ELSEVIER

Pattern Recognition Letters 20 (1999) 935–948

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

A fuzzy set-based accuracy assessment of soft classification

Elisabetta Binaghi [a,*], Pietro A. Brivio [b], Paolo Ghezzi [b], Anna Rampini [a]

[a] *Istituto per le Tecnologie Informatiche Multimediali – ITIM, C.N.R., Via Ampère 56, 20131 Milan, Italy*
[b] *Telerilevamento-IRRS, C.N.R., Via Bassini 15, 20133 Milan, Italy*

Received 19 June 1998; received in revised form 15 June 1999

Abstract

Despite the sizable achievements obtained, the use of soft classifiers is still limited by the lack of well-assessed and adequate methods for evaluating the accuracy of their outputs. This paper proposes a new method that uses the fuzzy set theory to extend the applicability of the traditional error matrix method to the evaluation of soft classifiers. It is designed to cope with those situations in which classification and/or reference data are expressed in multimembership form and the grades of membership represent different levels of approximation to intrinsically vague classes. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Soft classifiers; Accuracy measures; Fuzzy sets theory; Error matrix

1. Introduction

In many applications, it is desirable to have a "soft" classifier that, for a given input pattern vector, computes the "likelihood" that the pattern lies in any of a set of possible classes. In general, soft models for classification are rooted in specific representation frameworks within which the partial belongingness of a given pattern to several categories at the same time is explicitly modeled (Binaghi et al., 1996; Bouchon-Meunier et al., 1995).

Statistical classification models interpret a given pattern as fully contributing to a given class, and the computed probabilities are an expression of the frequency with which this full membership occurs.

In soft models for classification, non-probabilistic uncertainty due to vagueness and/or ambiguity should be modeled as partial belongingness to several categories at the same time (Klir and Folger, 1988).

Various approaches may be used to derive a soft classifier. These approaches are based on specific uncertainty representation frameworks such as the fuzzy set theory, Dempster–Shafer theory and certainty factors (Binaghi et al., 1996; Bloch, 1996). In addition to the use of specific representation frameworks, the output of "hard" classifiers, such as the maximum likelihood classifier and the multilayer perceptron, can be softened to derive measures of the strength of class membership (Schowengerdt, 1996; Wilkinson, 1996).

The most common solutions adopt a fuzzy set framework (Pedrycz, 1990; Binaghi and Rampini, 1993; Ishibuchi et al., 1993). The apparatus of the fuzzy set theory serves as a natural framework for modeling the gradual transition from membership to non-membership in intrinsically vague classes. Here the assumption is that the classification

[*] Corresponding author. Tel.: +39-2-70643287; fax: 39-2-70643292; e-mail: ebinaghi@itim.mi.cnr.it

PII: S0167-8655(99)00061-6

process is possibilistic in nature. The fuzzy set framework introduces vagueness, with the aim of reducing complexity, by eliminating the sharp boundary dividing the members of a class from non-members. In some situations, these sharp boundaries may be arbitrary, or powerless, as they cannot capture the semantic flexibility inherent in complex categories. The grades of membership correspond to the degree of compatibility with the concepts represented by the class concerned: the direct evaluation of grades with adequate measures is a significant stage for subsequent decision-making processes.

After the production of soft results, a hardening process is sometime performed to obtain final crisp assignments to classes. This is done by applying appropriate ranking procedures and decision rules based on the inherent uncertainty and total amount of information dormant within the data, information that is lost when conventional classifications are considered.

The capacity of soft models has been proven empirically in many applications (Foody, 1996; Pal and Dutta Majumder, 1977).

However, despite the sizable achievements obtained, the use of soft classifiers is still limited by the lack of well-assessed and adequate methods for the evaluation of the accuracy of their outputs, an element of primary concern, which must be considered an integral part of the overall classification procedure.

Accuracy is generally assessed empirically by selecting a sample of reference data and comparing their actual class assignments with those provided by the automated classifier. The measures of accuracy employed in the evaluation of a classification are usually those derived for application to "crisp" classification outputs.

One of the most common ways of representing accuracy assessment information is in the form of an error matrix, or contingency table (Congalton, 1991). Using an error matrix to represent accuracy has been recommended by many researchers, as it provides a detailed assessment of the agreement between the sample reference data and classification data at specific locations, together with a complete description of the misclassifications registered for each category. In addition to the valuable role of the full error matrix, a number of descriptive and analytical statistical techniques based on the error matrix have been proposed (Congalton, 1991) to summarize information and obtain accuracy measures that can meet specific objectives. The most commonly used are the "overall proportion of sample data classified correctly", user's and producer's accuracy, various forms of kappa (κ) coefficients of agreement, the τ coefficient (Stehman, 1997). Each of these provides a different summary of the information contained in the error matrix. Because all these indexes suffer from some limitation and no consensus has been reached on the most suitable measure for a given evaluation objective, the error matrix should be considered the basic descriptive tool for organizing and presenting accuracy information and should be reported whenever feasible.

Unfortunately, in their present form the error matrix and the derived accuracy measures are appropriate only for hard classification. The underlying assumption of these conventional measures is that each element of sample data is associated with only one class in the classification and only one class in the reference data. Consequently, a class assignment is judged exactly right, or exactly wrong.

In soft classification, gradual membership in several classes is allowed for each element of sample data and assignments to classes are judged correct, or incorrect in varying degrees. But to apply conventional measures of classification accuracy, these soft classification outputs must be hardened and the comparison limited to crisp reference data, causing a general loss of information. The accuracy derived does not necessarily reflect how correctly the strength of class membership has been partitioned among the classes. Consequently, what is needed are soft accuracy statements that can adequately keep track of the uncertainty expressed in reference and classification data by extending the notion of crisp matching to that of soft matching.

This paper proposes a new evaluation method which uses the fuzzy set theory to extend the applicability of the traditional error matrix method to the evaluation of soft classifiers. It is designed to cope with those situations in which classification and/or reference data are expressed in multimembership form and the grades of membership represent different levels of approximation to intrinsically vague classes. The method assumes that membership values in classes are known for the set of reference data. Real applications may call for different approaches to the derivation of these values. In the context of remote sensing image classification, for example, the grades of membership in a given land cover class are correlated with the percentages of coverage within pixels; the membership values for the reference data are subpixel land cover estimations of ground truth which can be obtained by labeling procedures based on images of varied resolution (Binaghi and Rampini, 1993). More generally, in supervised soft classification, as classes are intrinsically vague, the "hard" labeling, traditionally used in the construction of the reference data set, appears difficult to perform, artificial, or both. When dealing with vague classes, experts may more naturally use qualitative linguistic, scales to measure the strengths of membership. Grades of membership for reference data may then be obtained by defining these linguistic labels as labels of fuzzy sets and applying appropriate elicitation techniques for the definition of the corresponding membership functions (Hall et al., 1986).

