EuroVis Workshop on Visual Analytics (2020)
K. Vrotsou and C. Turkay (Editors)

SepEx: Visual Analysis of Class Separation Measures

Jürgen Bernard[1], Marco Hutter[2], Matthias Zeppelzauer[3], Michael Sedlmair[2], and Tamara Munzner[1]

[1]University of British Columbia, Canada
[2]University of Stuttgart
[3]St. Pölten University of Applied Sciences

Abstract

Class separation is an important concept in machine learning and visual analytics. However, the comparison of class separation for datasets with varying dimensionality is non-trivial, given a) the various possible structural characteristics of datasets and b) the plethora of separation measures that exist. Building upon recent findings in visualization research about the qualitative and quantitative evaluation of class separation for 2D dimensionally reduced data using scatterplots, this research addresses the visual analysis of class separation measures for high-dimensional data. We present SepEx, an interactive visualization approach for the assessment and comparison of class separation measures for multiple datasets. SepEx supports analysts with the comparison of multiple separation measures over many high-dimensional datasets, the effect of dimensionality reduction on measure outputs by supporting nD to 2D comparison, and the comparison of the effect of different dimensionality reduction methods on measure outputs. We demonstrate SepEx in a scenario on 100 two-class 5D datasets with a linearly increasing amount of separation between the classes, illustrating both similarities and nonlinearities across 11 measures.

1. Introduction

The separability of classes in datasets is an essential topic in many data science problems. Class separation measures aim at quantifying how well distributions of classes, clusters, or groups in datasets can be separated, as illustrated in Figure 1. Class separation can be measured, per instance, per class, or per dataset. In this work, we analyze separation measures for high dimensional data, two-dimensional data, and both. We focus on measures that give one value per dataset (coarsest granularity), allowing for the assessment of measures for hundreds of datasets in one analysis approach.

Class separation measures play an important role in Machine Learning (ML) in, e.g., the synthesis of datasets [ALM11], the selection of datasets for evaluations, data studies, or sensitivity analyses [RLSKB17], the selection of features [RFT18], the analysis of cluster quality [HBV01, AGM*13], or the analysis of classification quality, confusion, and problem complexity [HB02, SL09, RFT18].

Separation measures have also been explored in the visualization (VIS) community over the past decade as one type of visual quality measures [BTK11]. In VIS, class separability itself has many facets, with spatial overlap being the major defining characteristic for human observers [STMT12]. Previous work has investigated how to quantify the degree to which class separation is preserved when high-dimensional (nD) data is projected to an easily visualizable 2D space with dimensionality reduction (DR) methods [SNLH09, TAE*09, EMK*19]. These class separation measures have been assessed qualitatively with respect to human judgements [STMT12], which have been used as ground truth for quantitative learned predictions [SA15], leading to the design of many more separation measures [AS16].

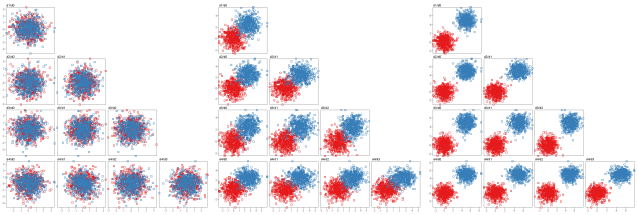

Figure 1: *Scatterplot matrices showing 3 out of 100 5D datasets with two classes (blue and red). Overall, these 100 synthetic datasets, discussed in our usage scenario, differ linearly by the degree of class separation, from total overlap to well-separated.*

Selecting a reasonable separation measure for a given problem remains a major challenge. It has long been argued that there is not a single best measure that serves all purposes [MC85]. While several specific studies exist that compare different separation measures [AGM*13], e.g., by using statistical comparison [GMA*11] or user studies [LAdS12], the questions of how to select an adequate measure for a given problem and how to compare different measures qualitatively remain open. These questions occur in both the ML literature with a focus on nD data [AGM*13, LAdS12] and in the VIS literature with a focus on DR-reduced 2D data [AS16].

