ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Fuzzy Fleiss-κ for Comparison of Fuzzy Classifiers

Dietlind Zuehlke[1], Tina Geweniger[2], Ulrich Heimann[3] and Thomas Villmann[4]

[1]Fraunhofer Institute for Applied Information Technology - Life Science Informatics
Schloss Birlinghoven, 53229 Sankt Augustin - Germany

[2]University of Leipzig - Working Group Computational Intelligence
Semmelweisstrasse 10, 04103 Leipzig - Germany

[3]Helmut Hund GmbH
Wilhelm-Will-Str. 7, 35580 Wetzlar - Germany

[4]University of Applied Sciences Mittweida
Department of Mathematics/Physics/Informatics, Computational Intelligence Group
Technikumplatz 17, 09648 Mittweida - Germany

Abstract. In this paper we show a straight forward extension of the fuzzy Cohen's-κ to Fleiss'-κ for the determination of classification agreements of fuzzy classifiers. In addition we investigate the influence of different interpretations of fuzzy intersection in terms of *t-norms*. These considerations are done for exemplary artificial data as well as for classification in image recognition for counting pollen grains.

1 Introduction

Classification problems take a major part in adaptive machine learning tasks. During the model generation and evaluation, frequently different classification models have to be assessed. Standard methods for crisp classification are the evaluation on training and test data or cross-validation approaches in terms of achieved accuracies. More advanced methods include significance statistics like conformal prediction [1]. Otherwise, methods from traditional statistics can be utilized too. Here χ^2-statistics as well as the Cohen's κ_{C}-statistics are the great players if two classifiers have to be compared [2],[3]. For the latter one, an extension for judgement of more than two classifiers exist, the Fleiss' κ_{F} [4],[5]. The advantage of these methods is that the comparison is evaluated in terms of classification agreement over the same data set and not obtained by achieved accuracies for given test data. In worst case two classifiers with nonzero misclassifications each could have disjunct subset of the data set were they fail. In this case accuracy based performance evaluations may suggest stronger agreement than happens.

For fuzzy classification, the number of available methods is reduced. Frequently, accuracy based methods are applied using the majority vote of a fuzzy classifier. Yet, this contradicts the aim of fuzzy classification. As shown in [1], conformal prediction can be applied to achieve information about the confidence of a fuzzy decision. For Cohen's κ_{C}-statistics, a extension for fuzzy classifiers was recently proposed [6]. It is based on the utilization of FUZZY-AND-*operators*. However, a respective extension of Fleiss' κ_{F} doesn't exist so far. In this paper we develop such an extension. For this purpose, we first consider the fuzzy variant of Cohen's κ. Subsequently we give the equivalent extension of Fleiss' κ_{F} for the fuzzy case. Moreover, we compare different realizations of the FUZZY-AND-*operator* in terms of *t-norms* [7], as it appear in both fuzzy measures with respect to the interpretability of κ_{C} and κ_{F} as suggested by Cohen.

2 Fuzzy Variants of Cohen's κ_{C} and Fleiss' κ_{F}

We look at a classification problem where N data points $\mathbf{x}_k$ have to be classified into C (disjoint) subsets. In the case of crisp classification one data point is classified to one class. The output of a crisp classifier is a vector $\mathbf{u}(\mathbf{x}_k) = (u_1(\mathbf{x}_k), \ldots, u_C(\mathbf{x}_k))$ with each $u_i \in \{0,1\}$ and $\sum u_i(\mathbf{x}_k) = 1$, i.e. $u_i(\mathbf{x}_k) = 1$ iff the data point $\mathbf{x}_k$ is classified into the ith class. In the case of a probabilistic fuzzy classifier the output is a vector $\boldsymbol{\mu}(\mathbf{x}_k) = (\mu_1(\mathbf{x}_k), \ldots, \mu_1(\mathbf{x}_k))$ with now $\mu_i \in [0,1]$ but constraint $\sum_{i=1}^{C} \mu_i(\mathbf{x}_k) = 1$. A possibilistic variant is obtained if the latter restriction is neglected.