Basing the evaluation method on the error matrix preserves the property of error "localization" consisting in "the capability of identifying the contribution of each category relative to the actual category as verified in the reference data" (Congalton, 1991). The derived descriptive techniques are reformulated in the light of the fuzzy set theoretical framework.

As with any fuzzy extension of traditional concepts and operators, when the range of membership grades is restricted to the set $(0, 1)$, the fuzzy error matrix performs as precisely as the corresponding traditional matrix and is a clear generalization of the latter.

2. Previous work

Various investigations have been made, and several approaches suggested in the literature in an attempt to overcome shortcomings in the evaluation of soft classification. Since soft classification output explicitly represents some kind of uncertainty in class assignment, measures based on information uncertainty seem the most appropriate, and particular emphasis has been placed on *fuzzy measures, measures of fuzziness* and *classical measures of uncertainty*, such as Shannon entropy (Klir and Folger, 1988).

Many authors, especially those operating in the field of soft land-cover mapping, measure the correlation between the proportions of corresponding memberships of reference and classification data by means of the coefficient of determination (r^2) or of correlation (r).

Another, more empirical approach proposed for the evaluation of soft classification accuracy is to simply measure the *distance* between classification and reference data without referring to specific uncertainty management frameworks (Foody, 1996). "*Fuzzy distances*", "*fuzzy similarity relations*", may be also appropriate in this context (Miyamoto, 1990).

Although all of the methods derived from these approaches have something to offer, none is at the moment universally applicable. Advantages and disadvantages coexist in each of them. Some methods, such as *entropy*, are only appropriate for situations in which the output of the classification is soft and the reference data are crisp. Inversely, other methods can be applied in those situations in which the uncertainty lies in reference data and not in the classification data (Gopal and Woodcock, 1994). Classical measures of uncertainty are rooted in a statistical framework based on assumptions that cannot be shared by soft classification. These measures may fail to appropriately represent the accuracy of results produced in non-probabilistic frameworks.

Measures of closeness may suffer from heuristic solutions and lack of information concerning the sampling design (Foody, 1996). Other, global approaches, such as the measure of determination, or of correlation do not necessarily reflect the reality of specific locations.

Finally, a limitation common to all these alternative approaches which go beyond the error matrix is that they do not provide "location preserving" accuracy. In a "non-location preserving" accuracy assessment where the amounts of a category are considered without regard for the location the accuracy assessment, if all the errors balance out, yields very high and misleading results (Congalton, 1991).

This situation has convinced many authors to organize hybrid evaluation strategies that contemplate the complementary use of a combination of evaluation methods (Binaghi et al., 1998), but this in turn means abandoning the advantages of an organic, easily interpreted method that expresses accuracy in the form of a single, or a few meaningful indexes.

3. The fuzzy error matrix

An error matrix is a square array of integer numbers set out in rows and columns that represent the number of sample units of the actual category assigned to a particular category. The columns usually represent the sample elements assigned to corresponding actual categories (reference data), while the rows indicate the sample elements assigned to corresponding classes by the classifier (classification data). In this matrix the diagonal elements show the number of sample elements which have been classified correctly, while off-the-diagonal elements represent misclassifications. We may formalize these concepts and the matrix building procedure in terms of classical set theory and derived set operations.

We let R_n be the set of reference data assigned to class n, and C_m the set of classification data assigned to class m, with $1 \leqslant n \leqslant Q$, $1 \leqslant m \leqslant Q$ and Q as the number of classes. $\{R_n\}$ and $\{C_m\}$ form two hard partitions of the sample data set X.

When dealing with conventional hard classification (crisp reference and classification data), R_n and C_m are assumed to be crisp sets. The process by which individuals from a given sample data set X are determined to be either members or nonmembers of the classes n and m is defined by the characteristic or discrimination function of the sets R_n and C_m,

$$\mu_{R_n} : X \to \{0, 1\}, \tag{1a}$$

$$\mu_{C_m} : X \to \{0, 1\}. \tag{1b}$$

For the given sets this function assigns values $\mu_{R_n}(x)$ and $\mu_{C_m}(x)$, respectively, to every $x \in X$ such that

$$\mu_{R_n}(x) = \begin{cases} 1 & \text{iff } x \in R_n, \\ 0 & \text{otherwise,} \end{cases} \tag{2a}$$

$$\mu_{C_m}(x) = \begin{cases} 1 & \text{iff } x \in C_m, \\ 0 & \text{otherwise.} \end{cases} \tag{2b}$$

The element of the error matrix M in row m and column n represents the cardinality of the intersection set $C_m \cap R_n$:

$$M(m, n) = |C_m \cap R_n| = \sum_{x \in X} \mu_{C_m \cap R_n}(x), \tag{3}$$

with the characteristic function:

$$\mu_{C_m \cap R_n}(x) = \begin{cases} 1 & \text{iff } x \in C_m \wedge x \in R_n, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The generic error matrix is shown in Table 1, where $p_{m,n}$ represents the cardinality of the intersection set $C_m \cap R_n$ computed according to (3); p_{i+} and p_{+i} are the total assignment to the ith class for classification and reference data, respectively.

Within the soft classification context, the vagueness conveyed by the grades of membership in classes leads us to conceive classification statements as *less exclusive* than in conventional hard classification and to compare them in the light of more relaxed, flexible conditions, which results in degrees of matching.

In fact, a conventional statistical framework and a fuzzy or soft classification framework represent classification outcomes, respectively, in the form:

the probability that the pattern x belongs to ith class is α where $\alpha \in [0, 1]$,

the grade of membership of pattern x in ith class is α where $\alpha \in [0, 1]$.

The first statement implies that the element x belongs totally to ith class and contributes to the entire set of patterns to a degree equal to α. The gradual value must be interpreted as the frequency of occurrence of this precise event, and the comparison must be made in terms of the total match or mismatch on final crisp memberships.

