Good Classification Measures and How to Find Them

Martijn Gösgens, Anton Zhiyanov, Alexey Tikhonov, Liudmila Prokhorenkova

NeurIPS 2021

Summary

- Theoretically analyze classification evaluation measures
- Formally define desirable properties and check them for each measure
- Impossibility theorem: three important properties cannot be simultaneously satisfied
- Propose new measures that satisfy all desirable properties except one

Notation

Assume that we are given a *true labeling* and a *predicted labeling* of some elements

- n — number of elements
- m — number of classes
- $\mathcal{C}$ — confusion matrix
- c_{ij} — the number of elements with true label i and predicted label j
- $a_i = \sum_{j=1}^{m} c_{ij}$ — size of i-th class in the true labeling
- $b_i = \sum_{j=1}^{m} c_{ji}$ — size of i-th class in the predicted labeling

Commonly used evaluation measures

	Binary	Multiclass
F-measure (F_β)	$\frac{(1+\beta^2)\cdot c_{11}}{(1+\beta^2)\cdot c_{11}+\beta^2\cdot c_{10}+c_{01}}$	—
Jaccard (J)	$\frac{c_{11}}{c_{11}+c_{10}+c_{01}}$	—
Matthews Coefficient (CC)	$\frac{c_{11}c_{00}-c_{01}c_{10}}{\sqrt{b_1\cdot a_1\cdot b_0\cdot a_0}}$	$\frac{n\sum_{i=1}^m c_{ii}-\sum_{i=1}^m b_i a_i}{\sqrt{\left(n^2-\sum_{i=1}^m b_i^2\right)\left(n^2-\sum_{i=1}^m a_i^2\right)}}$
Accuracy (Acc)	$\frac{\sum_{i=1}^m c_{ii}}{n}$	
Balanced Accuracy (BA)	$\frac{1}{m}\sum_{i=1}^m \frac{c_{ii}}{a_i}$	
Cohen's Kappa (κ)	$\frac{\sum_{i=1}^m c_{ii}-\frac{1}{n}\sum_{i=1}^m a_i b_i}{n-\frac{1}{n}\sum_{i=1}^m a_i b_i}$	
Confusion Entropy (CE)	see the paper	

Averaging

Micro averaging

Sum up binary confusion matrices corresponding to m one-vs-all classifications.

Averaging

Micro averaging

Sum up binary confusion matrices corresponding to m one-vs-all classifications.

Macro averaging

Average the values of a measure for m one-vs-all classifications.

Averaging

Micro averaging

Sum up binary confusion matrices corresponding to m one-vs-all classifications.

Macro averaging

Average the values of a measure for m one-vs-all classifications.

Weighted averaging

Average the values of a measure for m one-vs-all classifications with weights proportional to the class-sizes.

Are the measures consistent?

Are the measures consistent?

Table: Ranking algorithms according to different measures on SST-5: from 1 (best) to 7 (worst)

	Acc	BA	κ	CE	F_1	CC
Flair+ELMo	1	1	1	1	1	1
Flair+BERT	2	4	2	2	5	2
Svm	3	3	3	5	3	3
Logistic	4	5	5	3	4	5
FastText	5	2	4	6	2	4
VADER	6	6	6	7	6	6
TextBlob	7	7	7	4	7	7

Are the measures consistent?

Table: Ranking algorithms according to different measures on SST-5: from 1 (best) to 7 (worst)

	Acc	BA	κ	CE	F_1	CC
Flair+ELMo	1	1	1	1	1	1
Flair+BERT	2	4	2	2	5	2
Svm	3	3	3	5	3	3
Logistic	4	5	5	3	4	5
FastText	5	2	4	6	2	4
VADER	6	6	6	7	6	6
TextBlob	7	7	7	4	7	7

Are the measures consistent?

Inconsistency of top results on **ImageNet**:

- Take top-10 methods in the leaderboard (based on accuracy)
- Rank them according to other measures
- Observe that rankings differ
- Thus, the problem exists even for balanced data

Are the measures consistent?

Table: Inconsistency on weather forecasting data (precipitation prediction), %

	Acc	BA	κ	CE	F_1	CC
Acc	—	96.57	37.69	3.15	41.02	44.35
BA		—	58.89	99.72	55.56	52.22
κ			—	40.83	3.33	6.67
CE				—	44.17	47.50
F_1					—	3.43
CC						—

Are the measures consistent?

Table: Inconsistency on weather forecasting data (precipitation prediction), %

	Acc	BA	κ	CE	F_1	CC
Acc	—	96.57	37.69	3.15	41.02	44.35
BA		—	58.89	99.72	55.56	52.22
κ			—	40.83	3.33	6.67
CE				—	44.17	47.50
F_1					—	3.43
CC						—

How to choose a suitable measure?

