Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)

Volume 34 (2015), Number 3

Data-driven Evaluation of Visual Quality Measures (Supplemental Material)

M. Sedlmair[1] and M. Aupetit[2]

[1]Visualization and Data Analysis Group, University of Vienna, Austria
[2]Qatar Computing Research Institute, Doha, Qatar

Abstract

This supplemental material file provides additional information along with the EuroVis'15 paper "Data-driven Evaluation of Visual Quality Measures". We provide (1) further details on the study setup in terms of how we cleaned the base data, (2) mathematical descriptions of the separation measures we tested, (3) additional information on the number of bootstrap samples used, and (4) a sanity check pilot study that we conducted.

1. Data Cleaning

In the following, we provide additional information to the cleaning step that we needed to undertake in order to make the data from Sedlmair et al. [SMT13] applicable to our study. These details extend upon Section 4.1. of the paper.

We needed to filter and clean the data from Sedlmair et al. [SMT13]. This process included five steps: (i) filtering out 3D and multi-D scatterplots, (ii) aligning image and data space, (iii) removing occluded points, (iv) removing datasets that exceeded reliable human judgment, and (v) removing uncertain human judgments.

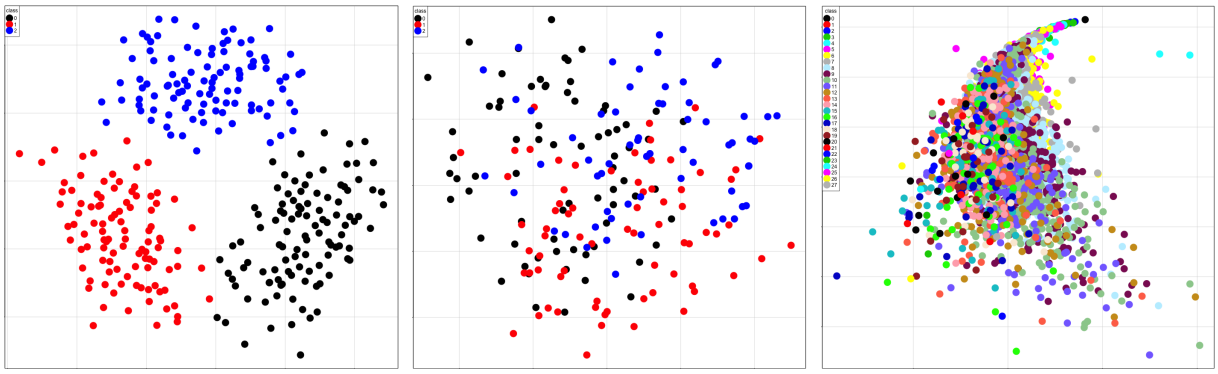
The overall cleaning process is summarized in the Figures 2 and 3.

i) 2D scatterplots only: We first discarded all the 3D and multi-D scatterplots and only considered the 2D scatterplots, as these are the visual encodings current separation measures operate on.

ii) Align image and data space: The base data we used came as 1008×1008 pixel scatterplot images consisting of labeled points rendered as 15-pixel-diameter colored discs. Two examples are shown in Figure 1 (a) and (b). As these scatterplot images did not come as normalized representations, we needed to ensure that the data space that the separation measures see aligns with the image space that the human coders saw. Towards that goal, we used simple image processing techniques that allowed us to detect the colored discs from the image space (See an example in the Figure 2(a)), and match those with the points from the data space. Visually inspecting the detected discs along with original abstract set of points, showed considerable misalignments and differences in number of points (due to occlusion). Hence, to ensure a robust alignment between image and data space, we used standard linear programming technique to compute an appropriate transformation matrix (Figure 2(b)). In theory, it would have been enough to select two discs maximally spaced along some diagonal in the picture and their matching points to find the alignment. However, in many cases, this matching between discs and points was not even possible due to the afore mentioned problems. Thus, we had to use a systematic exploration of all the possible pairs of discs and data and compute the mean square distances between each point and its nearest center as a matching score to get the best match automatically.

We visually checked all of the aligned sets of points (visual checking was far easier than manual finding of the good alignment) and although most of the set of points were successfully aligned with this automatic process, we discovered a dozen of failures to be manually aligned through a specific visual interface we designed for this sake. Note that the coordinates of the aligned set of points were rounded to the ones of the nearest pixel in the image space $\mathcal{I}$.

iii) Removing fully occluded points: We also had to decide which data point was visible to a human and which was not due to full occlusion. We rendered the aligned labelled data using a 15-pixel circle, one at a time. We then counted the



(a) (b) (c)

Figure 1: *Examples of two scatterplots (a) and (b) each as shown to the expert coder in the study by Sedlmair et al. [SMT13]. (c) An example of a scatterplot not included in the study due to a too large number of classes and points.*

number of pixels in the original picture falling within this circle and having the correct color. We delete every point whose rendering showed no pixel of the correct color in the correct place. We then kept only one of the multiple occurrences of aligned points in the same pixel.