Many separation measures have been proposed, with popular ones including Dunn [Dun74], Silhouette [Rou87], Davies–Bouldin [DB79], and Calinski–Harabasz [CH74]. Separation mea-

DOI: 10.2312/eurova.20201079

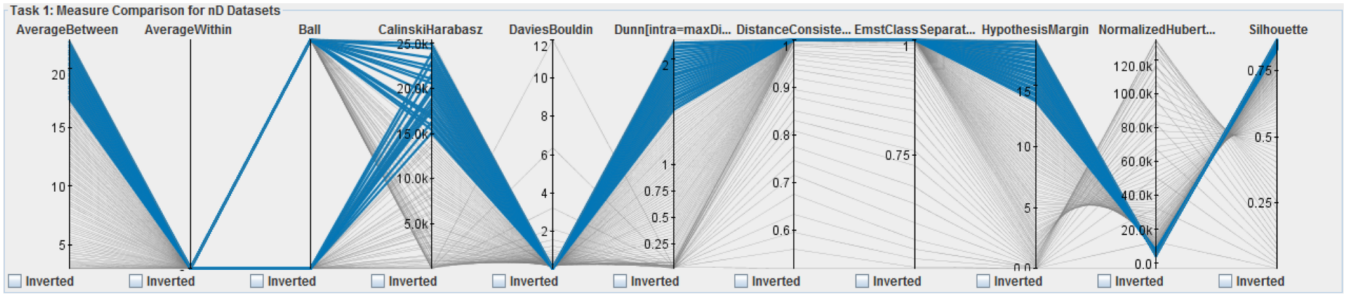

Figure 2: *T1: Parallel coordinates visualization used for the visual comparison of measure outputs across datasets. Each measure defines one axis, scores of datasets are represented as grey lines (blue when selected). The 100 datasets used in the usage scenario differ in a linear increase of class separation (cf. Figure 1). We analyze 11 measures and identify interesting behaviors: 7 measures asses high separability with high values, 3 with low values (Average Within, Davies Bouldin, and Normalized Hubert), Ball is binary. Average Between, Callinski Harabasz, Dunn, and Hypothesis Margin reflect the linearity of the controlled dataset very well. In contrast, Davis Bouldin, Distance Consistency, Emst Class Separation, and Silhouette show a non-linear behavior. Most measures preserve the order, except Ball and Normalized Hubert. Finally, the value domains of the measure outputs differ considerably. Distance Consistency and Silhuette are bound to [0..1], whereas some measures are open in one direction, some with very high values such as Callinski Harabasz or Normalized Hubert.*

sures capture different characteristics such as local neighborhoods, entropy, within-class and between-class distances, class diameters, class density and compactness, and minimum spanning trees [AS16]. The measures we picked range from early and well-established to newer and more exotic ones. Selection criteria have been their popularity, wide usage, and diversity. The output of separation measures are numeric scores; they may range along different scales, impeding their comparison. This heterogeneity makes selecting an appropriate separation measure difficult even with only a few datasets, and extremely challenging with dozens or hundreds.

Moreover, the outputs generated by separation measures very much depend on dataset characteristics, such as the number of instances, classes, and dimensions as well as class shape and class imbalance. The dependency between dataset characteristics and separation measures is a research line that has not yet been fully explored in a systematic way [GMA*11, LAdS12]. Another open question, which is particularly relevant in the VIS community, is how DR affects class separation scores. A measure applied to original nD datasets does not necessarily produce consistent output on the 2D projected equivalents. Finally, it is particularly difficult for people who are not experts in high-dimensional data characteristics, such as domain experts and students, to make informed decisions about selection measures for a given collection of datasets.

Interactive exploratory visual analysis tools could lead people to a deeper understanding of these rather abstract separation measures and their complex relationship to dataset characteristics. With high-dimensional data, the interpretation of nD distances is difficult and unintuitive for humans [BGRS99], and understanding their transformation into 2D distances is a challenge when working with DR-reduced data [SZS*16].

To provide better guidance in this analysis process, we contribute SepEx, the first interactive visual analysis tool for identifying and comparing characteristics of class separation measures. SepEx supports the interactive visual analysis of up to twelve measures of class separation in parallel for both direct nD data and DR-reduced 2D data using three different DR methods. It supports comparison across hundreds of datasets between measures, between DR methods, and measure consistency between nD and 2D. We illustrate SepEx in a usage scenario in which we analyze 11 measures for 100 synthetic 5D datasets, differing in a controlled parameter: distances of class centers. The dataset is publicly available for usage and replication purposes [†] Maximizing the number of controlled dataset parameters allowed us to validate SepEx with a full focus on separation measures: The tool helped us to quickly identify and compare expected characteristics of measures and DR methods, and also revealed some surprises. SepEx is designed to help many target audiences, including users who want to apply measures, developers who design and validate novel measures, and students who aim at learning and understanding these measures.

