8.07f/W(Jul162004) +model | NEUCOM : 10473 | Prod.Type:FTP pp.1–9(col.fig.:NIL) | ED:ArulselviU. PAGN:sree SCAN:

ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

NEUROCOMPUTING

ELSEVIER

Neurocomputing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

www.elsevier.com/locate/neucom

Fuzzy kappa for the agreement measure of fuzzy classifications ☆

Weibei Dou[a,b,*], Yuan Ren[a], Qian Wu[a], Su Ruan[c], Yanping Chen[d], Daniel Bloyet[b], Jean-Marc Constans[e]

[a]*Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China*
[b]*GREYC-CNRS UMR 6072, 6 Boulevard Maréchal Juin, 14050 Caen, France*
[c]*cCReSTIC, 9 Rue de Qubec,10026 Troyes, France*
[d]*Imaging Diagnostic Center, Nanfang Hospital Guangzhou, China*
[e]*Unité d'IRM, EA3916, CHRU, 14033 Caen, France*

Abstract

In this paper, we propose an assessment method of agreement between fuzzy sets, called fuzzy Kappa which is deduced from the concept of Cohen's Kappa statistic. In fuzzy case, the Cohen's Kappa coefficient can be calculated generally by transforming the fuzzy sets into some crisp α-cut subsets. While the proposed fuzzy Kappa allows to directly evaluate an overall agreement between two fuzzy sets. Hence, it is an efficient agreement measure between a given fuzzy "ground truth" or reference and a result of fuzzy classification or fuzzy segmentation. Based on membership function, we define its agreement function and its probability distribution to formulate the deduction of the expectation agreement. So the fuzzy Kappa is calculated from the proportion of the observed agreement and the agreement expected by chance. All the definitions and deductions are detailed in this paper. Both Cohen's Kappa and the fuzzy Kappa are then used to evaluate the agreement between a fuzzy classification of brain tissues on MRI images and its "ground truth". A comparison of the two types of Kappa coefficient is carried out and shows the advantage of the fuzzy Kappa and some limitations of Cohen's Kappa in the fuzzy case.

Keywords: Kappa statistic; Classification; Fuzzy; Agreement; Similarity; Assessment; Evaluation; Brain tissue; MRI

1. Introduction

Agreement measure is a very important issue just like the similarity measure for a decision of pattern recognition or information retrieval. Agreement measures are used frequently in reliability studies that involve categorical data [15] and also used to assess the quality of clustering classification [1]. The quality assessment of classification offers an explicit method to select a finite set of known objects from a potentially infinite set of unknowns [26]. It is a postclassification test to the underlying system for identifying the finite set of objects.

Because usually there is no "gold standard", or the truth of a given clinical classification system is not known, the diagnostic accuracy still remains a significant problem [27]. To assess the reliability of a classification system, the Kappa statistic was introduced by Cohen [7].

In 1960, Jacob Cohen [7] proposed firstly an agreement coefficient for nominal scales from the study of natural psychological measurement. This coefficient is called Kappa coefficient, and is a correlation-like coefficient of pairwise agreement [27], the observed proportion of agreement with agreement expected solely by chance. So this method of agreement measure is also called Kappa statistic. It was extended by Fleiss [10] as the weighted Kappa to assess the ordinal scale degrees of agreement (disagreement).

This concept of Cohen's proposition is being developed and being used more and more widely in various research domains. It is now a general approach for assessing the

☆Project NSFC60372023 supported by the National Natural Science Foundation of China.
*Corresponding author. Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China. Fax: +86 10 62770317.
E-mail addresses: douwb@tsinghua.edu.cn, wdou@greyc.ensicaen.fr (W. Dou).

0925-2312/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.neucom.2006.10.007

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

classification agreement and is applied in the field of electronics [17], Geographical information science [14], medical informatics [15] and clinic, bionomics [21], etc. Chen [5] uses Kappa statistic to measure the diversity/agreement between classifiers. A quantity serving this purpose is the measurement of the degree of agreement among dependent classifiers. Based on the concept of classifier fusion, the Kappa statistic is an informative measure of the strength of association (among dependent classifiers) in a number of different task domains and under varied conditions.

Carletta [3] has presented several variants of the Kappa coefficient in the literature: Scott's π [13] assess the agreement on move boundaries in monologues using action assembly theory; Krippendorff's α [19] is an extension of the argument from category data to interval and ratio scales; Siegel and Castellan's K [23] is used for category judgments in the assumptions under which an expected agreement is calculated. The advantages and disadvantages of different extensions of Kappa statistic have been discussed in many fields [2,12,18,24].

The chance-corrected agreement in the form of the Kappa statistic is easy to be calculated and used frequently based on its correspondence to an intraclass correlation coefficient, but its magnitude depends on the tasks and categories in the experiment [15].

Conventional approaches to accurately assess land cover maps use the Kappa statistic to quantify map quality by comparing classification results with independent ground-truth data [20]. For the similar application, Fritz et al. [11] propose a methodology for using fuzzy logic to capture the uncertainty in classification through the development of a fuzzy membership matrix which reflects the degree of difficulty in classifying different land cover types. The membership values are applied to a confusion matrix to produce a Kappa value that captures the uncertainty in classification, and a spatial representation of the uncertainty. But the Kappa statistic was for some tests that yielded numeric scores [8]. The fuzzy comparison yields a map for each cell the degree of similarity on a scale of [0,1]. Besides this spatial assessment of similarity an overall value for similarity is also derived [14].

Some methods of similarity measure have been proposed for fuzzy classification [16,4]. These methods focused on the similarity of element-to-element or fuzzy set-to-fuzzy set. Therefore, they cannot give an overall evaluation for some fuzzy classification that consist of multiple or especially uncountable fuzzy subsets. If we use directly the Kappa statistic for a fuzzy classification, a crisp set processing by selecting some thresholds, e.g. the α-cut, is necessary to granulate the result of fuzzy classification. The different granulation processing induces different Kappa coefficients, such as the measured agreement is dependent on the selection of thresholds. The multiple Kappa coefficients for the same fuzzy classification result in some difficulties or problems of evaluation. So an extension of Kappa statistic is needed for a fuzzy classification.