iv) Removing datasets: Subsequently, we excluded scatterplots for which only one class was visible. We then also filtered out all scatterplots that had more than 14 classes or more than 1000 (visible) points (Figure 1(c)). This filter step had several reasons. First, given our base data with color-coded discs as points and the limitations of human perception, reliable judgments cannot be guaranteed over a certain scale. Second, there is also a considerable computational burden involved in this step as the computation of some of the separation measures is quite expensive.

v) Removing uncertain human judgments: After this alignment and cleaning process, we generated the data $d_{\mathfrak{s},\mathcal{C}_t}$ to be used in our evaluation framework, that is, scatterplots $\mathfrak{s}$ focused on a specific target class $\mathcal{C}_t$. There are as many data generated for a scatterplot as there are classes in its original abstract set of points. For each of these data $d_{\mathfrak{s},\mathcal{C}_t}$, we had two attached human judgments $\mathfrak{h}_1(.)$ and $\mathfrak{h}_2(.)$ with values in $\{1,2,3,4,5\}$. In order to end up with a single ground-truth judgment $\underline{\mathfrak{h}}(.)$ that is reliable, we used a consensus-based aggregation function G where a doubt (judgment value 3) or too big a difference between the human judgments ($|\mathfrak{h}_1(.)-\mathfrak{h}_2(.)|>1$) were discarded:

$$\underline{\mathfrak{h}}(d_{\mathfrak{s},\mathcal{C}_t}) = G(\{\mathfrak{h}_1(d_{\mathfrak{s},\mathcal{C}_t}),\mathfrak{h}_2(d_{\mathfrak{s},\mathcal{C}_t})\})$$

where

$$G(u,v)=\begin{cases} 1 \quad (sep) & \Leftrightarrow (u,v)\in\{(5,5),(4,5),(5,4)\} \\ 0 \quad (nonsep) & \Leftrightarrow (u,v)\in\{(1,1),(2,1),(1,2)\} \\ -1 & \text{in any other case} \end{cases}$$

Any data $d_{\mathfrak{s},\mathcal{C}_t}$ ending up with $\underline{\mathfrak{h}}(d_{\mathfrak{s},\mathcal{C}_t})=-1$ were discarded from the evaluation.

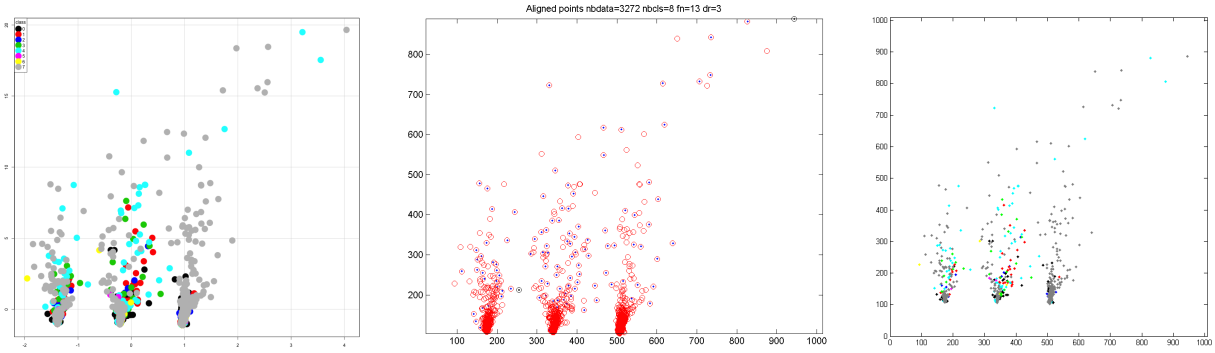
We finally ended up with the labeled dataset $\mathcal{D}_{\mathcal{S}}$ attached to the above aggregated judgment decision rule $\underline{\mathfrak{h}}$. Eventually, this process resulted in 828 data items $d_{\mathfrak{s},\mathcal{C}_t}$, that is, target class-specific scatterplots (Figure 3). These data items were derived from 224 multi-class scatterplots issued from using PCA, robust PCA, Glimmer MDS, and t-SNE on 56 of the 75 multidimensional datasets, originally taken from Sedlmair et al. [SMT13]. The process is visually illustrated in Figure 2 and 3.

2. Separation Measures

In the following, we provide additional information to the measures that we tested in our study. These details extend Section 4.2. that gives a high-level overview of the measures.