2. SepEx Abstractions and Interface

We describe the three analysis tasks that motivated the design of SepEx, provide an overview of its interface, and discuss each of its three views.

2.1. Analysis Tasks

We target three central analysis questions related to class separation measures, including high-dimensional data, 2D data, and both.

T1: Comparing multiple measures for multiple datasets. The design target is to provide analysts with an overview to compare the characteristics of up to a dozen separation measures for up to hundreds of datasets. Analysts will have a means to identify individual characteristics, and to compare commonalities and differences across measures. This eases the selection of the most applicable measures for a task as well as the removal of measures with redundant behavior.

T2: Assessing the consistency between nD and 2D data. The design target is to help analysts gain an understanding of effects introduced by DR methods on multiple class separation measures. Ideally, class separation would be consistent across the nD dataset

† http://juergen-bernard.de/paperPages/euroVA2020Bernard/paper.html
100 Synthetical Datasets for the Assessment of Class Separation Measures.

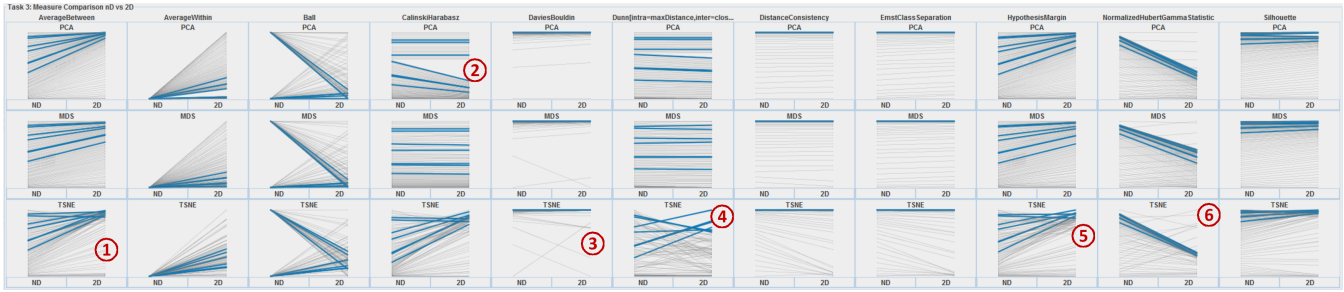

Figure 3: *T2: Comparison of pairs of measure outputs applied on nD data (left vertical axis) and DR-reduced 2D data (right vertical axis) using slope charts. 11 measures are aligned horizontally, 3 DR results in vertical direction (PCA, MDS, TSNE). Some interesting findings include: 1) & 5) For the Average Between measure, TSNE is inconsistent compared to PCA and MDS as it has larger slopes. 2) The PCA-based output shows an anomaly; after a detailed investigation, we found out that in WEKA's PCA implementation, the 2D PCA only returns one principal component, when the remaining variance is approaching zero. 3) & 4) For Davies Bouldin and Dunn, TSNE has some rank differences and is thus rather inconsistent. The datasets of finding 4) have been highlighted by selection. 6) Hubert Statistics has an interesting diagonal pattern across all three DRs. The projections seem to compress its value range but it remains mostly consistent.*

and its 2D projection, which should also be reflected by respective measures. T2 highlights inconsistencies and deviations from the ideal case and thereby helps to assess the robustness of separation measures with respect to DR, and the changes of intrinsic structures in data introduced by DR.

T3: Comparing measures for different DR methods. The design target is to support the analysis of 2D projections. T3 compares measures for multiple DR methods across multiple datasets, to identify and compare effects of DR on separation measures.

