A measure of expected agreement between independent classifiers*

Emil Bashkansky
Yariv N. Marmor

BRAUDE College of Engineering, Karmiel

University of Piraeus - ENBIS 2025
17/09/2025

*The topic of our report is essentially a continuation of the study metrological aspects of classification systems, the results of which were presented at the previous annual ENBIS conference and published earlier this year in
Gadrich, T.; Marmor, Y. N.; Bashkansky, E. (2025) Accuracy of Categorical Measurements: Nominal Scale. *Measurement,* 250, 117044

The Purpose of the Presentation

To propose the new measure of estimating inter-classifiers agreement based on metrological characteristics of classification system only, when classification is provided by several collaborators (classifiers) according to fixed and random model of their selection.

The Seven Sages (Latin: *Septem Sapientes*), depicted in the *Nuremberg Chronicle*

Brief *"roll back"* to the previous presentation

"Classification of the analyzed property value of the objects under study (OUS) into one of K exclusive categories forming a comprehensive spectrum (scale) of the studied property will be considered as categorical measurement." **(*)**

Note 1: The results of classification are presented by so-called categorical data. In cases where the spectrum of possible values consists only of two categories such data are binary, and the appropriate activity is also often called ***testing.***

Note 2: In this presentation categories are not ordered (*nominal* scale)

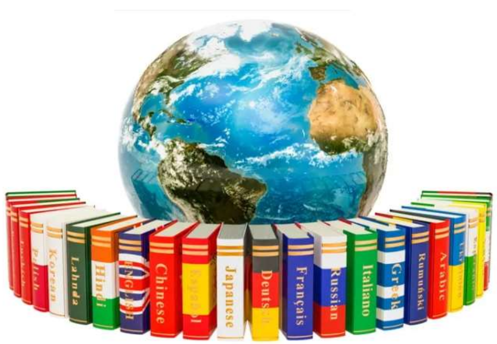
* T. Gadrich, E. Bashkansky, (2016) " A Bayesian approach to evaluating uncertainty of inaccurate categorical measurements", *Measurement* 91, 186–193.

Examples of *Classifiers* ($K > 2$)

Coins sorting machine

Egg classification machine

Plastic color sorting machine



Google language detector

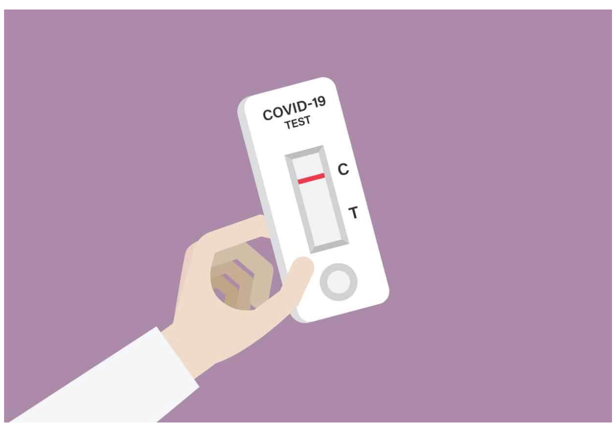
Musical tone recognition

Examples of *Binary Classifiers* (*K* = 2)

