University of Groningen

Category kappas for agreement between fuzzy classifications

Warrens, Matthijs J.

Published in:
Neurocomputing

DOI:
[10.1016/j.neucom.2016.02.038](https://doi.org/10.1016/j.neucom.2016.02.038)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Warrens, M. J. (2016). Category kappas for agreement between fuzzy classifications. *Neurocomputing*, 194, 385-388. <https://doi.org/10.1016/j.neucom.2016.02.038>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Brief Papers

Category kappas for agreement between fuzzy classifications

Matthijs J. Warrens*

GION, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 11 November 2014

Received in revised form

27 August 2015

Accepted 20 February 2016

Available online 10 March 2016

Keywords:

Cohen's kappa

Fuzzy classification

Category statistics

MRI

Brain tissue

Interobserver agreement

ABSTRACT

The kappa statistic is a widely used as a measure for quantifying agreement between two nominal classifications. The statistic has been extended to the case of two normalized fuzzy classifications. In this paper we define category kappas for quantifying agreement on a particular category of two normalized fuzzy classifications. The overall fuzzy kappa is a weighted average of the proposed category kappas. Since the value of the overall kappa lies between the minimum and maximum values of the category kappas, the overall kappa, in a way, summarizes the agreement reflected in the category kappas. The overall kappa meaningfully reflects the degree of agreement between the fuzzy classifications if the category kappas are approximately equal. If this is not the case, it is more informative to report the category kappas.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In various fields of science, classification instruments are used to classify individuals or objects into categories. The instruments can be people or algorithms, and the categories are often unordered and mutually exclusive (nominal). Examples are, an educational psychologist that classifies the arithmetic strategies used by children to solve a mathematics problem [1,2], a developmental psychologist that classifies the attachment style of adults [3], a physical therapist that rates the type of injury [4], a radiologist that identifies tissues [5], an algorithm that classifies voxels of magnetic resonance images [6], or an algorithm that classifies the areas of a land cover map [7,8].

An important property of a classification instrument is its accuracy or reliability. The accuracy of an instrument can be quantified by assessing the agreement between two classifications made with the same instrument. High agreement between the classifications can then be taken as an indicator of the quality of the category definitions and the classification instrument in general.

A widely used statistic for quantifying agreement between two nominal classifications is the kappa statistic proposed by Cohen [5,9–11]. The value of kappa is 1 if there is perfect agreement between the two classifications, and 0 when agreement is equal to that expected under independence. Cohen's kappa was originally proposed as a measure of interobserver agreement in the context

of psychological measurement, but the statistic is also used for map comparison in remote sensing [7,8] and content analysis [12]. Kappa depends on the boundary probabilities of the classifications. These quantities reflect how often a category was used in the classifications. Kappas from studies with different boundary probabilities are not comparable [13,14].

The popularity of the kappa statistic has led to the development of several extensions, e.g., kappas for three or more instruments [15], weighted kappas for classification instruments with ordinal categories [16–18], kappas for groups of observers [19] and kappas for clustered data [20,21]. In addition, extensions for the agreement between two fuzzy classifications [6] and two fuzzy categorical raster maps [7,8] have also been proposed.

The concept of a fuzzy classification is based on fuzzy set theory [22,23]. Fuzzy set theory is a well-developed mathematical system for addressing the degree of membership of objects to categories. In crisp classification objects are pigeonholed: an object either belongs or does not belong to a category. By contrast, fuzzy set theory permits the gradual assessment of the membership of an object into a category. The membership is described with the aid of a membership function valued in the real unit interval [0, 1].

Dou et al. [6] proposed a fuzzy kappa for quantifying agreement between two normalized fuzzy classifications. Their fuzzy kappa is an overall measure of agreement. If there is any disagreement between the classifications, the overall statistic does not tell us on which categories the agreement is almost perfect (if any) and on which categories agreement is less than perfect. Fuzzy kappas for the different categories are often more informative. The information provided by the category kappas can often be used to improve the accuracy of the classification instrument [24,25]. However,

* Tel.: +31 503636691.

E-mail address: m.j.warrens@rug.nl

category kappas have not been defined for fuzzy classifications. This will be done in this paper.

The paper is organized as follows. In the next section the notation is introduced and the fuzzy kappa statistics are defined. The fuzzy kappas are illustrated on hypothetical brain tissue data in Section 3. In Section 4 we consider the case of two categories. It is shown that in this case the two category kappas and the overall kappa coincide. An alternative definition of the category kappas is presented in Section 5. Section 6 contains a discussion.