We now briefly give the original variants of κ_{C} and κ_{F} following the the description in [6]. Thereafter we describe the fuzzy variant of κ_{C} and explain an analog approach for κ_{F}.

2.1 Original Cohen's κ_{C} and Fleiss' κ_{F}

Cohen's κ_{C} is a statistical measure of inter-rater agreement of *two* crisp classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ taking into account the agreement occurring by chance. It is given by

$$\kappa_{\mathrm{C}} = \frac{p_o - p_c}{1 - p_c} \tag{1}$$

where p_0 is the relative agreement among the classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ and p_c is the expected agreement by chance. p_c is the expected value of the joined event of classifier $\mathcal{C}_1$ and $\mathcal{C}_2$ classifying a data point to the same class. Under the assumption of independent classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, we can calculate p_c as follows:

$$p_c = \sum_{i=1}^{C} \sum_{u_i^{\mathcal{C}_1}=0}^{1} \sum_{u_i^{\mathcal{C}_2}=0}^{1} p_i^{\mathcal{C}_1} \cdot p_i^{\mathcal{C}_2} \left(u_i^{\mathcal{C}_1} \cdot u_i^{\mathcal{C}_2} \right) \tag{2}$$

The values $p_i^{\mathcal{C}_1}$ and $p_i^{\mathcal{C}_2}$ are the margin probabilities (densities) $p_i^{\mathcal{C}_j} = \frac{1}{N} \sum_{k=1}^{N} u_i^{\mathcal{C}_j}(x_k)$, $j = 1, 2$. p_0 can be counted given a contingency table of the outcomes of both raters:

$$p_0 = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{C} u_i^{\mathcal{C}_1}(\mathbf{x}_k) \cdot u_i^{\mathcal{C}_2}(\mathbf{x}_k) \tag{3}$$

Fleiss' κ_{F} is a direct expansion of Cohen's kappa for $M > 2$ classifiers. We formulate it for the crisp classification according to the above given description of Cohen's κ_{C}. Again, p_0^M is the counted relative agreement between now the M classifiers:

$$p_0^M = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{C} \prod_{j=1}^{M} u_i^{\mathcal{C}_j}(\mathbf{x}_k) \tag{4}$$

The respective value p_c^M calculates as follows:

$$p_c^M = \sum_{i=1}^{C} \sum_{u_i^{\mathcal{C}_1}=0}^{1} \cdots \sum_{u_i^{\mathcal{C}_M}=0}^{1} \prod_{j=1}^{M} p_i^{\mathcal{C}_j} \cdot u_i^{\mathcal{C}_j} \tag{5}$$

κ - value	meaning
$\kappa < 0$	poor agreement
$0 \le \kappa \le 0.2$	slight agreement
$0.2 < \kappa \le 0.4$	fair agreement
$0.4 < \kappa \le 0.6$	moderate agreement
$0.6 < \kappa \le 0.8$	substantial agreement
$0.8 < \kappa \le 1$	perfect agreement

Table 1: Interpretation of κ-values.

whereby we need the latter formulation in the following derivations. Then, the Fleiss' κ_{F} is also calculated according (1) replacing the respective values.

For both kappa the relation $\kappa \in [-1, 1]$ is valid and the values are interpreted according to the scheme given in Tab.1.

2.2 Fuzzy Variants of Cohen's κ_{C} and Fleiss' κ_{F}

In case of fuzzy classifiers the discrete values u_i are turned into continuous values μ_i. We start to derive a fuzzy variant for κ_{C} following [6]. We observe that we can replace the product $u_i^{\mathcal{C}_1} \cdot u_i^{\mathcal{C}_2}$ in (2) and (2) by a *logical* AND-operator without loss. This motivates the opportunity to do so also in case of fuzzy classifiers obtaining

$$P_0 = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{C} \left(\mu_i^{\mathcal{C}_1}(\mathbf{x}_k) \bigwedge \mu_i^{\mathcal{C}_2}(\mathbf{x}_k) \right) \tag{6}$$