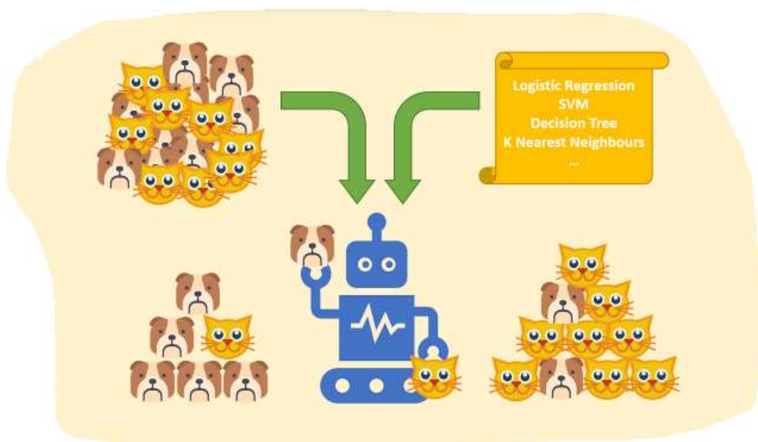
Pregnancy tester

Spam filtering

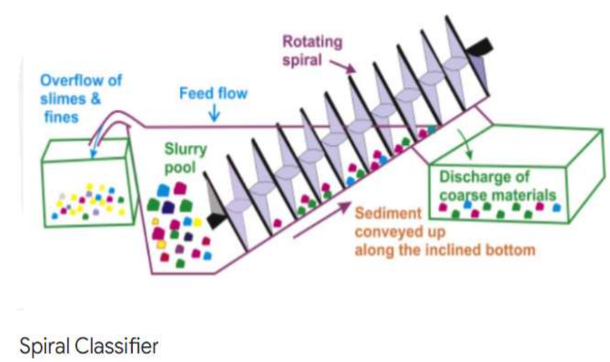
Counting machine: bank notes are classified as forged or accepted



Covid - 19 tester



Classification algorithm



Spiral classifier

***Ability* of a *single* classifier**

The conditional probabilities that an object will be classified as category ***k***, given that its actual/true category is ***i*** - $P_{k|i}$

$$1 \leq i, k \leq K; \qquad \sum_{k=1}^{K} P_{k|i} = 1$$

Ideal classifier: $every\ P_{k|i} = 0, except\ of\ P_{i|i} = 1$

Classification (Confusion) Matrix and Repeatability for the General Case of *K* Categories

$$\boldsymbol{P} = \begin{pmatrix} p_{1|1} & p_{2|1} & \cdots & p_{k|1} & \cdots & p_{K|1} \\ p_{1|2} & p_{2|2} & \cdots & p_{k|2} & \cdots & p_{K|2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{1|i} & p_{2|i} & \cdots & p_{k|i} & \cdots & p_{K|i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{1|K} & p_{2|K} & \cdots & p_{k|K} & \cdots & p_{K|K} \end{pmatrix} \qquad \boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

$$\boldsymbol{Repeatability}\ (variation) = \begin{bmatrix} Repeatability_1 \\ Repeatability_2 \\ \cdots \\ Repeatability_i \\ \cdots \\ Repeatability_K \end{bmatrix}$$ - (for the **same** classifier)

$$Repeatability_i = \frac{K}{K-1}\sum_{k=1}^{K}\left[p_{k|i}\cdot\left(1-p_{k|i}\right)\right] = \frac{K}{K-1}\left(1-\sum_{k=1}^{K}p_{k|i}^2\right) = VAR_{within}$$

$$0 \le VAR_{within} \le 1$$

The ***closeness between classifications abilities*** of **different and independent** ***classifiers*** participating in collaborative study (from here on - ***Classifiers effect***)

$$Classifiers\ effect\ (variation) = \begin{bmatrix} Classifier\ effect_1 \\ Classifier\ effect_2 \\ \cdots \\ Classifier\ effect_i \\ \cdots \\ Classifier\ effect_K \end{bmatrix}$$

Where: $Classifiers\ effect_i = \frac{K}{K-1} \sum_{k=1}^{K} VAR\left(p_{k|i}\right)$

$VAR\left(p_{k|i}\right)$ = classic variation of $p_{k|i}$ between ***L*** collaborators/classifiers

CATANOVA (*L* classifiers)

$$\text{Total Variation}_i = \frac{K}{K-1}\sum_{k=1}^{K}\overline{p}_{k|i}(1-\overline{p}_{k|i}) = \frac{K}{K-1}\left(1-\sum_{k=1}^{K}\overline{p}_{k|i}{}^{2}\right),$$

$$where\ \overline{p}_{k|i} = \frac{\sum_1^L \overline{p}_{k|i}^{(l)}}{L}$$

can be split to the *intra* and *inter* components:

$$\text{Total Variation}_i =$$
$$(\text{Mean repeatability variation})_i + \left(\text{Classidiers}'\text{effect variation}\right)_i$$