Theoretical approach:

- Formally define a list of desirable properties
- Check the properties for each measure
- Obtain recommendations on which measures are more appropriate than others

Properties

Maximal agreement

The measure has an upper bound c_{max} that is only achieved when the labelings are identical.

Properties

Maximal agreement

The measure has an upper bound c_{max} that is only achieved when the labelings are identical.

Minimal agreement

The measure has a lower bound c_{min} that is only achieved when $c_{ii} = 0$ for all i.

Properties

Maximal agreement

The measure has an upper bound $c_{\max}$ that is only achieved when the labelings are identical.

Minimal agreement

The measure has a lower bound $c_{\min}$ that is only achieved when $c_{ii} = 0$ for all i.

Symmetry

$M(\mathcal{C}) = M(\mathcal{C}^T)$ for all $\mathcal{C}$ — symmetry w.r.t. interchanging labelings.

Properties

Maximal agreement

The measure has an upper bound c_{max} that is only achieved when the labelings are identical.

Minimal agreement

The measure has a lower bound c_{min} that is only achieved when $c_{ii} = 0$ for all i.

Symmetry

$M(\mathcal{C}) = M(\mathcal{C}^T)$ for all $\mathcal{C}$ — symmetry w.r.t. interchanging labelings.

Class symmetry

Symmetry w.r.t. interchanging classes.

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties: monotonicity

Monotonicity

The value of a measure increases if we change one incorrect label to a correct label.

Properties: monotonicity

Monotonicity

The value of a measure increases if we change one incorrect label to a correct label.

Strong monotonicity

The value of a measure increases if we either increase a diagonal entry or decrease an off-diagonal entry of $\mathcal{C}$.

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties: constant baseline

Constant baseline (CB)

If predicted labels are random with probabilities $p_1, \ldots, p_m$, then the expected value of the measure is a constant c_{base} that does not depend on these probabilities.

Properties: constant baseline

Constant baseline (CB)

If predicted labels are random with probabilities $p_1, \ldots, p_m$, then the expected value of the measure is a constant c_{base} that does not depend on these probabilities.

Approximate constant baseline (ACB)

Filling in the expected value $c_{ij} = a_i p_j$ for each entry of the confusion matrix, should make the measure equal to a constant c_{base} that does not depend on $p_1, \ldots, p_m$.

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties: distance

Distance

A measure can be linearly transformed to a metric distance.

The following has to be satisfied for $d(A, B) = c_{\max} - M(A, B)$:

- Positive-definiteness $\Leftrightarrow$ maximum agreement property
- Symmetry $\Leftrightarrow$ symmetry property
- Triangle inequality: $d(A, C) \leq d(A, B) + d(B, C)$

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass)

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗

Impossibility result

Impossibility Theorem

For binary classification, there exists no measure that satisfies all of the three properties

1. Monotonicity
2. Constant baseline
3. Distance

Impossibility result

Impossibility Theorem

For binary classification, there exists no measure that satisfies all of the three properties

1. Monotonicity
2. Constant baseline
3. Distance

Several options for getting around this impossibility:

Impossibility result

Impossibility Theorem

For binary classification, there exists no measure that satisfies all of the three properties

1. Monotonicity
2. Constant baseline
3. Distance

Several options for getting around this impossibility:

- Discarding monotonicity is undesirable

Impossibility result

Impossibility Theorem

For binary classification, there exists no measure that satisfies all of the three properties

1. Monotonicity
2. Constant baseline
3. Distance

Several options for getting around this impossibility:

- Discarding monotonicity is undesirable
- Loosening constant baseline to *approximate* constant baseline

Impossibility result

Impossibility Theorem

For binary classification, there exists no measure that satisfies all of the three properties

1. Monotonicity
2. Constant baseline
3. Distance

Several options for getting around this impossibility:

- Discarding monotonicity is undesirable
- Loosening constant baseline to *approximate* constant baseline
- Discarding distance

Loosening CB to ACB: Correlation Distance

The *Correlation Distance (CD)* is the arccosine of Matthews coefficient:

$$\mathrm{CD} = \frac{1}{\pi}\arccos(\mathrm{CC})$$

Loosening CB to ACB: Correlation Distance

The *Correlation Distance (CD)* is the arccosine of Matthews coefficient:

$$\mathrm{CD} = \frac{1}{\pi}\arccos(\mathrm{CC})$$

Correlation Distance

CD satisfies all properties excluding CB, but including ACB.

Discarding distance

Matthews Correlation Coefficient

CC satisfies all properties except for being a distance (only in the binary case).