In our study, we compared the following separation measures from the visualization and machine learning communities: the Distribution Consistency measure (DC) and the Distance Consistency measure (DSC) [SNLH09]; the Class Density measure (CDM) and the 2D Histogram Density measure (HDM) [TAE*09] (we did not consider the 1D-HDM from the same authors because it is a heuristic approach similar to the well-grounded Linear Discriminant Analysis (LDA) that we also considered); the Silhouette (SIL) [Rou87], Dunn's index (DUNN) [Dun74], Gamma (GAM) [BH75], Calinski-Harabasz (CAL) [CH74] and Weighted Inter-Intra (WII) [Str02] were all compared in [LAdS12] as measures to evaluate cluster structures in terms of between-cluster separation and within-cluster homogeneity; the Hypothesis Margin (HM) [GBNT04], the Class Separation (CS) [MMdALO15] and the Linear Discriminant Analysis (LDA) [Fuk90] have been proposed in the machine learning community.

Table 1 shows all measures that we tested and gives a brief summary of how they operate. It also provides a rough high-level classification of these measures in terms of the elements they are used on, the locality criteria they are based



(a) (b) (c)

Figure 2: *(a) A scatterplot as seen by the human experts. The classes are displayed over each other in the same order than the colored spots appear within the legend from top to bottom. (b) Circular spots are automatically extracted from the original picture, and their centers (blue points) are automatically aligned with the data points (red circle) and their position rounded to the closest integer value. Each point of this aligned dataset is rendered as a 15-pixel width spot and kept in the final dataset (c) if at least one colored pixel of this spot is visible in the original picture. If multiple data points are in the same pixel position, then only one is kept, and if multiple colors are given to the same pixel position, then only the lowest one in the legend color sequence is kept (for instance, if yellow and red are over-plotted, then only yellow is kept).*

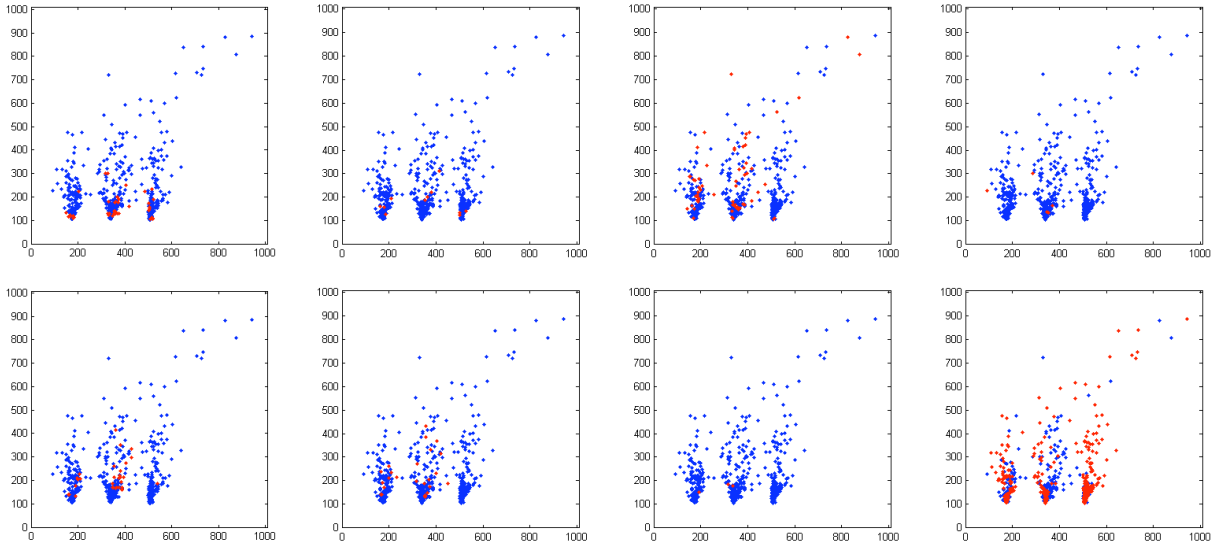


Figure 3: *The cleaned multi-class dataset from Figure 2 is transformed into as many 1-vs-all class datasets as there are classes in the cleaned dataset (Target class in red). The separation measures are evaluated on these one-vs-all cleaned datasets.*

on, the notion of discrepancy used, and their complexity. Some of the measures needed to be parameterized. For each of those, we tested between 5 and 10 different parameter settings. Formal descriptions of all measures are detailed below.