Hagen [14] proposed a fuzzy equivalent of Kappa statistic by using fuzziness of category. It assumes that each category definition exists in its intrinsic fuzziness, and some fuzzy classification results can be obtained by granulating a crisp classification. Then the agreement has been evaluated by using the fuzziness of location. Fundamentally, Hagen [14] proposed a comparison between crisp classifications. In addition, the fuzziness of location is not certainly evident in other domain, which limits its application.

Sousa et al. [25] have compared three assessment methods of agreement of fuzzy map (fuzzy classification at different resolutions): cell-by-cell, neighborhood hard (crisp) and soft comparison. In the cell-by-cell agreement between the two maps each cell is crisply classified, the measurement result contains information about only cell-by-cell agreement. Incorporating the neighborhood information into the comparison of categorical maps could be suitable for performing a hard or a fuzzy classification. But the hard classification or crisp process of a fuzzy classification has the disadvantage of modifying the maps before the comparison. However, by applying fuzzy classification for the comparison of categorical maps it is possible to obtain a special and gradual analysis of the similarity between two maps. The soft comparison would like to make a more accurate agreement of similarities. The choice of any of the three methods depends on applications and hence the less significance of choosing one of the them [25].

Our research aims to find a method of agreement measurement between two fuzzy clusters without using crisp process on fuzzy set. It will give an overall assessment about a fuzzy clustering by comparing with a reference cluster on the basis of one-by-one element of fuzzy set, e.g. pixel or voxel for image. It does not correspond to any crisp method. According to the concept of Kappa statistic, we find a deduction of the observed percentage of agreement P_o and the expected similarity P_e in the sense of fuzziness. In this paper, we firstly explain the meaning of Kappa statistic in the application of classification by the definition of an agreement function. Then we generalize the concept of the proportion of observed agreement and the proportion of random agreement by the definition of a fuzzy agreement function, which is based on membership function, to introduce an agreement assessment of fuzzy classification. This agreement assessment method is called fuzzy Kappa in this paper for knowing from Kappa statistic. Based on the proposed fuzzy Kappa, an overall assessment of the agreement of two fuzzy classifications can be obtained. A validation of the agreement measurement is given by the comparison of a fuzzy classification of brain tissues on MRI (magnetic resonance imaging) images with its reference fuzzy model.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

2. Agreement measurement by Cohen's Kappa

2.1. Meaning of agreement in classification

In the domain of traditional classification, a set $\mathbf{A}=\{x\}$ can be classified into N subsets noted as $\mathbf{A}=\bigcup_{i=1}^{N}A_i$ with $A_i\bigcap A_j=\phi$, $i,j=1,2,\ldots,N$, and $i\neq j$.

Let $u(x)$ in (1) be an eigenfunction of any element $x\in\mathbf{A}$ that represents the correlation of x and these subsets.

$$u_i(x)=\begin{cases}1 & \text{if } x\in A_i,\\ 0 & \text{otherwise.}\end{cases} \tag{1}$$

The property of $u_i(x)$ in (1) is

$$\sum_{i=1}^{N}u_i(x)=1,\quad x\in\mathbf{A}. \tag{2}$$

If $\mathbf{A}$ has been classified separately by two different classifiers $\mathscr{C}_1$ and $\mathscr{C}_2$, the eigenfunction of x is noted as $u_i^{\mathscr{C}_1}(x)$ and $u_i^{\mathscr{C}_2}(x)$. We define an agreement function for x, $f(x;u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2})$ to indicate the classification agreement of that any $x\in\mathbf{A}$ is classified in the A_i by the two different classifier. That is,

$$f(u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2})=\sum_{i=1}^{N}u_i^{\mathscr{C}_1}u_i^{\mathscr{C}_2}=\begin{cases}1 & \text{if } u_i^{\mathscr{C}_1}\neq 0\ u_i^{\mathscr{C}_2}\neq 0,\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

The properties of $f(u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2})$ are:

(1) $f(u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2})=1$, or 0;
(2) $f(u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2})=1$; if $\exists i$, $u_i^{\mathscr{C}_1}(x)=u_i^{\mathscr{C}_2}(x)=1$.

Thus, from (3) the proportion of observed agreement can be represented as P_{o}:

$$P_{\mathrm{o}}=\frac{1}{M}\sum_{m=1}^{M}f(x_m;u_i^{\mathscr{C}_1},u_i^{\mathscr{C}_2}), \tag{4}$$

where M denote the number of observed elements $x\in\mathbf{A}$ and m is the index of x, $m=1,2,\ldots,M$.

2.2. Representation of Cohen's Kappa in classification

The Cohen's Kappa is a measurement of agreement that compares the observed agreement to the expected agreement by chance if the observer ratings are independent. The Kappa coefficient as Eq. (5) indicates the proportionate reduction in error generated by a classification process, compared to the error of a completely random classification. Kappa coefficient $K=1$ means a perfect agreement and $K<1$ implies the proportion of error reduction compared with random classification.

$$K_{\mathrm{Cohen}}=\frac{P_{\mathrm{o}}-P_{\mathrm{e}}}{1-P_{\mathrm{e}}}, \tag{5}$$

where P_{o} is the proportion of observed agreement in (4), and P_{e} is the proportion of random agreement or the expectation of random classification.

