ProSeCo: Visual analysis of class separation measures and dataset characteristics

Bernard, Jürgen ; Hutter, Marco ; Zeppelzauer, Matthias ; Sedlmair, Michael ; Munzner, Tamara

Abstract: Class separation is an important concept in machine learning and visual analytics. We address the visual analysis of class separation measures for both high-dimensional data and its corresponding projections into 2D through dimensionality reduction (DR) methods. Although a plethora of separation measures have been proposed, it is difficult to compare class separation between multiple datasets with different characteristics, multiple separation measures, and multiple DR methods. We present ProSeCo, an interactive visualization approach to support comparison between up to 20 class separation measures and up to 4 DR methods, with respect to any of 7 dataset characteristics: dataset size, dataset dimensions, class counts, class size variability, class size skewness, outlieriness, and real-world vs. synthetically generated data. ProSeCo supports (1) comparing across measures, (2) comparing high-dimensional to dimensionally-reduced 2D data across measures, (3) comparing between different DR methods across measures, (4) partitioning with respect to a dataset characteristic, (5) comparing partitions for a selected characteristic across measures, and (6) inspecting individual datasets in detail. We demonstrate the utility of ProSeCo in two usage scenarios, using datasetsÂ 1 posted at https://osf.io/epcf9/.

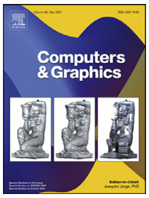
DOI: https://doi.org/10.1016/j.cag.2021.03.004

Computers & Graphics 96 (2021) 48–60

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag



Special Section on EuroVA 2020

ProSeCo: Visual analysis of class separation measures and dataset characteristics

Jürgen Bernard[a,d,*], Marco Hutter[b], Matthias Zeppelzauer[c], Michael Sedlmair[b], Tamara Munzner[d]

[a] University of Zurich, Switzerland
[b] University of Stuttgart, Germany
[c] St. Pölten University of Applied Sciences, Austria
[d] University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 13 November 2020
Revised 12 March 2021
Accepted 23 March 2021
Available online 28 March 2021

Keywords:
Computers and Graphics
Formatting
Guidelines

ABSTRACT

Class separation is an important concept in machine learning and visual analytics. We address the visual analysis of class separation measures for both high-dimensional data and its corresponding projections into 2D through dimensionality reduction (DR) methods. Although a plethora of separation measures have been proposed, it is difficult to compare class separation between multiple datasets with different characteristics, multiple separation measures, and multiple DR methods. We present ProSeCo, an interactive visualization approach to support comparison between up to 20 class separation measures and up to 4 DR methods, with respect to any of 7 dataset characteristics: dataset size, dataset dimensions, class counts, class size variability, class size skewness, outlieriness, and real-world vs. synthetically generated data. ProSeCo supports (1) comparing across measures, (2) comparing high-dimensional to dimensionally-reduced 2D data across measures, (3) comparing between different DR methods across measures, (4) partitioning with respect to a dataset characteristic, (5) comparing partitions for a selected characteristic across measures, and (6) inspecting individual datasets in detail. We demonstrate the utility of ProSeCo in two usage scenarios, using datasets [1] posted at https://osf.io/epcf9/.

1. Introduction

The separability of classes in datasets is an essential topic in many data science problems. Class separation measures aim at quantifying the extent that explicitly designated groups within datasets are distinguishable in terms of spatial proximity between instances, as illustrated in Fig. 1; these groups are typically called classes or clusters, and instances are sometimes called points or items.

Class separation measures play an important role in Machine Learning (ML), with applications that include dataset synthesis [2], feature selection [3], and cluster quality analysis [4]. Separation measures have also been explored in the visualization (VIS) community as visual quality measures to quantify characteristics important to human observers [5], such as the amount of spatial overlap between classes when high-dimensional data is projected to an easily visualizable 2D space with dimensionality reduction (DR) methods [6].

It has long been argued that no single best class separation measure serves all purposes [7,8]. The output of separation measures are numeric scores; these may be difficult to directly compare because of many differences. They may range along different scales, or have different distributions of their values. They may encode preferable states as either large or small values. Thus, in practice, it remains challenging to assess the suitability of a measure for a given problem or dataset.

In this work, we address analysis of class separation measures according to three central aspects: separation measures, dataset characteristics, and DR method.

First, separation measures are the most obvious and fundamental lens of analysis. Class separation can be measured per instance, per class, or per dataset. We focus on measures of the coarsest possible granularity, that give one value per dataset, because our goal is the assessment of measures for up to 1000 datasets simultaneously in a single analysis session.

[*] Corresponding author at: University of Zurich, Switzerland.
E-mail address: mail@juergen-bernard.de (J. Bernard).

https://doi.org/10.1016/j.cag.2021.03.004

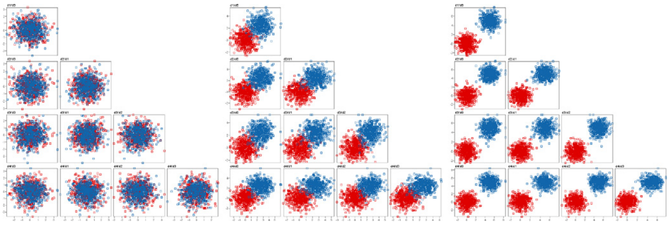


Fig. 1. Scatterplot matrices showing 3 out of 100 5D datasets with two classes (blue and red), synthetically generated to differ linearly by the degree of class separation, from total overlap to well separated. This controlled dataset collection is analyzed in depth in Section 6.1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Second, the characteristics of the datasets can dramatically affect analysis outcomes of separation measures and DR methods. We capture some interesting variations of this very general idea by focusing a small set of six continuous or discrete attributes: dataset size for both instances and dimensions, class counts and class size variability and skewness, and the number of outliers. We also consider a seventh binary attribute, whether datasets are synthetic or real-world.

Third, the DR method used can affect class separation scores. A measure applied to original nD datasets does not necessarily produce consistent output on the 2D projected (DR-2D) equivalents [9]. The interpretation of nD distances is difficult and unintuitive [10], so understanding their transformation into 2D is a challenge when working with dimensionality-reduced data [11].

Our primary contribution is the design and implementation of the interactive visual analysis tool ProSeCo, short for Probing Separation Comparison. It directly tackles the challenge of how to compare different separation measures qualitatively and quantitatively. It supports investigating the consistency of class separation measures in different contexts, such as analyzing the consistency of separation measures across large dataset collections, or studying the influence on DR on the estimates of class separation measures.

ProSeCo builds on a preliminary workshop paper [12] introducing SepEx, which covered only measures and DR methods. In this work, we also include dataset characteristics as a third aspect, and also scale to a larger set of measures, DR methods, and datasets. ProSeCo handles the interactive visual analysis of up to 20 class separation measures in parallel for both direct nD data and DR-2D data, for up to 4 DR methods, across a collection of up to 1000 datasets.

We provide evidence for the utility of ProSeCo through two usage scenarios, one with a *controlled* collection of datasets synthetically generated to differ linearly in a single parameter, and one with a *heterogeneous* collection of many datasets with variance across all seven dataset characteristics. We demonstrate different analysis stages including inversion of measure axes to align the valence of value domains for each measure, filtering outlier datasets based on specific measures, filtering and partitioning according to specific characteristics, and comparing nD data to its DR-2D analog. This scenario leads to a secondary contribution, a curated collection of 57 datasets where outliers across 12 different measures have all been culled.

2. Background

At the heart of our work is the visual analysis of class separation measures, which have been studied in both the machine learning and visualization communities. To provide context, we briefly review how these have been used in both literatures.

2.1. Classseparation measures in machine learning

The ML literature contains many synonyms for class separation measures including cluster separation measures, cluster separation factors, cluster separation indices, cluster quality indices, and cluster validity indices. These measures represent metrics that aim at quantifying how well distributions of classes, clusters, or groups in datasets are separated [4]. Separation measures quantify the separation and compactness of the groups and thereby help to assess the quality of a clustering or class assignment. A literature survey on cluster separation measures provides a more detailed discussion [13]. We focus here on measures that achieve one scalar value for the entire dataset.

Many separation measures have been proposed, with popular ones including Dunn [14], Silhouette [15], Davies–Bouldin [16], and Calinski–Harabasz [17]. Separation measures capture different characteristics such as local neighborhoods, entropy, within-class and between-class distances, class diameters, class density and compactness, and minimum spanning trees [18], see also Section 4.2.

Class separation measures play an important role in ML for problems such as the synthesis of datasets [2,19] and the selection of datasets for evaluations, data studies, or sensitivity analyses [20]. They further can be used to assess the complexity of a labeled dataset. Given the class labels for each data instance in the dataset, separation measures quantify the amount of overlap and class confusion [21] in the dataset and thereby the problem complexity of the dataset [3,22,23]. Separation measures can further be used to objectively compare the performance of clustering methods. The evaluation of clustering methods is difficult, because usually no ground truth exists. Given two clustering results, it is difficult to say which one is better. Separation measures enable to assess and compare cluster quality and clustering performance [4,24]. Another application of separation measures is feature selection [3]; that is, to assess how well a feature separates two given classes in a dataset. A classifier built upon this basic concept is the decision tree [25].

2.2. Class separation measures in visualization

Separation measures have also been explored in the VIS community over the past decade, as one type of visual quality measure [5,26]. The main difference to the ML community is that these measures usually focus on the visible 2D space, most commonly in the form of 2D scatterplots. Classical separation metrics for scatterplots took inspiration from ML and used, for instance, centroid-based or entropy-based approaches to model separability [27,28].