2.1. Notations

Each scatterplot $\mathfrak{s} \in \mathcal{S}$ is a set of points $X_\mathfrak{s} = \{x_1, \ldots, x_{N_\mathfrak{s}}\}$ such that $\forall x \in X_\mathfrak{s}, x \in \mathcal{I} \subset \mathbb{N}^2$ where $\mathcal{I}$ is the 2-dimension pixel space of the image rendering $\mathfrak{s}$. Let $|S|$ denote the number of elements in the set S. Let $d : \mathcal{I} \times \mathcal{I} \to \mathbb{R}^+$ be the Euclidean distance function and $D_{S_i,S_j} = \frac{1}{|S_i||S_j|} \sum_{x \in S_i} \sum_{y \in S_j} d_{x,y}$ be the average Euclidean distance between any point of $S_i \subseteq X_\mathfrak{s}$ and any point of $S_j \subseteq X_\mathfrak{s}$ and let us write $D_{S_i} = D_{S_i,S_i}$. Let $\mathcal{C}_\mathfrak{s} = \{\mathcal{C}_t, \mathcal{C}_o\}$ be a partition of $X_\mathfrak{s}$ into 2 disjoint subsets ($\mathcal{C}_t \cup \mathcal{C}_o = X_\mathfrak{s}$ and $\mathcal{C}_t \cap \mathcal{C}_o = \emptyset$) where t is the target class whose separation is to be evaluated by the human or the separation measure. Let $S_{btn} = \{(x,y) | x \in \mathcal{C}_t, y \in \mathcal{C}_o\}$ be the set of pairs of points assigned each to a different class and $S_{wtn} = \{(x,y) | \forall u \in \{t,o\}, (x,y) \in \mathcal{C}_u^2\}$ be the set of pairs of points assigned each to the same class. Let $N_i = |\mathcal{C}_i|$ be the number of elements in $\mathcal{C}_i$ so $N_t + N_o = N_\mathfrak{s}$ and let $\mathcal{N}_i = N_i(N_i - 1)/2$ be the num-

m	Short description	Elements	Locality	Discrepancy	Complexity	Param.	Reference
ABTN	between-class average distances	P	dist	B	$O(N^2)$	-	[LAdS12]
AWTN	within-class average distances	P	dist	W	$O(N^2)$	-	[LAdS12]
ABW	between-class ABTN over within-class AWTN average distances ratio	P	dist	B/W	$O(N^2)$	-	derived by us
WII	average between-class over average within-class distances weighted by the respective size of the classes	P	dist	B/W	$O(N^2)$	-	[Str02]
CAL	centers-of-mass between-class square distances over points-to-centers-of-mass within-class square distances	P, CoM	dist	B/W	$O(N)$	-	[CH74]
LDA	centers-of-mass between-class distances over points-to-centers-of-mass within-class distances ratio with optimal linear transformation of the points to maximize this ratio	P, CoM	dist	B/W	$O(N^2)$	-	[Fuk90]
DUNN	maximum within-class distance over the minimum between-class distance ratio	P	dist	$\min(B)/\max(W)$	$O(N^2)$	-	[Dun74]
GAM	normalized comparison of numbers of within-class distances smaller or greater than between-class distances	P	dist	$\frac{\lvert W<B\rvert-\lvert W>B\rvert}{\lvert W<B\rvert+\lvert W>B\rvert}$	$O(N^4)$	-	[BH75]
SIL	difference of between-class and within-class average distances normalized by the maximum of them	P	dist	$(B-W)/\max(B,W)$	$O(N^2)$	-	[Rou87]
HM	average differences between distances from each point to its other-class and its same-class nearest-neighbor	P	NN	$(B-W)/\max(B,W)$	$O(N^2)$	-	[GBNT04]
CS	average proportion of same-class neighbors of each point in minimum spanning tree	P	ExtMST	$W/(W+B)$	$O(N^2)$	-	[MMdALO15]
DSC	proportion of points x whose the nearest class-center-of-mass belongs to the same class as x	P, CoM	NN	$W/(W+B)$	$O(N)$	-	[SNLH09]
CDM	pixel-wise class-density differences with class-density estimated at pixel z as the inverse distance to its K^{th} nearest point of this class	P, Pix	KNN	$\Delta\rho$	$O(\lvert\mathcal{I}\rvert KN\log(N))$	K (10)	[TAE*09]
DC	average of the class entropy for each pixel computed over the classes of its ε-neighbors	P, Pix	ε NN	H	$O(\lvert\mathcal{I}\rvert N)$	ε (10)	[SNLH09]
HDM	entropy measure of the classes in each cell and their adjacent cells in a square-grid partition	P, Bn	β NN	H	$O(\lvert\mathcal{B}\rvert N)$	N_b, β (5)	[TAE*09]

Table 1: ***Summary of the separation measures used in the experiments****. Each separation measure aggregates a local measure of discrepancy between some elements of the classes. The* ***elements*** *can be the points (P), the center-of-mass (CoM) of each class, the image's pixels (Pix) or the histogram's bins (Bn). The* ***locality*** *can depend on distances (dist) or on some discrete neighborhood (KNN, ExtMST, εNN, βNN). And the* ***discrepancy*** *measure can depend on some aggregation of within-class (W) and between-class (B) quantities, on entropy (H) or a difference of densities ($\Delta\rho$). We also report the computational* ***complexity****, and free* ***parameters*** *if any (in parentheses we note how many different settings of these parameters we tested in our study). Measures are organized based on their technical similarities in terms of their elements, locality and discrepancy characteristics.*

ber of pairs of different elements in $\mathcal{C}_i$ so $|S_{btn}| = N_t N_o$ and $|S_{wtn}| = \mathcal{N}_t + \mathcal{N}_o$. Let us write $\overline{u} = t \Leftrightarrow u = o$ and $\overline{u} = o \Leftrightarrow u = t$. Finally let $c_i = \frac{1}{N_i}\sum_{x\in\mathcal{C}_i} x$ be the center-of-mass of the class $\mathcal{C}_i$.