Assume that $\mathscr{C}_1$ is independent to $\mathscr{C}_2$, for each observation element $x_m\in\mathbf{A}$, we can define the joined probability

$$p_{ij}^{\mathscr{C}_1,\mathscr{C}_2}=p_i^{\mathscr{C}_1}p_j^{\mathscr{C}_2}, \tag{6}$$

where $p_i^{\mathscr{C}_1}$ and $p_j^{\mathscr{C}_2}$ are boundary probability, $i,j=1,2,\ldots,N$, as shown:

$$p_i^{\mathscr{C}_1}=\frac{1}{M}\sum_{j=1}^{N}\sum_{m=1}^{M}u_i^{\mathscr{C}_1}(x_m)u_j^{\mathscr{C}_2}(x_m)=\frac{1}{M}\sum_{m=1}^{M}u_i^{\mathscr{C}_1}(x_m), \tag{7}$$

$$p_i^{\mathscr{C}_2}=\frac{1}{M}\sum_{j=1}^{N}\sum_{m=1}^{M}u_j^{\mathscr{C}_1}(x_m)u_i^{\mathscr{C}_2}(x_m)=\frac{1}{M}\sum_{m=1}^{M}u_i^{\mathscr{C}_2}(x_m). \tag{8}$$

In view of joined event $(\mathscr{C}_1,\mathscr{C}_2)$, the proportion of agreement expected by chance or random agreement P_{e} can be deduced from Eqs. (6), (7) and (8) as

$$P_{\mathrm{e}}=\sum_{i=1}^{N}\sum_{j=1}^{N}p_i^{\mathscr{C}_1}p_j^{\mathscr{C}_2}f(u_i^{\mathscr{C}_1},u_j^{\mathscr{C}_2})=\sum_{i=1}^{N}\sum_{u_i^{\mathscr{C}_1}=0}^{1}\sum_{u_i^{\mathscr{C}_2}=0}^{1}p_i^{\mathscr{C}_1}p_i^{\mathscr{C}_2}u_i^{\mathscr{C}_1}u_i^{\mathscr{C}_2}. \tag{9}$$

Some important properties of K_{Cohen} are:

(1) $K_{\mathrm{Cohen}}\leqslant 1$;
(2) $K_{\mathrm{Cohen}}=1$; if and only if, $u_i^{\mathscr{C}_1}(x)=u_i^{\mathscr{C}_2}(x)$, for $\forall i$ and $\forall x\in\mathbf{A}$;
(3) the symmetry property of K_{Cohen} is that $K(\mathscr{C}_1,\mathscr{C}_2)=K(\mathscr{C}_2,\mathscr{C}_1)$.

3. Fuzzy Kappa extended from Cohen's Kappa

3.1. Agreement of fuzzy classification

For a fuzzy classification, the observation spaces are fuzzy subclasses $A_i^{\mathscr{F}}\subset\mathbf{A}, i=1,2,\ldots,N$. These fuzzy subclasses are defined by using membership function $\mu_i(x)\in[0,1]$. It is a mapping of $\mu_i(x):\mathbf{A}\to[0,1]$.

If the membership function $\mu_i(x)$ is normalized as:

$$\sum_{i=1}^{N}\mu_i(x)=1,\ x\in\mathbf{A} \tag{10}$$

a fuzzy agreement function of two fuzzy classifications $\mu_i^{\mathscr{C}_1}(x)$ and $\mu_i^{\mathscr{C}_2}(x)$ for any element $x\in\mathbf{A}$, is introduced from (3)

$$f^{\mathscr{F}}(x)=\sum_{i=1}^{N}(\mu_i^{\mathscr{C}_1}(x)\wedge\mu_i^{\mathscr{C}_2}(x)). \tag{11}$$

The properties of the fuzzy agreement function $f^{\mathscr{F}}(x)$ are:

(1) $f^{\mathscr{F}}(x)\in[0,1]$;
(2) $f^{\mathscr{F}}(x)=1$, if and only if, $\forall i$, $\mu_i^{\mathscr{C}_1}(x)=\mu_i^{\mathscr{C}_2}(x)$.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

3.2. Fuzzy Kappa

An agreement assessment between two fuzzy sets named as fuzzy Kappa, is defined as follows which is extended from Kappa statistic or Cohen's Kappa. Let us firstly define the proportion of observed agreement in fuzzy classification, noted as $P_{\mathrm{o}}^{\mathscr{F}}$ which is introduced from Eqs. (4) and (11)

$$P_{\mathrm{o}}^{\mathscr{F}} = \frac{1}{M}\sum_{m=1}^{M} f^{\mathscr{F}}(x_m) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}(\mu_i^{\mathscr{C}_1}(x_m) \bigwedge \mu_i^{\mathscr{C}_2}(x_m)). \tag{12}$$

Assume that $\mu_i^{\mathscr{C}_1}(x_m)$ is independent to $\mu_i^{\mathscr{C}_2}(x_m)$. The expectation of random agreement, $P_{\mathrm{e}}^{\mathscr{F}}$ is the expectation of $f^{\mathscr{F}}(x)$ in (11), that is,

$$P_{\mathrm{e}}^{\mathscr{F}} = \sum_{i=1}^{N}\int_{\mu_i^{\mathscr{C}_1}=0}^{1}\int_{\mu_i^{\mathscr{C}_2}=0}^{1} p(\mu_i^{\mathscr{C}_1})p(\mu_i^{\mathscr{C}_2})(\mu_i^{\mathscr{C}_1} \bigwedge \mu_i^{\mathscr{C}_2})\,\mathrm{d}\mu_i^{\mathscr{C}_1}\,\mathrm{d}\mu_i^{\mathscr{C}_2}, \tag{13}$$

where $p(\mu_i^{\mathscr{C}_1})$ and $p(\mu_i^{\mathscr{C}_2})$ are the probability distribution of $\mu_i^{\mathscr{C}_1}(x)$, and $\mu_i^{\mathscr{C}_2}(x)$, respectively.

For comparing the $P_{\mathrm{o}}^{\mathscr{F}}$ and $P_{\mathrm{e}}^{\mathscr{F}}$ to (4) and (9), we can define the fuzzy Kappa as

$$K_{\mathrm{fuzzy}} = \frac{P_{\mathrm{o}}^{\mathscr{F}} - P_{\mathrm{e}}^{\mathscr{F}}}{1 - P_{\mathrm{e}}^{\mathscr{F}}}. \tag{14}$$

The fuzzy Kappa (14) has given the same meaning and formula as Cohen's Kappa (5). It takes the same properties as in Section 2.2. They are:

(1) $K_{\mathrm{Fuzzy}} \leqslant 1$;
(2) $K_{\mathrm{Fuzzy}} = 1$, if and only if $\mu_i^{\mathscr{C}_1}(x) = \mu_i^{\mathscr{C}_2}(x)$, for $\forall x \in \mathbf{A}$;
(3) the symmetry property is that $K_{\mathrm{Fuzzy}}(\mathscr{C}_1, \mathscr{C}_2) = K_{\mathrm{Fuzzy}}(\mathscr{C}_2, \mathscr{C}_1)$;
(4) if μ_i is a binary function, i.e. $\mu_i = 0$, or 1; the expectation of random agreement presented by Eq. (13) will retrogress to Eq. (9).