Discarding distance

Matthews Correlation Coefficient

CC satisfies all properties except for being a distance (only in the binary case).

Define *Symmetric Balanced Accuracy*: $\mathrm{SBA} = \frac{1}{2m}\sum_{i=1}^{m}\left(\frac{c_{ii}}{a_i} + \frac{c_{ii}}{b_i}\right)$

Symmetric Balanced Accuracy

SBA satisfies all properties except for being a distance (even for the multiclass case).

Discarding distance: Generalized Means Measure

Axiomization

All binary measures that satisfy all properties except distance must be of the form

$$M = s\left(\frac{a_0 a_1}{n^2}, \frac{b_0 b_1}{n^2}\right) \cdot \frac{c_{11} n - a_1 b_1}{n^2},$$

where the normalization factor $s(a, b)$ needs to satisfy some additional properties.

Discarding distance: Generalized Means Measure

Axiomization

All binary measures that satisfy all properties except distance must be of the form

$$M = s\left(\frac{a_0 a_1}{n^2}, \frac{b_0 b_1}{n^2}\right) \cdot \frac{c_{11} n - a_1 b_1}{n^2},$$

where the normalization factor $s(a, b)$ needs to satisfy some additional properties.

One interesting option is normalizing by the *generalized mean* $s(a, b)^{-1} = (\frac{1}{2}a^r + \frac{1}{2}b^r)^{1/r}$

Discarding distance: Generalized Means Measure

Axiomization

All binary measures that satisfy all properties except distance must be of the form

$$M = s\left(\frac{a_0 a_1}{n^2}, \frac{b_0 b_1}{n^2}\right) \cdot \frac{c_{11} n - a_1 b_1}{n^2},$$

where the normalization factor $s(a, b)$ needs to satisfy some additional properties.

One interesting option is normalizing by the *generalized mean* $s(a, b)^{-1} = (\frac{1}{2}a^r + \frac{1}{2}b^r)^{1/r}$

- This *Generalized Means (GM_r)* measure coincides with CC for $r \to 0$
- For $r = -1$, it coincides with SBA

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
SBA	✓	✓	✓	✓	✗	✓	✓	✓	✓
GM (binary)	✓	✓	✓	✓	✗	✓	✓	✓	✓
CD	✓	✓/✗	✓	✓	✓	✓	✓/✗	✗	✓
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Properties

Table: Properties of measures (binary/multiclass) and averagings

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
SBA	✓	✓	✓	✓	✗	✓	✓	✓	✓
GM (binary)	✓	✓	✓	✓	✗	✓	✓	✓	✓
CD	✓	✓/✗	✓	✓	✓	✓	✓/✗	✗	✓
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM_1]

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM$_1$]
- $n = 3$: cannot distinguish [Acc, BA, κ, CC, SBA, GM$_1$]

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM_1]
- $n = 3$: cannot distinguish [Acc, BA, κ, CC, SBA, GM_1]
- $n \in \{4, 5\}$: cannot distinguish [BA, κ, CC, SBA, GM_1]

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM_1]
- $n = 3$: cannot distinguish [Acc, BA, κ, CC, SBA, GM_1]
- $n \in \{4, 5\}$: cannot distinguish [BA, κ, CC, SBA, GM_1]
- $n \in \{6, 7\}$: cannot distinguish [CC, SBA, GM_1]

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM_1]
- $n = 3$: cannot distinguish [Acc, BA, κ, CC, SBA, GM_1]
- $n \in \{4, 5\}$: cannot distinguish [BA, κ, CC, SBA, GM_1]
- $n \in \{6, 7\}$: cannot distinguish [CC, SBA, GM_1]
- $n = 8$: cannot distinguish [CC, SBA]

Inconsistency of measures for small n

- Fix small n
- Check all pairs of non-degenerate labelings
- Find inconsistencies: $M_1(A, B_1) \geq M_1(A, B_2)$ but $M_2(A, B_1) < M_2(A, B_2)$

- $n = 2$: cannot distinguish [Acc, BA, F_1, κ, CE, CC, SBA, GM_1]
- $n = 3$: cannot distinguish [Acc, BA, κ, CC, SBA, GM_1]
- $n \in \{4, 5\}$: cannot distinguish [BA, κ, CC, SBA, GM_1]
- $n \in \{6, 7\}$: cannot distinguish [CC, SBA, GM_1]
- $n = 8$: cannot distinguish [CC, SBA]
- $n \geq 9$: can distinguish all measures

To sum up

If distance property is desirable:

- Choose CD

Otherwise:

- Binary classification $\Rightarrow$ choose GM_r with some r (e.g., CC or SBA)
- Multiclass classification $\Rightarrow$ choose SBA

If averaging is needed:

- Choose macro averaging