2.2. SepEx Overview

SepEx contains three views, each providing a different perspective on class separation measures that is directly aligned with one of the three analysis tasks. Two views use parallel coordinates and one uses strip plots, with several common themes. The class separation measures are always arranged side by side. The measure outputs are always shown along vertical axes, where class separation grows from bottom to top. Labels for the value domain of each measure can be shown, or suppressed to reduce clutter. Individual datasets are shown as grey lines, with transparency used to mitigate over-plotting occlusion. Showing the data at the granularity of individual items allows inspection of single item (with a click) or multiple items (with a rectangular drag or a lasso for an arbitrary shape). Selected dataset lines are blue, with linked highlighting across all views to enable seeing relationships between brushed data and comparing observations from different perspectives.

2.3. T1: Comparing Multiple Measures for Multiple Datasets

SepEx faces the challenge of visualizing up to a dozen measures for hundreds of multivariate datasets with parallel coordinates. Each of the coordinate axes depicts one measure, labelled at the top with its name. Every dataset is depicted with a gray line in the chart. With parallel coordinates, the active value domain of every measure is normalized in the visual space, which eases the visual comparison. With parallel coordinates, we face the trade-off between the available display space and the complex information to depict. Selecting datasets along the axes eases the comparison of non-adjacent measures. Figure 2 shows how 11 measures are made comparable, despite differing considerably in their value domains. Measures that are linearly correlated give rise to horizontal line segments. In contrast, sloped segments and crossed lines that show changes in rank indicate datasets for which the measures lead to inconsistent outputs and measure disagreement. In downstream steps, selected datasets can be analyzed in detail with scatterplot matrices for nD or scatterplots for DR data (figures in the supplemental material).

2.4. T2: Assessing Consistency Between nD and 2D Data

Figure 3 shows the view to support T2, i.e., to assess the effect of DR on the measure outputs. The visualization task at hand is showing ranks of two variables and rank changes between nD and 2D. This is why we use a slope chart, as opposed to a scatterplot, which would be preferable for correlation analysis. Again, separation measures are horizontally aligned columns, and each DR method is given a row, resulting in a grid of measures vs. DRs. With slope chart, analysts can compare the numerical outputs of both measures as well as rankings of datasets with respect to separability expressed by different measures. The results can be interpreted similarly as with T1, where slopes and crossings indicate inconsistencies between measure scores for nD and 2D.

2.5. T3: Comparing Measures for Different DR Methods

While T2 primarily focused on comparing nD to 2D, T3 focuses on comparing separation measures across DR methods directly. Figure 4 shows the interface for T3, again using PCA [Jol11], MDS [Kru64], and TSNE [MH08] as example DR methods. Vertical strip plots depict the distributions of every measure, aligned side-by-side horizontally, with the DR methods nested within each column. Labels for measures are placed on top, with labels for the DR methods at the bottom. Every dataset is represented by one grey line, normalized in the visual space to ease comparison, as in the other views. Selections are highlighted by blue triangles on the left side of the grey dataset lines. We chose strip plots as they show both distribution and visual density of the analyzed datasets. Furthermore, strip plots match the look-and-feel of our other designs. A user control allows switching to boxplots, as a variant for an aggregated result representation.