Uniform kappa index of agreement between two classifiers

$$\kappa = \frac{P_a - P_{a|Chance}}{1 - P_{a|Chance}}$$

$$P_{a|Chance} = \frac{1}{K}$$

In case of two classifiers, this means that one or both made maximally random, blind, and uninformative classifications.

$$1 = \kappa \geq -\frac{1}{K-1}$$

The agreement thus defined satisfies a very important superposition principle (*), i.e: "*the overall kappa is the weighted sum of partial categories kappa-s, where the "weight" of every category is the probability of an OUS to belong to this category (or its proportion in the classified population)*" .

(*) E. Bashkansky, T. Gadrich, "Some metrological aspects of the comparison between two ordinal measuring systems", *Accreditation and Quality Assurance*, Vol. 16, pp. 63-72, 2011

Uniform kappa index of agreement – dependence on K

$$\kappa = \frac{P_a - P_{a|Chance}}{1 - P_{a|Chance}}$$

$$P_a = 0.5;\ P_{a|Chance} = \frac{1}{K} \Rightarrow \kappa = \frac{0.5 - \frac{1}{K}}{1 - \frac{1}{K}}$$

When a half of all items are classified identically:

K	2	3	4	5	6	7	8	9	10	20	100
κ	0	0.25	0.33	0.38	0.4	0.42	0.43	0.44	0.44	0.47	0.5

"THE MORE CATEGORIES ARE IN THE SPECTRUM, THE HARDER IT IS TO GUESS"

Expected partial kappa index of agreement between two independent classifiers

Assuming that:

$$\overline{P}_{a|i} = \sum_{k=1}^{K} P_{k|i}^{(1)} \cdot P_{k|i}^{(2)} = \overrightarrow{P_i^{(1)}} \cdot \overrightarrow{P_i^{(2)}}$$

It is possible to prove, that:

$$\kappa_i = \frac{\overline{P}_{a|i} - P_{a|Chance}}{1 - P_{a|Chance}} = \frac{\overline{P}_{a|i} - \frac{1}{K}}{1 - \frac{1}{K}} =$$

$$= 1 - (\text{Total variation})_i - (\text{Classifiers}'\text{effect variation})_i$$

and by virtue of the superposition principle:

$$\kappa = 1 - (\text{Total variation}) - (\text{Classifiers}'\text{effect variation})$$

or alternatively:

$$\kappa = 1 - (\text{Repeatability variation}) - 2 \cdot (\text{Classifiers}'\text{effect variation})$$

Expected partial and general kappa index of agreement between L independent classifiers

Assuming that:

$$\overline{P}_{a|i} = \frac{2}{L(L-1)} \sum_{l'>l}^{L} \sum_{l}^{L} \overrightarrow{P_i^{(l\ )}} \cdot \overrightarrow{P_i^{(l')}}$$

It is possible to prove, that:

$$\kappa_i = 1 - (\text{Total variation})_i - \frac{1}{L-1}\,(\text{Classifiers}'\text{effect variation})_i \equiv$$
$$1 - (\text{Repeatability variation})_i - \frac{L}{L-1}\,(\text{Classifiers}'\text{effect variation})_i$$

and by virtue of the superposition principle which is valid for every mutual agreements:

$$\kappa = 1 - (\text{Total variation}) - \frac{1}{L-1}(\text{Classifiers}'\text{effect variation}) \equiv$$
$$1 - (\text{Repeatability variation}) - \frac{L}{L-1}(\text{Classifiers}'\text{effect variation})$$

```
Enter the number of Categories (K): 2
Enter the number of Classifiers (L): 2

Enter probabilities for Classifier number 1:
Enter probability for Category 1: 1
Enter probability for Category 2: 0
Probabilities for Classifier 1 entered successfully.

Enter probabilities for Classifier number 2:
Enter probability for Category 1: 1
Enter probability for Category 2: 0
Probabilities for Classifier 2 entered successfully.

Entered probability table:
[[1. 0.]
 [1. 0.]]

--- Results ---
Agreement value = 1.0000
Repeatability = 0.0000
ClassifiersEffect = 0.0000
Total_Variation = 0.0000
```

```
Enter the number of Categories (K): 2
Enter the number of Classifiers (L): 2

Enter probabilities for Classifier number 1:
Enter probability for Category 1: 1
Enter probability for Category 2: 0
Probabilities for Classifier 1 entered successfully.

Enter probabilities for Classifier number 2:
Enter probability for Category 1: 0
Enter probability for Category 2: 1
Probabilities for Classifier 2 entered successfully.

Entered probability table:
[[1. 0.]
 [0. 1.]]

--- Results ---
Agreement value = -1.0000
Repeatability = 0.0000
ClassifiersEffect = 1.0000
Total_Variation = 1.0000
```

```
Enter the number of Categories (K): 2
Enter the number of Classifiers (L): 2

Enter probabilities for Classifier number 1:
Enter probability for Category 1: 0.5
Enter probability for Category 2: 0.5
Probabilities for Classifier 1 entered successfully.

Enter probabilities for Classifier number 2:
Enter probability for Category 1: 0.5
Enter probability for Category 2: 0.5
Probabilities for Classifier 2 entered successfully.

Entered probability table:
[[0.5 0.5]
 [0.5 0.5]]

--- Results ---
Agreement value = 0.0000
Repeatability = 1.0000
ClassifiersEffect = 0.0000
Total_Variation = 1.0000
```

```
Enter the number of Categories (K): 2
Enter the number of Classifiers (L): 2

Enter probabilities for Classifier number 1:
Enter probability for Category 1: 1
Enter probability for Category 2: 0
Probabilities for Classifier 1 entered successfully.

Enter probabilities for Classifier number 2:
Enter probability for Category 1: 0.5
Enter probability for Category 2: 0.5
Probabilities for Classifier 2 entered successfully.

Entered probability table:
[[1.  0. ]
 [0.5 0.5]]

--- Results ---
Agreement value = 0.0000
Repeatability = 0.5000
ClassifiersEffect = 0.2500
Total_Variation = 0.7500
```

Random *Dirichlet* Model

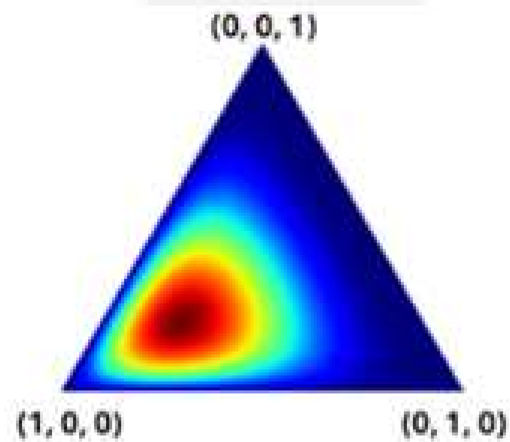
- L classifiers are randomly sampled from the population which classification abilities related to category i are distributed according to the Dirichlet distribution:

$$f\left(p_{1|i}, p_{2|i}, \cdots, p_{K|i}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_{k|i}{}^{\alpha_{k|i}-1} \qquad \left(\sum_{k=1}^{K} p_{k|i} = 1 \right)$$

where: $\boldsymbol{\alpha_i} = \left(\alpha_{1|i}, \alpha_{2|i}, \cdots, \alpha_{K|i}\right)$ - *parameters of the Dirichlet distribution characterizing the i-th category classification*;

$E\left(p_{k|i}\right) = \alpha_{k|i} / \alpha_{0|i}$;

$\alpha_{0|i} = \sum_{k=1}^{K} \alpha_{k|i}$ - *concentration parameter* (*)



Dir(4,2,2)

* Location is determined by repeatability, and dispersion ($\alpha_{0|i}$) is determined by degree of variation between classifiers.

