`classify` command: Over 260 measures of association and forecast evaluation for categorical data

Andrei Sirchenko (Nyenrode Business University)
Dragoș Bînzari (University of Amsterdam)
Konrad Wrębiak (University of Amsterdam)

June 19, 2026

To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it.

— K. A. C. Manderville, The Undoing of Lamia Gurdleneck

Measuring association & similarity

Contingency table

	Male	Female	Total
Blonde	8	16	24
Brunette	14	18	32
Total	22	34	56

Measuring associations and similarities

Contingency table

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Dice coefficient: $\frac{2n_{11}}{n_{1+}+n_{+1}} = 0.44$
- Yule association coefficient Q: $\frac{n_{11}n_{22}-n_{12}n_{21}}{n_{11}n_{22}+n_{12}n_{21}} = 0.26$
- Heidke skill score: $\frac{2(n_{11}n_{22}-n_{12}n_{21})}{n_{+1}n_{2+}+n_{1+}n_{+2}} = 0.12$

Evaluating categorical forecasts

Confusion matrix

Predicted values	Actual values: Positive	Actual values: Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

- Accuracy: $\frac{TP+TN}{n} = 0.57$
- Peirce skill score: $\frac{TP \times TN - FP \times FN}{(TP+FN)(FP+TN)} = 0.11$
- F_1-score: $\frac{2TP}{2TP+FP+FN} = 0.44$
- Yule correlation r_φ: $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(FP+TN)(FN+TN)}} = 0.12$

Evaluating probabilistic forecasts of categorical data

Diagnostic probability scores

- Brier score:

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\Pr\left(y_i=k\right)-\delta_{ik}\right)^2$$

- Logarithmic score:

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\ln(\Pr(y_i=k))+(1-\delta_{ik})\ln(1-\Pr(y_i=k))$$

- Ranked probability score:

$$\frac{1}{n(K-1)}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left(\sum_{j=1}^{k}\Pr\left(y_i=j\right)-\sum_{j=1}^{k}\delta_{ij}\right)^2$$

Literature on measures of association

is poorly integrated across different fields

- a wide variety of scalar statistics have been developed and used in different fields
- a similarly wide variety of nomenclature has appeared in relation to these statistics
- some of these measures have been reinvented, duplicated and misattributed on multiple occasions in other fields
- confusing terminology is confounded further by different notation

Reinvention and duplication

- Cohen kappa coefficient in psychometrics and inter-rater reliability (1960): $\frac{n\sum_{k=1}^{K} n_{kk} - \sum_{k=1}^{K} n_{k+}n_{+k}}{n^2 - \sum_{k=1}^{K} n_{k+}n_{+k}}$

Reinvention and duplication

- Cohen kappa coefficient in psychometrics and inter-rater reliability (1960): $\frac{n\sum_{k=1}^{K} n_{kk} - \sum_{k=1}^{K} n_{k+}n_{+k}}{n^2 - \sum_{k=1}^{K} n_{k+}n_{+k}}$

- Heidke skill score in meteorological forecast verification (1926): $\frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{+1}n_{2+} + n_{1+}n_{+2}}$

Reinvention and duplication

- Cohen kappa coefficient in psychometrics and inter-rater reliability (1960): $\frac{n\sum_{k=1}^{K} n_{kk} - \sum_{k=1}^{K} n_{k+}n_{+k}}{n^2 - \sum_{k=1}^{K} n_{k+}n_{+k}}$
- Heidke skill score in meteorological forecast verification (1926): $\frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{+1}n_{2+} + n_{1+}n_{+2}}$
- Doolittle association ratio in early weather-forecast verification (1887)

Reinvention and duplication

Year	Name / attribution	Field
1909	**Czekanowski**	Anthropology, taxonomy, numerical classification
1920	**Gleason**	Plant ecology, quadrat comparison
1945	**Dice**	Ecology, species association
1948	**Sørensen**	Plant sociology, ecological similarity
1948	**Burt**	Psychometrics / classification catalogues
1956	**Bray–Curtis**	Ecology, ordination, community ecology
1966	**Lance–Williams**	Numerical taxonomy, cluster analysis
1974	**F_1-score**	Information retrieval; machine learning.
1977	**Upholt *F***	Molecular genetics, restriction-fragment comparison
1979	**Nei–Li coefficient**	Molecular evolution, genetic similarity
1980s	**Pirlot/White**	Binary-similarity catalogues, classification literature
1994	**Zijdenbos**	Medical image segmentation, neuroimaging
2000s	**Fleiss–Levin–Paik**	Inter-rater reliability, agreement analysis