2. Fuzzy kappas

Let $X = \{x\}$ be a set of elements (voxels, individuals) of size $|X| = n$. In the classification process the objects are assigned to c unordered categories that are defined in advance, with $i \in \{1, 2, \dots, c\}$. The objects are classified twice, either with the same instrument or with different instruments. The two classifications are done independently.

We first define a fuzzy kappa for a single category. Let $u_i : X \rightarrow [0, 1]$ be the membership function of category i for the first classification, and let $v_i : X \rightarrow [0, 1]$ be the membership function of category i for the second classification. It is assumed that both classifications are normalized:

$$\sum_{i=1}^c u_i(x) = 1, \quad x \in X, \quad (1)$$

and

$$\sum_{i=1}^c v_i(x) = 1, \quad x \in X. \quad (2)$$

Normalized classifications describe a certain perfect case. In reality the requirement may be violated if the classification is done by human experts.

Let p_i and q_i denote the boundary probabilities of the first and second classification, respectively. The boundary probabilities are defined as

$$p_i := \frac{1}{n} \sum_{x \in X} u_i(x), \quad (3)$$

and

$$q_i := \frac{1}{n} \sum_{x \in X} v_i(x). \quad (4)$$

Quantities (3) and (4) reflect how often category i was used in the first and second classification, respectively. Following Dou et al. [6], the observed agreement between the classifications on category i is defined as

$$O_i := \frac{1}{n} \sum_{x \in X} \min(u_i(x), v_i(x)). \quad (5)$$

Let $p(u_i)$ and $q(v_i)$ denote the probability distribution of $u_i(x)$ and $v_i(x)$, respectively. Assuming that $u_i(x)$ is independent of $v_i(x)$ for all $x \in X$, the expectation of random agreement for category i is defined as

$$E_i := \int_{u_i=0}^1 \int_{v_i=0}^1 p(u_i) q(v_i) \min(u_i, v_i) du_i dv_i. \quad (6)$$

If u_i and v_i are binary functions, i.e., if u_i and v_i take on the values 0 and 1 only, then the expectation of random agreement becomes

$$E_i = p_i q_i. \quad (7)$$

A fuzzy kappa for category i can now be defined as

$$\kappa_i := \frac{O_i - E_i}{(p_i + q_i)/2 - E_i}. \quad (8)$$

If u_i and v_i are binary functions, statistic (8) becomes the category kappa for crisp classifications (see, e.g., [24–26]).

It turns out that the overall fuzzy kappa introduced in Dou et al. [6] is a weighted average of the category kappas in (8). Define for category i the weight

$$w_i := \frac{p_i + q_i}{2} - E_i. \quad (9)$$

An overall fuzzy kappa can be defined as a weighted average of the fuzzy category kappas:

$$\kappa := \frac{\sum_{i=1}^c w_i \kappa_i}{\sum_{i=1}^c w_i}. \quad (10)$$

Since

$$\sum_{i=1}^c \frac{p_i + q_i}{2} = 1 \quad (11)$$

statistic (10) is identical to

$$\kappa = \frac{\sum_{i=1}^c (O_i - E_i)}{1 - \sum_{i=1}^c E_i}, \quad (12)$$

which is the fuzzy kappa proposed in Dou et al. [6]. If u_i and v_i are binary functions, fuzzy kappa (12) becomes the ordinary kappa for crisp classifications [24,25]. Several properties of fuzzy kappa (12) are discussed by Dou et al. [6].

3. An illustration

Brain images can be decomposed into tiny cubes called voxels. A typical size of a voxel is $1 \times 1 \times 1 \text{ mm}^3$ and an image may consist of tens of thousands of voxels. To illustrate the fuzzy category kappas and the overall fuzzy kappa we consider a hypothetical example that consists of ten voxels. Table 1 presents two hypothetical fuzzy classifications of ten voxels into three tissue categories: gray matter (GM), white matter (WM) and cerebral spinal fluid (CSF). Each row of Table 1 corresponds to a voxel. The six membership functions of the categories take on values from the set $(0, 0.2, 0.4, 0.6, 0.8, 1)$. It can be verified that the fuzzy classifications are normalized, i.e., the values in each row of either classification sum up to 1.