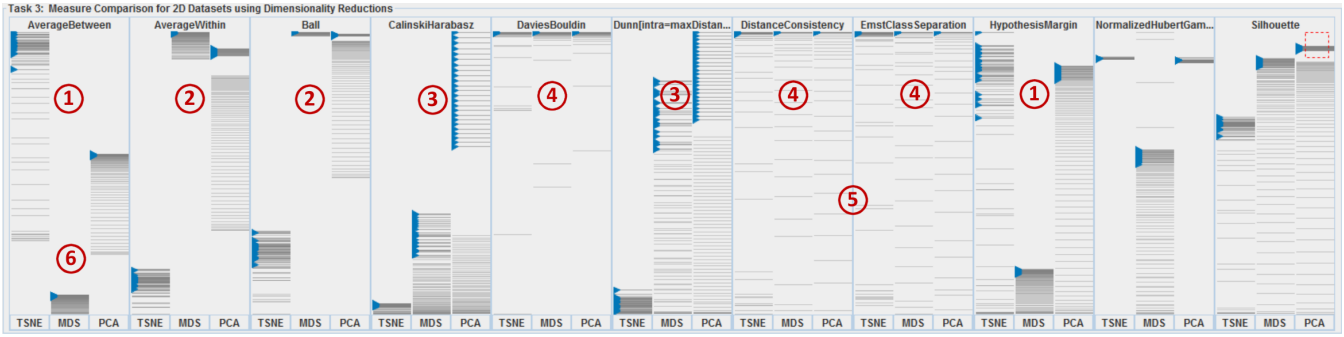

Figure 4: *T3: Strip plots for the visual comparison of 11 measures, each applied on 3 different DR-reduced datasets (PCA, MDS, TSNE). For both comparisons between measures as well as within measures (between DRs), we identify considerable differences. Some groups of measures with similar output patterns stand out: 1) [Average Between, Hypothesis Margin], 2) [Average Within, Ball], 3) [Calinski Habersz, Dunn], as well as 4) [Distance Consistency, Emst Class Separation, Davis Bouldin]. Focusing on the comparison of DR outputs reveals considerable differences as well. 5) Only for Distance Consistency and Emst Class Separation the DR results are similar. 6) An unexpected finding is how strongly measure outputs differ for different DRs: using Average Between as an example, TSNE yields average to high separability, all PCA-based results achieve medium separability. In contrast, Average Between assesses MDS results hardly separable. Finally, we select the most separated datasets according to the PCA-anomaly (cf. T2 in Figure 3) by using rectangle selection in the Silhouette measure at the upper right. It can be observed how differently this PCA-specific characteristics is reflected across the 11 measures.*

3. Usage Scenario

We demonstrate how SepEx can be used in a sensitivity analysis scenario. Our goals thereby are to validate SepEx by primarily studying measure characteristics, and excluding effects stemming from (uncontrolled) dataset characteristics (see future work). Therefore, we employ 100 synthetic datasets, all with 5 dimensions, 1000 instances, and two classes. The datasets differ by their class separation from overplotted to separated (cf. Figure 1, more details in the supplemental material). We analyze how consistent the estimates of 11 separation measures are for the differently separated datasets. The results of 3 DR methods further allow the analysis of consistency between nD and 2D data representations, followed the visual analysis of 11 measures applied on the different DR-reduced 2D datasets.

Figure 2 shows the different orientations, value domains, bounds, degrees of preserving the linearity, degrees of preserving order, and outlierness relevant for T1, affirming the great variety of measures and measure characteristics that we could identify with SepEx. Informed by this analysis, we invert the value domains of the measures Average Within, Davies Bouldin, and Hubert Statistics, to foster the comparability of measures for T2 and T3. In Figure 3, the analysis of consistency between the measure outputs of nD data and DR-reduced data are shown. We learn that TSNE-based projection leads to less consistent separability estimates compared to MDS and PCA, by observing slopes and line crossings (rank violations) across measures. These can be explained by the non-linearity of TSNE, its non-deterministic nature (randomizations), and the tendency to carve out cluster structures. In the supplemental material, we use scatterplots of dimensionality-reduced data, to further investigate this inconsistency of TSNE, using finding 4) in Figure 3 as a starting point. In Figure 4, we compare the 2D data representations of the 3 DR methods across the 11 measures. It stands out that combinations of measures and DR methods lead to considerably different separation results. While we were able to group measures with similar behavior, the output of the 3 DR methods leads to very different and individual separation results for most measures.

4. Conclusions & Research Opportunities

SepEx is the first interactive analysis tool to support the visual assessment of class separation measures for nD data, DR-reduced 2D data, and both. It supports the comparison of nD data to DR-reduced 2D projections for up to twelve class separation measures, using three or more DR methods across hundreds of datasets, supporting three analysis tasks. In a usage scenario, we demonstrated how SepEx enabled us to identify a series of measure characteristics, as well as commonalities and differences across measures. In summary, SepEx helps to better understand the interactions between separation measures, datasets, and DR methods and thereby can help in the deeper understanding of separation measures. SepEx eases the informed selection of measures for downstream tasks, such as data studies and sensitivity analyses.