In addition to such heuristic-based measures, newer approaches have proposed to directly model the human perception of class separability in scatterplots [18,29]. In other words, the metric should reflect in how far a human observer perceives classes to be separable or not. Perceptual class separability is a non-trivial process which is affected by many visual dataset characteristics, with spatial overlap being the major defining characteristic for human observers [6].

These ideas have further been used to optimize DR and visual encoding methods according to how humans would perceive them. Espadoto et al. conducted a quantitative study comparing different DR methods based on separation measure results [30]. Wang et al. went one step further and used the best performing perceptual separation measures directly within a linear DR method [31]. Their method is similar to the venerable Linear Discriminant Analysis (LDA) approach [32] but optimizes separability according to human-perceived instead of heuristic-based separability. Perceptual separation measures were also used to optimize visual encoding techniques. An example of their use is to automatically assign colors to classes in scatterplots in a way that makes them easy to separate for humans [33], or simply to create completely new and tailored color palettes for that purpose [34].

3. Related work

We now discuss previous approaches and tools for the systematic analysis and comparison of class separation measures.

Existing endeavors into this direction focus on theoretical and empirical studies comparing different separation measures. In ML, there are only a few broader and systematic comparisons of separation measures [4,7,8,35–37]. These existing comparative studies focus mostly on the comparison of achievable performance, for example in terms of performance metrics or in estimating an appropriate number of clusters for a dataset. In contrast, our approach with ProSeCo is to focus on the consistency of estimates from different separation measures and the influence of dataset characteristics on separation measure outcomes.

The VIS literature does contain some previous analyses of separation measures. Visual assessment of cluster separation with the particular focus on dimensionality-reduced datasets was carried out by Bernard et al. for self-organizing maps (SOM) [38] and for outlier analysis methods [39]. Tatu et al. conducted a user study comparing four separation measures on a single dataset [28]. Expanding both on separation measures and dataset characteristics, Sedlmair and Aupetit evaluated 15 existing separation measures based on 828 two-class scatterplots [40].

The results of such studies are helpful as they give generic advice on which measures perform generally well and which not. However, it is hard to derive insights that are tailored to a specific problem at hand. For such individualized analysis approaches, a visualization tool that allows people to directly analyze separation measures would seem to be an obvious option. While a number of tools have been proposed for the analysis and selection of DR methods [9,41,42], there is surprisingly little previous work on interactive visual tools for separation measures. The goal of our work is to fill this gap.

Our previous workshop paper presented SepEx, a preliminary version of ProSeCo that was the first interactive visual data analysis system to handle the two aspects of separation measures and DR methods. However, it did not handle dataset characteristics at all. It did support the first three tasks discussed below in Section 4.1, but not the last three. It had more limited scalability than ProSeCo, handling 3 rather than 4 DR methods, 12 rather than 20 separation measures, and several hundred rather than 1000 datasets in a collection.

4. ProSeCo: abstractions

We first present the analysis tasks that motivated the design of ProSeCo, and state its scalability targets. We then present the specific choices for the three targets of the analysis tasks: separation measures, dataset characteristics, and DR methods.

4.1. Analysis tasks

We articulate six analysis tasks as the primary design concerns for ProSeCo, summarized in Table 1.

Table 1
The six analysis tasks supported by ProSeCo, targeting measures, DR methods, and dataset characteristics.

Task	Description
T1	Compare across measures
T2	Compare nD to DR-2D data across measures
T3	Compare between DR-2D data across measures
T4	Partition within dataset characteristic
T5	Compare dataset partitions across measures
T6	Inspect within individual dataset

T1: Compare across measures. The most basic task is to compare the output of each separation measure for each dataset in the collection, in search of an overview of commonalities and differences across these measures. This comparison may lead to identifying measures with redundant behavior, or identifying datasets that are outliers for particular measures.

T2: Compare nD to DR-2D data across measures. This task is to understand the effects introduced by DR methods on multiple class separation measures. Ideally, class separation would be consistent across the nD dataset and its 2D projection, which should also be reflected by respective measures.

T3: Compare DR-2D data across measures. This task is to directly compare the effect of *different* DR methods through analysis of their DR-2D projection results.

T4: Partition within dataset characteristic. This task is to partition dataset collections into meaningful bins according to user-selectable thresholds for a specific dataset characteristic, so that aggregate numerical values can be computed for each bin.

T5: Compare partitions across measures. This task is to compare the distributions of measure outputs for each bin of a partitioned dataset collection, to develop and test hypotheses about dependencies and relationships between the dataset characteristic used to partition the dataset (T4) and the behavior of measure outputs.

T6: Investigate individual dataset. This task is to see all relevant details of a single dataset, including the distribution of data elements in the nD space, the distribution of DR-2D data elements in the reduced spaces, and the distribution of classes in either data space. This auxiliary task supports exploration of why certain datasets behave differently than others in the other five primary tasks.

Considering these tasks as action-target pairs [43], the actions are *compare*, *partition*, and also *inspect*. The targets are separation measures, DR methods, and dataset characteristics.

As with many systems, scalability is a goal. In our case, scalability primarily relates to the number of datasets, measures, and DR methods that our tool can manage without visually overloading the interface. To that end, our overarching requirement is to support tasks T1 through T5 in a single screen, to harness the benefits of a navigation-free overview. In contrast, we deem the T6 task of inspecting an individual dataset to be sufficiently modular that it is not subject to that requirement. We consider an ambitious yet still manageable information density for that single overview screen, and work backwards from that constraint to identify scalability targets. Our analysis of the tradeoffs between the competing goals of the five primary tasks led us to identify the following scalability aims: 1000 datasets, 20 measures, 7 dataset characteristics, and 4 DR methods.

We address these scalability aims in the design and implementation choices of ProSeCo. We note that the exact choices are not the central aspect of our contribution, and could most certainly be changed. The crucial point is the design of an interface that supports the six tasks listed above and scales to roughly these cardinalities for the three target types and the dataset collection.

4.2. Separation measures

ProSeCo supports the analysis of up to 20 instances of separation measures at the same time. For the purpose of readability, we simply refer to all measure instances as measures. The 12 main class separation measures that we use in the usage scenarios range from early and well-established to newer and more exotic ones, for a mix of popularity and diversity. In addition, we chose one main measure (Dunn) to explore in detail through 9 different parameterizations, for a total of 20 measures that can be investigated at once. All of these measures are at the coarsest possible granularity, providing one value for an entire dataset. The 12 main supported measures are:

- Average Within [44]: The average within measure estimates the within-class homogeneity based on average within-class distances.
- Average Between [44]: The average between measure is similar to average within and measures the between-class separation via average between-class distances.
- Ball Index [45]: The ball index is the mean of the dispersions per cluster. The dispersion per cluster is computed as the mean of the squared distances between the cluster members and the barycenter of the cluster.
- Calinski–Harabasz Index [17]: The Calinski–Harabasz measure estimates the concentration of the classes around their center-of-mass using squared Euclidean distances.
- Davies–Bouldin Index [16]: The Davies–Bouldin index is computed by the average Euclidean distance between the centroid of the class and its individual members.
- Dunn Index [14]: The Dunn index is computed as the ratio of the minimum between-class distances over the maximum within-class distance in order to identify compact and well separated clusters or classes. The nine parameterizations that we explored arise from two parameters: within-class comparison (*avgCentroid*, *avg*, and *max*) and between-class comparison (*closest*, *centroids*, *furthest*), always referring to the distance relations used.
- Distance Consistency [27]: The distance consistency measure is computed as the portion of data points whose nearest class center (center-of-mass) belongs to the same class.
- Extended Minimum Spanning Tree (Emst) Class Separation [46]: The Emst separation measure is the average portion of the neighbors of each sample, which belong to the same class as the sample. The neighborhood is estimated by the extended minimum spanning tree of all samples.
- Hypothesis Margin [47]: The hypothesis margin measure is the average of the differences between distances from each sample to its nearest neighbor from another class and distances to its own class nearest neighbor.
- Normalized Hubert Statistics [48]: The normalized Hubert statistics measure is computed from two pairwise similarity matrices: one of all input samples and one of all class centers. Both matrices are multiplied with each other and z-standardized. The result is averaged to obtain one separation value for over all classes.
- Point-Biserial Index [49]:The Point-Biserial Index builds upon the point-biserial correlation coefficient and is computed as a product of two quantities: first, the difference of the mean within-class distances and the mean of the between-class difference. Second, the square root of the product of the number of within-class samples and the number of between-class samples divided by the total number of samples.
- Silhouette [15]: The silhouette measure represents the separation as the difference between the average between class distances and the average within-class distances, normalized by the maximum of these two quantities.

4.3. Dimensionality reduction techniques

From the large class of DR methods proposed in previous work [50], we selected a mix of longstanding approaches in widespread use and newer methods gaining in popularity. From oldest to newest, the 4 supported DR methods are:

- PCA [51]: Principal Components Analysis is a linear DR approach that identifies orthogonal directions (principal components) in the data. Each component represents one direction in the data along which variance in the data can be explained. For visualization the data is projected onto the first two principal components.
- MDS [52]: Multidimensional Scaling is a family of linear and nonlinear DR methods that attempt to preserve pairwise distances between objects. Our implementation takes the pairwise distances of all points as input and maps the points to 2D by preserving the between-point distances using Kruskal's stress optimization criterion.
- t-SNE [53]: t-distributed Stochastic Neighbor Embedding is a nonlinear DR method that emphasizes cluster structure using divergence of probability distributions between pairs of high dimensional objects to keep similar instances close and keep dissimilar distances far apart.
- UMAP [54]: Uniform Manifold Approximation and Projection first builds a high-dimensional graph representation of the data and then tries to find a lower-dimensional graph with as similar as possible structure. UMAP is a nonlinear approach with better preservation of global structure and faster computation time than t-SNE.