Repeatability* (*within*) and *Classifiers* (*between-classifiers*) *effect

$$\boldsymbol{Repeatability_i} = \frac{K}{K-1}\sum_{k=1}^{K} E\left[p_{k|i}\cdot\left(1-p_{k|i}\right)\right] = \frac{K}{K-1}\frac{\alpha_{0|i}}{\alpha_{0|i}+1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right]$$

$$\boldsymbol{Classifier\ effect_i} = \frac{K}{K-1}\sum_{k=1}^{K} VAR\left(p_{k|i}\right) = \frac{K}{K-1}\frac{1}{\alpha_{0|i}+1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right] = \frac{\boldsymbol{Repeatability}}{\alpha_{0|i}}$$

$$\boldsymbol{Disagreement\ value_i} = \frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right]\left(1+\frac{1/(L-1)}{(\alpha_{0|i}+1)}\right) =$$

$$= \boldsymbol{Total\ precision\ variation}\cdot\left[1+\frac{1}{(L-1)}\frac{1}{(\alpha_{0|i}+1)}\right]$$

$\alpha_{0\|i} \to 0$	$\alpha_{0\|i} \to \infty$
Repeatability = 0	Classifiers effect =0
$\boldsymbol{Disagreement\ value_i} = \frac{L}{L-1}\frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k\|i}^2}{\alpha_{0\|i}^2}\right]$	$\boldsymbol{Disgreement\ value_i} = \frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k\|i}^2}{\alpha_{0\|i}^2}\right]$

How will the results of a war with Iran affect the Abraham Accords?

(a poll conducted shortly after the end of the 12-day war)

1. The circle of countries that have signed the Abraham Accords will expand: **54%**
2. The circle of countries that have signed the Abraham Accords will shrink: **3%**
3. Nothing will change: **37%**
4. I find it difficult to answer: **6%**

Total Votes: **6935**

***Agreement value* = 1 - (4/3)[1- (0.54^2+0.03^2+0.37^2+0.06^2) = 0.244**

(Fleiss' kappa – 0.282)

Comparison

	Question	Fleiss' score	Proposed score
1	Is expected agreement expressed directly through the precision characteristics of the classification system?	**No**	**Yes**
2	Whether total score is a weighted sum of partial (category-specific) ones?	**No** (prevalence paradox)	**Yes** (no prevalence paradox)
3	If we add one more classifier to a group of classifiers whose classifications coincide with the average values of this group, how will the expected agreement change?	Kappa score will **decrease**	Proposed will **increase**
4	What is the agreement score for 100 classifiers, 99 of which systematically point to the first category and only one systematically points to the second?	Approximately **0.0101**	Proposed is approximately **0.96**

Summary

1. The agreement between classifiers is important whenever the consistency, reliability and trustworthy of judgment are crucial for data quality and decision-making, especially where the cost of false output is high.
2. This agreement is directly related to the metrological characteristics of the precision of the classification system, its *intra* and *inter* (R&R) components.
3. The proposed measure of agreement satisfies the superposition principle, i.e. the overall measure is the weighted sum of partial categories measures.
4. For a sufficiently large number of classifiers, the proposed measure of *disagreement* simply coincides with the total precision variation of the classification system.
5. In the absence of general agreement, the authors plan to investigate the possibility of using the proposed measure to solve the problems of classifiers' clustering.

Thank You for Your Attention!

E-mail: *ebashkan@braude.ac.il*

Link to the Tool:
https://drive.google.com/file/d/1WpqpgcBJ_QhqCgVnlc1yaBQYbKEYIXnd/view?usp=sharing