For demonstration the fuzzy Kappa's effect, a comparison experiment between Cohen's Kappa and the fuzzy Kappa is shown in the following section by an application of fuzzy classification of brain tissues on MR images.

4. Comparison experiment of the two types of Kappa

For answering the question of how to evaluate the performance of a fuzzy classification (or a fuzzy cluster), the fuzzy Kappa is perhaps a good solution to assess the agreement between the estimated fuzzy model and a standard fuzzy model. As an application example of the fuzzy Kappa, we present an experiment of fuzzy classification of brain tissues on MR images.

In this section, both the Cohen's Kappa and the proposed fuzzy Kappa are used to assess the agreement of the tested classifier and a reference. By comparing the two assessment procedures, we show the fuzzy Kappa's ability of generalization and some advantages in the application of fuzzy classification.

The simulated MRI volumes, available at the online BrainWeb [6], are used for our study. Each volume set consists of $181 \times 217 \times 181$ voxels with a cubic resolution $1 \times 1 \times 1\,\mathrm{mm}^3$.

The observed space $\mathbf{B} = \{v\}$ is the brain images of MRI, where $v = (x, y, z)$ is the coordinate of voxel. It will be classified in three fuzzy subclasses $A_i \subset \mathbf{B}, i = 1, 2, 3$, corresponding to the three tissues, cerebral spinal fluid (CSF), gray matter (GM), and white matter (WM).

As a reference, the membership functions $\mu_{A_i}^{\mathrm{Std}}(v)$: $\mathbf{B} \to [0, 1]$, $i = 1, 2, 3$, have been provided by the BrainWeb [6] as three anatomic fuzzy models $\mu_{\mathrm{CSF}}^{\mathrm{Std}}(v)$, $\mu_{\mathrm{GM}}^{\mathrm{Std}}(v)$ and $\mu_{\mathrm{WM}}^{\mathrm{Std}}(v)$ that are shown in Fig. 1. The membership functions $\mu_{A_i}(v)'$:$\mathbf{B} \to [0, 1]$, $i = 1, 2, 3$, obtained by the fuzzy classifier of [9], considered as a tested classifier. The three classification results $\mu_{\mathrm{CSF}}(v)$, $\mu_{\mathrm{GM}}(v)$ and $\mu_{\mathrm{WM}}(v)$ are shown in Fig. 2. The agreement between $\mu_{A_i}(v)$ and $\mu_{A_i}^{\mathrm{Std}}(v)$ are separately assessed by using Cohen's Kappa and the fuzzy Kappa.

4.1. Experiment of agreement assessment using Cohen's Kappa

Considering the assessment pairs $(\mu_{A_i}(v), \mu_{A_i}^{\mathrm{Std}}(v))$, for $i = 1, 2, 3$, we have three pairs of components $(\mu_{\mathrm{CSF}}, \mu_{\mathrm{CSF}}^{\mathrm{Std}})$, $(\mu_{\mathrm{GM}}, \mu_{\mathrm{GM}}^{\mathrm{Std}})$ and $(\mu_{\mathrm{WM}}, \mu_{\mathrm{WM}}^{\mathrm{Std}})$ for the agreement assessment between the tested classifier and the reference.

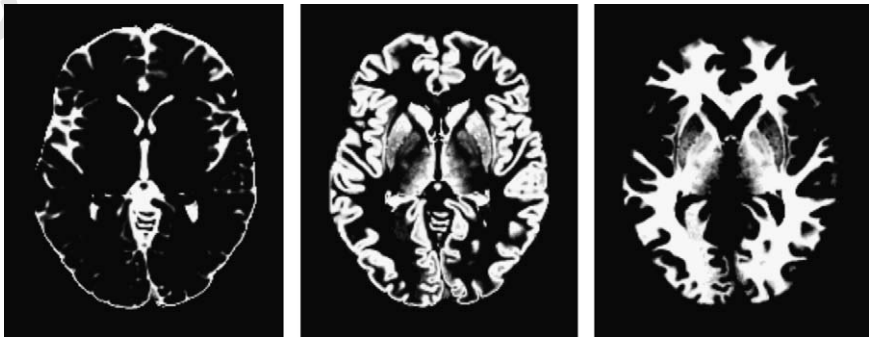

Fig. 1. Anatomic fuzzy model of CSF ($\mu_{\mathrm{CSF}}^{\mathrm{Std}}$), GM ($\mu_{\mathrm{GM}}^{\mathrm{Std}}$) and WM ($\mu_{\mathrm{WM}}^{\mathrm{Std}}$), available at the BrainWeb [6].

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

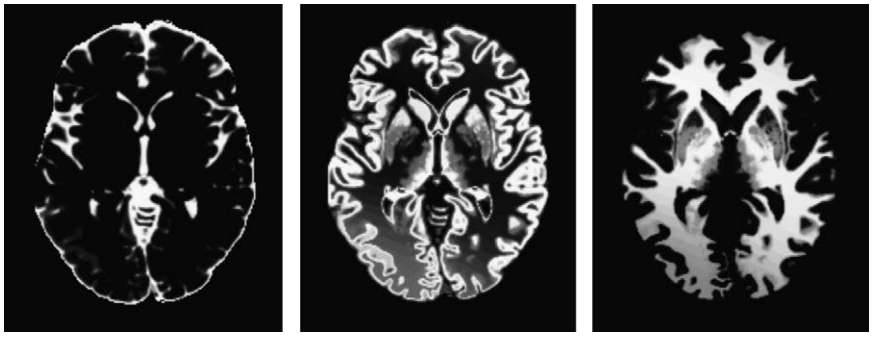

Fig. 2. Fuzzy classification results of CSF (μ_{CSF}), GM (μ_{GM}) and WM (μ_{WM}) by using the fuzzy classifier proposed in [9].