Alternative terminology

- Accuracy
- (Crude) agreement
- Causal support
- Classification rate
- Count R^2
- Hit score
- Holsti $C.R.$ coefficient
- Kendall coefficient
- Osgood coefficient
- Proportion correct
- Rand coefficient
- Ratio test discriminant
- Simple matching coefficient
- Sokal-Michener coefficient

A new Stata command classify

Output

- A contingency table (confusion matrix)
- 256 measures of association and 11 diagnostic scores for probabilistic forecasts
- The global ranges for each measure
- The class-specific measures for each class as well as their simple and weighted averages
- An Excel file with all computed measures

Output of command classify

Input: contingency table (confusion matrix)

```
. matrix Confusion = (11,2,1 \ 4,16,1 \ 3,8,7)

. classify, mat(Confusion)
```

Contingency Table

Actual	1	2	3
Predicted			
1	11	2	1
2	4	16	1
3	3	8	7

Selected association & similarity measures

Measure	Range	Value
1. Accuracy	[0 → 1]	**0.64**
11. Balanced accuracy	[0 → 0.5 → 1]	**0.67**
42. Clayton skill score	[-1 → 0 → 1]	**0.46**
54. Cramer concordance	[0 → 1]	**0.24**
129. Goodman-Kruskal lambda weighted	[0 → 1]	**0.43**
131. Gorodkin Rk	[-1 → 0 → 1]	**0.47**
136. Hamann	[-1 → 1]	**0.28**
139. Heidke skill score	[-1 → 0 → 1]	**0.46**
190. Peirce skill score	[-1 → 0 → 1]	**0.49**
211. Scott pi	[-1 → 1]	**0.45**
216. Sokal-Sneath 1	[0 → 1]	**0.78**
232. Theil	[0 → 1]	**0.21**
235. Tschuprow T bias corrected	[0 → 1]	**0.46**

Output of command classify

Input: values of two variables

```
. classify x2 y2
```

Contingency Table

x2=	1	0
y2=		
1	58	127
0	40	54

Selected association & similarity measures

Measure	Range	Value	Class 1	Class 0	Macro avg	Weighted avg
1. Accuracy	[0 → 1]	0.40				
11. Balanced accuracy	[0 → 0.5 → 1]	0.45				
42. Clayton skill score	[-1 → 0 → 1]	-0.11				
56. Czekanowski	[0 → 1]	0.41	0.41	0.39	0.40	0.40
84. G-mean	[0 → 1]	0.42	0.42	0.42	0.42	0.42
130. Goodman-Kruskal tau	[0 → 1]	0.01				
136. Hamann	[-1 → 1]	-0.20				
139. Heidke skill score	[-1 → 0 → 1]	-0.09				
140. Hit rate	[0 → 1]	0.59	0.59	0.30	0.45	0.40
190. Peirce skill score	[-1 → 0 → 1]	-0.11				
197. Precision	[0 → 1]	0.31	0.31	0.57	0.44	0.48
211. Scott pi	[-1 → 1]	-0.20				
253. Yule phi	[-1 → 0 → 1]	-0.11	-0.11	-0.11	-0.11	-0.11
254. Yule Q	[-1 → 0 → 1]	-0.24	-0.24	-0.24	-0.24	-0.24
255. Yule Y	[-1 → 0 → 1]	-0.12	-0.12	-0.12	-0.12	-0.12

Output of command classify

Input: observed values & predicted probabilities

```
. quietly oprobit y bias house gdp spread

. predict p1 p2 p3
(option pr assumed; predicted probabilities)

. classify y, probs(p1 p2 p3) metrics(p1 p7 p11 1 3 56 139 140 190 211 )
```

Confusion Matrix

Actual	-1	0	1
Predicted			
-1	30	9	0
0	25	163	26
1	0	9	17

Selected probabilistic scores

Score	Range	Value
Brier score	[0 ← 1]	**0.1679**
Ranked probability score (ORD)	[0 ← 1]	**0.0847**
Zero-one score	[0 ← 1]	**0.2473**

ORD: suited for ordinal data only.

