

Visual Quality Assessment of Subspace Clusterings

Michael Hund¹, Ines Färber², Michael Behrisch¹,
Andrada Tatu^{1*}, Tobias Schreck³, Daniel A. Keim¹, Thomas Seidl⁴

¹ University of Konstanz, Germany {lastname@dbvis.inf.uni-konstanz.de}

² RWTH Aachen University, Germany {faerber@informatik.rwth-aachen.de}

³ Graz University of Technology, Austria {tobias.schreck@cgv.tugraz.at}

⁴ Ludwig-Maximilians-University, Munich, Germany {seidl@dbvis.lmu.de}

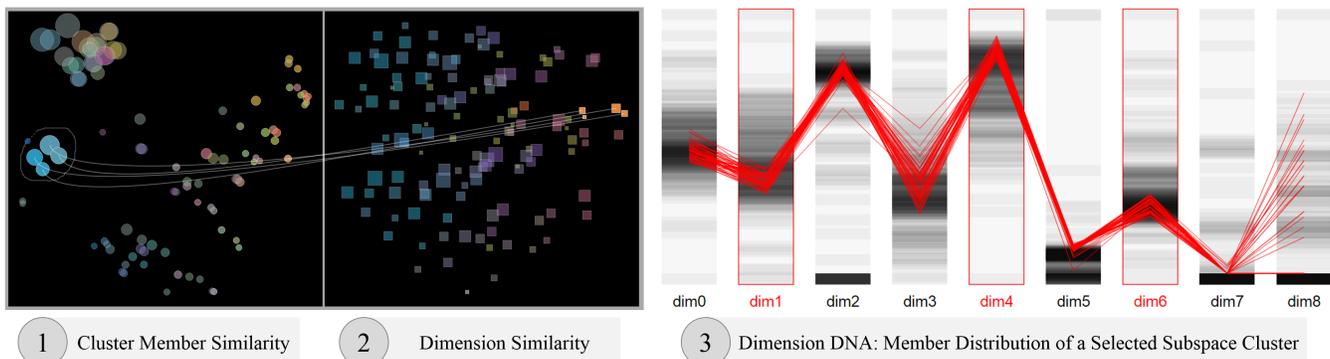


Figure 1: A comparative overview of 132 detected subspace clusters generated by the CLIQUE [2] algorithm: The two inter-linked MDS projections in the SubEval analysis framework show simultaneously the cluster member- (1) and dimension similarity (2) of subspace clusters. While the cluster member similarity view focuses on the object-wise similarity of clusters, the dimension similarity view highlights similarity aspects w.r.t. their common dimensions. The coloring encodes the redundancy of clusters in the opposite projection. Both views together allow to derive insights about the redundancy of subspace clusters and the relationships between subspaces and cluster members. The DimensionDNA view (3) shows the member distribution of a selected subspace cluster in comparison to the data distribution of the whole dataset.

ABSTRACT

The quality assessment of results of clustering algorithms is challenging as different cluster methodologies lead to different cluster characteristics and topologies. A further complication is that in high-dimensional data, *subspace clustering* adds to the complexity by detecting clusters in multiple different lower-dimensional projections. The quality assessment for (subspace) clustering is especially difficult if no benchmark data is available to compare the clustering results.

In this research paper, we present SUBEVAL, a novel subspace evaluation framework, which provides visual support for comparing quality criteria of subspace clusterings. We identify important aspects for evaluation of subspace clustering results and show how our system helps to derive quality assessments. SUBEVAL allows assessing subspace cluster quality at three different granularity levels: (1) A global overview of similarity of clusters and estimated redundancy in cluster members and subspace dimensions. (2) A view of

a selection of multiple clusters supports in-depth analysis of object distributions and potential cluster overlap. (3) The detail analysis of characteristics of individual clusters helps to understand the (non-)validity of a cluster. We demonstrate the usefulness of SUBEVAL in two case studies focusing on the targeted algorithm- and domain scientists and show how the generated insights lead to a justified selection of an appropriate clustering algorithm and an improved parameter setting. Likewise, SUBEVAL can be used for the understanding and improvement of newly developed subspace clustering algorithms. SUBEVAL is part of SUBVA, a novel open-source web-based framework for the visual analysis of different subspace analysis techniques.

CCS Concepts

•Human-centered computing → Visualization design and evaluation methods;

Keywords

Subspace Clustering; Evaluation; Comparative Analysis; Visualization; Information Visualization; Visual Analysis

*Former member.

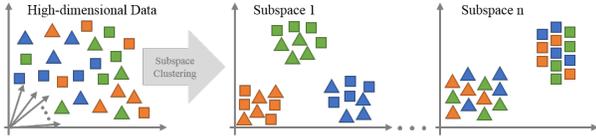


Figure 2: Subspace clustering in high-dimensional data: the same objects are grouped differently in different combinations of dimensions (=subspaces).