In order to use the Cohen's Kappa presented in Section 2.2, we have to build some crisp subclasses as the results from that of fuzzy classification. So we disassemble the each component into several independent subclasses $\mathbf{H}_l$ by using α-cut like

$$\mathbf{H}_l = \mathbf{A}_i; \quad \text{if } (\alpha \cdot (l-1) < \mu_{A_i} \leqslant \alpha \cdot l), \tag{15}$$

where

$$\mathbf{H}_l \bigcap \mathbf{H}_k = \phi; \quad l,k = 1,2,\ldots,L, \quad \text{and} \quad l \neq k \tag{16}$$

and $\alpha \in (0,1]$, so that

$$\mathbf{A}_i = \bigcup_{l=1}^{L} \mathbf{H}_l. \tag{17}$$

From Eq. (15), if we select $\alpha = 0.1$, $\mu_{A_i} \in [0,1]$ are disassembled into 10 subclasses and so $L = 10$ in this instance. In the same way, if $\alpha = 0.5$, $\mu_{A_i} \in [0,1]$ are disassembled into two subclasses and $L = 2$.

As Eq. (1), for each subclass, we have the eigenfunction for one voxel v:

$$u_l(v) = \begin{cases} 1 & \text{if } v \in \mathbf{H}_l, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The agreement function of $\mu_{A_i}(v)$ and $\mu_{A_i}^{\text{Std}}(v)$ is

$$f(u_l^{\mu_{A_i}}, u_l^{\mu_{A_i}^{\text{Std}}}) = \sum_{l=1}^{L} u_l^{\mu_{A_i}} u_l^{\mu_{A_i}^{\text{Std}}} = \begin{cases} 1 & \text{if } u_l^{\mu_{A_i}} \neq 0 \; u_l^{\mu_{A_i}^{\text{Std}}} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Put $f(u_l^{\mu_{A_i}}, u_l^{\mu_{A_i}^{\text{Std}}})$ in Eqs. (4) and (9), we get the observed agreement P_o and the proportion random agreement P_e for A_i. The Kappa coefficients of $(\mu_{A_i}(v), \mu_{A_i}^{\text{Std}}(v))$ are calculated by Eq. (5) and shown in Table 1.

There are two existent problems by using this assessment method:

(1) For one classifier, we have three Kappa coefficients $K_{\text{Cohen}}(\mu_{CSF}, \mu_{CSF}^{\text{Std}})$, $K_{\text{Cohen}}(\mu_{GM}, \mu_{GM}^{\text{Std}})$ and $K_{\text{Cohen}}(\mu_{WM}, \mu_{WM}^{\text{Std}})$. How can we entirely evaluate the classifier?

(2) For different selection of α or L, we get various Kappa coefficients (see Table 1). How can we objectively evaluate the classifier?

Table 1
Agreement of $\mu_A(v)$ and $\mu_A^{\text{Std}}(v)$ assessed by Cohen's Kappa statistic

Kappa coefficient	$\alpha = 0.1$, $L = 10$ (10 crisp subsets)	$\alpha = 0.5$, $L = 2$ (2 crisp subsets)
$K_{\text{Cohen}}(\mu_{CSF}, \mu_{CSF}^{\text{Std}})$	0.75	0.97
$K_{\text{Cohen}}(\mu_{GM}, \mu_{GM}^{\text{Std}})$	0.67	0.96
$K_{\text{Cohen}}(\mu_{WM}, \mu_{WM}^{\text{Std}})$	0.73	0.97
Average K_{Cohen}	0.72	0.97

An average of the three Kappa coefficients may be a solution of the first problem. But for the second problem, conventional methods fail to solve it. The proposed fuzzy Kappa provides an alternative to cater for the problem.

Section 4.2 presents an application and some advantages of the assessment by using fuzzy Kappa.

In Table 1, the assessment result is diverse versus the number of crisp subsets. For the 10 crisp sets, the average K_{Cohen} is 0.72, but for the 2 crisp sets, it is 0.97. It is not appropriate to evaluate this fuzzy classifier by using Cohen's Kappa.

4.2. Experiment of agreement assessment using the fuzzy Kappa

In the case of the fuzzy Kappa, the properties of the tested classifier and the reference are that for $v_m \in \mathbf{B}$,

$$\sum_{i=1}^{3} \mu_{A_i}(v_m) = 1, \quad m = 1,2,\ldots,M, \tag{20}$$

as well as

$$\sum_{i=1}^{3} \mu_{A_i}^{\text{Std}}(v_m) = 1, \quad m = 1,2,\ldots,M, \tag{21}$$

where m is the index of voxel in $\mathbf{B}$, and M is the total number of voxels in $\mathbf{B}$; $A_i \subset \mathbf{B}$ and $i = 1,2,3$.

The probability distribution of μ_i^{Std}, $p(\mu_i^{\text{Std}})$ and that of μ_i, $p(\mu_i)$ have been estimated by the normalized histogram of the membership degree images shown in Figs. 1 and 2, respectively. Fig. 3 is an example of $p(\mu_i^{\text{Std}})$ and Fig. 4 is that of $p(\mu_i)$.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