One limitation we observed is the algorithmic scalability with respect to multiple datasets and DRs, which we resolve with precomputation to achieve interactive rates. For 100 datsets, 3 DRs and 11 measures, pre-computation took five hours on a standard notebook. Inspired by the findings of the usage scenario, another aspect is preprocessing of measure outputs, regarding the inversion of value domains, coping with outliers, and normalizations to cope with special or even unique value domains. Another idea comes with the parameters of DRs or DR variants, which opens another space for detailed analysis. To further enhance analytical capabilities, ideas would be to strengthen the focus on rank-based analyses of measure outputs as well as guidance concepts, to support the scalability of the approach for dozens of measures and thousands of datasets. Finally, our results show that even with a very simple dataset, surprising results are revealed. A next step is to analyze higher-dimensional datasets, datasets with different class shapes, and real-world data, to see which observations can be generalized, and how measures behave with respect to such data characteristics.

Acknowledgements

Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161.

References

[AGM*13] ARBELAITZ O., GURRUTXAGA I., MUGUERZA J., PÉREZ J. M., PERONA I.: An extensive comparative study of cluster validity indices. *Pattern Recognition 46*, 1 (2013), 243–256. 1

[ALM11] ALBUQUERQUE G., LOWE T., MAGNOR M.: Synthetic generation of high-dimensional datasets. *Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2317–2324. 1

[AS16] AUPETIT M., SEDLMAIR M.: SepMe: 2002 new visual separation measures. In *Pacific Visualization Symposium (PacificVis)* (2016), IEEE, pp. 1–8. 1, 2

[BGRS99] BEYER K., GOLDSTEIN J., RAMAKRISHNAN R., SHAFT U.: When is "nearest neighbor" meaningful? In *International Conference on Database Theory* (1999), Springer, pp. 217–235. 2

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2203–2212. 1

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

[DB79] DAVIES D. L., BOULDIN D. W.: A cluster separation measure. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) PAMI-1*, 2 (1979), 224–227. 1

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

[EMK*19] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Towards a quantitative survey of dimension reduction techniques. *Transactions on Visualization and Computer Graphics (TVCG)* (2019). doi:10.1109/TVCG.2019.2944182. 1

[GMA*11] GURRUTXAGA I., MUGUERZA J., ARBELAITZ O., PÉREZ J. M., MARTÍN J. I.: Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters 32*, 3 (2011), 505–515. 1, 2

[HB02] HO T. K., BASU M.: Complexity measures of supervised classification problems. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24*, 3 (2002), 289–300. 1

[HBV01] HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M.: On clustering validation techniques. *Journal of Intelligent Information Systems 17*, 2-3 (2001), 107–145. 1

[Jol11] JOLLIFFE I.: *Principal Component Analysis*. Springer Berlin Heidelberg, 2011, pp. 1094–1096. 3

[Kru64] KRUSKAL J. B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27. 3

[LAdS12] LEWIS J., ACKERMAN M., DE SA V.: Human cluster evaluation and formal quality measures: A comparative study. In *Annual Meeting of the Cognitive Science Society* (2012), vol. 34. 1, 2

[MC85] MILLIGAN G. W., COOPER M. C.: An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*, 2 (1985), 159–179. 1

[MH08] MAATEN L. V. D., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, Nov (2008), 2579–2605. 3

[RFT18] RAUBER P. E., FALCÃO A. X., TELEA A. C.: Projections as visual aids for classification system design. *Information Visualization 17*, 4 (2018), 282–305. doi:10.1177/1473871617713337. 1

[RLSKB17] RAMIREZ-LOAIZA M. E., SHARMA M., KUMAR G., BILGIC M.: Active learning: an empirical study of common baselines. *Data Mining and Knowledge Discovery 31*, 2 (2017), 287–313. 1

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

[SA15] SEDLMAIR M., AUPETIT M.: Data-driven evaluation of visual quality measures. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 201–210. 1

[SL09] SOKOLOVA M., LAPALME G.: A systematic analysis of performance measures for classification tasks. *Information Processing & Management 45*, 4 (2009), 427–437. 1

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 831–838. 1

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1335–1344. 1

[SZS*16] SACHA D., ZHANG L., SEDLMAIR M., LEE J. A., PELTONEN J., WEISKOPF D., NORTH S. C., KEIM D. A.: Visual interaction with dimensionality reduction: A structured literature analysis. *Transactions on Visualization and Computer Graphics (TVCG) 23*, 1 (2016), 241–250. 2

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNORK M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Symposium on Visual Analytics Science and Technology (VAST)* (2009), IEEE, pp. 59–66. 1