Interpretation of the Fleiss' kappa Computation

Tianyi Sun

1 BACKGROUND

When assigning categorical ratings to a number of (classifying) items, **Cohen's kappa** assesses the reliability of the agreement between two raters or the intra-rater reliability[1] , and **Fleiss' kappa** assesses the reliability of the agreement between $n \in \mathbf{Z}$ number of raters. Both Fleiss' kappa and Cohen's kappa are statistical measures, which calculates the degree of agreement in classification. Fleiss' kappa specifically allows that different items could be rated by a fixed number of different individuals[2] .

1.1 Definition

Let

- N be the total number of subjects;
- n be the number of ratings per subject;
- k be the number of categories into which assignments are made.

The subjects are indexed by $i = 1, ..., N$ and the categories are indexed by $j = 1, ..., k$. Let n_{ij} represent the number of raters who assigned the i-th subject to the j-th category.

First calculate the proportion of all assignments which were assigned to the j-th category, p_j:

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, \qquad 1 = \sum_{j=1}^{k} p_j$$

Second calculate the extent to which raters agree for the i-th subject[3] , P_i:

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1) = \frac{1}{n(n-1)}\sum_{j=1}^{k}(n_{ij}^2 - n_{ij}) = \frac{1}{n^2-n}\left[\left(\sum_{j=1}^{k} n_{ij}^2\right) - n\right].$$

Third compute $\bar{P}$, the mean of the P_i's, and $\bar{P_e}$, then plugging both into the formula (1.1) for κ:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{N(n^2-n)}\left[\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2\right) - Nn\right]$$

$$\bar{P_e} = \sum_{j=1}^{k} p_j^2$$

The Fleiss' kappa, κ, is defined as,

$$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}$$

The factor $1 - \bar{P_e}$ gives the degree of agreement that is attainable above chance. In other words, the maximum degree of agreement.
The factor $\bar{P} - \bar{P_e}$ gives the degree of agreement actually achieved above chance.

1.2 Value interpretation

Wikipedia[1] gives an interpretation of Fleiss' kappa, κ, values for a 2-annotator 2-class example, Figure 1.

1.3 Application in Normal Case

Please refer to the Worked example in Wikipedia[1], here we call it the Normal Case.

[1] One rater versus him/herself.
[2] For example, Item 1 is rated by Raters A, B, and C; but Item 2 could be rated by Raters D, E, and F
[3] For example, compute how many rater–rater pairs are in agreement, relative to the number of all possible rater–rater pairs

Condition	κ	Interpretation
Subjective example: ***only*** *for two annotators, on two classes. See Landis & Koch 1977*	< 0	Poor agreement
	0.01 – 0.20	Slight agreement
	0.21 – 0.40	Fair agreement
	0.41 – 0.60	Moderate agreement
	0.61 – 0.80	Substantial agreement
	0.81 – 1.00	Almost perfect agreement

Figure 1: An interpretation of the Fleiss' kappa value [1].

2 DEVELOPMENT

2.1 Application in our case

The difference between our case, Table 2, and the Normal Case, is that one annotator could assign multiple categories to a single SQL hypothesis, however, in Normal Case, one annotator is only allowed to assign one category for one SQL hypothesis. So for matrices constructed for computing the Fleiss' kappa, in our case, each row sum is not a fixed number, however, in Normal Case, each row sum is a fixed number, which equals to the number of annotators, n.

	A	B	L	C	K	D	F	N	O	P	E
0	2	0	0	0	0	0	0	0	0	0	0
1	0	2	2	2	0	0	0	0	0	0	0
2	2	0	0	0	2	0	0	0	0	0	0
3	0	2	2	0	0	0	0	0	0	0	0
4	0	0	2	2	0	2	0	0	0	0	0
5	0	0	0	0	0	2	0	0	0	0	0
6	0	2	2	0	0	0	0	1	0	0	0
7	0	0	0	0	0	2	0	0	0	0	0
8	0	0	0	0	0	2	0	0	0	0	0
9	0	0	0	0	1	2	0	0	0	0	0
10	0	2	0	0	0	0	2	0	0	0	0
11	0	0	0	0	0	0	0	2	2	0	0
12	2	0	0	0	0	0	0	0	0	0	0
13	0	2	2	0	0	0	1	2	0	0	0
14	0	2	2	1	1	0	0	1	0	2	1
15	0	0	0	2	0	0	0	0	0	0	2
16	0	0	0	2	0	0	0	0	0	0	2
17	0	2	2	0	0	0	0	0	0	0	0
18	2	0	0	0	2	0	0	0	0	0	0
19	0	0	0	0	0	0	2	0	0	0	0
20	0	0	2	2	0	0	0	0	0	0	0

Figure 2: This is the first 21 rows of a matrix constructed for computing the Fleiss' kappa based on the first batch of the second round annotation. Each column represents an error category. Each row represents a SQL hypothesis. There are two annotators. Each annotator are allowed to assign multiple error categories to a single SQL hypothesis.

2.2 Batches 4-7 of the first round annotation and all batches of the second round annotation

Take the first batch of the second round annotation as an example. We have two annotators, forty-sixth SQL hypothesis, and eleven error categories. We first construct a matrix, Table 2 shows a part of it, which is the same as what the Normal Case did. To address the problem mentioned above, for each SQL hypothesis/row we construct a matrix, Table 3, and compute the Fleiss' kappa value for it. Each subject/row represents an error category. The elements in the matrix represents how many annotators have selected or not selected the error category for the corresponding SQL hypothesis. Since the number of annotators is two and they are either selected an error category or not for each SQL hypothesis, each row sum is two, which is fixed. Then we compute the Fleiss' kappa value for those matrix as the Normal Case. At last, we take the average of the sum of Fleiss' kappa values computed for each of the SQL hypothesis as the Fleiss' kappa value for that batch.

SQL hypothesis 0

	selected	not_selected
A	2	0
B	0	2
L	0	2
C	0	2
K	0	2
D	0	2
F	0	2
N	0	2
O	0	2
P	0	2
E	0	2

SQL hypothesis 1

	selected	not_selected
A	0	2
B	2	0
L	2	0
C	2	0
K	0	2
D	0	2
F	0	2
N	0	2
O	0	2
P	0	2
E	0	2

SQL hypothesis 2

	selected	not_selected
A	2	0
B	0	2
L	0	2
C	0	2
K	2	0
D	0	2
F	0	2
N	0	2
O	0	2
P	0	2
E	0	2

SQL hypothesis 3

	selected	not_selected
A	0	2
B	2	0
L	2	0
C	0	2
K	0	2
D	0	2
F	0	2
N	0	2
O	0	2
P	0	2
E	0	2

Figure 3: For each SQL hypothesis/row in Table 2, we construct a matrix, where each subject/row is an error category. The elements represents that how many annotators have selected or not selected a error category for the corresponding SQL hypothesis. Each row sum is two which is fixed. This figure shows four matrices corresponding to the first four rows in Table 2.

2.3 Value interpretation

As those sub matrices, Figure 3, are 2-annotator 2-class examples, the interpretation in Figure 1 could be applied in our case.

2.4 Batches 1-3 of the first round annotation

Those batches are computed similar with the last section except that the number of annotators, n, is 4.

2.5 Overall batches

We make a big matrix similar with Table 2. The rows are the overall SQL hypothesis. We use the same approach as illustrated above.

REFERENCES

[1] 2021. Fleiss' kappa. https://en.wikipedia.org/wiki/Fleiss_kappa