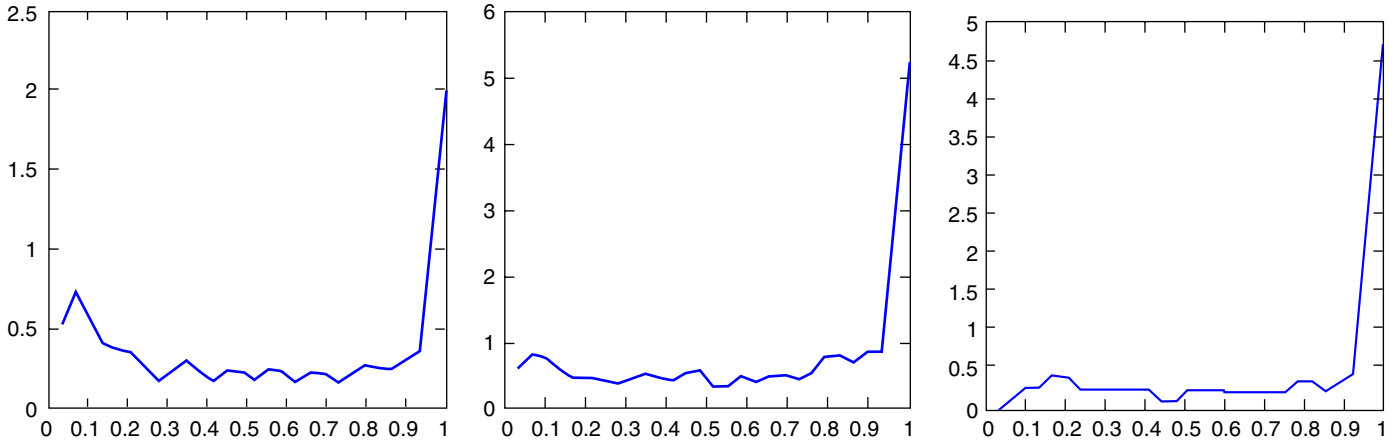

Fig. 3. Probability distributions of $\mu_{\mathrm{CSF}}^{\mathrm{Std}}$ ($p(\mu_{\mathrm{CSF}}^{\mathrm{Std}})$), $\mu_{\mathrm{GM}}^{\mathrm{Std}}$ ($p(\mu_{\mathrm{GM}}^{\mathrm{Std}})$) and $\mu_{\mathrm{WM}}^{\mathrm{Std}}$ ($p(\mu_{\mathrm{WM}}^{\mathrm{Std}})$) estimated by using the histogram of the image shown in Fig. 1.

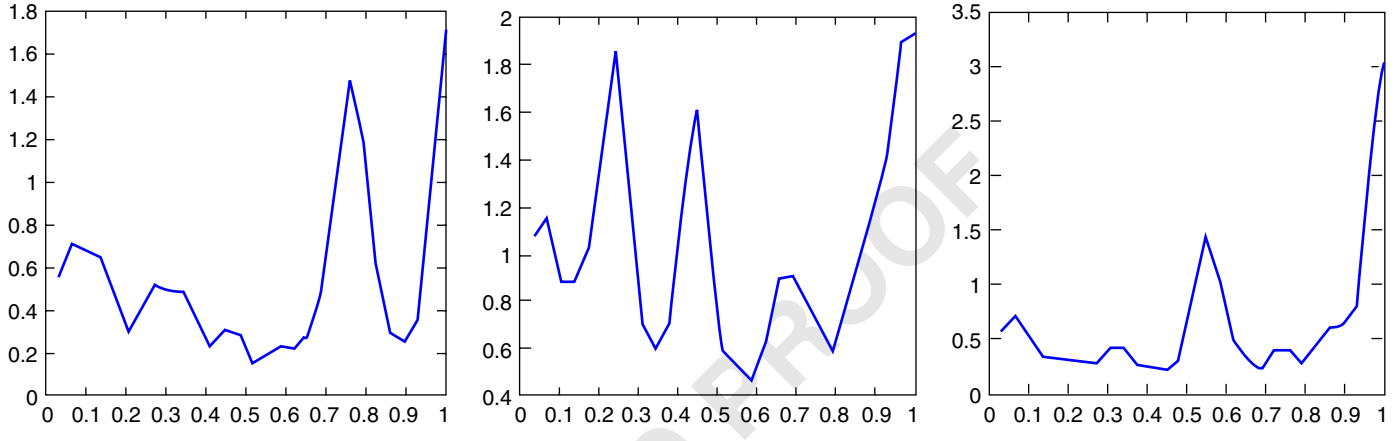

Fig. 4. Probability distribution of μ_{CSF} ($p(\mu_{\mathrm{CSF}})$), μ_{GM} ($p(\mu_{\mathrm{GM}})$) and μ_{WM} ($p(\mu_{\mathrm{WM}})$) estimated by using the histogram of the image shown in Fig. 2

To evaluate the agreement between μ_i^{Std} and μ_i by using the fuzzy Kappa introduced in Section 3.2, we calculate firstly the fuzzy agreement function $f^{\mathscr{F}}(v_m)$ according to Eq. (11), such that

$$f^{\mathscr{F}}(v_m) = \sum_{i=1}^{3}(\mu_i^{\mathrm{Std}}(v_m) \wedge \mu_i(v_m)), \quad m = 1, 2, \ldots, M. \tag{22}$$

The calculation result of $f^{\mathscr{F}}(v_m)$ is shown in Fig. 5.

Then the proportion of observed agreement $P_{\mathrm{o}}^{\mathscr{F}}$ is calculated according to Eq. (4), that is,

$$P_{\mathrm{o}}^{\mathscr{F}} = \frac{1}{M}\sum_{m=1}^{M} f^{\mathscr{F}}(v_m) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{3}(\mu_i(v_m) \wedge \mu_i^{\mathrm{Std}}(v_m)) = 0.8449. \tag{23}$$

The expectation of random agreement, $P_{\mathrm{e}}^{\mathscr{F}}$ is calculated by using (13). The probability distribution of μ_i^{Std}, $p(\mu_i^{\mathrm{Std}})$ and that of μ_i, $p(\mu_i)$ are shown in Figs. 3 and 4, respectively. So we get

$$P_{\mathrm{e}}^{\mathscr{F}} = \sum_{i=1}^{3}\int_{\mu_i^{\mathrm{Std}}=0}^{1}\int_{\mu_i=0}^{1} p(\mu_i^{\mathrm{Std}})p(\mu_i)(\mu_i^{\mathrm{Std}} \wedge \mu_i)\,\mathrm{d}\mu_i^{\mathrm{Std}}\,\mathrm{d}\mu_i = 0.3829. \tag{24}$$

Finally, the fuzzy Kappa K_{fuzzy} is calculated by according to Eq. (14), we have

$$K_{\mathrm{fuzzy}} = \frac{P_{\mathrm{o}}^{\mathscr{F}} - P_{\mathrm{e}}^{\mathscr{F}}}{1 - P_{\mathrm{e}}^{\mathscr{F}}} = 0.7486. \tag{25}$$

The value 0.7486, of the fuzzy Kappa gives an overall evaluation of the tested classifier which means that the tested classifier and the reference are similar about 75% agreement. We can refer to the fuzzy agreement function to analyze the existent problem of the tested classifier.

