Accuracy Evaluation of Soft Classifiers using Interval Type-2 Fuzzy Sets Framework

Elisabetta Binaghi
Dept. of Theoretical and Applied Sciences
University of Insubria
Varese, Italy
Email: elisabetta.binaghi@uninsubria.it

Alberto A. Vergani
Dept. of Theoretical and Applied Sciences
University of Insubria
Varese, Italy
Email: aavergani@uninsubria.it

Valentina Pedoia
Dept. of Radiology
UCSF School of Medicine
San Francisco, California
Email: Valentina.Pedoia@ucsf.edu

***Abstract*—This paper proposes a new accuracy evaluation method within a behavioral comparison strategy which uses interval type-2 fuzzy sets and derived operations to model reference data and define soft accuracy indexes. The method addresses the case in which grades of membership, collected by surveying experts, will often be different for the same reference pattern, because the experts will not necessarily be in agreement. The approach is illustrated using simple examples and an application in the domain of biomedical image segmentation.**

I. Introduction

Soft models for classification have had an enormous impact in many fields of application. The most common solutions adopt fuzzy or hybrid, neuro-fuzzy frameworks for modeling the gradual transition from membership to non-membership in intrinsically vague classes [1]–[4]. In some situations, a hardening process is performed, after a soft stage, for the production of final useful results. In other cases, the grades of membership are directly used as final results or as partial, but semantically defined outcomes, furtherly processed in subsequent decision-making phases. 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 and constitute the final result of unmixing procedures [5]. In MRI brain tumor segmentation, the final grade of membership in infiltrative tumor classes, represents the extent of the infiltration and is used in further analysis to select the most appropriate treatment and to plan the best surgical approach [6]. Furthermore, content based filtering techniques that make use of soft text classifiers, provide list of grades that can be used directly by the subsequent phases of the filtering processes [7]. The accuracy evaluation procedures for all these cases must be based on soft accuracy statements to keep track properly of the uncertainty expressed in classification and/or reference data [6], [8]. Even though a large number of evaluation criteria and accuracy indexes for classification have been proposed in the literature, the extension of the notion of crisp matching to that of soft matching has received lot less attention so far. In most cases the measures of accuracy employed in the evaluation of a soft classification, are usually those derived for the evaluation of "crisp" classification outputs. But when conventional measures of classification accuracy are applied, soft classification outputs are hardened and the comparison limited to crisp reference data. This procedure causes a loss of information, unacceptable for those applications that consider grades of memberships their final outputs. The accuracy derived does not necessarily reflect how correctly the strength of class membership has been partitioned among the classes. 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. In many relevant contexts, a direct objective measure of reference standard representative of true classification is not logistically feasible and accuracy assessment is performed according to behavioral comparison strategies [9], [10]. Automated results are then compared with those produced by experts to decide if the automated classifier can be considered an acceptable substitute of human operators. Basing on a behavioral comparison strategy, the evaluation involves two important components: first the choice of appropriate accuracy indexes and second, the definition of a suitable reference standard through combining expert classification judgments. The most appropriate way to carry out both these aspects is still under investigation [10] especially in case of soft classifications applied to complex domains with stringent requirements of reliability. Soft accuracy has been studied by some researchers [11], [8] [12], [6], but the results of their studies have not been widely adopted. Fuzzy accuracy indexes derived from confusion matrix [13] and a fuzzified version of Jaccard Similarity Coefficient oriented to spatial overlap evaluation in image segmentation studies are proposed in our previous works [8], [6]. The second aspect involved in behavioral comparison concerns the creation of a reference standard representative of a common agreement and involves fusion strategies of expert reference data. The underlying assumption is that the merged result would be more accurate and representative of a common agreement, with respect to the exact sought classification than each individual inputs. Majority Voting is the widespread used rule to fuse classification results [14], but more sophisticated methods have also been proposed and applied in many context such as biomedical image segmentation studies. A well-known technique based on the Expectation Maximization framework, called STAPLE, is initially proposed by Warfield et al. [9], and used in a variety of studies. However, these solutions naturally apply to crisp classification results and force a unified reference data representation without considering the inherent uncertainty at the base of the disagreement among expert judgments.