1. INTRODUCTION

In data analysis the selection and parametrization of clustering algorithms is usually a trial-and-error task requiring appropriate methods and analyst experience to assess the quality of the results. Furthermore, the selection of an appropriate algorithm design has a direct impact on the expected results. For example, k -Means-type clustering will likely favor voronoi-shape spaced partitions, while a density-based clustering (e.g., DBSCAN [12]) usually results in arbitrarily shaped clusters. The parameter setting, the underlying data topology and -distribution usually influence the clustering results, too. For varying applications, different cluster characteristics can be of interest for a user. Therefore, there is a need for efficient and effective evaluation methods to reliably assess the usefulness of a clustering result.

In high-dimensional data, clustering computation is influenced by the so-called *curse of dimensionality*. Noise, correlated, irrelevant, and conflicting dimension may detriment meaningful similarity computation for the input data [7]. Experiments show that the application of full-space clustering, i.e., a clustering that considers all dimensions, is often not effective in datasets with a large number of dimensions ($\geq 10 - 15$ dimensions) [20]. To overcome these problems, the notion of subspaces must be taken into consideration. *Subspace clustering* [21] aims to detect clusters in different, lower-dimensional projections of the original data space, as illustrated in Figure 2. The challenge is to simultaneously select meaningful subsets of objects and subsets of dimensions (=subspaces). In existing subspace cluster algorithms, the number of reported clusters is typically large and may contain substantial redundancy w.r.t. clusters and/or subspaces.

Quality assessment of subspace clustering shows to be particularly challenging as, besides the more complex result interpretation, evaluation methods for full-space clustering are not directly applicable. Generally, (subspace) clustering strives to group a given set of objects into clusters, such that objects within clusters are similar (*cluster compactness*), while objects of different clusters are dissimilar (*cluster separability*). This abstract goal leads to various *different*, yet valid and useful, cluster definitions [17]. Due to these diverging definitions it is challenging, if not impossible, to design or commonly agree on a single evaluation measure for (subspace) clustering results.

It is therefore desirable to have a unified approach for an objective quality assessment of (subspace) clustering based on different clustering methodologies, the data distribution and -topology and variety of application- and domain-dependent quality criteria. We tackle this multi-faceted analysis problem with a visual analysis process by which the computer’s processing power and the human’s skills in interpretation and association can be effectively combined. Numeric performance measures alone are not effective enough to give an all-embracing picture, as they are typically abstract and

heuristic in nature, and defined in an application-independent way. Several application fields can benefit from such a user-supported evaluation approach: (1) selection of an appropriate clustering algorithm, (2) selection of appropriate parameter settings and (3) the design of new data mining algorithms, where algorithm scientists continuously evaluate the algorithm’s results against original assumptions.

In this paper, we tackle the problem of visually evaluating the quality of one subspace clustering result. We present a novel open-source evaluation framework, called SUBEVAL. It enhances standard evaluation approaches with effective visualizations to support the assessment of (subspace) clustering algorithms. Our contributions are as follows: (1) We present a summary of subspace cluster evaluation approaches, point to their specific foci and contrast their benefits and disadvantages. (2) We systematically structure the major evaluation criteria for subspace clustering results. (3) We discuss design- and user interaction requirements for visualizations to provide deep insights into the different quality criteria and (4) make the open-source tool SUBEVAL available.

Compared to existing subspace visualization techniques like CODA [15] or CLUSTNAILS [29], focusing on the knowledge extraction of subspace clusters, SUBEVAL targets primarily the quality aspect of a clustering result. Our novel framework uses two interlinked MDS plots to simultaneously represent cluster member and subspace similarity and provides different tools for in-depth analysis of different quality criteria.

2. BACKGROUND

This section introduces definitions, terminology, concepts and related work that we rely upon to describe our approach.

2.1 Definitions and Terminology

Data record/object are used synonymously for a data instance of the dataset, i.e., $r_i \in \mathcal{R}$. A **subspace** s_l is defined as a subset of dimensions of the dataset: $s_l = \{d_i, \dots, d_j\} \in \mathcal{D}$.

A **cluster** $c_j \subseteq \mathcal{R}$ contains a set of objects which are similar to each other based on a similarity function. A **clustering** result $\mathcal{C} = \{c_1, \dots, c_n\}$ comprises the set of all clusters detected by an algorithm.

Crucial for the understanding of this paper is to differentiate between **full-space** and **subspace clustering**. Full-space clustering considers all dimensions (\mathcal{D}) for the similarity computation of its cluster members (e.g., k -Means).