Table 1
Error matrix with p_{mn} representing the cardinality of the intersection between classification data (rows) and reference data (columns)

	Reference data				Total assignment
	1	2	···	Q	
Classification data					
1	p_{11}	p_{12}	···	p_{1q}	p_{1+}
2	p_{21}	p_{22}	···	p_{2q}	p_{2+}
⋮	⋮	⋮	···	⋮	⋮
Q	p_{q1}	p_{q2}	···	p_{qq}	p_{q+}
Total assignments					
	p_{+1}	p_{+2}	···	p_{+q}	

The second statement conveys information on the degree to which pattern x approximates the prototype of ith class. Comparison too is a matter of degrees.

Proceeding from these considerations, we now attempt to extend the applicability of the error matrix to the case of soft classifications. Since the concept of multiple and partial class memberships is fundamental to fuzzy set-based techniques, the extension is made according to the fuzzy set theory (Zadeh, 1965).

Within the soft classification context R_n and C_m may be considered fuzzy sets (we denote them as $\tilde{R}_n$ and $\tilde{C}_m$) having the membership function

$$\mu_{\tilde{R}_n} : X \to [0, 1], \tag{5a}$$

$$\mu_{\tilde{C}_m} : X \to [0, 1], \tag{5b}$$

where $[0, 1]$ denotes the interval of real numbers from 0 to 1 inclusive.

$\mu_{\tilde{R}_n}(x)$ and $\mu_{\tilde{C}_m}(x)$ represent the gradual membership of the sample element x in classes n and m as indicated in the reference and classification data, respectively. $\{\tilde{R}_n\}$ and $\{\tilde{C}_m\}$ form two fuzzy partitions of the sample data set X. In fuzzy classification the condition of *orthogonality* or *sum-normalization* is sometimes introduced (exemplified for the reference data set: $\sum_{n=1}^{Q} \mu_{R_n}(x) = 1$), requiring that membership functions sum up to one for each element of the sample data set (Pedrycz, 1990).

We use fuzzy set operators within the error matrix building procedure to provide a fuzzy error matrix $\tilde{M}$. The assignment to the element $\tilde{M}(m, n)$ involves the computation of the degree of membership in the fuzzy intersection set $\tilde{C}_m \cap \tilde{R}_n$.

For the intersection operation, several different classes of functions have been proposed in the literature (Dubois and Prade, 1985), and each can be considered in our context. However, despite the variety of fuzzy set operators, the *standard operations* of the fuzzy set theory still possess particular significance (Klir and Folger, 1988). We use here the “min” operator introduced in the original formulation of the theory of fuzzy sets (Zadeh, 1977):

$$\mu_{\tilde{C}_m \cap \tilde{R}_n}(x) = \min(\mu_{\tilde{C}_m}(x), \mu_{\tilde{R}_n}(x)). \tag{6}$$

The assignment to element $\tilde{M}(m, n)$ is an extension of Eq. (4).

The cardinality of the fuzzy set intersection $\tilde{C}_m \cap \tilde{R}_n$ provides the global value of the generic element in row m and column n computed on the overall sample data set:

$$\tilde{M}(m, n) = \left|\tilde{C}_m \cap \tilde{R}_n\right| = \sum_{x \in X} \mu_{\tilde{C}_m \cap \tilde{R}_n}(x). \tag{7}$$

In the case of multimembership, the generic element $p_{m,n}$ in Table 1 represents the cardinality of the intersection set $\tilde{C}_m \cap \tilde{R}_n$ computed according to (7). The element $p_{m,n}$ in the fuzzy error matrix denotes a crisp number, due to the fact that in (7) scalar cardinality is applied to the fuzzy set $\tilde{C}_m \cap \tilde{R}_n$. As in the conventional case, p_{i+} and p_{+i} represent the total grades of membership assigned to the ith class for classification and reference data, respectively.

The fuzzy error matrix can be used as the starting point for descriptive techniques, in the same way as the conventional error matrix. The simplest index in both cases is *overall accuracy* (OA). This is conventionally computed by dividing the sum of the major diagonal by the total number of sample elements. In the fuzzy case we proceed by dividing the sum of the major diagonal by the total of grades of membership found in reference data, interpreting OA as a measure of the total match between reference and classification data. When the condition of orthogonality holds, the total of grades of membership in the reference data coincides with the total number of sample elements. As in the conventional case, the accuracy of the individual categories is computed by dividing the corresponding element of the major diagonal by the total of grades of membership found in reference and classification data in either the corresponding column, or the corresponding row. For each category we obtain the *producer's accuracy* (PA), related to *errors of omission*, together with the *user's accuracy* (UA), related to *errors of commission*. All these measures, OA, PA and UA, are limited to the range $[0, 1]$ and assume the value of 1 in the case of a complete match between the gradual membership of reference and of classification data.

Some very simple examples suffice to illustrate the effects of applying the above operators to the construction of the error matrix (operating in multimembership it is possible to build a matrix for one sample element). The error matrices in Table 2 compare the class assignment provided in $\tilde{R}$ and that provided in $\tilde{C}$ for an element x and for a three class (q_1, q_2, q_3) problem.

- Case a:

$$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4,$$
$$\mu_{\tilde{C}_1}(x) = 0.4, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4.$$

- Case b:

$$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4,$$
$$\mu_{\tilde{C}_1}(x) = 0.2, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4.$$

- Case c:

$$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4,$$
$$\mu_{\tilde{C}_1}(x) = 0.6, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4.$$

Table 2
Fuzzy error matrices with accuracy descriptive measures for the three cases of (a) coincidence, (b) underestimation and (c) overestimation