This paper proposes a new accuracy evaluation method within a behavioral comparison strategy which uses interval type-2 fuzzy sets (IT2 FSs) and derived operations [15], [16] to both model reference data derived from experts that not all agree, and define soft accuracy indexes. The proposed framework extends the applicability of the traditional confusion matrix method and derived indexes to the evaluation of soft classifiers. The solutions investigated in this paper are an extension of those presented in a previous work [8] in which an accuracy assessment method based on type-1 fuzzy sets (T1 FSs) was illustrated. The extended method, proposed in the present work, specifically addresses the case in which grades of membership, collected by surveying experts, will often be different for the same reference pattern, because the experts will not necessarily be in agreement. The high level of inter, intra-variability in expert judgments creates limitations in using T1 FS framework that not have enough degrees of freedom to model disagreement in reference data and to compute unified similarity indexes. Using T1 FS-based accuracy assessment, arbitrary fusion procedures or multiple evaluations based on individual inputs must be necessarily managed. Expert inter, intra- variability must be conceived as a further level of uncertainty that can be properly modeled using IT2 FS-based framework.

II. IT2 Fuzzy confusion matrix

A confusion or error matrix [13] is a square array of numbers set out in rows and columns which express the number of sample units assigned to a particular category relative to the actual category (Table I). It provides a detailed assessment of the agreement between reference data and classification data at specific locations, together with a complete description of the misclassifications registered for each category.

TABLE I: Error matrix with p_{mn} representing the number of sample units assigned to *m* category relative to the actual category *n*

Class. Data	Reference Data			
	1	2	...	q
1	p_{11}	p_{12}	...	p_{1q}
2	p_{21}	...	...	p_{2q}
...	p_{31}	...	...	...
q	p_{q1}	p_{q2}	...	p_{qq}

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). The diagonal elements show the number of sample elements which have been classified correctly, while off-the-diagonal elements represent misclassifications.

When dealing with conventional hard classification, reference and classification data are assumed to be crisp sets. 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 $1 \leq n \leq Q, 1 \leq m \leq Q$). The element of the error matrix in row m and column n, $p_{m,n}$, represents the cardinality of the intersection set $R_n \cap C_m$. We now consider soft classifiers and assume that during the accuracy evaluation process, gradual class assignments are extracted from a group of experts who may not all agree in all the steps of the reference estimation process and may provide different gradual class memberships for the same pattern. In the most general case, even the soft classifier may produce more than one gradual class assignment for each pattern. It is the case, for example, of a semi-automated classifier that bases classification on multiple ambiguous user defined seeds. We address these situations by modelling reference and classification data as IT2 fuzzy sets (we denote them $\widetilde{R}_n$ and $\widetilde{C}_m$ respectively). Letting X the sample data set (discrete universe of discourse), selected for the accuracy evaluation procedure and Q the number of classes, the process by which experts assign the element $x \in X$ to actual category n and the soft classifier assign it to classes m, produces two IT2 Fuzzy sets $\widetilde{R}_n$ (reference data) and $\widetilde{C}_m$ (classification data) respectively, expressed as follows [15] :

$$\widetilde{R}_n = \sum_{x\in X} \sum_{u \in J_x^{\widetilde{R}^n}} \mu_{\widetilde{R}^n}(x,u)/(x,u) = \sum_{x\in X} \sum_{u \in J_x^{\widetilde{R}^n}} 1/(x,u) \tag{1}$$

$$\widetilde{C}_m = \sum_{x\in X} \sum_{u \in J_x^{\widetilde{C}^m}} \mu_{\widetilde{C}^m}(x,u)/(x,u) = \sum_{x\in X} \sum_{u \in J_x^{\widetilde{C}^m}} 1/(x,u) \tag{2}$$

where $u \in J_x^W \subseteq [0,1], W = \left\{\widetilde{R}^n, \widetilde{C}^m\right\}$.