Table 2 presents the values of the observed agreement, expectation of random agreement, and the kappa statistics for the data in Table 1. The value of the overall kappa is 0.77, indicating a good level of agreement between the two fuzzy classifications. The values of the three category kappas are quite different. The values for WM and CSF are 0.75 and 0.66, respectively, indicating a good level of agreement. However, the value for GM is 0.92, indicating an excellent level of agreement. The category kappas show that there is virtually no disagreement on GM, while there is some

Table 1
Two hypothetical fuzzy classifications of ten voxels.

Classification 1			Classification 2		
GM	WM	CSF	GM	WM	CSF
0.2	0.4	0.4	0.2	0.6	0.2
0.6	0.4	0	0.6	0.4	0
0.4	0.4	0.2	0.4	0.6	0
0.2	0.4	0.4	0.2	0.4	0.4
0.4	0.4	0.2	0.4	0.2	0.4
0	0.8	0.2	0	0.8	0.2
0.6	0.4	0	0.6	0.4	0
0.2	0.2	0.6	0	0.2	0.8
0	0.8	0.2	0	0.8	0.2
0.2	0.8	0	0.2	0.8	0

Table 2

Values of various statistics for the data in Table 1.

Tissue category	Statistics		
	O	E	κ
GM	0.26	0.15	0.92
WM	0.48	0.39	0.75
CSF	0.18	0.10	0.66
Overall	0.92	0.65	0.77

disagreement between the classifications on the other two categories.

Since the overall kappa is a weighted average of the category kappas, its value, like an ordinary average, lies between the minimum and maximum value of the category kappas. The overall kappa is an useful summary statistic if the values of the category kappas are approximately the same. Since the category kappas have quite different values in this example the overall kappa is not a good summary statistic. More information is obtained by considering the individual category kappas.

4. The case of two categories

In this section we show that if there are only $c=2$ categories then the category kappas and the overall kappa coincide. The subscripts 1 and 2 will be used to denote the two categories. Let $x \in X$. Expressing the identity

$$\max(u_1(x), v_1(x)) + \min(u_1(x), v_1(x)) = u_1(x) + v_1(x), \quad (13)$$

in terms of $\max(u_1(x), v_1(x))$, and using the result in the identity

$$\min(u_2(x), v_2(x)) = \min(1 - u_1(x), 1 - v_1(x)) = 1 - \max(u_1(x), v_1(x)) \quad (14)$$

we obtain the identity

$$\min(u_2(x), v_2(x)) = \min(u_1(x), v_1(x)) + 1 - u_1(x) - v_1(x). \quad (15)$$

Summing (15) over all $x \in X$ and dividing the result by n we obtain

$$O_2 = O_1 + 1 - p_1 - q_1. \quad (16)$$

Next, we consider the expectation of random agreement. For category 2 we have the identities

$$p_2 = \int_{u_2=0}^1 p(u_2) u_2 du_2 \int_{v_2=0}^1 q(v_2) dv_2, \quad (17)$$

and

$$q_2 = \int_{u_2=0}^1 p(u_2) du_2 \int_{v_2=0}^1 q(v_2) v_2 dv_2. \quad (18)$$

Consider (6) for $i=2$. If we substitute the identity

$$\min(u_2, v_2) = u_2 + v_2 - \max(u_2, v_2) \quad (19)$$

in (6) for $i=2$ we obtain, using (17) and (18), the identity

$$E_2 = p_2 + q_2 - E_2^*, \quad (20)$$

where

$$E_2^* = \int_{u_2=0}^1 \int_{v_2=0}^1 p(u_2) q(v_2) \max(u_2, v_2) du_2 dv_2. \quad (21)$$

Making the change of variables $u_2 = 1 - u_1$ and $v_2 = 1 - v_1$ in (21), and using the identity

$$\max(1 - u_1, 1 - v_1) = 1 - \min(u_1, v_1) \quad (22)$$

in the result, we obtain

$$E_2^* = 1 - \int_{u_1=1}^0 \int_{v_1=1}^0 p(1 - u_1) q(1 - v_1) \min(u_1, v_1) du_1 dv_1. \quad (23)$$

Since (23) contains a double integration, reversing the limits of integration in (23) yields

$$E_2^* = 1 - \int_{u_1=0}^1 \int_{v_1=0}^1 p(u_1) q(v_1) \min(u_1, v_1) du_1 dv_1 = 1 - E_1. \quad (24)$$

