumar and Dr. A. J. Pinz for allowing them to use the satellite images Satimage1 and Satimage2 respectively to test the proposed methods.

References

- [1] A. Bárdossy and L. Samaniego, "Fuzzy rule-based classification of remotely sensed imagery", *IEEE Trans. Geosci. Remote Sensing*, vol. 40, no.2 pp. 362-374, 2002.
- [2] J. C. Bezdek, J. Keller, R. Krishnapuram and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing* Kluwer, Massachusetts, 1999.
- [3] H. Bischof, W. Schneider and A. J. Pinz, "Multispectral Classification of Landsat-Images Using Neural Networks", *IEEE Trans. on Geosci. Remote Sensing*, vol. 30, no. 3, pp. 482-490, 1992.
- [4] A. S. Kumar, S. Chowdhury and K. L. Majumder, "Combination of neural and statistical approaches for Classifying space-borne multispectral data," *Proc. of ICAPRDT99*, pp. 87-91, 1999.
- [5] 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. on Syst. Man and Cybern: B*, vol 31, no. 6, pp. 881-890, 2001.
- [6] 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. on Geosci. Remote Sensing*, vol. 33, pp. 981-996, July, 1995.
- [7] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, 1976.
- [8] P. Smets and R. Kennes, "The transferable belief model", *Artificial Intelligence*, vol. 66, pp. 191-234, 1994.