We can also express (1) and (2) as

$$\widetilde{R}_n = 1/FOU(\widetilde{R}_n) \tag{3}$$

$$\widetilde{C}_m = 1/FOU(\widetilde{C}_m) \tag{4}$$

The *footprint of uncertainty* (FOU) of R_n and C_m conveys the uncertainty of their primary memberships and it is defined as

$$FOU(\widetilde{R}_n) = \bigcup_{\forall x \in X} J_x^{\widetilde{R}_n} = \{(x,u) : u \in [0,1]\} \tag{5}$$

$$FOU(\widetilde{C}_m) = \bigcup_{\forall x \in X} J_x^{\widetilde{C}_m} = \{(x,u) : u \in [0,1]\} \tag{6}$$

$J_x^{\widetilde{R}_n}$ and $J_x^{\widetilde{C}_m}$ can be also expressed as

$$J_x^{\widetilde{R}_n} = \left\{\underline{\mu}_{\widetilde{R}_n}(x), ..., \overline{\mu}_{\widetilde{R}_n}(x)\right\} \forall x \in X \tag{7}$$

$$J_x^{\widetilde{C}_m} = \left\{\underline{\mu}_{\widetilde{C}_m}(x), ..., \overline{\mu}_{\widetilde{C}_m}(x)\right\} \forall x \in X \tag{8}$$

where $\underline{\mu}_W(x)$ and $\overline{\mu}_W(x)$ (with $W = \{R^n, C^m\}$) are *upper membership function* (UMF) and *lower membership function* (LMF) respectively.

FOU($\widetilde{R}_n$) provides a unified representation of expert judgments, preserving at the same time their inherent inter-, intra-variability. This representation creates the premise to overcome the difficulty of performing accuracy evaluation separately, one for each expert, and to eliminate the drawback of managing multiple evaluation scores or fusing expert judgments arbitrarily as in the case of T1 FS-based representation. We use operators and uncertainty measures of IT2 FSs to formalize the notions of IT2 fuzzy error matrix $\widetilde{M}$ and to derive accuracy indexes [15] [16]. Consistently with the conventional crisp case and with the T1 FS-based method, the assignment to the element $\widetilde{M}(m,n)$ involves the computation of the cardinality of the intersection set $\widetilde{R}_n \cap \widetilde{C}_m$. According to Mendel et al. [15] intersection between reference and classification data set is defined, $\forall x \in X$, as

$$\widetilde{R}_n \cap \widetilde{C}_m = 1/\left\{\underline{\mu}_{\widetilde{R}_n}(x) \wedge \underline{\mu}_{\widetilde{C}_m}(x), ..., \overline{\mu}_{\widetilde{R}_n}(x) \wedge \overline{\mu}_{\widetilde{C}_m}(x)\right\} \tag{9}$$

The global value of the generic element in row m and column n of the IT2 fuzzy error matrix M is obtained by computing the cardinality of the IT2 FS intersection. The cardinality of T2 FSs has been proposed with a variety of operations [16], and each can be considered in our context. However, we orient our choice to formulations that are conceived as extension of the cardinality adopted in the previously proposed T1 FS-based error matrix [8]. According to De Luca and Termini [17] the cardinality of a fuzzy set A, also called the power of a T1 FS, is defined as the sum of all membership grades, i.e.,

$$p_{DT}(A) = \sum_{i=1}^{N} \mu_A(x_i) \tag{10}$$

where N is the number of elements in the domain of A. Assuming this formulation, Szmidt and Kacprzyks interval cardinality [18] for an IT2 FS $\widetilde{A}$ is

$$P_{SK} = \left[p_{DT}(\underline{\mu}_{\widetilde{A}}), p_{DT}(\overline{\mu}_{\widetilde{A}})\right] \tag{11}$$

A useful concept in our context is the *average cardinality* (AC) of $\widetilde{A}$, proposed by Vlachos and Sergiadis [19], which is defined as the average of its minimum and maximum cardinalities, i.e.,

$$AC(\widetilde{A}) = \frac{p_{DT}(\underline{\mu}_{\widetilde{A}}) + p_{DT}(\overline{\mu}_{\widetilde{A}})}{2} \tag{12}$$