Using (24) in identity (20) we obtain

$$E_2 = E_1 + p_2 + q_2 - 1. \quad (25)$$

Substitution of $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$ in (25) yields

$$E_2 = E_1 + 1 - p_1 - q_1. \quad (26)$$

The category kappa of category 2 is given by

$$\kappa_2 = \frac{O_2 - E_2}{(p_2 + q_2)/2 - E_2}. \quad (27)$$

Using (16) and (26) and the identity

$$\frac{p_2 + q_2}{2} - 1 + p_1 + q_1 = \frac{p_1 + q_1}{2}, \quad (28)$$

statistic (27) can be written as

$$\frac{O_1 - E_1}{(p_1 + q_1)/2 - E_1} = \kappa_1. \quad (29)$$

Thus, with $c=2$ categories we have $\kappa_1 = \kappa_2$. Furthermore, since the overall kappa is a weighted average of κ_1 and κ_2 , we also have $\kappa = \kappa_1 = \kappa_2$. Thus, in the case of two categories the two category kappas and the overall kappa coincide.

5. An alternative definition

It follows from the results in the previous section that there is an alternative definition of a category kappa. A category kappa can also be obtained by reducing both classifications to two membership functions, one for the category of interest and one for all other categories, and then calculate the overall fuzzy kappa in (12) for the reduced classifications. The membership function of the “all other” category is obtained by summing the values of the membership functions over the corresponding categories. Summing the values of two or more membership functions is not a standard binary operation in fuzzy set theory. In the current application the operation makes sense because the classifications are normalized. If we sum the values of the membership functions over the categories the normalization is preserved.

Tables 3 and 4 present an illustration of the alternative definition. The fuzzy classifications in Table 3 are obtained by combining the categories WM and CSF in Table 1 into an “all other”

Table 3

Two hypothetical fuzzy classifications of ten voxels.

Classification 1		Classification 2	
GM	WM/CSF	GM	WM/CSF
0.2	0.8	0.2	0.8
0.6	0.4	0.6	0.4
0.4	0.6	0.4	0.6
0.2	0.8	0.2	0.8
0.4	0.6	0.4	0.6
0	1	0	1
0.6	0.4	0.6	0.4
0.2	0.8	0	1
0	1	0	1
0.2	0.8	0.2	0.8

Table 4
Values of various statistics for the data in Table 3.

Tissue category	Statistics		
	<i>O</i>	<i>E</i>	κ
GM	0.26	0.15	0.92
WM/CSF	0.72	0.61	0.92
Overall	0.98	0.76	0.92

category. The category of interest here is GM. The fuzzy classifications in Table 3 are again normalized. Table 4 presents the values of the observed agreement, expectation of random agreement, and the kappa statistics for the data in Table 3. The value of the two category kappas and the overall kappa is 0.92, indicating an excellent level of agreement.

6. Conclusion

In this paper we defined category kappa statistics for summarizing agreement on a particular category of two normalized fuzzy classifications. The overall kappa presented in Dou et al. [6] is a weighted average of the proposed category kappas. Because it is a weighted average of the category kappas, the value of the overall kappa, like an ordinary average, lies between the minimum and maximum values of the category kappas. The overall kappa is thus a summary statistic of the category kappas.

With two categories the values of the category kappas and the overall kappa are identical. From this it follows that the proposed category kappa can be obtained by reducing both fuzzy classifications to two membership functions, one for the category of interest and one for all other categories, and then calculate the overall fuzzy kappa for the reduced classifications. The membership function of the “all other” category is obtained by summing the values of the membership functions over the associated categories.

With three or more categories it would be good practice to report both the overall kappa and the category kappas. The overall kappa meaningfully reflects the degree of agreement between the fuzzy classifications if the category kappas are approximately equal. If this is not the case, it is more informative to report the different category statistics.

Acknowledgments

The author thanks two anonymous reviewers for their helpful comments and valuable suggestions on a previous version of the paper.

References

- [1] M. Hickendorff, W.J. Heiser, C.M. van Putten, N.D. Verhelst, Solution strategies and achievement in dutch complex arithmetic: latent variable modeling of change, *Psychometrika* 74 (2009) 331–350.
- [2] M. Hickendorff, C.M. van Putten, N.D. Verhelst, W.J. Heiser, Individual differences in strategy use on division problems: mental versus written computation, *J. Educ. Psychol.* 102 (2010) 438–452.