4.4. Dataset characteristics

Datasets can vary considerably in arbitrary aspects and characteristics. We formalized a small subset of frequently used dataset characteristics to make them available for analysis. We considered expressibility and intuitiveness, and only use characteristics that are robust, easy to re-implement, and computable at interactive rates. We were also inspired by the VizNet approach where similar data characteristics have been used to provide an overview of millions of datasets [55]. We selected a set of 7 basic dataset characteristics for integration in ProSeCo:

- **Dataset Size** is measured in number of data points (instances) in the dataset.
- **Dataset Dimensions** refers to the number of attributes (feature dimensions) in the dataset.
- **Class Counts** provides the number of distinct classes in the dataset.
- **Class Sizes Variation** measures the variation of the individual class sizes, from a value of zero for completely balanced to larger values that indicate more imbalance. We measure imbalance by the *coefficient of variation* of the class cardinalities (class sizes).
- **Class Sizes Skewness** measures the asymmetry of the distribution of class sizes. Balanced datasets have a value of 0, many small classes and few large ones yields a negative value, and the inverse is positive. Class sizes skewness is the third moment of the distribution of class cardinalities.
- **Outlierness** measures the amount of outliers in the dataset. We chose the MAD outlier detection criterion [56], following the rationale of VizNet [55].
- **Real-World** is a binary characteristic, assigning a dataset either as real-world or synthetic. Real-world is a meta-information provided manually, as the characteristics cannot be derived from the dataset content directly.

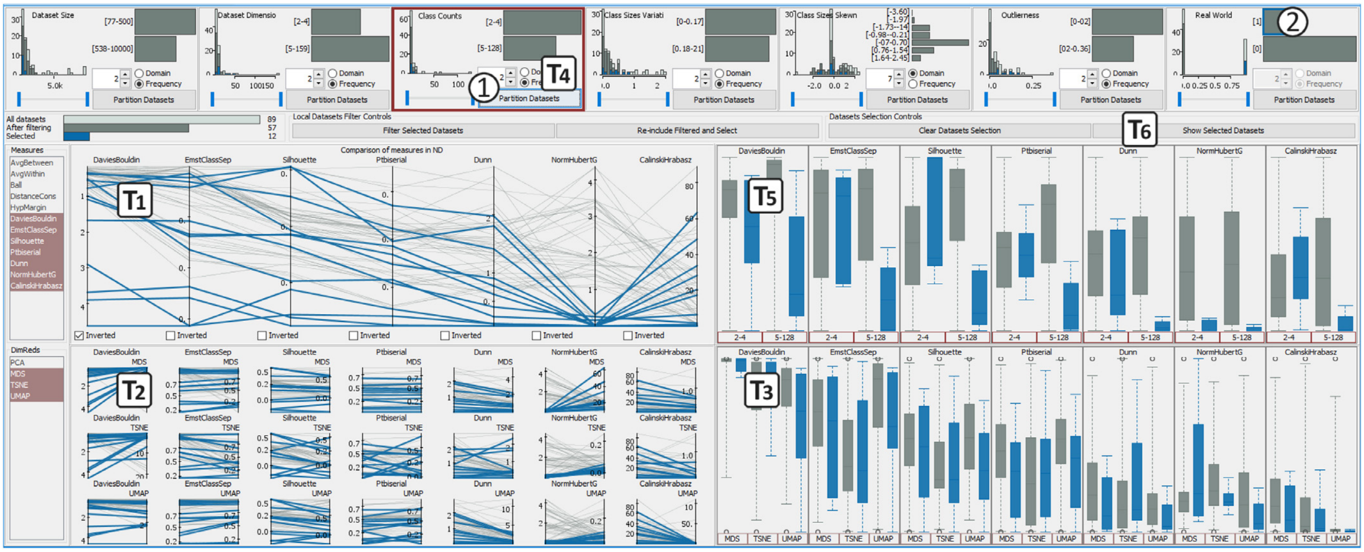


Fig. 2. ProSeCo Interface. Top view supports analysis and partitioning of dataset characteristics (T4). The four main views support comparing across measures of nD data (T1, middle left closeup), nD to DR-reduced 2D datasets (T2, bottom left closeup), DR-2D data across measures (T3, bottom right closeup), and partitions of datasets (T5, middle right closeup). The top T4 view shows how the selected data characteristic Class Counts is used to partition the datasets into two subsets (1, dark red outline). It also shows the selection of a bar (bin) that contains all real-world datasets (2, blue outline). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. ProSeCo interface

The ProSeCo interface is shown in Fig. 2. Each of the five views is designed to support one of the primary analysis tasks T1 - T5, described in Section 4.1. The figure shows ProSeCo applied to 89 heterogeneous datasets, as described in Usage Scenario 2 (Section 6.2). For the illustration of the different techniques in this section, we use the collection with 100 synthetic datasets, as used in Usage Scenario 1 (Section 6.1).

5.1. ProSeCo overview

At the center, four views are shown in a 2*x*2 grid, each supporting one individual analysis task. There are several common themes between these views, two of which use parallel coordinates (T1 and T2) and two of which use strip plots or box plots (T3 and T5). In each view, 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 (or vice versa if the value domain is inverted). Labels for the value domain of each measure can be shown, or suppressed to reduce clutter. Datasets are shown as gray lines, with transparency used to mitigate over-plotting occlusion.

Selected dataset lines are blue, with linked highlighting across all views to enable seeing relationships between brushed data and comparing observations from different perspectives. The selected measures, DR methods, and dataset characteristic are indicated with red.

Showing the data at the granularity of individual items with strip plots allows inspection of a single item (with a click) or multiple items (with a rectangular drag or a lasso for an arbitrary shape). The strip plots handle collections of hundreds of datasets, and for larger collections these two views can be switched to boxplots to support the analysis of statistical distributions of measure outputs in aggregate. The top view is a strip of small multiples, with vertical histograms showing data characteristic distributions and horizontal histograms of counts for the partitioned bins (T4). The control panel below it supports filtering and selection, including a button on the right to open popup panes for the detailed analysis of selected datasets (T6).

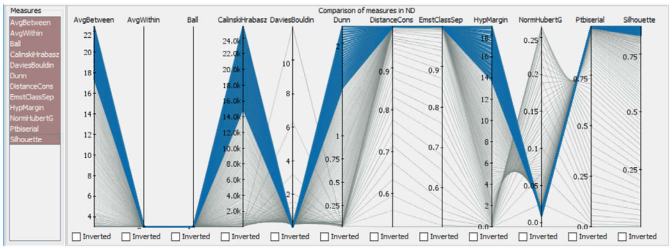


Fig. 3. Middle Left Closeup (T1): Parallel coordinates visualization used for comparing measure outputs across datasets, with one axis per measure and a grey line for each dataset (blue when selected). The 100 datasets shown are generated to have a linear increase of class separation (cf. Fig. 1). In the example, DistanceCons and EmstClassSep produce consistent results for the entire dataset collection, while measures AvgWithin and Ball show little sensitivity to the datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. T1: Compare across measures

The middle left view supports T1 with parallel coordinates, where each axis corresponds to one measure. This technique provides a compact representation to align with our goal of substantial information density, considering the trade-off between the available display space and the complex information to depict. The active value domain of every measure is normalized in the visual space, which eases the visual comparison. Interactive selection of contiguous groups of lines (representing datasets) aids in the comparison of non-adjacent measures.

The list-based control panel on the far left allows measures to be filtered out or reordered, updating all four of the middle views including the axis ordering in this view. Fig. 3 shows a closeup of this view where 12 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 highlight changes in rank and indicate that two measures yield inconsistent separability estimates for different datasets. An advantage of parallel coordinates is that they immediately show such inconsistencies prominently by crossings.

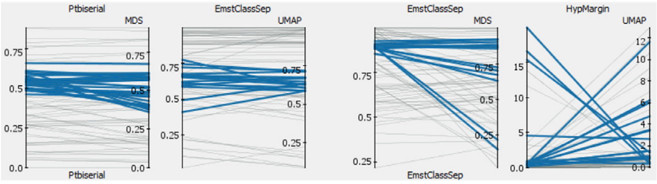


Fig. 4. Bottom Left Closeup (T2): Comparison of pairs of measure outputs applied on nD data (left vertical axis) and DR-reduced 2D data (right vertical axis) using slope charts. Four examples are shown: the left two charts show highly consistent mappings from nD to 2D, but the two on the right are rather inconsistent.

5.3. T2: Compare nD to DR-2D data across measures

The lower left view of ProSeCo supports T2, comparing nD to DR-2D data to show the effect of DR on the selected measure outputs (cf. Fig. 2). Fig. 4 shows a closeup of a few of the individual slope charts of the view.

Again, separation measures are horizontally aligned columns, and each DR method is given a row, resulting in a grid of measures vs. DR methods. The grid in this view supports three types of comparisons. The first is measure-centered: *how do measures compare?*, supported by comparing columns. The second is DR-centered: *how do DR methods compare?*, supported by considering rows. The third is mapping-centered: *how does the mapping of a single measure and DR from nD to 2D behave?*, supported by investigating individual cells in the matrix interface.