Selected association & similarity measures

Measure	Range	Value	Class -1	Class 0	Class 1	Macro avg	Weighted avg
1. Accuracy	[0 → 1]	**0.75**					
3. Adjusted noise-to-signal ratio	[0 ← 1 ← 278]		**0.07**	**0.58**	**0.10**	**0.25**	**0.40**
56. Czekanowski	[0 → 1]		**0.64**	**0.83**	**0.49**	**0.65**	**0.74**
139. Heidke skill score	[-1 → 0 → 1]	**0.46**					
140. Hit rate	[0 → 1]		**0.55**	**0.90**	**0.40**	**0.61**	**0.75**
190. Peirce skill score	[-1 → 0 → 1]	**0.41**					
211. Scott pi	[-1 → 1]	**0.46**					

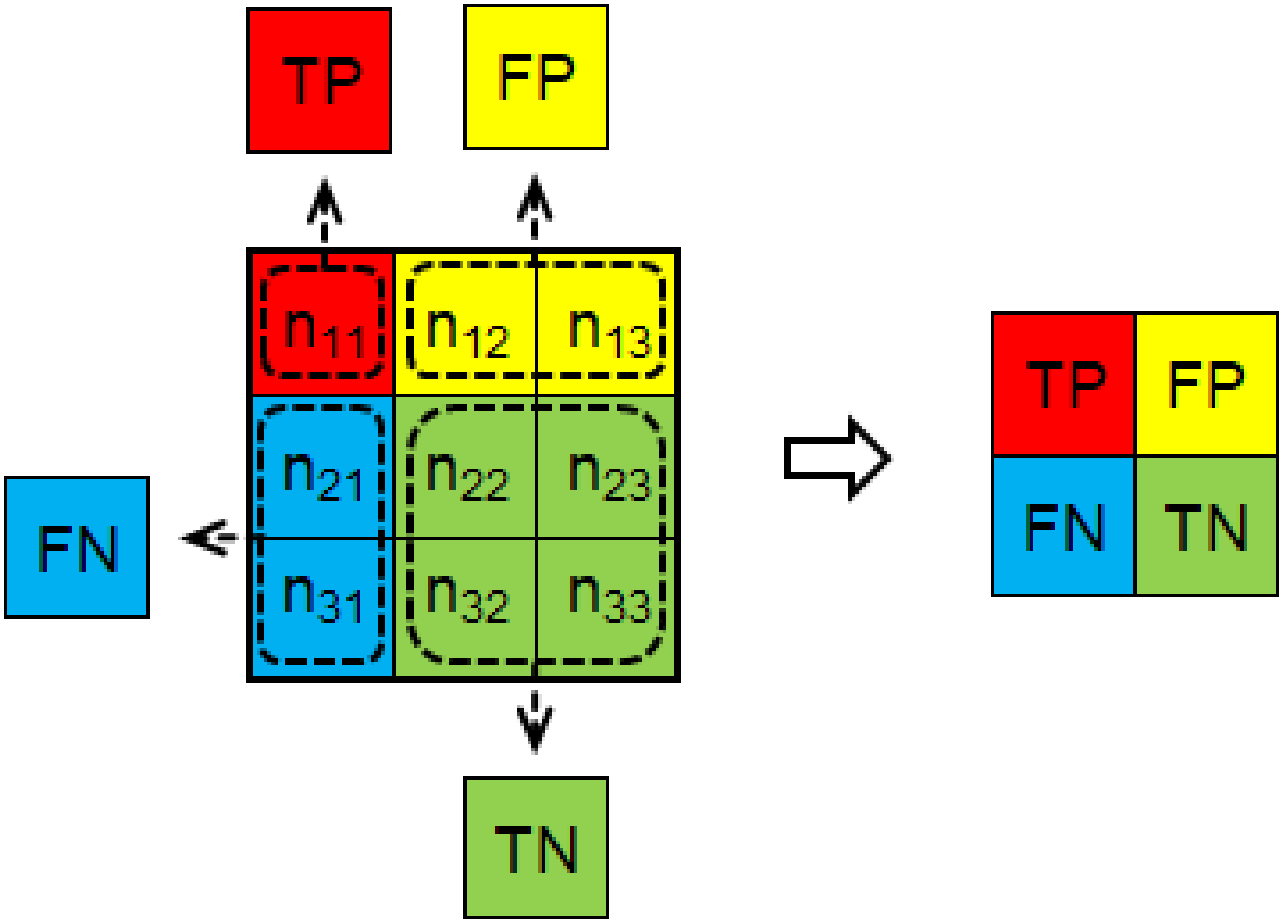
Multy-class confusion matrix

	$x=1$	...	$x=K$
$y=1$	n_{11}	...	n_{1K}
...	...	...	...
$y=K$	n_{K1}	...	n_{KK}

- Accuracy: $\frac{1}{n}\sum_{k=1}^{K} n_{kk}$
- Gorodkin R_K: $\frac{n\sum_{k=1}^{K} n_{kk} - \sum_{k=1}^{K} n_{k+}n_{+k}}{\sqrt{\left(n^2 - \sum_{i=1}^{K} n_{i+}^2\right)\left(n^2 - \sum_{j=1}^{K} n_{+j}^2\right)}}$
- Tschuprow T bias-corrected: $\sqrt{\frac{1}{K-(K-1)^2/(n-1)-1}\max\left(0, \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(n_{ij}-n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}} - \frac{(K-1)^2}{n-1}\right)}$