Turning from discrete values u_i to continuous values μ_i, the sums over the discrete values for the u_i in (2) become integrals over μ_i. Thus we can calculate the analog P_c as

$$P_c = \sum_{i=1}^{C} \int_{\mu_i^{\mathcal{C}_1}=0}^{1} \int_{\mu_i^{\mathcal{C}_2}=0}^{1} p\left(\mu_i^{\mathcal{C}_1}\right) \cdot p\left(\mu_i^{\mathcal{C}_2}\right) \left(\mu_i^{\mathcal{C}_1} \bigwedge \mu_i^{\mathcal{C}_2}\right) d\mu_i^{\mathcal{C}_1} d\mu_i^{\mathcal{C}_2} \tag{7}$$

where the values $p\left(\mu_i^{\mathcal{C}_1}\right)$ and $p\left(\mu_i^{\mathcal{C}_2}\right)$ are the probability densities of $\mu_i^{\mathcal{C}_1}(x)$ and $\mu_i^{\mathcal{C}_2}(x)$ respectively. In this way, both needed quantities for the (1), can be computed for fuzzy classifiers.

We now derive the fuzzy variant for Fleiss' κ_{F}, too. For this purpose, we first rewrite (5) as

$$p_c^M = \sum_{i=1}^{C} \sum_{u_i^{\mathcal{C}_1}=0}^{1} \cdots \sum_{u_i^{\mathcal{C}_M}=0}^{1} \left(\prod_{j=1}^{M} p_i^{\mathcal{C}_j} \right) \cdot \left(\prod_{k=1}^{M} u_i^{\mathcal{C}_k} \right).$$

Now, replacing the product by an AND-operator and changing the sums into integrals over the continuous values $\mu_i^{\mathcal{C}_1}(x)$ as done for fuzzy κ_{F} we get

$$P_c^M = \sum_{i=1}^{C} \int_{\mu_i^{\mathcal{C}_1}=0}^{1} \cdots \int_{\mu_i^{\mathcal{C}_M}=0}^{1} \left(\prod_{j=1}^{M} p\left(\mu_i^{\mathcal{C}_j}\right) \right) \cdot \left(\bigwedge_{k=1}^{M} \mu_i^{\mathcal{C}_k} \right) d\mu_i^{\mathcal{C}_1} \dots d\mu_i^{\mathcal{C}_M}. \tag{8}$$

Straight forward we obtain analogously

$$P_0^M = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{C} \bigwedge_{j=1}^{M} \mu_i^{\mathcal{C}_j} \left(\mathbf{x}_k \right). \tag{9}$$

Again, the kappa vale is computed inserting P_c^M and P_0^M into (1).

Yet, in the derivation of the fuzzy variants we equivalently replaced the multiplication by the binary AND-operator in the crisp case. Thereafter, we formally interpreted this operator as being valid also for the fuzzy case. However, the AND-operation for fuzzy values is not uniquely determined. There exist many possibilities. The respective theoretic basis is the definition by *t-norms* [7].

2.3 Fuzzy intersections and and-operator based on *t-norms*

To formalize the intersection/AND-operator of fuzzy sets the concept of *t-norms* was introduced. A function $\top : [0,1]^2 \to [0,1]$ is called a *t-norm* if the following holds:

- $\top(a,1) = a$ (neutral element)
- $a \leq b \Rightarrow \top(a,c) \leq \top(b,c)$ (monotonicity)
- $\top(a,b) = \top(b,a)$ (commutativity)
- $\top(a, \top(b,c)) = \top(\top(a,b), c)$ (associativity)

Obviously, the definition doesn't determine uniquely a norm. Examples are

- the min-norm $\top_{\text{min}}(a,b) = \min\{a,b\}$
- the product norm $\top_{\text{prod}}(a,b) = a \cdot b$
- the Lukasiewicz norm $\top_{\text{Luka}}(a,b) = \max\{0, a+b-1\}$

Using different norms for the calculation of κ will lead to different values. Therefore, the interpretation according to Tab. 1 may be misleading and has to be analyzed.