Our choice of slope chart was driven by the needs of T2, to show ranks of two variables and rank changes between nD and 2D. We considered scatterplots, which would be preferable for correlation analysis, but are less suited for handling ranks. With slope charts, analysts can compare the numerical outputs of both measures as well as rankings of datasets with respect to separability expressed by different measures. Moreover, they can be interpreted similarly to the view supporting T1, where slopes and crossings indicate inconsistencies between measure scores for nD and DR-reduced 2D data. Few or small ordering inconsistencies have minor visual impact, but a substantial number will be highly visually salient, to emphasize where the transformation leads to discrepancies. With the list-based interface on the lower left of ProSeCo, the filtering and reordering of DR methods can be done interactively, for both this view and its neighbor to the right. In Fig. 2, MDS, t-SNE, and UMAP were selected, whereas PCA was filtered out.

5.4. T3: Compare DR-2D data across measures

The bottom right view addresses task T3, comparing separation measures across DR methods directly with only the DR-2D data, in contrast to the comparison of nD to DR-2D data focus of T2. Vertical 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.

There are several choices for the visual representation in this view, shown in Fig. 5. In the fine-grained depiction, each individual dataset is shown through strip plots, which are normalized in the visual space to ease comparison as in the other views. Selections can be shown with blue triangles on the left of each line for further emphasis. In the coarse-grained depiction, distributions of many datasets are shown in aggregate via box plots, a design variant scaling up to thousands of datasets or even more. To further enhance visual comparison, selected datasets can also be highlighted within a second blue boxplot side-by-side to the original boxplot, which is the most simplified design variant. A benefit of strip plots is that they show both distribution and visual density of the analyzed datasets, and moreover closely match the look and feel of the parallel coordinates and slope plots. All three of these representations provide the ability to both highlight and directly select individual items, which is not available through aggregate representations such as histograms or boxplots. A benefit of boxplots is of course scalability. Although alternative to boxplots such as violin plots [57] have the benefits of visually indicating multi-modal distributions, they have the cost of requiring more horizontal screen space per plot to be legible.

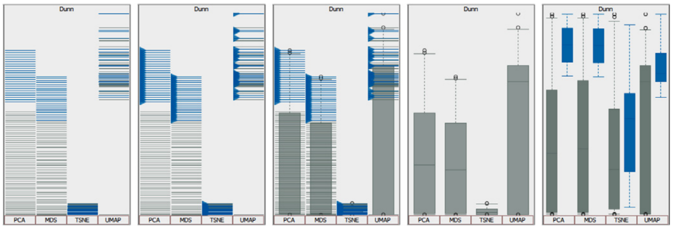


Fig. 5. Bottom Right Closeup (T3): Comparison of measures for DR-reduced datasets. Users can adjust the visual representation, leading to the five different variants of the interface. In the first three variants every dataset is visualized. In the last two variants boxplots are shown, thus being agnostic to the number of datasets. In example two and three, selected datasets are emphasized with blue triangles. The last variant also shows boxplots for current dataset selections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5. T4: Partition within dataset characteristics

The top view, shown in detail in Fig. 8, supports T4, providing an overview of the seven dataset characteristics (cf. Section 4.4) and allowing one of them to be selected as the partitioning to use in the view supporting T5.

There are seven small multiples, each with a vertical histogram to show the distribution of the characteristic across the value domain. Light gray bars in the histogram represent the distribution of the entire dataset collection, dark gray bars depict the current filter status, and blue bars highlight the currently selected subset of datasets. The horizontal sliders below the histograms enable filtering by restricting a characteristic's value range to select only a subset of datasets.

Each small multiple also contains partitioning controls and displays. The respective value distribution of each characteristic can be partitioned into a user-specified number of bins. Using the *Class Count* characteristics in Fig. 2 as an example, the interface allows analysts to apply binning operations to the dataset collection, specified and triggered in the visual interface. The first parameter that the user can interactively control is the number of bins of the partition, which can be adjusted using a numeric up-down control at the right of the interface. The second controllable user parameter is the choice between a domain-preserving or a frequency-preserving binning variant, enabled with a radio button below [58]. These two binning options either partition the given value range into *equally spaced* bins (domain) or partitions the underlying population of datasets into *equally sized* bins (frequency); both variants can be highly beneficial [59]. The result of the binning is displayed at the top right of the interface with a labeled horizontal bar chart. It is horizontal in order to (a) avoid confusion with the histogram and (b) to make labels more readable.

By default, ProSeCo uses two bins with a frequency-preserving strategy. Analysts can use the binning support to adjust the partition of the dataset collection interactively.

5.6. T5: Compare dataset partitions across measures

The middle right view supports task T5, comparing a single selected partition for one of the dataset characteristics across multi-

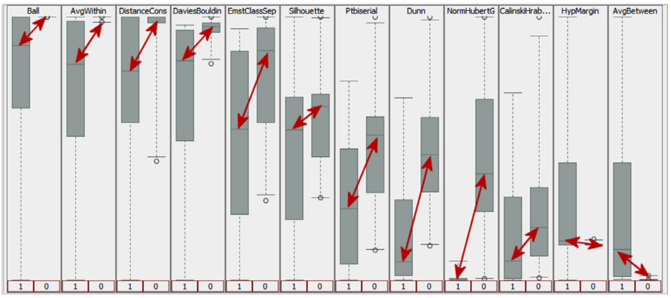


Fig. 6. Middle Right Closeup (T5): Comparison of measures across the different bins of a partition of the dataset collection. In the example, the real world ("1") versus synthetic ("0") dataset characteristic was chosen to partition datasets into two bins for each measure. We identify an interesting pattern across most measures: in the dataset collection, real-world seem to be less separable than synthetic datasets, highlighted with red arrows (manual annotation). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ple measures (cf. Fig. 2). This view supports identifying potential dependencies between these two targets. It relies on the choices made for T4 and is linked to the settings made in the small multiples in the top view.

As this task has very similar underlying requirements to T3, we re-use the same visualization design in this view as in the corresponding view below it. Fig. 6 shows a closeup example comparing real-world to synthetic datasets.

5.7. T6: Inspect within individual dataset

Task T6 is supported through a popup pane that appears only on demand. A button control on the upper right (cf. Fig. 2) enables users to open multiple popups at once, one for every selected dataset, also allowing the detailed visual comparison of multiple datasets on demand. This pane contains multiple views: a scatterplot matrix showing all pairwise combinations of dimensions to show full details about the nD data, individual scatterplots of the dimensionally reduced DR-2D instances for all of the active DR methods, and a parallel coordinates view showing all instances. In all of the views, the instances are color-coded by their class. Fig. 7 shows one example dataset with 22 dimensions and two classes using this interface. Figs. 35–122 in the supplemental materials document show this view for all datasets in Usage Scenario 2 (Section 6.2).

5.8. View coordination

ProSeCo uses extensive linkage between the views, with linked highlighting to show selections and linked filtering to arrive at more meaningful subsets of dataset collections. Datasets are the analysis objects that are shown for all the three aspects: separation measures, characteristics of dataset collections, and DR methods and DR-2D results; they can be selected in any of those views and filtered in all of them. There is a consistent color coding across all five primary views, with all datasets in light gray, the current filter status in dark gray, selected datasets in blue, and selected measures, dataset characteristics, and DR methods in red.

Highlighting naturally allows the comparison of selected and unselected subsets, a mechanism that is particularly relevant in tools with linked views such as ProSeCo where analysis objects are shown from different perspectives. Using data selection, analysts can gain insights into differences and commonalities between selected and unselected datasets, especially in linked views. The linked highlighting uses a global selection model that is connected with each individual view via appropriate event handling.

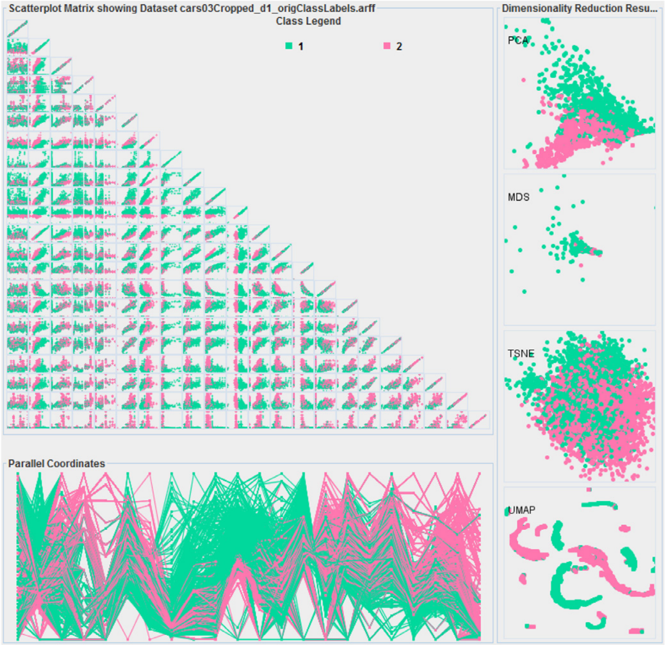

Scatterplot Matrix showing Dataset cars03Cropped_d1_origClassLabels.arff
Class Legend
1
2
Dimensionality Reduction Resu...
PCA
MDS
TSNE
UMAP
Parallel Coordinates

Fig. 7. Popup (T6): Interface for the detailed analysis of datasets including, scatterplot matrix, parallel coordinates, scatterplots for DR results and colored class labels. For the data set in the example DR results are quite different.

The global selection model is linked to all views in the system wherever datasets are (a) shown and (b) interactively selectable.

Individual datasets may be outlying or less relevant for the analysis. Filtering out these datasets helps to make space in the interface for the remaining dataset subset of interest, by individualizing the dataset collection to subsets that are appropriate for specific analysis tasks. As with selection, there is a global filter model that combines filter operations of the individual views.