Specializing the above concepts to our context we define the assignment of the generic element in row m and column n of the IT2 fuzzy error matrix $\widetilde{M}$ as

$$\widetilde{M}(m,n) = AC(\widetilde{R}_n \cap \widetilde{C}_m) = \frac{p_{DT(\underline{\mu}_{\widetilde{R}_n \cap \widetilde{C}_m})} + p_{DT}(\overline{\mu}_{\widetilde{R}_n \cap \widetilde{C}_m})}{2} \tag{13}$$

III. Accuracy Indexes

The confusion matrix can be used as a starting point for a series of descriptive statistical techniques. From the IT2 confusion matrix, accuracy indexes that meet specific objectives may be derived. The simplest index is *overall accuracy* $\widetilde{OA}$ which is conventionally computed by dividing the sum of the major diagonal by the total number of elements in the error matrix. Proceedings by extending the above concept to the IT2 FS-based framework, we obtain the following formulation for the corresponding $\widetilde{OA}$ index:

$$\widetilde{OA} = \frac{\sum_{i=1}^{Q}(\widetilde{M}(i,i))}{\sum_{j=1}^{Q}\widetilde{AC}(\widetilde{R}_j)} \tag{14}$$

For individual categories, we obtain, the *producer accuracy* $\widetilde{PA}$, related to errors of omission (underestimation), together with the *user accuracy* $\widetilde{UA}$, related to errors of commission (overestimation) as follows:

$$\widetilde{PA}_i = \frac{\widetilde{M}(i,i)}{\widetilde{AC}(\widetilde{R}_i)} \tag{15}$$

$$\widetilde{UA}_i = \frac{\widetilde{M}(i,i)}{\widetilde{AC}(\widetilde{C}_i)} \tag{16}$$

All these measures, $\widetilde{OA}$, $\widetilde{PA}$ and $\widetilde{UA}$, are limited to the range [0,1] and assume the value of 1 in case of complete match.

A. Illustrative Examples

We compare class assignments provided in reference and classification data for elements $(x_1; x_2)$, for a two class $(q_1; q_2)$ problem. Reference data are provided by two experts $(e_1; e_2)$. Two cases are considered distinguished by a different level of matching.

- Case A Perfect Matching

$$FOU(\widetilde{R}_1) = J_{x_1}^{\widetilde{R}_1} + J_{x_2}^{\widetilde{R}_1} = 0.5, 0.7 + 0.0, 0.5$$
$$FOU(\widetilde{R}_2) = J_{x_1}^{\widetilde{R}_2} + J_{x_2}^{\widetilde{R}_2} = 0.3, 0.5 + 0.5, 1.0$$

$$FOU(\widetilde{C}_1) = J_{x_1}^{\widetilde{C}_1} + J_{x_2}^{\widetilde{C}_1} = 0.5, 0.7 + 0.0, 0.5$$
$$FOU(\widetilde{C}_2) = J_{x_1}^{\widetilde{C}_2} + J_{x_2}^{\widetilde{C}_2} = 0.3, 0.5 + 0.5, 1.0$$

The symbol $+$ denotes the union.

The above expressions formalize the fact that expert e_1 provides the following gradual assignments:

- for the element $x_1 \rightarrow q_1$: 0.5; q_2: 0.3;
- for the element $x_2 \rightarrow q_1$: 0.0; q_2: 0.5.

Expert e_2 provides the following gradual assignments:

- for the element $x_1 \rightarrow q_1$: 0.7; q_2: 0.5;
- for the element $x_2 \rightarrow q_1$: 0.5; q_2: 1.0.

The classifier produces for x_1 and x_2 two classification results (we may hypothesize that this is due to two different parameter setting) that show a perfect match with the expert assignments.

