Data-driven Evaluation of Visual Quality Measures

Michael Sedlmair & Michaël Aupetit
Visual Quality Measures

• Scagnostics
  [Wilkinson & Anand, 2005]

• Paragnostics
  [Dasgupta & Kosara 2010]

• Visual class separation measures
  [e.g., Sips et. al., 2009]

• Visual correlation measures
  [e.g., Tatu et. al., 2010]

• etc…
Visual Quality Measures

e.g., visual class separation measures

imitate human perception of a low-level perceptual task

good!

bad!
What can you do with them?

Good!
Measure evaluation

1. Poor generalizability over dataset characteristics
2. A lot of manual inspection work
Contributions

• Framework for data-driven evaluation
• Instantiation of framework for class separation measures
• Evaluation of 15 class separation measures
• Guidelines for visual quality measure evaluation
Framework for Data-driven Evaluation (overview)

1. Evaluation based on how measures would perform on previously unseen data

2. Automatic evaluation

- define
  - visual encoding
  - task
  - representative sample
  - reliable judgments

- prepare
  - make them comparable
  - predict human “ground truth”

- machine learning
  - Predict!
Illustration with visual class separation
define the basic setting
large sample of pre-classified scatterplots
make them comparable
integrated judgments
class-wise judgments

separable

not separable!
1-vs-all scatterplots

i.e.

binary classification problem

separable

not separable

99

56
define
visual encoding

prepare
make them comparable

machine learning
predict human “ground truth"

use ML pipeline
e.g., ROC & bootstrapping
ROC / AUC
(Receiver Operating Characteristic / Area Under the Curve)
ROC / AUC
(Receiver Operating Characteristic / Area Under the Curve)

Decision Threshold?

False Positives

False Negatives
ROC / AUC
(Receiver Operating Characteristic / Area Under the Curve)

TP rate = TP/P
FP rate = FP/P

AUC
Area Under the Curve

good classifiers

ROC Curve
Evaluate different measures …

Note: 50% means random guess

… cross validation

… bootstrapping

70%

55%

85%

so far descriptive

initial goal:

predict performance on unseen data (inferential)

Michael Sedlmair

EuroVis 2015
Bootstrapping

1. bootstrap sample

2. bootstrap sample

...
Evaluate different measures …

average & percentiles

representing the predictive performance of measure

50% 100%

avg. 69%

57%

82%
Evaluation of 15 measures
Setup:

Data

• from a previous study [Sedlmair et al., InfoVis 2013]
  • 272 pre-classified 2D scatterplots
  • from 75 real and synthetic datasets

• 828 1-vs-all scatterplots
  (420 real / 408 synthetic)
Human Judgements

• from a previous study [Sedlmair et al., InfoVis 2013]
  • judged by two expert coders
  • 5-point scale: 1 - not separable … 5 - fully separable
  • Krippendorff’s alpha = 0.85

• Aggregation into binary judgments
  • (1,1) (1,2) (2,1) —> not separable
  • (5,5) (4,5) (5,4) —> separable
Separation Measures

- 15 measures
  - from Visualization & ML community
  - 12 non-parametric
  - 3 parametric (different parameterization)
- 35 measure instances

→ 10,000 bootstrap samples
Results
The Winner

DSC
Distance Consistency

Sips et al.: Selecting good views of high-dimensional data using class consistency [EuroVis 2009]

82.5% average AUC
DSC (Distance Consistency)

82.5% average AUC
... but still ~20%
room for improvement
Synthetic vs. real

![Graph showing synthetic vs. real average AUC](image)
Parameterizations

![Graph showing parameterizations of different models: DSC, HDM, CDM, and DC. The graph plots average AUC synthetic vs. average AUC real.]
Discussion
Guidelines

• Generalize over datasets
• Separate human judgment studies from measure studies (reuse instead of redo!)

Data-driven evaluation framework, not just for visual separation measures
Summary

• Framework for data-driven evaluation

• Study of 15 measures: DSC $\rightarrow$ 82.5% AUC avg.

• Guidelines: Generalization & Separation
Data-driven Evaluation of Visual Quality Measures

Michael Sedlmair & Michaël Aupetit

Thanks!