In all four center views for the tasks T1, T2, T3, and T5, selection of datasets is enabled via rectangle selection or simple click selection. In the top view supporting T4, users can also select subsets of the dataset collection with a click on the vertical or horizontal histograms in the 7 small multiples. For the vertical histograms multi-selection is also enabled via rectangle selection.

In the top view showing dataset characteristics (T4), filtering is provided with a range slider at the bottom of every view, and filter events automatically update the filter status in every view.

There is an additional view coordination region, just below the top view (T4). It shows the counts of all datasets, after filtering, and selected through a bar chart (with color-coded bars, to serve double duty as a legend). It has buttons to move the current set of selections into the filter set, show all of the filtered sets again and select them, clear filters, and show the selected datasets with popup inspection views (T6).

5.9. Implementation, performance, and scalability

ProSeCo is implemented in Java, and is built on top of many external libraries that implement the separation measures and DR methods. To achieve interactive response times for large dataset collections and multiple DR methods, we do extensive precomputation of DR results for all datasets under the assumption that these are known a priori and remain static during analysis. For 100 datsets and four DRs, pre-computation took six hours on a standard notebook. The pre-computation of measure results for the nD and the DR-reduced 2D dataset collection only takes seconds to minutes, depending on the separation measure. Accordingly, in situations where the set of measures is adapted during the analysis session, such as when measure parameters are modified, the ProSeCo architecture takes only minutes to calculate new measure outputs for the nD dataset collection and the 2D dataset collections. In situations where dataset collections or the set of DRs are modified, ProSeCo is not real-time capable. It remains an open challenge how approaches like ProSeCo can be applied in scenarios where dataset collection are synthesized as an upstream process.

The two usage scenarios feature collections of around 100 datasets, but we have also tested ProSeCo at larger scales. The computational performance assessment with the largest tested collection was based on an experiment conducted by Sedlmair et al. [40] and included 828 datasets. ProSeCo maintained reasonable information density, and interactive frame rates with aggregate boxplots rather than fine-grained strip plots.

We thus consider that ProSeCo has achieved the scalability target of 1000 datasets. From a design point of view, scaling up from 100 to 1000 datasets is quite viable because each new dataset only adds another line to existing view, and we have aggregate visual encodings that further support this large scale. In contrast, it would be difficult to scale to substantially more than the current 20 measures, 7 dataset characteristics, and 4 DR methods. A core design decision is to have a single-screen overview that supports tasks T1 through T5, which imposes a very strong constraint on what numbers are viable. Scaling up beyond 20 measures would lead to some challenges, since it would require adding more axes to some of the views. Our choices for the design of the bottom two views lead to very strong constraints on the number of DR methods that could be used at once, since adding more has a combinatorial effect on the number of axes needed within them. However, the interface supports selecting only a subset of the measures and the DR methods, so it would be straightforward to add several more alternatives. A few more dataset characteristics could simply be added with the current design, perhaps up to one dozen, but going beyond that would require adding some kind of selection mechanism for those as well.

6. Usage scenarios

We present two usage scenarios, showing that ProSeCo fosters revealing a series of characteristics of measures for class separation, both expected and unexpected. The first scenario describes a highly synthetic dataset, which offers us full control to illustrate the ideas and potentials of ProSeCo. The second scenario focuses on the analysis of metrics using real-world data. Please note that both scenarios are realistic in the sense that they could support an analyst in understanding metrics. The equivalent in empirical science would be the duality of controlled and field experiments [60]. Similar approaches of studying synthetic vs. real-world datasets have been used in previous data studies as well [6,61]. Below, we summarize the major aspects of the analysis. The supplemental materials contain considerably more details, context information, and enlarged figures for both workflows. The two datasets [1] used in this analysis are posted at https://osf.io/epcf9/.

6.1. *Controlled datasets*

In the first scenario we fully control the dataset collection, so that the variations in the measured outputs can definitely be attributed to the measures, rather than dataset characteristics. Some parts of this scenario draw on sensitivity analysis, where an independent variable is changed while the output of a dependent variable is observed.

6.1.1. *Controlled dataset collection*

We create a collection of synthetic datasets, the *controlled dataset collection*, where we keep the values of all synthesis parameter constant, except the one that we vary in a controlled way. We create 100 synthetic datasets, all of which have 5 dimensions, 1000 instances, two perfectly balanced classes (500 instances per class), and the same distribution of instances of both classes. The variable parameter was the distance between centroids (centers of gravity) of the two classes (cf. Fig. 1). We went for constant change of distances between any two datasets, resulting in a set of datasets with a linear increase of class distances from zero to 10 times the size of diameter of the classes (which was identical for both classes). A more detailed overview of the controlled dataset collection is provided in the supplemental materials.

6.1.2. *Select measures*

We first reduce the number of separation measures from 20 down to 12, by using ProSeCo to conduct a parameter analysis for the nine different instances of the Dunn measure (cf. Section 4.2). We find out that between-class comparison seems to have a stronger impact on measure results as the within-class comparison criterion. The between-class parameter value which preserves the linearity of the datasets best is *centroids*. The supplemental materials contain the details of this parameter analysis. Informed by these findings, we decide to choose Dunn[*avgCentr*, *centroids*].

6.1.3. *Invert measure axes*

We continue the analysis by comparing measures for nD data (T1). We investigate valences, inverting individual measures as needed to ensure that all measure valences have the same orientation and are thus visually comparable. To make the identification of valence easy, we select a subset of datasets that all have high separability (measured with Silhouette), as shown in Fig. 3. Eight measures show positive valence, with value distribution clearly shifted vertically to the top. However, four measures have negative valence, assessing high separability with low values (AverageWithin, Ball, Davies Bouldin, and Normalized Hubert), so we invert them. We further observe that most measures preserve the order of class separation, except Normalized Hubert where multiple changes in rank occur shown as crossings in the parallel coordinates.

6.1.4. *Order measures*

We then determine groups of measures with similar value domains and re-order them accordingly, with the result shown in Fig. 9. Our analysis of the value domains of the 12 separation measures reveals considerable differences. The Distance Consistency, Point Biserial, and Silhouette measures are bound to [0.1], whereas some measures are open in one direction, some with very high values such as Calinski-Harabasz or Normalized Hubert. We check whether the linear increase of class distances in the controlled dataset, is reflected by the individual measures. We move the four measures that reflect the linearity together to the far right side, separating them from the group of five with non-linear behavior (from Davies-Bouldin to Silhouette). In between these groups is a measure that stands out, Normalized Hubert. It is the only one which violates the order of separation measure estimates: the least separated dataset receives the highest score, whereas the most separated dataset receives the second highest separation score (see red highlight in Fig. 9).

6.1.5. *Compare nD to DR-2D across measures*

The next step in the workflow is the comparison of measure results for nD data vs. DR-reduced 2D data (T2). Overall, the 12 measures and 4 DR methods yield 48 mappings from nD to 2D. In Fig. 10, we show eight out of 12 measures to emphasize on most apparent findings. In this figure, these findings are manually annotated with rectangles of different colors. From left to right, a first finding (left yellow rectangle) describes the 8 mappings of

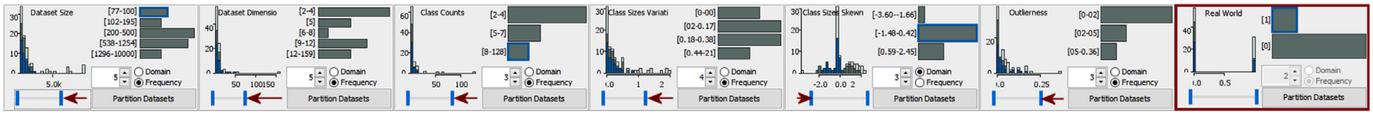


Fig. 8. Top Closeup (T4: Overview of dataset characteristics as provided with the dynamic query interface of ProSeCo. Overall, seven characteristics build the basis for analysis, filtering, partition, and selection operations. Red arrows mark the outlier filtering operations applied in the usage scenario on the heterogeneous dataset collection using the range sliders. The last characteristics (Real World) was used to partition the dataset collection, indicated by a dark red outline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

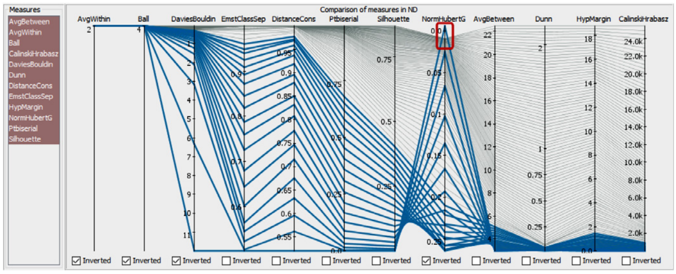


Fig. 9. Interactive re-ordering of measures according to the value domains. The selection of the 20 least separable datasets help to identify characteristics of the value domains and to validate the measure ordering.

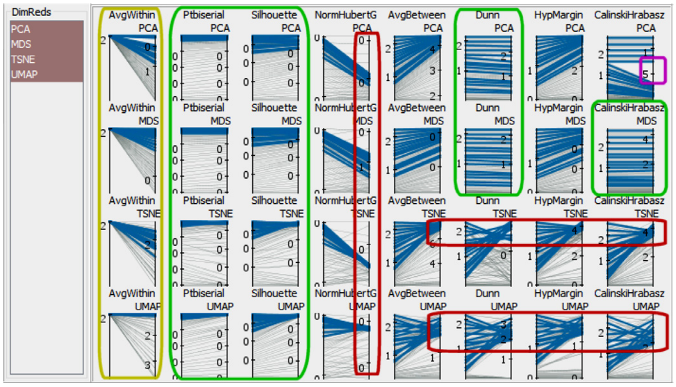


Fig. 10. Comparison of pairs of measure outputs applied on nD data and DR-reduced 2D data (T2). Eight out of 12 measures (aligned horizontally) and 4 DR methods (vertically) yield a grid with 32 slope charts for the detailed nD to 2D comparison. Many charts show interesting patterns, most expressive findings are hand-labeled with rectangles of different colors (see text for explanations).