	Reference data				Overall accuracy (OA)	
Class data	$\tilde{R}_1$	$\tilde{R}_2$	$\tilde{R}_3$	Total grades	Producer's acc.	User's acc.
(a) *Perfect matching* (OA = 1)						
$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4$						
$\mu_{\tilde{C}_1}(x) = 0.4, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4$						
$\tilde{C}_1$	**0.4**	0.4	0.4	0.4	$PA_1 = 1$	$UA_1 = 1$
$\tilde{C}_2$	0.4	**0.4**	0.4	0.4	$PA_2 = 1$	$UA_2 = 1$
$\tilde{C}_3$	0.4	0.4	**0.4**	0.4	$PA_3 = 1$	$UA_3 = 1$
Total grades	0.4	0.4	0.4			
(b) *Underestimation* (OA = 0.833)						
$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4$						
$\mu_{\tilde{C}_1}(x) = 0.2, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4$						
$\tilde{C}_1$	**0.2**	0.2	0.2	0.2	$PA_1 = 0.50$	$UA_1 = 1$
$\tilde{C}_2$	0.4	**0.4**	0.4	0.4	$PA_2 = 1$	$UA_2 = 1$
$\tilde{C}_3$	0.4	0.4	**0.4**	0.4	$PA_3 = 1$	$UA_3 = 1$
Total grades	0.4	0.4	0.4			
(c) *Overestimation* (OA = 1)						
$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4$						
$\mu_{\tilde{C}_1}(x) = 0.6, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4$						
$\tilde{C}_1$	**0.4**	0.4	0.4	0.6	$PA_1 = 1$	$UA_1 = 0.67$
$\tilde{C}_2$	0.4	**0.4**	0.4	0.4	$PA_2 = 1$	$UA_2 = 1$
$\tilde{C}_3$	0.4	0.4	**0.4**	0.4	$PA_3 = 1$	$UA_3 = 1$
Total grades	0.4	0.4	0.4			

In the fuzzy set classification framework membership grades are not necessarily constrained to sum up to 1 and represent the degree of compatibility of the element with each class.

Examining the error matrix values, we see that even when gradual memberships coincide for a given element of reference and classification data, non-null degrees of mismatch are still contemplated and represented in the matrix in the off-the-diagonal cells.

The fuzzy accuracy measures computed for the three cases of a, b and c are reported in Table 2. In case a the individual grades coincide, implying that the measures OA, PA and UA are equal to 1 for all classes. In case b a condition of underestimation is introduced; the OA is less than 1, as is the PA corresponding to the underestimated class ($\mathrm{PA}_1 = 0.50$).

In case c a condition of overestimation is introduced. Both the error matrix and the OA values are the same as those of case a: case c can be distinguished from case a only by evaluating the UA, which keeps track of the overestimation condition.

In Table 3 the fuzzy error matrices and derived accuracy measures are shown for cases in which the condition of orthogonality is introduced: three cases, respectively, of coincidence (a), underestimation (b) and overestimation (c) are considered.

When the orthogonality hypothesis holds, both the fuzzy error matrix and the OA can distinguish situations of coincidence and overestimation. The OA is the same ($\mathrm{OA} = 0.9$) in both cases b and c consistent with the interpretation of OA as the "overall proportion of grades classified correctly": the amount of underestimated grades in case b is equal to the amount of overestimated grades in case c. Category measures PA and UA capture the corresponding errors of omission and commission in cases b and c, respectively.

The results obtained in these hypothetical examples demonstrate that the accuracy measures derived from the fuzzy error matrix are an extension of the conventional measures based on crisp

Table 3
Fuzzy error matrices with accuracy descriptive measures for the three cases of (a) coincidence, (b) underestimation and (c) overestimation, assuming the orthogonality hypothesis

	Reference data				Overall accuracy	
Class data	$\tilde{R}_1$	$\tilde{R}_2$	$\tilde{R}_3$	Total grades	Producer's acc.	User's acc.
(a) *Perfect matching* ($\mathrm{OA} = 1$)						
$\mu_{\tilde{R}_1}(x) = 0.7,\quad \mu_{\tilde{R}_2}(x) = 0.2,\quad \mu_{\tilde{R}_3}(x) = 0.1$						
$\mu_{\tilde{C}_1}(x) = 0.7,\quad \mu_{\tilde{C}_2}(x) = 0.2,\quad \mu_{\tilde{C}_3}(x) = 0.1$						
$\tilde{C}_1$	**0.7**	0.2	0.1	0.7	$\mathrm{PA}_1 = 1$	$\mathrm{UA}_1 = 1$
$\tilde{C}_2$	0.2	**0.2**	0.1	0.2	$\mathrm{PA}_2 = 1$	$\mathrm{UA}_2 = 1$
$\tilde{C}_3$	0.1	0.1	**0.1**	0.1	$\mathrm{PA}_3 = 1$	$\mathrm{UA}_3 = 1$
Total grades	0.7	0.2	0.1			
(b) *Underestimation* ($\mathrm{OA} = 0.9$)						
$\mu_{\tilde{R}_1}(x) = 0.7,\quad \mu_{\tilde{R}_2}(x) = 0.2,\quad \mu_{\tilde{R}_3}(x) = 0.1$						
$\mu_{\tilde{C}_1}(x) = 0.6,\quad \mu_{\tilde{C}_2}(x) = 0.3,\quad \mu_{\tilde{C}_3}(x) = 0.1$						
$\tilde{C}_1$	**0.6**	0.2	0.1	0.6	$\mathrm{PA}_1 = 1$	$\mathrm{UA}_1 = 0.86$
$\tilde{C}_2$	0.3	**0.2**	0.1	0.3	$\mathrm{PA}_2 = 0.67$	$\mathrm{UA}_2 = 1$
$\tilde{C}_3$	0.1	0.1	**0.1**	0.1	$\mathrm{PA}_3 = 1$	$\mathrm{UA}_3 = 1$
Total grades	0.7	0.2	0.1			
(c) *Overestimation* ($\mathrm{OA} = 0.9$)						
$\mu_{\tilde{R}_1}(x) = 0.7,\quad \mu_{\tilde{R}_2}(x) = 0.2,\quad \mu_{\tilde{R}_3}(x) = 0.1$						
$\mu_{\tilde{C}_1}(x) = 0.8,\quad \mu_{\tilde{C}_2}(x) = 0.1,\quad \mu_{\tilde{C}_3}(x) = 0.1$						
$\tilde{C}_1$	**0.7**	0.2	0.1	0.8	$\mathrm{PA}_1 = 0.87$	$\mathrm{UA}_1 = 1$
$\tilde{C}_2$	0.1	**0.1**	0.1	0.1	$\mathrm{PA}_2 = 1$	$\mathrm{UA}_2 = 0.50$
$\tilde{C}_3$	0.1	0.1	**0.1**	0.1	$\mathrm{PA}_3 = 1$	$\mathrm{UA}_3 = 1$
Total grades	0.7	0.2	0.1			

matches and mismatches: they register the gradual strengths in class assignment and express the way in which the strength of class membership is partitioned between the classes and how closely this represents the partitioning of class membership found in the reference data. The conventional question of "how coincident are classification and reference data" must be reformulated as "how close are the grades in class assignments for classification and reference data".