The meaning of agreement can be also observed by $f^{\mathscr{F}}(\mu_{A_i}, \mu_{A_i}^{\mathrm{Std}})$, the agreement function illustrated in Fig. 5. The brighter region corresponds to a high agreement. The regions with low brightness show a low agreement of the two fuzzy classifications. The brightness of whole image in Fig. 5 is very high. It means a higher agreement between the fuzzy classification $\mu_{A_i}(v)$ and the reference $\mu_{A_i}^{\mathrm{Std}}(v)$. The same result can be taken by comparing directly Figs. 2 and 1.

By compared with an atlas of brain, we know that these regions with low agreement are the crossing regions of the three main tissues, CSF, GM and WM. They perhaps correspond to another tissue type, such as glia. Because the

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

Fig. 5. Result of fuzzy agreement function $f^{\mathcal{F}}(v_m)$ in Eq. (22). The brighter regions correspond to higher agreement and the darker regions show lower agreement of the two fuzzy classifications.

reference only provide the fuzzy anatomic models of three main tissues of brain and the results of tested classification are obtained from entire MRI images. So some small regions that correspond to the other tissues may exist in the results mapping but not in the reference. Moreover, the lower brightness regions indicate that the tested classifier is not a perfect one, and some work is needed to improve its performance.

5. Discussion

In fact, there are two level indexes to assess the fuzzy classifier of brain tissues on MR images:

- tissue class A_i;
- membership degree of each class μ_{A_i}.

It means that for a voxel v_m, it is classified to class A_i with a membership degree $\mu_{A_i}(v_m)$.

For each class A_i, the agreement measurement can be done by using conventional Kappa statistic, such as the methods of Hyung et al. [16] or the methods of Chen [4]. In this case, the result of the agreement measurement is the function of the number of crisp subsets L, because the result $K(\mu_{A_i} \geqslant \alpha_l)$, $l = 1, 2, \ldots, L$ is diverse versus L.

For a given L, the agreement measurement is in fact used as same as for a hard classifier. These conventional Kappa coefficients give results versus number of classes, i.e. $K(\mu_{A_i} \geqslant \alpha_L)$, $i = 1, 2, \ldots$.

An average operation can be used to combine the measurement results of different classes for providing a final agreement evaluation of a classifier. But it is difficult to combine the measurement results $K(\mu_{A_i} \geqslant \alpha_l)$, $l = 1, 2, \ldots, L$, which correspond to different L for a given class A_i, because the number of crisp subsets L is elective and not exclusive.

For the same evaluation task, Ruan et al. [22] use the measurement of the absolute average error ξ

$$\xi = \frac{\sum_{v \in S} |a_v - a_v^f|}{Card(S)}, \tag{26}$$

where a_v^f denotes the proportion of tissue at the voxel v of one class of reference image, and a_s denotes the same quantity at each voxel of the classified image, $Card(S)$ denotes the number of voxels in the reference image S.

This is an absolute difference measurement for a classifier assessment. The values of ξ corresponding to different classes A_i can also be combined for an overall evaluation. But in the case of relative agreement, for example, in the case of linear relative agreement (such as $a_v = a_v^f + b$ or $a_v = a_v^f \times c$, where b and c are any constants) the values of ξ are different. So this is not a general assessment method and limited to some specific applications. However, the fuzzy Kappa presented in Section 4.2 is a general extension of agreement assessment.

6. Conclusion

The fuzzy Kappa proposed in this paper is an extension of Cohen's Kappa. As the same concept as Cohen's Kappa, the fuzzy Kappa expresses the proportional reduction of a classification error that generated by a classification process, relative to the error that generated by a completely random classification. But the meaning of the fuzzy agreement function is different from Cohen's Kappa, although they have the similar formula and properties. The fuzzy Kappa takes all the advantages of Cohen's Kappa, and also is extended to give exclusively an overall agreement measurement for evaluating the fuzzy classifier.

An application of fuzzy classification of brain tissues on MRI images shows that the fuzzy agreement function can be used to analyze the problem of a tested classifier and give some suggestions for improvement.

References

[1] A. Baraldi, L. Bruzzone, P. Blonda, Quality assessment of classification and cluster maps without ground truth knowledge, IEEE Trans. Geosci. Remote Sensing 43 (4) (2005) 857–873.

[2] C.C. Berry, The Kappa statistic, J. Am. Med. Assoc. 268 (18) (1992) 2513.

[3] J. Carletta, Assessing agreement on classification tasks: the Kappa statistic, Comput. Linguist. 22 (2) (1996) 249–254.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