2.2. Measures

We now formally define all measures that we tested, in alphabetical order.

Notice that in some cases (HM, WII), we reverse the measure as originally written in order to have greater values for greater separation. We also ignored the penalization factor used in some measures (CAL, WII) related to the number of classes, as we always have only two classes ($\mathcal{C}_t$ and $\mathcal{C}_o$) for all the data in our framework.

2.2.1. ABTN, AWTN and ABW

The Average Between and Average Within measures evaluate the between-class separation and within-class homogeneity, respectively:

$$ABTN(\mathcal{C}_\mathfrak{s}) = D_{\mathcal{C}_t,\mathcal{C}_o}$$

$$AWTN(\mathcal{C}_\mathfrak{s}) = \frac{\sum_{u\in\{t,o\}} \mathcal{N}_u D_{\mathcal{C}_u}}{\sum_{u\in\{t,o\}} \mathcal{N}_u}$$

.

We also used the ratio between these two measures:

$$ABW(\mathcal{C}_\mathfrak{s}) = \frac{ABTN(\mathcal{C}_\mathfrak{s})}{AWTN(\mathcal{C}_\mathfrak{s})}$$

2.2.2. CAL

The Calinski-Harabasz measure is related to the concentration of the classes around their center-of-mass:

$$CAL(\mathcal{C}_\mathfrak{s}) = \frac{\sum_{u\in\{t,o\}} N_u d^2_{c_u,c_\mathfrak{s}}}{\sum_{u\in\{t,o\}}\sum_{x\in\mathcal{C}_u} d^2_{c_u,x}}$$

2.2.3. CDM

The density of each class is estimated in each image pixel as the inverse distance to its K^{th} nearest points of this class. The

sum over the pixels of the differences between these class-density images give the CDM measure. K is a parameter of this measure.

$$CDM(\mathcal{C}_\mathfrak{s}) = \sum_{z\in\mathcal{I}} |\rho_{\mathcal{C}_t}(z,K) - \rho_{\mathcal{C}_o}(z,K)|$$

where

$$\forall z \in \mathcal{I}, \forall \mathcal{C}_i \in \mathcal{C}_\mathfrak{s}, \rho_{\mathcal{C}_i}(z,K) = \frac{1}{\max_{x\in KNN_{\mathcal{C}_i}(z,K)}(d_{z,x})}$$

and

$$KNN_S(z,K) = \{x \in S | K \geq |\{y \in S | d_{z,y} < d_{z,x}\}|\}$$

is the set of the K nearest neighbors of z.

2.2.4. CS

CS is the average proportion of the neighbors of each point x with respect to the ExtMST graph spanning $X_\mathfrak{s}$, which belong to the same class as x.

$$CS(\mathcal{C}_\mathfrak{s}) = \frac{1}{N_\mathfrak{s}} \sum_{u\in\{t,o\}} \sum_{x\in\mathcal{C}_u} \frac{|ExtMST(x)\cap\mathcal{C}_u|}{|ExtMST(x)|}$$

where $ExtMST(x)$ is the set of neighbors of x in the ExtMST graph. The algorithm to compute the ExtMST graph is given in [MMdALO15].

2.2.5. DC

DC is the average of the class entropy for each pixel z computed over the classes of its ε-neighbors $\varepsilon NN(z,\varepsilon)$. ε is a parameter of this measure.

$$DC(\mathcal{C}_\mathfrak{s}) = 1 - \frac{\sum_{z\in\mathcal{I}} |\varepsilon NN(z)| H_{\varepsilon NN}(z)}{\sum_{z\in\mathcal{I}} |\varepsilon NN(z)|}$$

where $H_{\varepsilon NN}(z,\varepsilon)$ is the entropy in the ε-neighborhood of the pixel z:

$$H_{\varepsilon NN}(z,\varepsilon) = -\sum_{u\in\{t,o\}} \frac{|\varepsilon NN(z,\varepsilon)\cap\mathcal{C}_u|}{|\varepsilon NN(z,\varepsilon)|} \log \frac{|\varepsilon NN(z,\varepsilon)\cap\mathcal{C}_u|}{|\varepsilon NN(z,\varepsilon)|}$$

and $\varepsilon NN(z,\varepsilon) = \{x \in X_\mathfrak{s} | d_{x,z} \leq \varepsilon\}$.