From a theoretical point of view, the comparison of grades of membership for classification and reference data could address another, independent question: "at what level of fuzziness/vagueness is the gradual matching performed". Consider for example the following cases:

- Case a:

$$\mu_{\tilde{R}_1}(x) = 0.4, \quad \mu_{\tilde{R}_2}(x) = 0.4, \quad \mu_{\tilde{R}_3}(x) = 0.4,$$

$$\mu_{\tilde{C}_1}(x) = 0.4, \quad \mu_{\tilde{C}_2}(x) = 0.4, \quad \mu_{\tilde{C}_3}(x) = 0.4.$$

- Case b:

$$\mu_{\tilde{R}_1}(x) = 0.8, \quad \mu_{\tilde{R}_2}(x) = 0.8, \quad \mu_{\tilde{R}_3}(x) = 0.8,$$

$$\mu_{\tilde{C}_1}(x) = 0.8, \quad \mu_{\tilde{C}_2}(x) = 0.8, \quad \mu_{\tilde{C}_3}(x) = 0.8.$$

The global and category measures are the same for the two cases and are all equal to 1. However, the data in case b are less fuzzy than in case a. The level of *fuzziness* can be quantified by introducing the *index of fuzziness* (Klir and Folger, 1988) defined for both classification and reference data in terms of the metric distance of fuzzy sets $\tilde{R}_i$ and $\tilde{C}_i$ from the nearest crisp set S, if any.

As an example, for $\tilde{R}_i$ we have

$$\mu_S(x) = \begin{cases} 0 & \text{if } \mu_{\tilde{R}_i}(x) \leqslant \frac{1}{2}, \\ 1 & \text{if } \mu_{\tilde{R}_i}(x) > \frac{1}{2}. \end{cases} \tag{8}$$

Using the Hamming distance, we express the normalized *index of fuzziness* (IF) of $\tilde{R}_i$, $\mathrm{IF}_{\tilde{\mathrm{R}}_i}$, by the function

$$\mathrm{IF}_{\tilde{R}_i} = \frac{\sum_{x \in X} \left|\mu_{\tilde{R}_i}(x) - \mu_S(x)\right|}{\left|\tilde{R}_i\right|}. \tag{9}$$

To have a global measure of fuzziness for the overall reference and classification data sets we introduce a *mean index of fuzziness* ($\overline{\mathrm{IF}}$) which, exemplified for the reference data set, has the following form:

$$\overline{\mathrm{IF}}_{\tilde{R}} = \frac{\sum_{i=1}^{Q} \mathrm{IF}_{\tilde{R}_i}}{Q}. \tag{10}$$

The quantification of the levels of fuzziness of the classification and reference data can serve as a concise indication of the behavior of a classifier and introduce additional criteria of evaluation particularly useful in comparative studies, as shown in the following example.

3.1. A real example

To verify the applicability of the method in a real domain and evaluate the effectiveness of the measures proposed when applied to real data sets, where cumulative and compensatory effects occur, we conducted a remote sensing study on a highly complex real scene of the Venice lagoon (Italy) where water and wetland merge into one another, at sub-pixel level, in an intricate and complex pattern. A Landsat TM image (30 m ground resolution) of the Venice lagoon (Fig. 1) was selected to study and compare the capabilities of both the fuzzy and neural network classifiers in estimating the mixture of water and wetland. The study area was a 15 km×12 km rectangular window in which water and wetland mainly represented the lagoon environment and were the land cover types of interest. Two other class vegetation and bare-soil were also present in the mainland part of the study area.

In this context the gradual membership values of a given pixel in land cover classes were drawn from the proportions of water and wetland occurring within that pixel. The condition of orthogonality was therefore automatically introduced. The membership grades of the reference data set were then generated with a specific procedure consistent with the above interpretation: color aerial photographs on a scale of 1:20 000 taken at the same time as the satellite overpass, and covering a part of the scene were digitized at 10 m ground resolution and geometrically registered to the satellite image using the nearest