[4] S.-M. Chen, Measures of similarity between vague sets, Fuzzy Sets and Systems 74 (1995) 217–223.
[5] D. Chen, K. Sirlantzis, D. Hua, X. Ma, On the relation between dependence and diversity in multiple classifier systems, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05).
[6] A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans, Brain Web: Online interface to a 3D MRI simulated brain database, Available at http://www.bic.mni.mcgill.ca/brainweb.
[7] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[8] J. Cohen, This week's citation classic—a coefficient of agreement for nominal scales, Current contents/Number 3 January 20, (1986) 18.
[9] W. Dou, S. Ruan, D. Bloyet, J.-M. Constans, Y. Chen, Segmentation based on information fusion applied to brain tissue on MRI, SPIE-IST Electronic Imaging, vol. 5298, 2004, pp. 492–503.
[10] J.L. Fleiss, Statistical Methods for Rates and Proportions, second ed., Wiley, New York, 1981.
[11] S. Fritz, L. See, Improving quality and minimising uncertainty of land cover maps using fuzzy logic, In: Proceedings of the 12th Annual Conference on GIS Research (GISRUK 2004), University of East Anglia, Norwich, UK, 28–30 April 2004.
[12] L.R. Goldman, The Kappa statistic—in reply, J. Am. Med. Assoc. 268 (18) (1992) 2513–2514.
[13] J.O. Greene, J.N. Cappella, Cognition and talk: the relationship of semantic units to temporal patterns of fluency in spontaneous speech, Lang. Speech 29 (2) (1986) 141–157.
[14] A. Hagen, Fuzzy set approach to assessing similarity of categorical maps, Int. J. Geogr. Inf. Sci. 17 (2003) 235–249.
[15] G. Hripcsak, D.F. Heitjan, Measuring agreement in medical informatics reliability studies, J. Biomed. Inf. 35 (2002) 99–110.
[16] L.K. Hyung, Y.S. Song, K.M. Lee, Similarity measure between fuzzy sets and between elements, Fuzzy Sets and Systems 62 (1994) 291–293.
[17] H.-W. Jung, Evaluating interrater agreement in SPICE-based assessments, Comput. Stand. Interfaces 25 (2003) 477–499.
[18] H.C. Kraemer, Extension of the Kappa coefficient, Biometrics 36 (1980) 207–216.
[19] K. Krippendorff, Content Analysis: An Introduction to its Methodology, Sage Publications, Beverley Hills, CA, 1980.
[20] D.K. McIver, M.A. Friedl, Estimation pixel-scale land cover classification confidence using nonparametric machine learning methods, IEEE Trans. Geosci. Remote Sensing 39 (9) (2001) 1959–1968.
[21] M.P. Robertson, M.H. Villet, A.R. Palmer, A fuzzy classification technique for predicting species' distributions: applications using invasive alien plants and indigenous insects, Diversity Distrib. 10 (2004) 461–474.
[22] S. Ruan, B. Moretti, J. Fadili, D. Bloyet, Fuzzy merkovian segmentation in application of magnetic resonance images, Comput. Vision Image Understanding 85 (2002) 54–69.
[23] S. Siegel, N.J. Castellan Jr., Nonparametric Statistics for the Behavioral Sciences, second ed., McGraw-Hill, New York, 1988.
[24] K. Soeken, P. Prescott, Issues in the use of Kappa to assess reliablility, Med. Care 24 (1986) 733–743.
[25] S. Sousa, S. Caeiro, R.G. Pontius Jr., M. Painho, Sado estuary management areas: hard versus soft classification maps comparison, Conference Proceedings of Coastal GIS 2003, Fifth International Symposium on GIS and Computer Cartography for Coastal Zone Management, Genova, Italy, Geographical Information Systems International Group and ISSOPS International Center of Coastal and Ocean Policy Studies.
[26] R.W. Taylor, A.P. Reeves, Classification quality assessment for a generalized model-based object identification system, IEEE Trans. Sys. Man Cybern. 19 (4) (1989) 861–866.
[27] N.O.B. Thomsen, L.H. Olsen, S.T. Nielsen, Kappa statistics in the assessment of observer variation: the significance of multiple observers classifying ankle fractures, J. Orthop. Sci. 7 (2002) 163–166.

W. Dou, associate professor, received the Bachelor's degree of Radio Technology from UESTC (University of Electronic Science and Technology of China) in 1984, and the DEA (Diplôme d'Etude Approfondi) on Signal Telecommunications Image Radar, from Université de RENNES I of France in 1993. She joined the Electronic Engineering Department of Tsinghua University of China in 1995 and is an Associate Professor since 1999. Her research interests span the area of digital signal processing, audio and video signal processing, digital signal processor's application and design. From 2001, she works also on fuzzy information fusion for application of biomedical information, such as segmentation of tumorous brain tissues in MRI images.

Y. Ren, undergraduate student of Electronic Engineering Department of Tsinghua University. Research interests: Pattern Recognition and Neural Network, Image Process, MEMS/Robot, Bioinformatics.

Q. Wu, Ph.D. student, received the Bachelor's degree from Ningbo University in 2002, and the Master's degree from Tsinghua University in 2005. She is now a Ph.D. student in Imperial College and works in the visual information group (VIP) for research. Her research is mainly in the fields of medical image computing, imaging system and machine learning.

Y. Chen, associate professor, received the Bachelor's degree of medicine from Southern Medical University (the former First Military Medical University) of China in 1985, and received Master's degree of medicine from the same University in 1990. She has been working in the imaging diagnostic center, Nanfang hospital, the first affiliated hospital of Southern University as a doctor since 1986, and as associate professor since 2000. Her major research work is in medical Imaging Diagnose, especially in CT and MRI diagnose. Her research is mainly in head and neck diseases.

S. Ruan, professor, received the Ph.D. degree in image processing from l'Université de Rennes 1 in 1993. She was assisted professor at l'Université de Caen from 1993 to 2003. She is now professor at l'Université de Reims, and works in the CReSTIC laboratory for the research. Her research is manly in the fields of segmentation and pattern recognition applied for brain images.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007

D. Bloyet received in 1970 the Ph.D. in electrical engineering from the University Paris 11, France. Since 1979 he has been Professor in electronics at ENSICAEN Caen France. His research activities deal with the design of low noise sensors and systems: very low noise amplifiers, SQUID magnetometers, study of excess low frequency noise in high frequency BICMOS technologies. Part of his activities is related to image acquisition and preprocessing (neuroscience, cytology, cytometry). He is author of about 65 papers in international periodicals and 80 communications in international congress with extended proceedings.

J.-M. Constans, Maître de Conférences Universitaire et Praticien Hospitalier, after almost 3 years of research in magnetic resonance (MR; especially in spectroscopy) at UCSF and VAMC San Francisco came back to the university hospital of Caen at the end of 1993. He is Maître de Conférences Universitaire et Praticien Hospitalier since 1998 and is being doing research at the MR Unit and in Equipe d'Accueil 3916 ''Imagerie Fonctionnelle et M'etabolique en Oncologie''. His research consists of development, evaluation and application of MR techniques and methods in segmentation and in proton spectroscopy on brain diseases.

Please cite this article as: W. Dou, et al., Fuzzy kappa for the agreement measure of fuzzy classifications, Neurocomputing (2006), doi:10.1016/j.neucom.2006.10.007