2.2.6. DSC

DSC is the proportion of points x whose the nearest class-center-of-mass belongs to the same class as x:

$$DSC(\mathcal{C}_\mathfrak{s}) = \frac{\sum_{u\in\{t,o\}} |\{x \in \mathcal{C}_u | \arg\min_{v\in\{t,o\}}(d_{x,c_v}) = u\}|}{N_\mathfrak{s}}$$

2.2.7. DUNN

Dunn's index compares the maximum within-class distance to the minimum between-class distances:

$$DUNN(\mathcal{C}_\mathfrak{s}) = \frac{\min_{(x,y)\in S_{btn}} d_{x,y}}{\max_{(x,y)\in S_{wtn}} d_{x,y}}$$

2.2.8. GAM

The Gamma measure is defined as follows. Let $d^+ = |\{\{x,y,x',y'\} | (x,y) \in S_{wtn}, (x',y') \in S_{btn}, d_{x,y} \leq d_{x',y'}\}|$ be the number of times that a pair of points that belong to the same class has distance smaller than two points assigned to different classes, and let $d^- = |\{\{x,y,x',y'\} | (x,y) \in S_{wtn}, (x',y') \in S_{btn}, d_{x,y} \geq d_{x',y'}\}|$ be the opposite, then:

$$GAM(\mathcal{C}_\mathfrak{s}) = \frac{d^+ - d^-}{d^+ + d^-}$$

.

2.2.9. HDM

The Histogram Density Measure is an entropy measure of the classes in each cell and their adjacent cells in a square-grid partition $\mathcal{B}$ of the image $\mathcal{I}$. The number of bins N_b is a parameter of this measure.

$$HDM(\mathcal{C}_\mathfrak{s}) = 1 - \frac{\sum_{z\in\mathcal{B}} |\beta NN(z)| H_{\beta NN}(z)}{\sum_{z\in\mathcal{B}} |\beta NN(z)|}$$

where $H_{\beta NN}(z,\beta)$ is the entropy in the β-neighborhood of the bin z:

$$H_{\beta NN}(z,\beta) = -\sum_{u\in\{t,o\}} \frac{|\beta NN(z,\beta)\cap\mathcal{C}_u|}{|\beta NN(z,\beta)|} \log \frac{|\beta NN(z,\beta)\cap\mathcal{C}_u|}{|\beta NN(z,\beta)|}$$

and $\beta NN(z,\beta) = \{x \in X_\mathfrak{s} | \|x-z\|_\infty \leq \beta\}$ which for $\beta = 1$ is the neighborhood of the bin z defined as all (up to 8) the adjacent bins to z in the grid $\mathcal{B}$.

2.2.10. HM

The Hypothesis Margin measure is the average of the differences between distances from each point to its other-class nearest-neighbor and to its same-class nearest-neighbor:

$$HM(\mathcal{C}_\mathfrak{s}) = \frac{1}{N_\mathfrak{s}} \sum_{u\in\{t,o\}} \sum_{x\in\mathcal{C}_u} \frac{\min_{y\in\mathcal{C}_{\bar{u}}} d_{x,y} - \min_{y\in\mathcal{C}_u, y\neq x} d_{x,y}}{\max(\min_{y\in\mathcal{C}_{\bar{u}}} d_{x,y}, \min_{y\in\mathcal{C}_u, y\neq x} d_{x,y})}$$

2.2.11. LDA

The Linear Discriminant Analysis between-scatter over within-scatter ratio as been proposed as a separation measure by Fukunaga [Fuk90]. The between-scatter M_{btn} and within-scatter M_{wtn} are 2×2 matrices defined as:

$$M_{btn} = \frac{1}{N_\mathfrak{s}} \sum_{u\in\{t,o\}} N_u (c_u - c_\mathfrak{s})(c_u - c_\mathfrak{s})^T$$