A **subspace cluster** $sc_i = (s_i, c_i)$ considers only the subspace s_i for the similarity computation of the cluster members of c_i . As shown in Figure 2, a subspace clustering $\mathcal{SC} = \{sc_1, \dots, sc_n\}$ consists of multiple clusters which are defined in specific subspaces. Based on the algorithm, cluster members and/or dimensions of any two clusters sc_i and sc_j may overlap, i.e., $|c_i \cap c_j| \geq 0$ and $|s_i \cap s_j| \geq 0$. The number of detected subspace clusters is typically large. For a dataset with d dimensions, there are $2^d - 1$ possible subspaces of which many may contain useful, but highly similar/redundant clusters. Same as for full-space clustering, there is a variety of different methodologies to compute useful subspace clusters [21]. However, there is no formal definition of a valid and useful (subspace) clustering result which has been accepted thoroughly by the community.

2.2 Visualization of (Subspace) Clusterings

Several techniques exist to visualize (subspace) clusters and allow users to extract semantics of the cluster structures.

The visual analysis and comparison of full-space clustering is a problem in high-dimensional data. Standard techniques like Parallel Coordinates, Dimension Stacking or Projection Techniques are applicable as a baseline [32]. Multidimensional glyphs can help to represent clusters in a 2D layout to support cluster comparison [31]. In [10], a Treemap-based glyph was designed to represent clusters and associated quality measures for visual exploration. In previous work, we considered a comparisons of hierarchical clusterings in a Dendrogram representation [9], and a comparison of Self-Organizing Map clusterings using a color-coding [8].

Visual comparison of subspace clusters is an even more difficult problem. In addition to full-space cluster visualization, also set-oriented information pertaining to subspace dimensions and possibly, multi-set membership of elements in clusters needs to be reflected. The first approaches in subspace cluster comparison is VISA [3] which visualizes subspace clusters in a MDS projection based on their cluster member similarity. Further approaches to visually extract knowledge of detected subspace clusters are CLUSTNAILS [29], SUBVIS [20], and an approach by TATU et al. [28].

Visual redundancy analysis of subspace clusters is presented for example by CODA [15] and MCEXPLORER [16]. Both, however, comprise only a single aspect, either dimension or cluster member redundancy. As discussed by TATU et al. [28] clusters are only true redundant if the cluster member and the subspace topology are similar.

While the existing visualizations focus mainly on the extraction of knowledge for domain experts, SUBEVAL changes the point of view and targets the depiction of quality criteria of subspace clusterings, such as non-redundancy, compactness, and the dimensionality of clusters.

2.3 Evaluation of Full-Space Clustering

In the following we summarize classical approaches for the evaluation of full-space clustering. We carefully investigate the advantages and drawbacks of the presented methods and highlight why visual interactive approaches are beneficial in many scenarios. As subspace clustering is a special instance of full-space clustering, the same challenges apply.

Evaluation Based on Internal Criteria.

Internal quality measures evaluate clusters or clustering results purely by their characteristics, e.g., cluster density. The literature provides a large variety of commonly used measures [22], each treating the cluster characteristics differently but usually focusing on compactness and separability. Internal measures, designed for evaluating full-space clustering, assume a single instance-to-cluster assignment and have not yet been adapted for (partially) overlapping clusters, as in subspace clustering. The criticism of this evaluation method, which does not qualify it for general performance quantification, is its subjectivity. Each measure usually favors a more particular cluster notion (e.g., RMSSTD [22] favors voronoi-shaped clusters). For each quality measure one could design an algorithm to optimize the clustering w.r.t. this particular quality measure, making comparisons to other approaches inappropriate.

External Evaluation Based on Ground Truth.

External quality measures compare the topology of a clustering result with a given ground truth clustering. Although benchmark evaluation is well accepted in the community

and allows an easy comparison of different algorithms and parameter settings, the criticism to this evaluation method is manifold: The main problem of external quality measures lies in the use of a ground truth clustering itself. In most (real-world) applications and datasets with unknown data a ground truth is not available. Even if a ground truth labeling exists, it is either synthetically generated with specific clustering characteristics (c.f. criticism in Section 2.3), or it is providing a classification labeling instead of a clustering label [13]. Consequently, an algorithm, which does not correctly retrieve an already known categorization, cannot generally be regarded as bad result, as the fundamental task of clustering is to find previously unknown patterns.

Evaluation by Domain Experts.

The actual usefulness of a clustering for a certain application domain can only be assessed with a careful analysis by a domain expert. However, in many (higher-dimensional) real-world applications, the cluster result complexity is overwhelming even for domain experts. Accordingly, domain expert-based evaluation is not suited for a comparison of different clusterings, since (1) a domain expert cannot evaluate a large number of algorithms and/or parameter setting combinations, and (2) the evaluation depends on the expert and the application and does therefore not result in quantitative performance scores.