3 Comparison of different *t-norms* for κ_{C} and κ_{F}

3.1 Artificial data set

As explained above the choice for the AND-operator in case of fuzzy classifiers is not unique for the calculation of κ_{C} and κ_{F}. Therefore, we compare the different realizations of the *t-norm* on artificial classifier decisions in dependence of the deviation of these classification from a related crisp decision. We exemplary consider the set of three parametrized two-class-classifiers given in Tab2 with different degrees of fuzziness. We start with investigations for κ_{F}. The experiments a), b) and c) have the same degree of agreement of the classifiers but with increasing level of fuzziness, see Tab2. Hence, the fuzzy κ_{C} should yield approximately the same value as for the original crisp $\kappa_{\text{C}} = 0.6$. The lines 1–3 in Tab.3 report the results. We can conclude that $\top_{\text{min}}$ is the only *t-norm* which judges the agreement adequately. In experiment d) one decision slowly changes.

Classifier	e_0	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
$\mathcal{C}_1$	a	a	a	a	a	$1-b$	$1-a$	$1-a$	$1-a$	$1-a$
$\mathcal{C}_2$	a	a	a	a	$1-a$	a	$1-a$	$1-a$	$1-a$	$1-a$
$\mathcal{C}_3$	a	a	a	a	a	b	$1-a$	$1-a$	$1-a$	$1-a$

Table 2: Description of the classifiers for a two-class problem. Depicted are the fuzzy values for the decision for the repective classifiers for the first class. The second class is chosen accordingly. Example cases: a) *crisp* $a = 1$, $b = 0$ b) *slightly fuzzy* $a = 0.95$, $b = 0$ c) *moderate fuzzy* $a = 0.75$, $b = 0$ d) *single fuzzy* $a = 1$, $b = 0.95$, 0.75, 0.5.

	$\top_{\text{min}}$	$\top_{\text{prod}}$	$\top_{\text{Luka}}$	$\top_{\text{min}}$	$\top_{\text{prod}}$	$\top_{\text{Luka}}$
$a = 1.00$, $b = 1$	0.5988	0.5992	0.5992	0.7308	0.7323	0.7323
$a = 0.95$, $b = 1$	0.6000	0.4860	0.4909	0.7263	0.5924	0.5921
$a = 0.75$, $b = 1$	0.6000	0.1500	0.2000	0.6842	0.1800	0.1394
$a = 1$, $b = 0.95$	0.6089	0.6093	0.6093	0.7308	0.7323	0.7323
$a = 1$, $b = 0.75$	0.6490	0.6493	0.6494	0.7308	0.7323	0.7323
$a = 1$, $b = 0.5$	0.6991	0.6994	0.6995	0.7309	0.7323	0.7324

Table 3: Fuzzy κ_{C} (left part) and κ_{F}-values (right part) for the agreement of the classifiers in table 2. In case of κ_{C}, the classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ were used for the lines 1–3, for the lines 4–6 these are the classifiers $\mathcal{C}_1$ and $\mathcal{C}_3$.

The results correspond to line 4–6 in Tab.3. We observe increasing κ_{C}-values with increasing parameter b in $\mathcal{C}_3$ which is in agreement that the decision of $\mathcal{C}_3$ for class 2 becomes more and more uncertain.

The same experiments were processed for κ_{F} but here taking all three classifiers into account. In the crisp case one obtains $\kappa_{\mathrm{F}} = 0.7333$. The fuzzy results are given in Tab.3 emphasizing the observation for κ_{C}: The best results are obtained for $\top_{\text{min}}$.