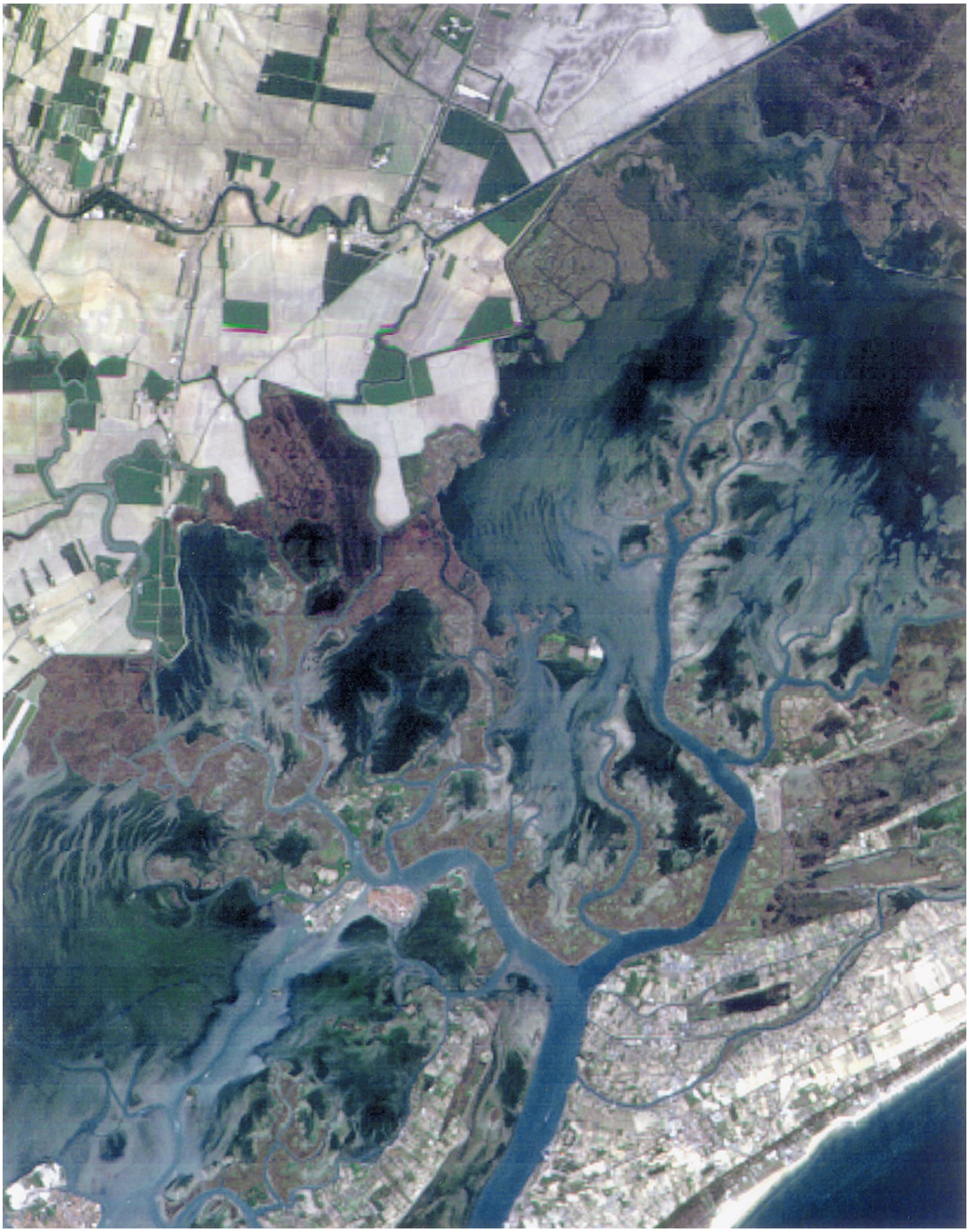

Fig. 1. North-eastern part of Venice lagoon showing the complex environment of the transitional zones between the mainland and the open sea, represented by wetlands. Landsat Thematic Mapper image of 8 May 1987 (RGB: TM3, TM2, TM1).

neighborhood resampling function. This allowed us to produce sub-pixel land cover information in terms of a 3-by-3 pixel grid. Reference data, including pure pixels (full membership in one land cover class) and water and wetland mixed pixels (gradual memberships in water and wetland), were produced for training and testing during a machine-aided session by a photointerpreter who assigned the land cover class proportions within the TM pixels, working from the corresponding 3-by-3 pixel grid of the aerial photograph (Binaghi et al., 1999).

The error matrices and the accuracy measures derived for the fuzzy-statistical and neural classifiers are shown in Table 4. To simplify the interpretation of the data in the error matrices and rapidly identify the allocation of gradual memberships in the classes involved in the mixture, that is water and wetland, the columns and rows related to vegetation and bare-soil have been combined in a single column and row labeled *other.* The fuzzy error matrix and accuracy measures for the ideal case in which reference data and classification data would compare exactly are stated first to facilitate the interpretation of the new measures. The OA of the neural classifier (OA = 0.74) is higher than that of the fuzzy statistical classifier (OA = 0.59). Examining the category measures, we note that the producer accuracy of the class water is very low for the fuzzy statistical classifier ($PA_1 = 0.22$). This indicates a consistent underestimation of the class water in favor of the class wetland which is consequently significantly overestimated ($UA_2 = 0.57$). Looking at the matrix values, we see that both the classifiers have misclassified pixels: the fuzzy statistical classifier limited the misclassification between the two classes involved in the mixture water and wetland, the neural classifier assigns degrees of membership either to the class *other.* These results, obtained by analyzing the accuracy measures, are consistently reflected in the indexes of fuzziness. There is less of a difference between the IF values of the reference data and classification data for the neural classifier ($\Delta = +25\%$) than for the fuzzy classifier ($\Delta = -84\%$). The IF provides additional information about the behavior of the two classifiers. The low IF values for the fuzzy statistical classifier

Table 4
Comparison of the results obtained from two classifiers, neural network and fuzzy statistical, in the pixel unmixing problem between water and wetland classes for a satellite image of the Venice lagoon

	Reference data				Overall accuracy		Index of fuzziness	
Class data	Water	Wetland	Other	Total grades	Producer's acc.	User's acc.	Per class IFs	
Complete matching (OA = 1) $IF_{\tilde{R}} = 0.155$, $IF_{\tilde{C}} = 0.155$								
$\tilde{C}_1$	**111.58**	74.36	0	111.58	$PA_1 = 1$	$UA_1 = 1$	$IF_{\tilde{R}_1} = 0.310$	$IF_{\tilde{C}_1} = 0.310$
$\tilde{C}_2$	74.36	**128.53**	0	128.53	$PA_2 = 1$	$UA_2 = 1$	$IF_{\tilde{R}_2} = 0.309$	$IF_{\tilde{C}_2} = 0.309$
$\tilde{C}_3$	0	0	**0**	0	$PA_3 = \#$	$UA_3 = \#$	$IF_{\tilde{R}_3} = \#$	$IF_{\tilde{C}_3} = \#$
Total grades	111.58	128.53	0					
Neural network classification (OA = 0.74) $IF_{\tilde{R}} = 0.155$, $IF_{\tilde{C}} = 0.194$								
$\tilde{C}_1$	**90.78**	86.59	0	116.25	$PA_1 = 0.81$	$UA_1 = 0.78$	$IF_{\tilde{R}_1} = 0.310$	$IF_{\tilde{C}_1} = 0.324$
$\tilde{C}_2$	71.48	**87.16**	0	97.12	$PA_2 = 0.68$	$UA_2 = 0.90$	$IF_{\tilde{R}_2} = 0.309$	$IF_{\tilde{C}_2} = 0.341$
$\tilde{C}_3$	26.34	26.21	**0**	26.75	$PA_3 = \#$	$UA_3 = 0$	$IF_{\tilde{R}_3} = \#$	$IF_{\tilde{C}_3} = 0.055$
Total grades	111.58	128.53	0					
Fuzzy statistical classification (OA = 0.59) $IF_{\tilde{R}} = 0.155$, $IF_{\tilde{C}} = 0.025$								
$\tilde{C}_1$	**24.34**	20.88	0	35.34	$PA_1 = 0.22$	$UA_1 = 0.69$	$IF_{\tilde{R}_1} = 0.310$	$IF_{\tilde{C}_1} = 0.051$
$\tilde{C}_2$	97.12	**117.53**	0	204.66	$PA_2 = 0.91$	$UA_2 = 0.57$	$IF_{\tilde{R}_2} = 0.309$	$IF_{\tilde{C}_2} = 0.051$
$\tilde{C}_3$	0	0	**0**	0	$PA_3 = \#$	$UA_3 = \#$	$IF_{\tilde{R}_3} = \#$	$IF_{\tilde{C}_3} = \#$
Total grades	111.58	128.53	0					

indicate that gradual class assignments tend definitely towards the extremes, 0 and 1, of the membership scale. The neural network classifier, better preserves the fuzziness of the mixed data. These results are confirmed by the visual inspection performed by photointerpreters who were experts of the scene of the soft maps produced by the two classifiers (Fig. 2).