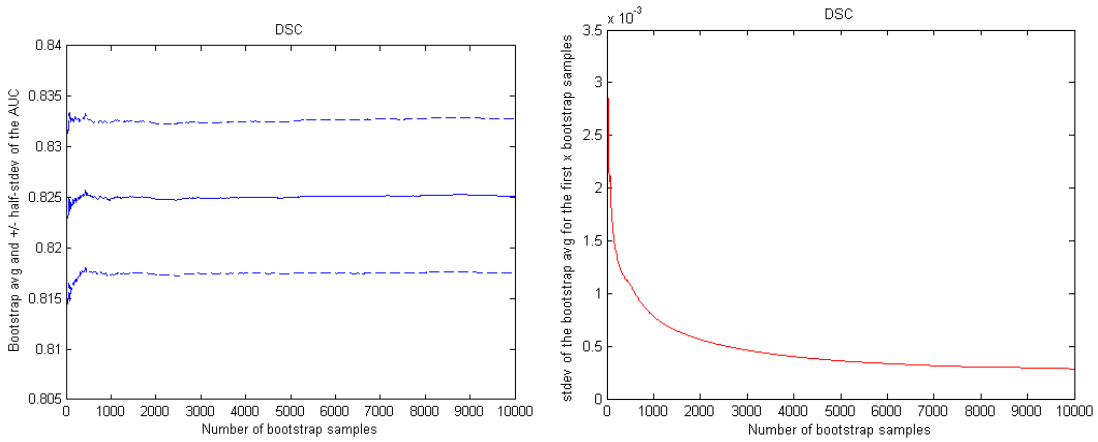


Figure 4: *Analyzing the number of necessary bootstrap samples illustrated by the performance of the DSC measure. The x-axis in both graphs shows the increasing number of bootstrap samples. (left) The y-axis encodes the value of the AUC, the plain line shows the average AUC value, the dashed lines the standard deviation. (right) The y-axis shows the standard deviation of the AUC average over the x first bootstrap samples. Both graphs start to converge at around 1,000. At 10,000—the number we selected—we found them to be very stable.*

and

$$M_{wtn} = \frac{1}{N_{\mathfrak{s}}} \sum_{u \in \{t,o\}} \sum_{x \in \mathcal{C}_u} (c_u - x)(c_u - x)^T$$

The LDA seeks to find the optimal matrix W such as to maximize the between-scatter over within-scatter ratio:

$$LDA(\mathcal{C}_{\mathfrak{s}}) = \max_W \left(\frac{tr(W^T M_{btn} W)}{tr(W^T M_{wtn} W)} \right)$$

This is equivalent to maximizing the average pairwise distance between class means and minimizing the average within-class pairwise distance over all classes.

2.2.12. SIL

The Silhouette measure quantifies the separation as the difference between the average between-class distances and the average within-class distances, normaliwed by the maximum of these two quantities. It is defined as:

$$SIL(\mathcal{C}_{\mathfrak{s}}) = \sum_{u \in \{t,o\}} \sum_{x \in \mathcal{C}_u} \frac{D(x, \mathcal{C}_{\overline{u}}) - D(x, \mathcal{C}_u)}{\max(D(x, \mathcal{C}_{\overline{u}}), D(x, \mathcal{C}_u))}$$

2.2.13. WII

The Weighted Inter-Intra measure is the average between-class over average within-class distances weighted by the respective size of the classes [Str02] (eq. (3.5) p. 61):

$$WII_{orig}(\mathcal{C}_{\mathfrak{s}}) = \frac{N_{\mathfrak{s}} D(\mathcal{C}_t, \mathcal{C}_o)}{N_t D(\mathcal{C}_t) + N_o D(\mathcal{C}_o)}$$

2.3. Parameterization

Three of the measures came with additional parameters, which we also were curious to test.

We set the number of bins N_b along one axis in the HDM partition to $\{5, 10, 20, 40, 80\}$, with the neighborhood size set to $\beta = 1$. The bins are square-shaped and their width is equal to the largest range of values among the two axes divided by N_b. The lower-left corner of the lower-left bin starts at the minimum value onto both axes.

The number K of K-Nearest Neighborhood in the CDM measure has been set to $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

We compute DC for the following values of ε: $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\} \times \Delta(X_{\mathfrak{s}})$ where $\Delta(X) = \max_{(x,y) \in X^2}(d_{x,y})$ is the diameter of the set of points, which means that the radius ε of the neighborhood was set between 0.1% and 20% of this diameter.

3. Number of Bootstrap Samples

As described at the beginning of Section 4.3. in the paper, we had to decide on a number of bootstrap samples to use in our study. To do so, we studied how the AUC average values as well as the standard deviation for different measures varied with increasing numbers of bootstrapping samples. We stopped at 10,000 bootstrap samples, a number that we found highly sufficient, given that average and standard deviation values already started to converge at about 1,000 samples. Figure 4 illustrates this analysis for the *DSC* measure.