AvgWithin which look like a fan: while the measures for nD are clumped together, the 2D results distribute quite well. In the second pattern (center green rectangle) two measures (Ptbiserial, and Silhouette) have consistent mappings from nD to 2D for all four DR methods, with many parallel lines and few rank changes in the slope charts. On the right, we identify more mappings of high consistency, also outlined in green. Normalized Hubert (red vertical rectangle) is among the measures with value distributions for 2D that are compressed by outliers. Particularly weak consistency can be observed for AvgBetween, Dunn, Hypothesis Margin, and Calinski-Hrabasz in combination with tSNE and UMAP (red horizontal rectangles). Especially tSNE seems to produce many inconsistencies, which may originate from its inability to preserve the global structure in the data. An anomaly can be identified for the PCA-based output of Calinski-Hrabasz (purple rectangle). We found out that in the PCA implementation (WEKA), the 2D PCA only returns one principal component, when the remaining variance is approaching zero.

6.1.6. Comparison of measures for DR-reduced datasets

Informed by the analysis of nD vs. 2D mapping consistency, we select the seven best-performing measures and continue with T3.

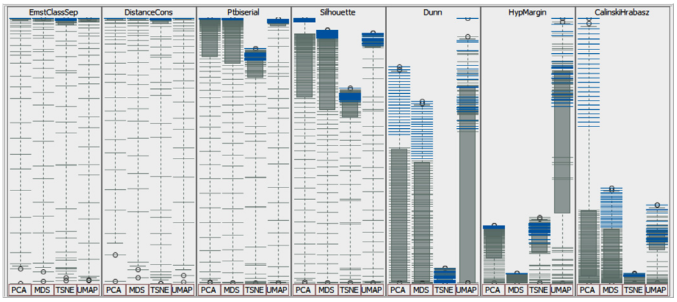


Fig. 11. Strip plots in combination with boxplots for the visual comparison of 7 measures and 4 DRs (PCA, MDS, TSNE, and UMAP), applied on the controlled dataset collection. The left four measures have similar behavior, but the three on the right show very individual measure distributions.

Fig. 11 shows the distribution of separability scores of the different DR-reduced datasets. We select the 24 most separable datasets, exactly those which stood out due to the PCA anomaly in the last analysis step. Interestingly, the left three measures form clusters with very similar measure results, whereas the results for the four on the right are all different in their own way. For tSNE, Dunn yields low separability for all datasets, whereas the other DRs lead to scores that are up to ten times higher. The problem here cannot be the 2D projections obtained by tSNE, because other measures like Distance Consistency yield high values for tSNE projected data. The index seems to be unreliable in this case. Hypothesis Margin has a similar problem when applied with MDS, whereas when applied with UMAP particularly high class separability is measured. Calinski-Hrabasz assigns barely separable classes for datasets projected with tSNE, yet particularly high values to PCA, especially for the 24 datasets where we identified the PCA anomaly earlier. Overall, we observe strongly varying behavior between class separation measures for different dimensionality-reduced datasets, showing how important the well-informed selection of a separation measure is for a given task.

6.2. Heterogeneous datasets

In the second scenario we maximize the heterogeneity of datasets in the collection, to assess effects that can be observed when measures and DRs are applied on a great variety of dataset characteristics.

6.2.1. Heterogeneous dataset collection

For the analysis of separation measures with heterogeneous datasets, we build upon the comprehensive dataset collection by Sedlmair et al. [6]. Sources for these datasets are UCI [62], umass [63], xmdv [64], VisuMap [65], sap [66], dataset syntheses [6], and datasets from colleagues [27,28,67].

We also add seven datasets from our controlled dataset collection [12] to increase the overall variety of datasets, and provide a link back to our analyses under controlled conditions. We would like to see if these well-understood datasets stand out in our analyses on the heterogeneous dataset collection.

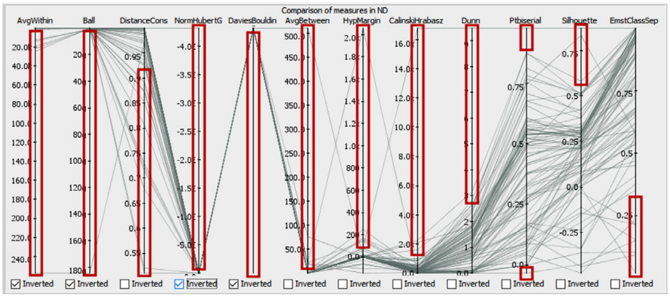


Fig. 12. Overview of the 12 measures for the heterogeneous dataset collection. Large parts of the value domains are only allocated by single outlier datasets, made visible in T1 view, highlighted with hand-labeled red rectangles. While this provides interesting indications about individual measure behaviors, we remove outlier datasets to shed light on the remaining datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

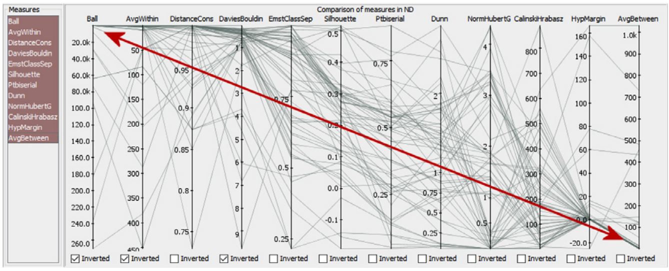


Fig. 13. Ordering of measures with an emphasis on their value distributions. Distributions skewed towards the top are assigned to the left, distributions skewed towards the bottom are assigned to the right. In the center are measures with more balanced distributions.

6.2.2. *Filter datasets*

First, we use ProSeCo to gain an overview of the distributions of measure outputs for all datasets (T1). Fig. 12 shows that most measures yield a highly skewed distribution of scores due to outliers, leading to large portions of unused space in the value domain (highlighted by red rectangles, manually annotated). Datasets that produce outliers in the separability scores might be special cases (worth a separate investigation), so we remove them to shed light on the remaining datasets.

We undertake an iterative filtering strategy that always removes the dataset which causes the most violations in several measures at the same time, reducing the number of datasets from 89 to 62 (details in supplemental materials). The result can be seen in Fig. 13.

6.2.3. *Order measures*

We modify the order of measures to enhance utility of the parallel coordinate plots, where distributions of measure outputs are aligned side-by-side. The two driving principles for the interactive ordering of the measure axis are the local criterion of aligning measures with similar output distributions next to each other, and the global criterion of aligning (groups of) measures so that their distributions show a continuum.

Fig. 13 shows the results of the interactive ordering process. The four measures on the far left have a skewed distribution where the majority of values is high. The next six measures have distributions with the majority of values at the center of the axis, or at least distributed equally across the axis. Finally, the right two measures have value distributions with many low and few high values. Some individual measure behaviors stand out: Hypothesis Margin maps almost all datasets to the same value and is thus not a very promising measure candidate. Similar observations can be made for Ball, Average Within, and Average Between; these concentration points are also the extreme values of the measures. Finally, Silhouette, Point Biserial and Dunn yield broad distributions which shows that they are sensitive to the different data characteristics.

6.2.4. *Filter characteristics ranges*

The interfaces for the analysis of data characteristics reveal additional information that can be used for dataset filtering. As shown with the red arrow annotations in the Fig. 8 T4 view, we use the range sliders to filter out datasets with extreme values. By filtering out 5 more datasets with extreme dataset characteristics, we achieve much more control of the dataset collection. The result of the filtering process is shown in the supplemental materials document.

6.2.5. *Comparepartitioned characteristic across measures*

As also shown in Fig. 8, we partition (T4) the dataset collection into two bins: real-world datasets (bottom) and synthetic datasets (top). The overall goal of partitioning dataset collections is the identification of interesting dependencies between measures and dataset characteristics.

Fig. 6 shows the distribution of measure values according to the partition (T5).

We can observe that most of the real-world datasets seem to be less separable than the synthetic datasets used, by noticing that the left bin has a lower distribution than the right one for for most measures. The effect is most visible with the measures at the center that have less skewed value distributions.

6.2.6. *Compare nD to DR-2D across measures*

The analysis of the consistency between measure results in nD and 2D (T2) can be seen in in Fig. 14. Overall, we analyzed all 48 (12 measures x 4 DR) slope charts and manually annotated the consistency of the mappings with the color-coded rectangles: green for good consistency, yellow for moderate, and red for weak. To assess consistency, we took outliers in both nD and 2D into account. In addition, we analyzed the number of parallel and horizontal lines, as well as the number of rank changes (line crossings). We noticed patterns in both vertical (measure-centred) and horizontal (DR-centered) orientation. Two findings stand out. First, a block of four measures at the center performs particularly well (Emst Class Separation, Silhouette, Point Biserial, and Dunn), and a fifth also shows acceptable consistency (Calinski-Hrabasz). The remaining measures on the left and right have rather weak performance. Second, t-SNE seems to be the least applicable DR method for this class separation task: none of the 12 measures yields good results when t-SNE is used (third row of charts).