Alternative evaluation procedures have been developed for this application in order to demonstrate the value and the advantages of the proposed measures as compared with other approaches. Table 5 shows the standard error matrices for the neural and fuzzy-statistical classifiers. The matrices have been constructed by hardening classification and reference data sets, and then performing traditional crisp matches. The conventional accuracy measures register values consistent with those of the new measures. However the differences in accuracy between the classifiers are reduced significantly: their OAs are quite similar. The loss of information on the distribution of the gradual strengths in class assignments has led to misleading results.

For a second comparison we applied the *standard errors of estimate* to the results obtained by the neural and fuzzy-statistical classifiers. This evaluation procedure preserves the gradual membership of the data in the classes: substantially what it does measure is the differences in the allocation of gradual memberships of reference data (y^r) and classification data (y^c) in classes:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i^r - y_i^c)^2}{n-2}}, \tag{11}$$

where n is the cardinality of the data set. If n represents the number of sample data belonging to a jth class, we obtain the *standard error of estimate* per class S_{e_j}.

This analysis has confirmed the superiority of the neural classifier in quantifying land cover proportions in mixed pixels. Table 5 lists the values, which for the neural network are in all cases half those for the fuzzy classifier. With this evaluation tool the differences between the two classifiers are once again significant, but, in contrast with the new evaluation method we cannot *localize* the misclassification, having lost information regarding *omissions* and *commissions* at the category level.

4. Conclusions

The new evaluation method we propose for soft classifiers is based on the fuzzy set theory and is a generalization of the traditional confusion matrix method. It is designed for those situations in which classification and/or reference data are expressed in multimembership form. Even when the membership grades of classification and/or reference data are crisp, i.e. restricted to the set $(0, 1)$, the method continues to be applicable. When both the reference and classification data are crisp, the new method performs precisely like the traditional error matrix method. We have demonstrated the suitability of fuzzy sets in accuracy assessment by defining a fuzzy error matrix and deriving global and category measures.

The approach has been illustrated using simple examples of the effects of applying the above measures. An application in the evaluation of remote sensing image classification has been reported to illustrate the tractability and the effectiveness of the new approach in a real domain. To better demonstrate their advantages, the new evaluation tools have been compared with other, conventional approaches. The results show that the accuracy information the proposed procedure provides, consistently reflects how correctly the strength of class membership is partitioned among classes.

As the method is based on the error matrix, it inherits the advantages and disadvantages of this evaluation tool.

The important property of localizing misclassification errors, on the basis of the contributions per-class distributed within the matrix and summarized in the category measures, is maintained. The comparative analysis performed shows that the OA index obtained from the fuzzy error matrix, which keeps track of the multimembership grades, provides a more appropriate measure of accuracy than the conventional OA values based on hardened data. Although, as illustrated in the