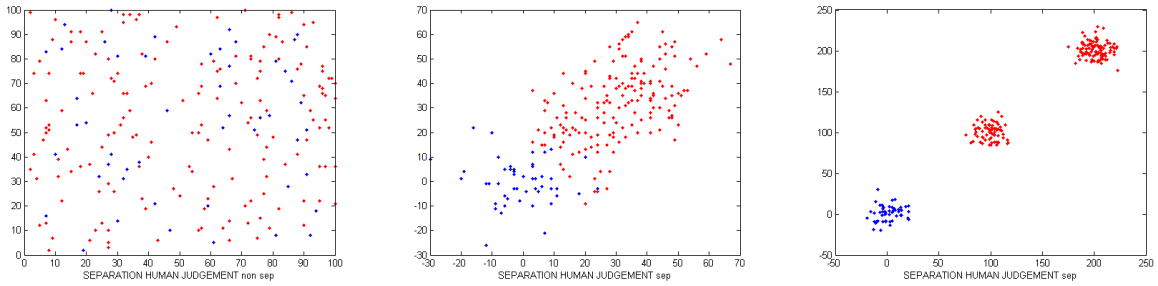


Figure 5: *The* 3 *families of datasets used in the pilot study. There are twice as many red points as blue ones in all families. (left) Both classes are drawn from the same uniform distribution, their human judgment is 'non-sep'. (center) Classes are drawn from three equal variance Normal along a diagonal, two of them being red, the other being blue. The blue class overlap slightly the red class. (right) Classes are drawn from three clearly separated Normal. The human judgment to be predicted is 'sep' for the last two families.* 10 *datasets where drawn randomly from each of these* 3 *families to get the* 30 *datasets we used in this pilot study.*

4. Sanity Check Pilot Study

Here, we provide additional information regarding the sanity check pilot study that we mention in Section 4.3 of the paper.

In this pilot study, we tested all the measures on 3 families of simplistic, synthetically-generated datasets. One example from each family is shown in Figure 5. The obvious human class separability in these three families was pre-specified by the authors, with two families generating clearly separable classes, and one family non-separable classes.

We generated 10 datasets from each of these families, ending up with 30 datasets for which all the separation measures were computed. Their AUC bootstrap distribution is displayed in Figure 6. The measures are ranked in decreasing order of their AUC bootstrap average. This experiment shows that 26 over 35 measures are perfect in predicting the human judgment with a 100% AUC bootstrap median. This result is in contrast to the one we get with the same measures onto the realistic and larger sets of data we used in the main experiment of this work.

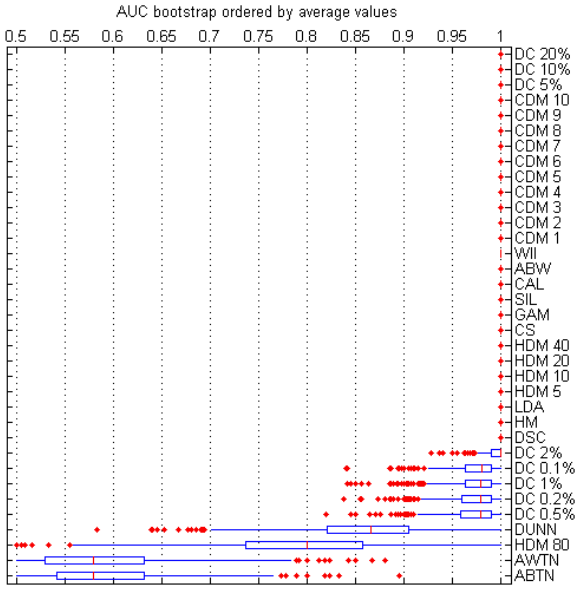


Figure 6: *Results of our pilot study. Each row represents a separation measure with a box plot encoding the AUC bootstrap distribution: median (red center line); interquartile range, IQR, i.e., from 25 to 75 percentile (box);* 25/75 *percentile +/- 1.5 times the IQR, including* 99.3% *of the data (whiskers); and outliers (red points). The measures are ranked in decreasing order of the AUC bootstrap average.*

References

[BH75] BAKER F., HUBERT L.: Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association 70*, 349 (1975), 31–38. 2, 4

[CH74] CALIŃSKI T., HARABASZ J.: A dendrite method for cluster analysis. *Communications in Statistics Simulation and Computation 3*, 1 (1974), 1–27. 2, 4

[Dun74] DUNN J. C.: Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics 4*, 1 (1974), 95–104. 2, 4

[Fuk90] FUKUNAGA K.: *Introduction to statistical pattern recognition*, second ed. Computer Science and Scientific Computing. Academic Press, 1990. 2, 4, 5

[GBNT04] GILAD-BACHRACH R., NAVOT A., TISHBY N.: Margin based feature selection – theory and algorithms. In *Proc. 21st Int. Conf. on Machine Learning (ICML)* (2004), Brodley C. E., (Ed.), ACM, pp. 43–50. 2, 4

[LAdS12] LEWIS J. M., ACKERMAN M., DE SA V.: Human cluster evaluation and formal quality measures: A comparative study. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)* (2012), pp. 1870–1875. 2, 4

[MMdALO15] MOTTA R., MINGHIM R., DE ANDRADE LOPES A., OLIVEIRA M. C. F.: Graph-based measures to assist user

assessment of multidimensional projections. *Neurocomputing* (2015), 583 – 598. preprint. 2, 4, 5

[Rou87] ROUSSEEUW P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 1 (1987), 53–65. 2, 4

[SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 19*, 12 (2013), 2634–2643. 1, 2

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis) 28*, 3 (2009), 831–838. 2, 4

[Str02] STREHL A.: *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, University of Texas at Austin, 2002. 2, 4, 6

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)* (2009), pp. 59–66. 2, 4