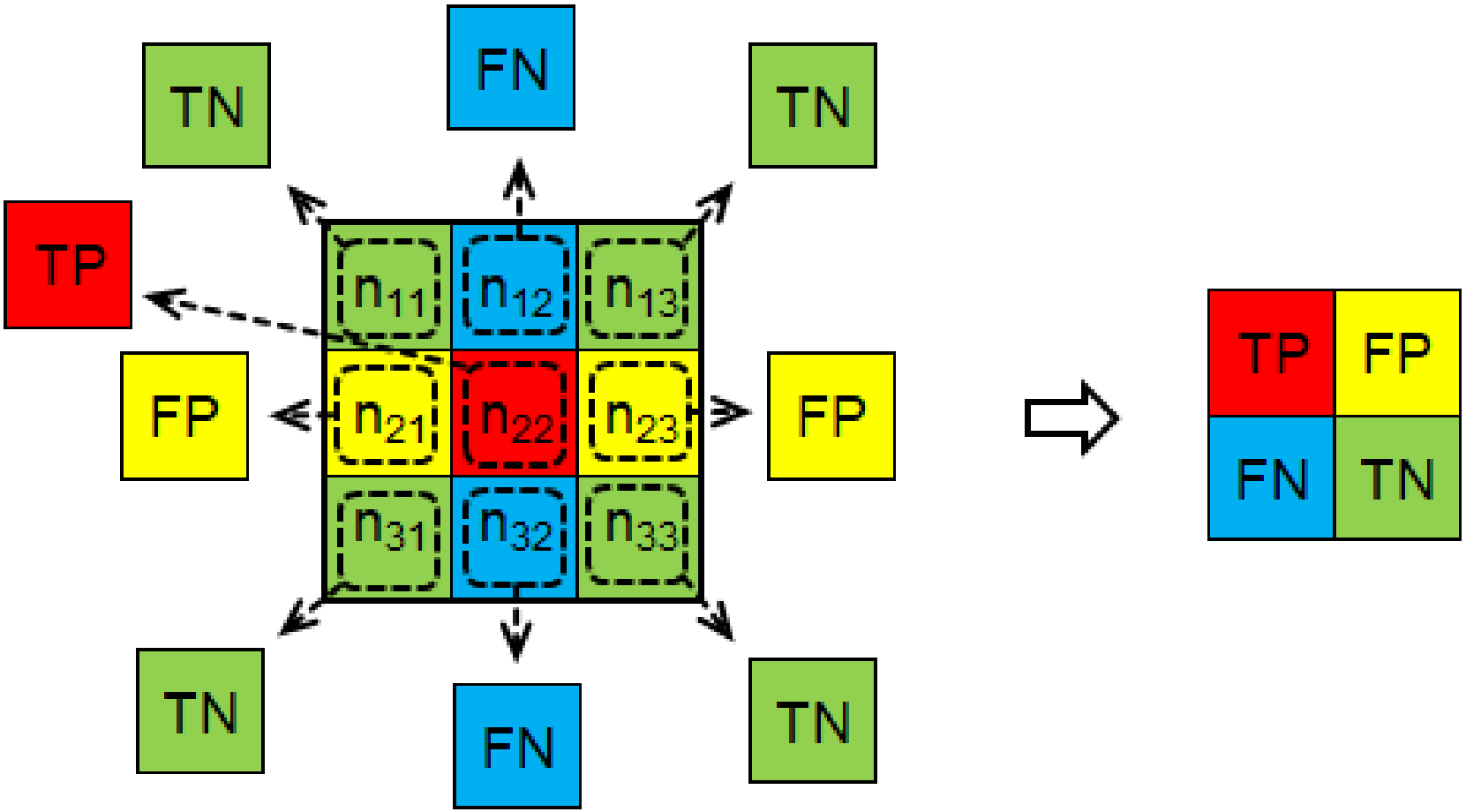
Class-specific measures

Class 1



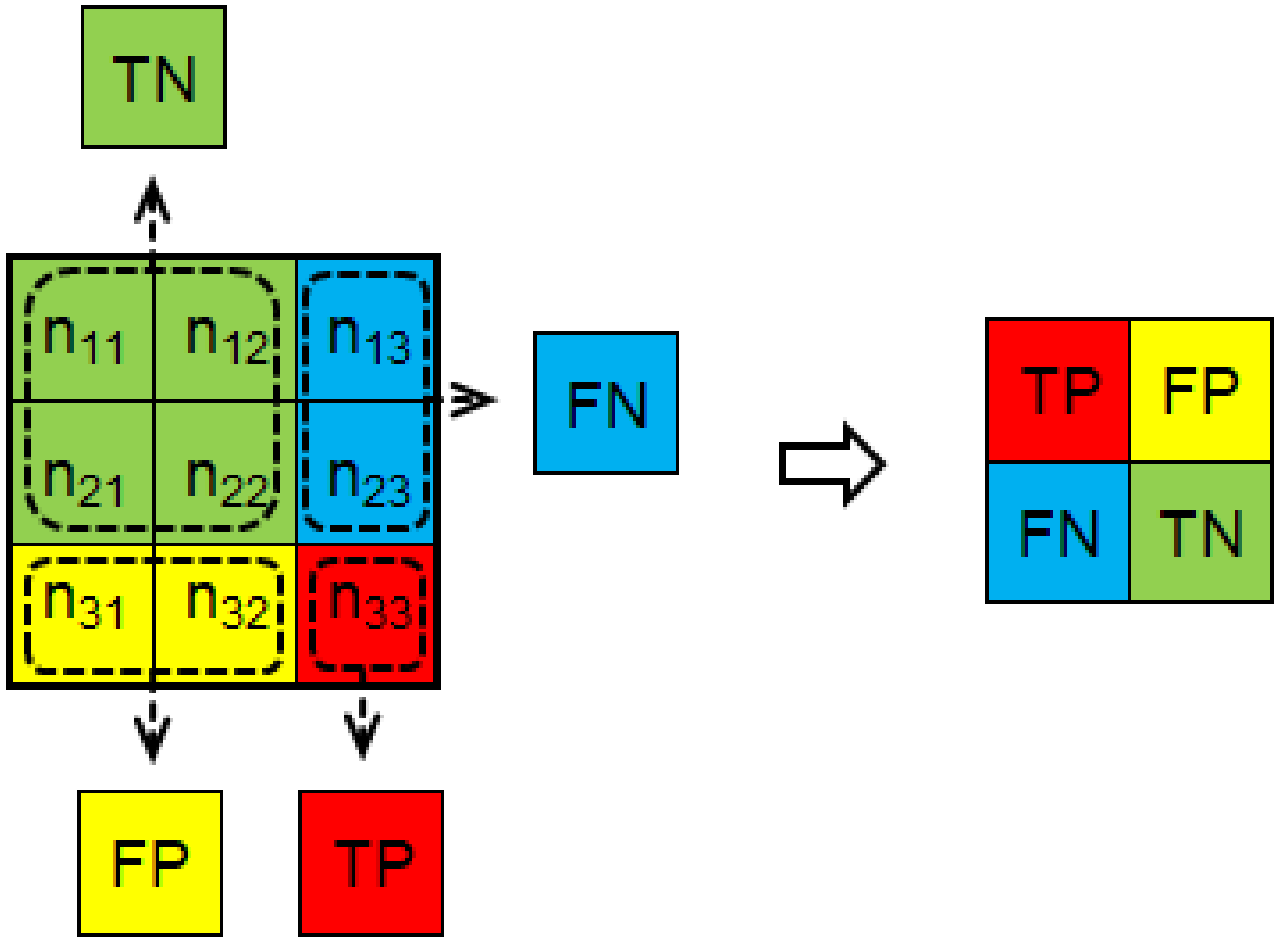
Class-specific measures

Class 2



Class-specific measures

Class 3



Class-specific measures

Arithmetic and weighted averages

- The `classify` command also computes the simple arithmetic and weighted arithmetic averages of all class-specific measures as:

$$Measure_{macro} = \frac{1}{K}\sum_{k=1}^{K} Measure_k$$

$$Measure_{weighted} = \sum_{k=1}^{K} Measure_k \frac{n_{+k}}{n}$$

- The macro-averaged measure calculates unweighted mean of class-specific coefficients.
- The weighted-averaged measure makes a weighted mean.

Classify but use common sense

" . . . there is no absolutely general measure of the degree of dependence. Every attempt to measure a conception like this by a single number must necessarily contain a certain amount of arbitrariness and suffer from certain inconveniences."

— Cramér (1924)