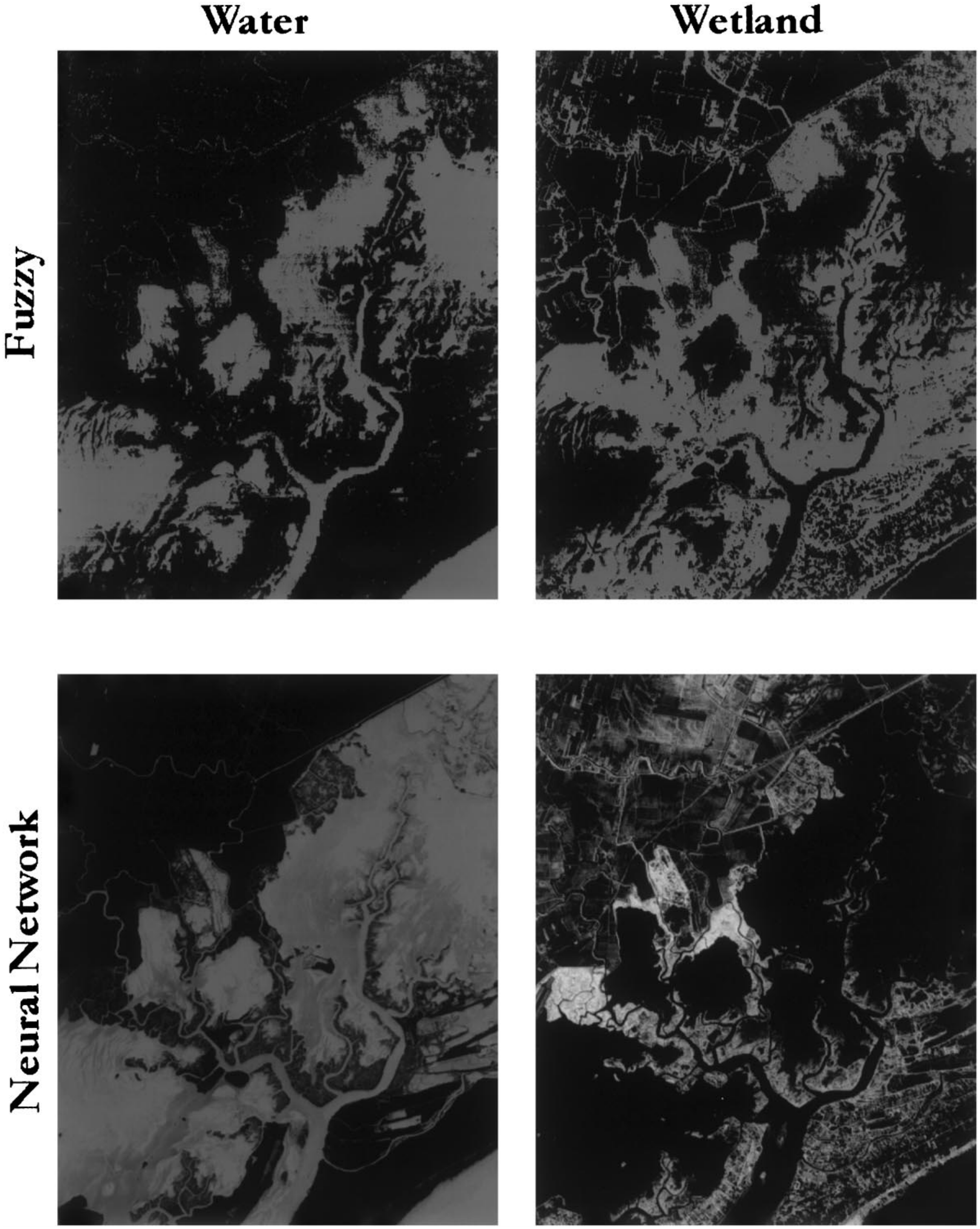

Fig. 2. Venice lagoon study area: maps of component proportions for the classes water and wetland obtained by the fuzzy statistical and neural network classifications. The grey levels represent the percentages of the proportions within each pixel: black to white = 0–100%.

Table 5
Standard error matrices, obtained with hardening, for neural network and fuzzy statistical classifiers in the pixel unmixing problem between water and wetland classes for a satellite image of the Venice lagoon and standard errors of estimate between actual versus predicted land cover proportions

	Reference data			Total assignments	Overall accuracy	
Class data	Water	Wetland	Other		Producer's acc.	User's acc.
Neural network classification (OA = 0.64)						
$\tilde{C}_1$	**69**	51	0	120	$PA_1 = 0.67$	$UA_1 = 0.57$
$\tilde{C}_2$	33	**86**	0	119	$PA_2 = 0.63$	$UA_2 = 0.72$
$\tilde{C}_3$	1	0	**0**	1	$PA_3 = \#$	$UA_3 = 0$
Total assignments	103	137	0			
Fuzzy statistical classification (OA = 0.61)						
$\tilde{C}_1$	**21**	12	0	33	$PA_1 = 020$	$UA_1 = 0.64$
$\tilde{C}_2$	82	**125**	0	207	$PA_2 = 0.91$	$UA_2 = 0.60$
$\tilde{C}_3$	0	0	**0**	0	$PA_3 = \#$	$UA_3 = \#$
Total assignments	103	137	0			
Classification	Network	Fuzzy				
Standard errors of estimate						
Water	24.85	46.42				
Wetland	26.68	46.39				

examples, when the condition of orthogonality is relaxed, the OA measure may not be able to represent situations of overestimation, these are, however, always correctly represented in the corresponding category measures.

As in the conventional procedure, neither the overall or the category indexes give equal consideration to the information contained in all of the cells of the error matrix.

At present, the best course of action to obtain all the accuracy information is to support the interpretation of the descriptive measures with a detailed inspection of the full fuzzy error matrix.

The proposed measure of fuzziness, computed in terms of *index of fuzziness*, has been proved a useful tool for investigating the behavior of a classifier in partitioning gradual class assignment, providing concise evaluation criterion that is especially useful in comparative studies and that can otherwise be obtained only by prolonged analysis (Binaghi et al. 1999).

We now plan to develop other measures derived from the fuzzy error matrix to keep track of all the information contained in the matrix, so that, in computing and comparing the various accuracy measures, we can examine the nature of the differences.

References

Binaghi, E., Rampini, A., 1993. Fuzzy decision making in the classification of multisource remote sensing data. Optical Engineering 6, 1193–1203.

Binaghi, E., Rampini, A., Brivio, P.A., Schowengerdt, R.A. (Eds.), 1996. Special Issue on Non-conventional Pattern Analysis in Remote Sensing. Pattern Recognition Letters 17 (13).

Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. Investigating the behaviour of neural and fuzzy-statistical classifiers in sub-pixel land cover estimations. Canad. J. Remote Sensing, to appear.

Bloch, I., 1996. Information combination operators for data fusion: a comparative review with classification. IEEE Trans. Systems Man. Cybernet. 26, 52–67.

Bouchon-Meunier, B., Yager, R., Zadeh, L.A. (Eds.), 1995. Fuzzy Logic and Soft Computing. World Scientific, Singapore.

Congalton, R.G., 1991. A review of assessing the accuracy of classification of remotely sensed data. Remote Sensing Environ. 37, 35–46.

Dubois, D., Prade, H., 1985. A review of fuzzy set aggregation connectives. Inform. Sci. 36, 85–121.

Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. Internat. J. Remote Sensing 17 (7), 1317–1340.

Gopal, S., Woodcock, C., 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogrammetric Engineering & Remote Sensing 60 (2), 181–188.

Hall, L.O., Szabo, S., Kandel, A., 1986. On the derivation of memberships for fuzzy sets in expert systems. Inform. Sci. 40, 39–52.

Ishibuchi, H., Nozaki, K., Tanaka, H., 1993. Efficient fuzzy partition of pattern space for classification problems. Fuzzy Sets and Systems 59, 295–304.

Klir, J.G., Folger, T.A., 1988. Fuzzy Sets, Uncertainty and Information. Prentice Hall, Englewood Cliff, NJ.

Miyamoto, S., 1990. Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer Academic Publishers, The Netherlands.

Pal, S.K., Dutta Majumder, D., 1977. Fuzzy sets and decision-making approaches in vowel and speaker recognition. IEEE Trans. Systems Man. Cybernet. 7, 625–629.

Pedrycz, W., 1990. Fuzzy sets in pattern recognition: methodology and methods. Pattern Recognition 23, 121–146.

Schowengerdt, R.A., 1996. On the estimation of spatial-spectral mixing with classifier likelihood functions. Pattern Recognition Letters 17, 1379–1387.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing Environ. 62, 77–89.

Wilkinson, G.G., 1996. Classification algorithms – where next?. In: Binaghi, E., Brivio, P.A., Rampini, A. (Eds.), Soft Computing in Remote Sensing Data Analysis. Series in Remote Sensing, Vol. 1, pp. 93–100.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.

Zadeh, L.A., 1977. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–28.