- Case B Under/Overestimation

$$FOU(\widetilde{R}_1) = J_{x_1}^{\widetilde{R}_1} + J_{x_2}^{\widetilde{R}_1} = 0.3, 0.5 + 0.5, 0.7$$
$$FOU(\widetilde{R}_2) = J_{x_1}^{\widetilde{R}_2} + J_{x_2}^{\widetilde{R}_2} = 0.5, 0.7 + 0.3, 0.5$$

$$FOU(\widetilde{C}_1) = J_{x_1}^{\widetilde{C}_1} + J_{x_2}^{\widetilde{C}_1} = 0.0, 0.3 + 0.3, 0.5$$
$$FOU(\widetilde{C}_2) = J_{x_1}^{\widetilde{C}_2} + J_{x_2}^{\widetilde{C}_2} = 0.7, 1.0 + 0.5, 0.7$$

The above expressions means that expert e_1 provides

- for the element $x_1 \rightarrow q_1$: 0.3; q_2: 0.5;
- for the element $x_2 \rightarrow q_1$: 0.5; q_2: 0.3.

Expert e_2 provides

- for the element $x_1 \rightarrow q_1$: 0.5; q_2: 0.7;
- for the element $x_2 \rightarrow q_1$: 0.7; q_2: 0.5.

The classifier working with the first parameter's configuration, provides

- for the element $x_1 \rightarrow q_1$: 0.0; q_2: 0.7;
- for the element $x_2 \rightarrow q_1$: 0.3; q_2: 0.5.

Using the second configuration, the classifier provides

- for the element $x_1 \rightarrow q_1$: 0.3; q_2: 1.0;
- for the element $x_2 \rightarrow q_1$: 0.5; q_2: 0.7.

The fuzzy accuracy measures computed for *Case A* and *Case B* are reported in Table II. Matching values in the error matrix are computed using equation (9) and (13). Accuracy indexes are obtained by applying equations (14), (15), (16).

In *Case A* the individual grades coincide, implying that the measures, $\widetilde{OA}$, $\widetilde{PA}$ and $\widetilde{UA}$ for all classes, are equal to 1. In *Case B* a condition of mismatch is introduced; as expected, the $\widetilde{OA}$ is less than 1, as is the $\widetilde{PA}$ measure corresponding to the underestimated classes ($\widetilde{PA}_1 = 0.55$) and $\widetilde{UA}$ measure corresponding to the overestimated class ($\widetilde{PA}_2 = 0.68$). The results obtained in these simple examples demonstrate that the accuracy measures derived from the IT2 fuzzy confusion matrix adequately reflect how closely reference and classification data distribute gradual strength in class membership.

TABLE II: IT2 fuzzy error matrices with accuracy descriptive measures for two cases of (a) perfect match, (b) mismatch.

	Reference Data		Total Grades	Accuracy Indexes	
Class. Data	$\widetilde{R}_1$	$\widetilde{R}_2$		$\widetilde{OA}$=1	
<u>*Casa A*</u>				$\widetilde{PA}_1$=1	$\widetilde{UA}_1$=1
$\widetilde{C}_1$	0.85	0.65	0.85	$\widetilde{PA}_2$=1	$\widetilde{UA}_2$ =1
$\widetilde{C}_2$	0.65	1.15	1.15		
Total Grades	0.85	1.15			
<u>*Case B*</u>				$\widetilde{OA}$=0.77	
$\widetilde{C}_1$	0.55	0.55	0.55	$\widetilde{PA}_1$=0.55	$\widetilde{UA}_1$=1
$\widetilde{C}_2$	1.0	1.0	1.45	$\widetilde{PA}_2$=1	$\widetilde{UA}_2$=0.68
Total Grades	1.0	1.0			

IV. An application in real domain

To verify the effectiveness of the IT2 fuzzy accuracy measures proposed when applied to real data, we perform the accuracy evaluation of a soft classifier in the domain of biomedical image segmentation. Because of the difficulty in establishing a standard for this type of data, accuracy assessment is usually addressed using a group of experts to manually trace the boundaries of the regions of interest [10]. Automated results are then compared with those produced by manual labeling to decide whether the automated segmentation algorithm can be considered an acceptable substitute for human operators. Considering first the comparison of hard segmentations, the minimal problem of assessing the agreement between two binary maps R and S, which indicate reference and segmented data respectively, is done by measuring the number of pixels (or voxels) that both in R and S are labeled $Object$ or $Background$, the number of voxels in R labeled $Object$ and in S labeled $Background$ and vice-versa. These measures can be accommodated in a $2X2$ confusion matrix shown in Table III.