6.2.7. *Select measures*

Reflecting on the previous steps in this scenario reveals that many measures struggled with this heterogeneous dataset collection. The dataset filtering process reveals many outlier-prone measures (Average Between, Average Within, Ball, Davies-Bouldin, Distance Consistency, and Hypothesis Margin). The measure ordering step revealed many measures with value domains considerably more skewed than the others (Ball, Average Within, Distance Consistency, Hypothesis Margin, and Average Between). With Normalized Hubert, we had problems with assessing the orientation, as the measure shows a rather unpredictable behavior for many datasets. Focusing on positive examples, the assessment of measures in this usage scenario reveals that Point Biserial was a particularly usable and useful measure, without major drawbacks and shortcomings. Alternative measures that also prove useful are Emst Class Separation, Silhouette, and Dunn.

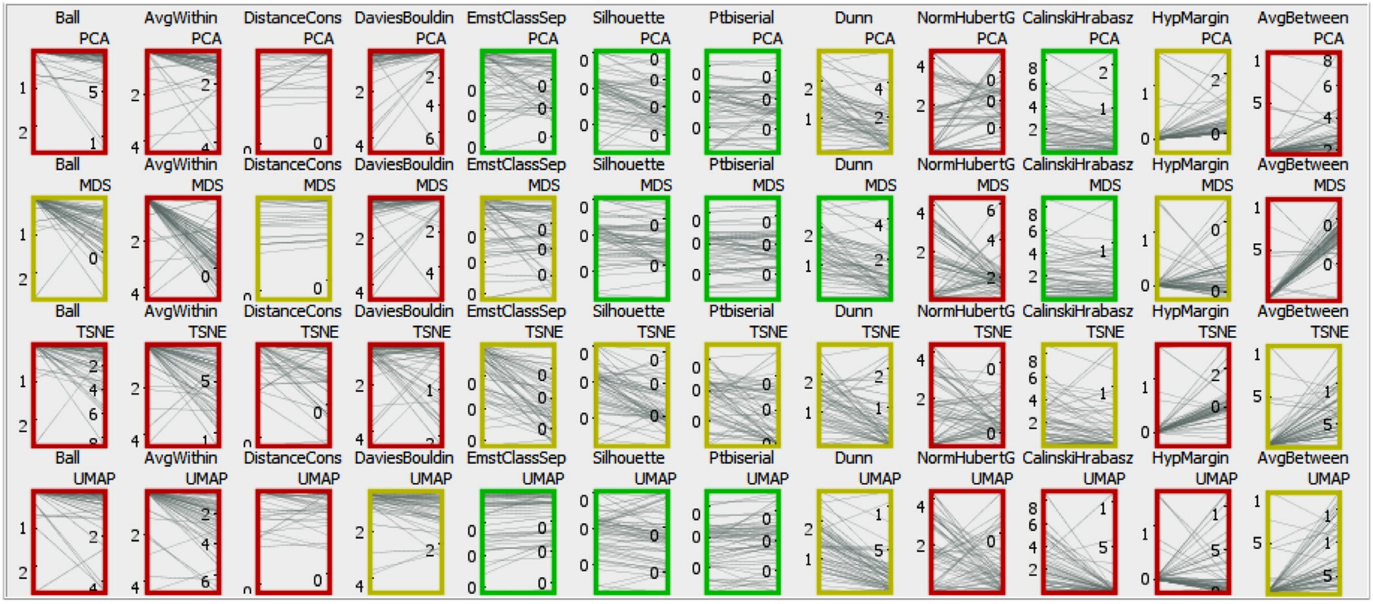


Fig. 14. Detailed analysis of the consistency of measures between nD and DR-2D datasets, manually annotated with rectangles colored green (good consistency), yellow (moderate), and red (weak). The block of measures at the center perform well, the t-SNE DR method (third row) performs poorly for class separation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Discussion and research opportunities

The usage scenarios showcase the power and flexibility of ProSeCo. In both, we observe surprising and unexpected results. For example, although the fully controlled dataset collection has a rather simple structure (two clusters of points with different distances), many separation measures do not yield consistent results and were distorted when projecting the data from nD to 2D. Furthermore, the value ranges of different separation measures are utilized very differently across the separation measures, making their direct comparison difficult. These observations are fully confirmed in the second usage scenario on the heterogeneous dataset collection containing numerous real-world datasets. Furthermore, we observe in this scenario that individual dataset characteristics can lead to unexpected, inconsistent, and problematic measure outputs in very different ways. Before ProSeCo, we could only speculate about the influences that datasets and DR methods have on separation measure outputs. We now provide a tool to make these influences visible and measurable through interactive investigation.

ProSeCo can support analysis by experts in multiple roles. One group of users are analysts who study high-dimensional data and need to make informed decisions about which selection measures they use for a given collection of datasets. Typically an analyst may favor measures which yield consistent results and a differentiated (balanced) distribution of scores. Another group of users are developers of new separation measures. With ProSeCo they can compare their measure to other measures and see how it behaves in cases where other measures show biases in their scores. A third group of users may be designers of new benchmark datasets who want to find datasets that challenge existing class separation measures (and thereby for example also clustering algorithms). Such a benchmark dataset may, for example, be designed in a way that existing measures yield inconsistent results on it. With ProSeCo this can be immediately be verified and made visible. Developers of DR methods may use ProSeCo to assess how strong the distortions are that are introduced by their DR method. ProSeCo enables them to visualize the distortions in terms of inconsistent separation measure estimates for hundreds of datasets at the same time. Thus, developers can immediately evaluate if a certain feature improves results on a sound and representative data basis, to significantly accelerate the evaluation and iterative improvement of DR methods. To provide such functionality we plan to integrate an importer for custom DR results into ProSeCo to enable in-depth analysis and comparison with other DR methods. Finally, ProSeCo can be a means for students who aim at learning and understanding these measures. A useful extension of ProSeCo would be to add automated analytics features to generate highlights for potentially interesting observations, outliers, and characteristic patterns, in the spirit of the manually added highlights in Fig. 14. This future work could also include the integration of previously proposed methods for automatically ordering parallel coordinate axes [68]

In this paper, we have shown results for two very different usage scenarios, one with a simple and strongly controlled dataset collection and one with a heterogeneous collection containing mostly real-world datasets. One straight-forward step for further research is to analyze other so far unseen collections of high-dimensional datasets to see which observations made in the two presented usage scenarios might be generalized. There are many additional avenues of future work, such as analyzing the parameters of DR methods or extending the analysis to more DR methods. Another direction is the analysis of inconsistencies in measure outputs between different datasets. From such an in-depth study, guidelines and recommendations could be derived for which types of datasets (i.e. characteristics) which separation measure are best suited and which separation measures yield the most robust and consistent separability scores.

ProSeCo provides a powerful interface. While the consistent color coding and the full linking of all views supports users in finding patterns and comparing data, we envision concepts for guiding users towards interesting measures, datasets, and DRs. Such concepts could be defined via heuristics or learned from user behavior iteratively over time. Such mechanisms would allow to model user preferences and support the analysis in cases where users are overwhelmed by the large number of datasets, DR methods or separation measures.

8. Conclusions

We presented ProSeCo, an interactive analysis tool to support the visual assessment of class separation measures. ProSeCo enables the analysis of three aspects: class separation measures, dimensionality reduction (DR) methods, and dataset characteristics. It supports the comparison of nD data to its 2D projections for up to 20 class separation measures and 4 DR methods, the investigation of 7 dataset characteristics, and the concurrent and comparative analysis of collections of up to 1000 datasets. We formulated six different tasks that can be performed with ProSeCo. In two usage scenarios, we demonstrated how ProSeCo enabled us to identify a series of measure characteristics, as well as commonalities and differences across measures. We analyzed dataset characteristics and drew connections between these characteristics and measure behaviors. Finally, we were able to assess effects of DR on different class separation measures.

In summary, ProSeCo helps expert analysts to better understand the interactions between separation measures, datasets, and DR methods, to gain a deeper understanding of separation measures and their selection for a given task at hand. The six tasks realized by ProSeCo are beneficial for several user groups in the ML and VIS communities: users selecting separability measures, developers of measures and DR methods, and even students who would like to further understand these topics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jürgen Bernard: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Marco Hutter:** Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization. **Matthias Zeppelzauer:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Michael Sedlmair:** Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Tamara Munzner:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition.

Acknowlgedgments

Funded by the Deutsche Forschungsgemeinschaft (DFG) – Project-ID 251654672 – TRR 161. Further funding was received from the Project IMREA funded by the Austrian Research Promotion Agency (FFG) under grant no. 880546 and by the Project DiPh-Spec funded by the Lower Austrian Research and Education Agency (NFB) under grant no. FTI18-005.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cag.2021.03.004.