3.2 Real world data

To test the findings for exemplary artificial data we analyzed classifications of different classifiers for a 13-class problem. The data set was established together with the Helmut Hund GmbH, Wetzlar (Germany) in the development of a fully automated system for the recognition of pollen concentrations in the ambient air. The Bio Aerosol Analyzer (BAA 1000) collects pollen from the ambient air. They are prepared as a probe that afterwards is scanned using a transmitted light microscope. Single objects are segmented from the image and 75 features are extracted for every single object. For the establishment of the data set we randomly selected 300 samples for each of 13 pollen classes. The data set was then randomly divided up into balanced training and test set. In the results we will show comparisons on the test set.

We applied 3 different classifiers, which all give fuzzy decisions. Namely, these are a multi class Linear Discriminant Analysis (*mLDA*) and a one-vs.-all approach for all pollen classes using a simple LDA (*OVA-LDA*) [8]. Both are linear classifiers. We compare these results with a FLSOM-classifier, which can be seen as a non-linear semi-supervised fuzzy-classifier but restricted to the apriori fixed lattice structure [9]. The balancing factor between unsupervised and supervised learning was set such that unsupervised learning was dominating. Crisp classification accuracies for the classifiers are 91.6%, 88.5% and 76.3% respectively (majority vote) which yields a $\kappa_{\mathrm{F}} = 0.2988$. The κ_{F}-values for different *t-norms* are given in Tab.4. We observe also for this application that the several norms lead to differences for κ_{F}. We can conclude with high certainty from the above considerations for artificial data that $\top_{\mathrm{min}}$ will lead to the best result in case of fuzzy classifiers and, therefore, $\kappa_{\mathrm{F}} = 0.0133$ approximates best the total agreement recommending only a slight one. This is the consequence of the disagreement between FLSOM and the both LDA-approaches: *mLDA* vs. *OVA-LDA*: $\kappa_{\mathrm{C}} = 0.8107$, *mLDA* vs. *FLSOM*: $\kappa_{\mathrm{C}} = -0.0281$, *FLSOM* vs. *OVA-LDA*: $\kappa_{\mathrm{C}} = -0.0269$ (κ_{C}-values calculated using $\top_{\mathrm{min}}$), i.e. semi-supervised learning with an only small amount of supervision is not sufficient to learn the classification.

t-norm	$\top_{\mathrm{min}}$	$\top_{\mathrm{prod}}$	$\top_{\mathrm{Luka}}$
fuzzy κ_F	0.0113	0.0450	0.0155

Table 4: Fuzzy κ_{F}-values for the agreement of the three classifiers *mLDA*, *OVA-LDA*, *FLSOM* with respect to several *t-norms* in case of the pollen data.

4 Discussion

In this paper we develop a fuzzy extension of Fleiss' κ_{F} following an approach for fuzzy Cohen's κ_{C}. Furthermore, we discuss for both measures the dependency on the utilized *t-norm* with respect to the interpretability. It turns out that the minimum norm $\top_{\mathrm{min}}$ seems to be most appropriate.

References

[1] F.-M. Schleif, M. Ongyerth, and T. Villmann. Supervised data analysis and reliability estimation for spectral data. *Neurocomputing*, page submitted, 2009.

[2] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[3] J. Cohen. Weighted chi square: An extension of the kappa method. *Educational and Psychological Measurement*, 32:61–74, 1972.

[4] J.L. Fleiss. *Statistical Methods for Rates and Proportions.* Wiley, New York, 2nd edition, 1981.

[5] An alternative to cohen's $\kappa. EuropeanPsychologist, 11 : 12 - -24, 2006.$

[6] W. Dou, Y. Ren, Q. Wu, S. Ruan, Y. Chen, and D. Bloyet AndJ.-M. Constans. Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing*, 70:726–734, 2007.

[7] B. Hammer and Th. Villmann. How process uncertainty in machine learning? In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2007)*, pages 79–90, Brussels, Belgium, 2007. d-side publications.

[8] L. Sachs. *Angewandte Statistik.* Springer Verlag, 7-th edition, 1992.

[9] T. Villmann, F.-M. Schleif, M. Kostrzewa, A. Walch, and B. Hammer. Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.