TABLE III: 2X2 Confusion Matrix assessing the agreement between two binary maps whose elements are labeled Object (*Obj*) and Background (*Bkg*)

	Obj	*Bkg*
Obj	p_{11}	p_{12}
Bkg	p_{21}	p_{22}

Basing on matrix in Table III, a large number of similarity indexes have been proposed in the literature to compare image segmentations [10] and measure their spatial overlap. Due to the large number of zeros in the binary maps, the *Jaccard coefficient* (JC) [20] is preferred to the simple overall accuracy index (OA) as it only evaluates the amount of overlap of the foreground (object) component. The JC index can be derived as follows:

$$\text{JC} = \frac{p_{11}}{p_{11} + p_{12} + p_{21}} \tag{17}$$

Reformulating the JC index as derived from the IT2 fuzzy confusion matrix $\widetilde{M}$ that accommodates object and background values as in the matrix in Table III, we obtain:

$$\widetilde{JC} = \frac{\widetilde{M}(1,1)}{\widetilde{M}(1,1) + \widetilde{M}(1,2) + \widetilde{M}(2,1)} \tag{18}$$

We use this similarity index together with the above defined $\widetilde{PA}$ and $\widetilde{UA}$ indexes to evaluate the accuracy of an automated segmentation of magnetic resonance (MR) brain images based on Fuzzy Connectedness [21]. The dataset used in our experiment, is composed of four FLAIR MR grey-scale volumes with the following acquisition parameters:

- grey-scale: 12 bit depth
- volume size: [432 x 432 x 300]
- slice thickness: 0.6 [mm]
- spacing between slices: 0.6 [mm]
- pixel spacing: (0.57, 0.57) [mm]

- repetition time: 8000 [ms]
- ECHO time: 282.89 [ms]

All volumes are altered by the presence of glial tumors. The objective of the segmentation is the delineation within the image data of pathological tissues (*Object*) and healthy tissues (*Background*). A team of five medical experts is asked to segment axial slices of these volume data by employing an image annotator [22]. The manual delineation process is affected by several levels of uncertainty related to the infiltrative nature of the pathology that causes smooth boundaries, imprecision in the annotation tool and different expert advices. For each image data, we obtain two sets of five different manually labeled soft maps for *Object* and *Background* respectively. Basing on them, two IT2 FSs are constructed, representing the reference data for the two classes concerned. Figure 1 shows the support of the five soft maps delineating the *Object* area, obtained by manual labeling and the corresponding soft map obtained by the automated *Fuzzy Connectedness* segmentation when processing one slices of an MRI volume.

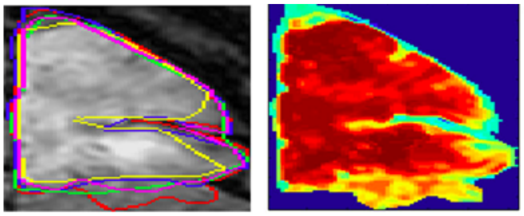

Fig. 1: From left to right: support of the *Object* maps manually traced by five experts and *Object* map produced by the automated segmentation.