References

[1] Bernard J, Hutter M, Zeppelzauer M, Sedlmair M, Munzner T. ProSeCo: Probing separation comparison; 2021. https://osf.io/epcf9/.
[2] Albuquerque G, Lowe T, Magnor M. Synthetic generation of high-dimensional datasets. IEEE Trans Vis Comput Gr (TVCG) 2011a;17(12):2317–24.
[3] Rauber PE, Falcão AX, Telea AC. Projections as visual aids for classification system design. Inf Vis 2018;17(4):282–305. doi:10.1177/1473871617713337.
[4] Arbelaitz O, Gurrutxaga I, Muguerza J, PéRez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit 2013;46(1):243–56.
[5] Bertini E, Tatu A, Keim D. Quality metrics in high-dimensional data visualization: an overview and systematization. IEEE Trans Trans Vis Comput Gr (TVCG) 2011;17(12):2203–12.
[6] Sedlmair M, Tatu A, Munzner T, Tory M. A taxonomy of visual cluster separation factors. In: Computer graphics forum, 31. Wiley Online Library; 2012. p. 1335–44.
[7] Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell (TPAMI) 2002;24(12):1650–4.
[8] Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika 1985;50(2):159–79.
[9] Cutura R, Aupetit M, Fekete J-D, Sedlmair M. Comparing and exploring high--dimensional data with dimensionality reduction algorithms and matrix visualizations. In: Advanced visual interfaces (AVI); 2020.
[10] Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful. In: Proceedings of the conference on database theory (ICDT). Springer; 1999. p. 217–35.
[11] Sacha D, Zhang L, Sedlmair M, Lee JA, Peltonen J, Weiskopf D, et al. Visual interaction with dimensionality reduction: a structured literature analysis. IEEE Trans Trans Vis Comput Gr (TVCG) 2016;23(1):241–50.
[12] Bernard J, Hutter M, Zeppelzauer M, Sedlmair M, Munzner T. SepEx: visual analysis of class separation measures. In: Proceedings of the EuroVis workshop on visual analytics (EuroVA). The Eurographics Association; 2020.
[13] Deborah LJ, Baskaran R, Kannan A. A survey on internal validity measure for cluster validation. Comput Sci Eng Survey 2010;1(2):85–102.
[14] Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybern 1974;4(1):95–104.
[15] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[16] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell (TMAPI) 1979(2):224–7.
[17] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat-Theory Methods 1974;3(1):1–27.
[18] Aupetit M, Sedlmair M. SepMe: 2002 new visual separation measures. In: Proceedings of the IEEE symposium pacific visualization (PacificVis); 2016. p. 1–8.
[19] Sánchez-Monedero J, Gutiérrez PA, Pérez-Ortiz M, Hervás-Martínez C. An n-spheres based synthetic data generator for supervised classification. In: Artificial Neural Networks. Springer; 2013. p. 613–21.
[20] Ramirez-Loaiza ME, Sharma M, Kumar G, Bilgic M. Active learning: an empirical study of common baselines. Data Min Knowl Discov 2017;31(2):287–313.
[21] Hinterreiter A, Ruch P, Stitz H, Ennemoser M, Bernard J, Strobelt H, et al. ConfusionFlow: A model-agnostic visualization for temporal analysis of classifier confusion. IEEE Trans Trans Vis Comput Gr (TVCG) 2020:1. doi:10.1109/TVCG.2020.3012063.
[22] Ho TK, Basu M. Complexity measures of supervised classification problems. IEEE Trans Pattern Anal Mach Intell (TPAMI) 2002;24(3):289–300.
[23] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45(4):427–37.
[24] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst 2001;17(2-3):107–45.
[25] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660–74.
[26] Wilkinson L, Anand A, Grossman R. Graph-theoretic scagnostics. In: Proceedings of the IEEE symposium information visualization (InfoVis); 2005. p. 157–64.
[27] Sips M, Neubert B, Lewis JP, Hanrahan P. Selecting good views of high-dimensional data using class consistency. Comput Graph Forum 2009;28(3):831–8. doi:10.1111/j.1467-8659.2009.01467.x.
[28] Tatu A, Albuquerque G, Eisemann M, Schneidewind J, Theisel H, Magnork M, et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In: Visual analytics science and technology (VAST). IEEE; 2009. p. 59–66.
[29] Albuquerque G, Eisemann M, Magnor M. Perception-based visual quality measures. In: Proceedings of the IEEE conference visual analytics science and technology (VAST); 2011b. p. 13–20.
[30] Espadoto M, Martins RM, Kerren A, Hirata NST, Telea AC. Towards a quantitative survey of dimension reduction techniques. IEEE Trans Trans Vis Comput Gr (TVCG) 2019a. doi:10.1109/TVCG.2019.2944182.
[31] Wang Y, Feng K, Chu X, Zhang J, Fu C-W, Sedlmair M, et al. A perception-driven approach to supervised dimensionality reduction for visualization. IEEE Trans Trans Vis Comput Gr (TVCG) 2017;24(5):1828–40.
[32] Cacoullos T. Discriminant analysis and applications. Academic Press; 2014.
[33] Wang Y, Chen X, Ge T, Bao C, Sedlmair M, Fu C-W, et al. Optimizing color assignment for perception of class separability in multiclass scatterplots. IEEE Trans Visualization & Computer Graphics (TVCG) 2018;25(1):820–9.

[34] Lu K, Feng M, Chen X, Sedlmair M, Deussen O, Lischinski D, et al. Palettailor: Discriminable colorization for categorical data. IEEE Trans Trans Vis Comput Gr (TVCG) 2020. To appear

[35] Legány C, Juhász S, Babos A. Cluster validity measurement techniques. In: Artificial intelligence, knowledge engineering and data bases (AIKED); 2006. p. 388–93.

[36] Rendón E, Abundez I, Arizmendi A, Quiroz EM. Internal versus external cluster validation indexes. Comput Commun 2011;5(1):27–34.

[37] Guerra L, Robles V, Bielza C, Larrañaga P. A comparison of clustering quality indices using outliers and noise. Intell Data Anal 2012;16(4):703–15.

[38] Bernard J, von Landesberger T, Bremm S, Schreck T. Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In: Proceedings of the SPIE conference on visualization and data analysis; 2011. p. 78680N.1–78680N.12. doi:10.1117/12.872545.

[39] Bernard J, Dobermann E, Sedlmair M, Fellner DW. Combining Cluster and Outlier Analysis with Visual Analytics. In: EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association; 2017. ISBN 978-3-03868-042-0.

[40] Sedlmair M, Aupetit M. Data-driven evaluation of visual quality measures. In: Computer Graphics Forum, 34. Wiley Online Library; 2015. p. 201–10.

[41] Ingram S, Munzner T, Irvine V, Tory M, Bergner S, Möller T. Dimstiller: workflows for dimensional analysis and reduction. In: Proceedings of the IEEE conference visual analytics science and technology (VAST); 2010. p. 3–10.

[42] Cutura R, Holzer S, Aupetit M, Sedlmair M. Viscoder: A tool for visually comparing dimensionality reduction algorithms. In: Proceedings of the European symposium on artificial neural networks (ESANN); 2018.

[43] Munzner T. Visualization analysis & design. CRC Press; 2014.

[44] 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, 34; 2012.

[45] Ball G, Hall D. A novel method of data analysis and pattern classification. Stanford Research Institute; 1965. (NTIS No AD 699616).

[46] Motta R, Minghim R, de Andrade Lopes A, Oliveira MCF. Graph-based measures to assist user assessment of multidimensional projections. Neurocomputing 2015;150:583–98.

[47] Gilad-Bachrach R, Navot A, Tishby N. Margin based feature selection-theory and algorithms. In: Proceedings of the conference on machine learning (ICML); 2004. p. 43.

[48] Hubert L, Schultz J. Quadratic assignment as a general data analysis strategy. Br J Math Stat Psychol 1976;29(2):190–241.

[49] Milligan GW. A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika 1981;46(2):187–99.

[50] Espadoto M, Martins RM, Kerren A, Hirata NS, Telea AC. Towards a quantitative survey of dimension reduction techniques. IEEE Trans Vis Comput Gr 2019b.

[51] Jolliffe I. Principal component analysis. Springer Berlin Heidelberg; 2011. p. 1094–6.

[52] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1–27.

[53] Maaten Lv d, Hinton G. Visualizing data using t-SNE. J Mach Learn Res (JMLR) 2008;9(Nov):2579–605.

[54] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction; 2020. arXiv: 1802.03426[stat.ML]

[55] Hu K, Gaikwad S, Hulsebos M, Bakker MA, Zgraggen E, Hidalgo C, et al. Viznet: Towards a large-scale visualization learning and benchmarking repository. In: Proceedings of the CHI conference on human factors in computing systems; 2019. p. 1–12.

[56] Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Exp Soc Psychol 2013;49(4):764–6.

[57] Wickham H, Stryjewski L. 40 years of boxplots. Tech. Rep. hadconz; 2012.

[58] Bernard J, Steiger M, Widmer S, Lücke-Tieke H, May T, Kohlhammer J. Visual-interactive exploration of interesting multivariate relations in mixed research data sets. Comput Gr Forum (CGF) 2014;33(3):291–300. doi:10.1111/cgf.12385.

[59] Mühlbacher T, Piringer H. A partition-based framework for building and validating regression models. IEEE Trans Vis Comput Gr (TVCG) 2013;19(12):1962–71.

[60] McGrath JE. Methodology matters: Doing research in the behavioral and social sciences. In: Readings in human–computer interaction. Elsevier; 1995. p. 152–69.

[61] Sedlmair M, Munzner T, Tory M. Empirical guidance on scatterplot and dimension reduction technique choices. IEEE Trans Trans Vis Comput Gr (TVCG) 2013;19(12):2634–43.

[62] Frank A, Asuncion A. University of California Irvine (UCI) Machine Learning Repository; 2010. http://archive.ics.uci.edu/ml.

[63] University of Massachusetts. Statistical data and software help; 2011. http://www.umass.edu/statdata/statdata/, last accessed 11/2020.

[64] Ward MO. Xmdv data repository; 2011. http://davis.wpi.edu/xmdv/datasets.html, last accessed 11/2020.

[65] VisuMap Technologies Inc. VisuMap Data Repository; 2011 http://www.visumap.net/.

[66] SAP. HANA; 2010 http://www.sap.com/hana/.

[67] Holt C, Bradford M. Evaluating benchmarks of population status for Pacific salmon. North Am J Fisher Manag 2011;31(2):363–78.

[68] Heinrich J, Weiskopf D. State of the Art of Parallel Coordinates. In: Eurographics 2013 - State of the Art Reports. The Eurographics Association; 2013.