- [3] C. Cozzarelli, J.A. Karafa, N.L. Collins, M.J. Tagler, Stability and change in adult attachment styles: associations with personal vulnerabilities, life events and global construals of self and others, *J. Soc. Clin. Psychol.* 22 (2003) 315–346.
- [4] J. Sim, C.C. Wright, The kappa statistic in reliability studies: use, interpretation, and sample size requirements, *Phys. Ther.* 85 (2005) 257–268.
- [5] P.E. Crewson, Fundamentals of clinical research for radiologists, *Am. J. Roentgenol.* 184 (2005) 1391–1397.
- [6] W. Dou, Y. Rena, Q. Wu, S. Ruan, Y. Chen, D. Bloyet, J.-M. Constans, Fuzzy kappa for the agreement measure of fuzzy classifications, *Neurocomputing* 70 (2007) 726–734.
- [7] A. Hagen, Fuzzy set approach to assessing similarity of categorical maps, *Int. J. Geogr. Inf. Sci.* 17 (2003) 235–249.
- [8] A. Hagen-Zanker, An improved fuzzy kappa statistic that accounts for spatial autocorrelation, *Int. J. Geogr. Inf. Sci.* 23 (2009) 61–73.
- [9] J. Cohen, A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* 20 (1960) 37–46.
- [10] L.M. Hsu, R. Field, Interrater agreement measures: comments on kappa, Cohen's kappa, Scott's π and Aickin's α , *Understand. Stat.* 2 (2003) 205–219.
- [11] M.J. Warrens, Inequalities between kappa and kappa-like statistics for $k \times k$ tables, *Psychometrika* 75 (2010) 176–185.
- [12] K. Krippendorff, Reliability in content analysis some common misconceptions and recommendations, *Hum. Commun. Res.* 30 (2004) 411–433.
- [13] J.S. Uebersax, Diversity of decision-making models and the measurement of interrater agreement, *Psychol. Bull.* 101 (1987) 140–146.
- [14] W. Vach, The dependence of Cohen's kappa on the prevalence does not matter, *J. Clin. Epidemiol.* 58 (2005) 655–661.
- [15] M.J. Warrens, Inequalities between multi-rater kappas, *Adv. Data Anal. Classif.* 4 (2010) 271–286.
- [16] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychol. Bull.* 70 (1968) 213–220.
- [17] M.J. Warrens, Some paradoxical results for the quadratically weighted kappa, *Psychometrika* 77 (2012) 315–323.
- [18] M.J. Warrens, Conditional inequalities between Cohen's kappa and weighted kappas, *Stat. Methodol.* 10 (2013) 14–22.
- [19] S. Vanbelle, A. Albert, Agreement between two independent groups of raters, *Psychometrika* 74 (2009) 477–491.
- [20] C. Kang, B. Qaqish, J. Monaco, S. Sheridan, J. Cai, Kappa statistic for clustered dichotomous responses from physicians and patients, *Stat. Med.* 32 (2013) 3700–3719.
- [21] Z. Yang, M. Zhou, Kappa statistic for clustered matched-pair data, *Stat. Med.* 33 (2014) 2612–2633.
- [22] L. Zadeh, Fuzzy sets, *Inf. Control* 8 (1965) 338–353.
- [23] M. Smithson, J. Verkuilen, *Fuzzy Set Theory Applications in the Social Sciences*, Sage, Thousand Oaks, 2006.
- [24] M.J. Warrens, Cohen's kappa is a weighted average, *Stat. Methodol.* 8 (2011) 473–484.
- [25] M.J. Warrens, New interpretations of Cohen's kappa, *J. Math. ID* 203907, 2014, <http://dx.doi.org/10.1155/2014/203907>.
- [26] J.L. Fleiss, B. Levin, M.C. Paik, *Statistical Methods for Rates and Proportions*, Wiley, Hoboken, New Jersey, 2003.

Matthijs J. Warrens received his Ph.D. degree in Methodology and Statistics in 2008 from Leiden University. He is currently an assistant professor in the Department of Educational Science at University of Groningen. His research interests include similarity measures and agreement indices, including, Cohen's kappa, weighted kappa, Cronbach's alpha and the adjusted Rand index.