Table IV(a) shows the values of the $\widetilde{JC}$ index and $\widetilde{PA}$ and $\widetilde{UA}$ indexes related to the Object category. These values are obtained by comparing automated results with reference data produced by experts. The accuracy values are computed by using equations (18), (15) and (16) respectively. 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 IV(b) shows the *Jaccard* (*JC*), *Producer* (*PA*) and *User* (*UA*) accuracy indexes computed according to the accuracy assessment method based on T1 FSs as proposed in our previous work [8]. By applying this method, the comparison between reference and automated results produces five accuracy values, one for each expert involved in the behavioral comparison. In Table IV(c) conventional crisp indexes (JC, PA and UA) are listed. These values have been computed by hardening classification and reference data sets, and then performing traditional crisp matches. Conventional procedure delimits mismatch to the comparison with reference data produced by the expert 3, detecting only a condition of underestimation (JC=0.66; PA=0.66; UA=1.0). T1 FS-based procedure finds misclassifications in the comparison with reference data produced by the experts 3, 4 and 5. In more detail the T1 FS-based procedure registers underestimation errors for the comparison with expert 3 (*JC*=0.55, *PA*= 0.69, *UA*=1.0) and both underestimations and overestimations for the comparison with experts 4 and 5 (*OA*=0.69, *PA*=0.69, *UA*=0.92). Category indexes indicate that the underestimated areas are larger than those overestimated. The IT2 fuzzy evaluation procedure unifies the five individual comparison in a single comparison and assigns just one score to quantify global similarity ($\widetilde{JC}$ = 0.60), overestimations ($\widetilde{UA}$ = 0.98) and underestimations ($\widetilde{JC}$ = 0.83) respectively. The accuracy indexes obtained by the three accuracy assessment procedures have been evaluated and compared by the experts that investigated additional clinical findings for the cases under study and acquired new insights about the distribution of healthy and pathological tissues in MR images. The comparative analysis of the accuracy evaluation indexes concluded that best results were obtained by the IT2 FS-based procedure. Conventional crisp measures are unable to detect completely the misclassifications occurred due to the loss of information on the distribution of the gradual strengths in class assignments caused by the hardening process. T1 FS-based accuracy evaluation framework works better detecting both underestimation and overestimation errors, but it is unable to fuse indexes derived from each expert in a single representative common index. This makes the interpretation of results difficult. The IT2 FS-based evaluation produces accuracy results that the experts judged reliable. Furthermore, single accuracy values representative of a common agreement, have been obtained and this facilitates the interpretation of the results for subsequent analysis and the application of ranking procedures when validation procedures involve several classifiers.

(a) IT2 accuracy indexes

$\widetilde{JC}$	$\widetilde{PA}$	$\widetilde{UA}$
0.60	0.83	0.98

(b) T1 accuracy indexes

	JC	*PA*	*UA*
Expert 1	1.0	1.0	1.0
Expert 2	1.0	1.0	1.0
Expert 3	0.55	0.69	1.0
Expert 4	0.69	0.69	0.92
Expert 5	0.69	0.69	0.92

(c) crisp evaluation

	JC	PA	UA
Expert 1	1.0	1.0	1.0
Expert 2	1.0	1.0	1.0
Expert 3	0.66	0.66	1.0
Expert 4	1.0	1.0	1.0
Expert 5	1.0	1.0	1.0

TABLE IV: Accuracy evaluation performed by comparing the Object map produced by the automated segmentation with Object reference maps, using Jaccard, Producer and User accuracy indexes derived from IT2 FS-based (a), T1 FS-based (b) and crisp (c) procedures.

V. Conclusion

In this paper, we proposed a new accuracy evaluation method within a behavioral comparison strategy. It is based on the IT2 fuzzy set theory and is conceived as generalization of the conventional and T1 fuzzy set-based method of accuracy evaluation based on confusion matrix. It is designed for those situations in which classification and/or reference data are expressed in multi-membership form and histogram of values are associated with single class assignments. We have demonstrated the suitability of the IT2 fuzzy sets and derived operations in accuracy assessment by defining a IT2 fuzzy confusion matrix and deriving global and category indexes by uncertainty measures for IT2 fuzzy sets. As seen in our experimental context, the proposed method consistently reflects how close are gradual memberships in reference and classification data. Ambiguity and/or disagreement in the assignment of grades are efficiently managed by using IT2 fuzzy sets. This representation allows to formally derive concise evaluation indexes that are especially useful in comparative studies and that can be otherwise obtained by prolonged and/or arbitrary analysis. Future work contemplates the use of other measures, such as measure of fuzziness, to address complementary questions in the accuracy evaluation process such as at what level of fuzziness/vagueness is the IT2 fuzzy set-based comparison performed. We also plan to better investigate the solutions proposed in biomedical image segmentation and in other domains characterized by the use of soft classifiers such as content-based filtering and text classification. The present work focused on the evaluation of soft classifications. The idea of using IT2 FSs to model uncertainty in the evaluation procedures may have interesting application in other domains such as control, information retrieval and decision making. These domains adopt specific accuracy indexes and it is interesting to investigate how they may be reformulated in an attempt to model inherent uncertainties in the validation processes.

ACKNOWLEDGMENT

The authors would like to communicate a message of gratitude to Pietro Alessandro Brivio (IREA-CNR, Italy) who stimulated and shared the birth of the first ideas on the use of fuzzy sets in the evaluation of soft classifiers

REFERENCES

[1] W. Pedrycz, "Fuzzy sets in pattern recognition: methodology and methods," *Pattern recognition*, vol. 23, no. 1-2, pp. 121–146, 1990.

[2] L. Kuncheva, *Fuzzy classifier design*. Springer Science & Business Media, 2000, vol. 49.

[3] E. Binaghi, P. Brivio, A. Rampini, and R. Schowengerdt, "Special issue on non-conventional pattern analysis in remote sensing," *Pattern Recognition Letters*, vol. 13, no. 17, pp. 1323–1324, 1996.

[4] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[5] A. Baraldi, E. Binaghi, P. Blonda, P. A. Brivio, and A. Rampini, "Comparison of the multilayer perceptron with neuro-fuzzy techniques in the estimation of cover class mixture in remotely sensed data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 5, pp. 994–1005, 2001.

[6] E. Binaghi, V. Pedoia, and S. Balbi, "Collection and fuzzy estimation of truth labels in glial tumour segmentation studies," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 4, no. 3-4, pp. 214–228, 2016.

[7] M. Vanetti, E. Binaghi, E. Ferrari, B. Carminati, and M. Carullo, "A system to filter unwanted messages from osn user walls," *IEEE Transactions on Knowledge and data Engineering*, vol. 25, no. 2, pp. 285–297, 2013.

[8] E. Binaghi, P. A. Brivio, P. Ghezzi, and A. Rampini, "A fuzzy set-based accuracy assessment of soft classification," *Pattern recognition letters*, vol. 20, no. 9, pp. 935–948, 1999.

[9] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[10] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M.-S. Koo, R. W. McCarley, and M. E. Shenton, "On evaluating brain tissue classifiers without a ground truth," *Neuroimage*, vol. 36, no. 4, pp. 1207–1224, 2007.

[11] S. Gopal and C. Woodcock, "Theory and methods for accuracy assessment of thematic maps using fuzzy sets," *Photogrammetric Engineering and Remote Sensing;(United States)*, vol. 60, no. 2, 1994.

[12] N. Ramli and D. Mohamad, "A function principle approach to jaccard ranking fuzzy numbers," in *Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of.* IEEE, 2009, pp. 324–328.

[13] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote sensing of environment*, vol. 37, no. 1, pp. 35–46, 1991.

[14] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[15] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 6, pp. 808–821, 2006.

[16] D. Wu and J. M. Mendel, "Uncertainty measures for interval type-2 fuzzy sets," *Information Sciences*, vol. 177, no. 23, pp. 5378–5393, 2007.

[17] A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory," *Information and control*, vol. 20, no. 4, pp. 301–312, 1972.

[18] E. Szmidt and J. Kacprzyk, "Entropy for intuitionistic fuzzy sets," *Fuzzy sets and systems*, vol. 118, no. 3, pp. 467–477, 2001.

[19] I. K. Vlachos and G. D. Sergiadis, "Subsethood, entropy, and cardinality for interval-valued fuzzy setsan algebraic derivation," *Fuzzy Sets and Systems*, vol. 158, no. 12, pp. 1384–1396, 2007.

[20] P. Jaccard, "The distribution of the flora in the alpine zone." *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[21] J. K. Udupa and P. K. Saha, "Fuzzy connectedness and image segmentation," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1649–1669, 2003.

[22] V. Pedoia, A. De Benedictis, G. Renis, E. Monti, S. Balbi, and E. Binaghi, "Manual labeling strategy for ground truth estimation in mri glial tumor segmentation," in *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*. ACM, 2012, p. 8.