2.4 Evaluation of Subspace Clustering

In the following, we discuss current approaches for the evaluation of subspace clusterings and highlight why novel human-supported evaluation methods, such as provided by SUBEVAL, are required for a valid quality analysis.

External Evaluation Measures.

The most commonly used method to assess the quality of a subspace clustering algorithm are external quality measures. As discussed above, the synthetically created ground-truth clusters are typically generated with particular clustering characteristics, and, for subspace clustering also with subspace characteristics. For real-world data the ground truth is not very expressive [13] and potentially varies depending on the used measure or data set [14, 24].

Internal Evaluation Measures.

The internal measures used for traditional (full-space) clustering are not applicable to subspace clustering results as (1) the existing methods do not allow for overlapping cluster members, (2) clusters need to be evaluated in their respective subspace only, i.e., it is not valid to assess the separability of two clusters which exist in different subspaces.

Domain Experts.

Often authors justify a new subspace clustering approach by exemplarily discussing the semantic interpretation of selected clusters, i.e., evaluation by domain scientists, which seems to be the only choice for some real-world data, e.g., [20]. Quite a few visualization techniques exist to support domain experts in the knowledge extraction of subspace clusters (c.f. Section 2.2). However, in subspace clustering we have to tackle three major challenges: (1) the subspace concept is complex for most domain experts, especially for non-computer-scientists, (2) the large result space and the redundancy makes it often practically unfeasible to investi-

gate all detected clusters and retrieve the most relevant ones, and (3) it is almost impossible to manually decide whether all relevant clusters have been detected or not.

Summarizing, existing quality measures for subspace clustering comprise the evaluation by external measures and/or a careful investigation by domain experts. Although both approaches have their advantages and disadvantages, they are valid and accepted in the community. Besides these techniques, we need novel methods which do not rely on ground-truth data and/or domain experts, but rather complement existing evaluation approaches. Therefore, our aim is to visualize the quality of a clustering for different clustering definitions. Furthermore, our approach supports the user in interpreting given subspace clustering result in terms of object groups and dimension sets, hence supports interactive algorithm parameter setting.

3. VISUAL QUALITY ASSESSMENT

In the following we summarize the most important quality criteria indicating a useful and appropriate *subspace* clustering result. Our quality criteria (*C1-C3*) are compiled from a literature review on objective functions for subspace clustering algorithms. Our coverage is not exhaustive, but targeted towards the major quality “understandings” in this field. For many applications, not all aspects need to be full-filled.

3.1 Quality Criteria for Subspace Clusterings

Non-Redundancy Criteria (C1).

One –if not the major– challenge in subspace clustering, is redundancy. It negatively influences a knowledge extraction due to highly similar, but not identical cluster results.

C1.1 Dimension Non-Redundancy. A useful subspace clustering algorithm emphasizes distinctive dimension/membership characteristics and avoids subspace clusters with highly similar subsets of dimensions.

C1.2 Cluster Member Non-Redundancy. A useful subspace clustering result focuses on important global groupings, avoiding clusters with many similar cluster members.

As elaborated in [28], subspace clusters are only *true redundant*, if they share most of their dimensions *and* most of their cluster members. Therefore, dimension- and cluster member redundancy have to be analyzed in conjunction.

C1.3 No Cluster-Splitup in Subspaces. Similar clusters occurring in different, non-redundant subspaces should be avoided. Generally, cluster-splitups cannot be considered redundant, as each cluster may contain new information. Yet, it provides reasons to suspect that the cluster members form a common cluster in a single, higher-dimensional subspace.

Object and Dimension Coverage Criteria (C2).

We define *object coverage* as the proportion of objects and *dimension coverage* as the proportion of dimensions of the datasets which are part of at least one subspace cluster. A high coverage of both objects and dimensions helps to understand the global patterns in the data.

C2.1 Object Coverage. To reason about all data objects, a useful subspace clustering algorithm extracts –not mandatorily a full– but obligatory high object coverage.

C2.2 Dimension Coverage. To reason about all dimension characteristics, a useful subspace clustering algorithm covers each dimensions in at least one subspace cluster.

Clustering Characteristics Criteria (C3).

Cluster characteristics are related to internal cluster evaluation measures. Although the following aspects are not summarized into a common measure for subspace clustering, most algorithms try to optimize the following properties:

C3.1 Cluster Compactness. Objects belonging to a cluster need to be similar in all dimensions of their respective subspace. Non-compact clusters represent dependencies between the cluster members which are not very strong.

C3.2 Cluster Separability. A useful algorithm assigns similar objects to the same cluster. Objects belonging to different clusters in the same subspace need to be dissimilar. A separability definition of clusters existing in different subspaces does not exist yet.

C3.3 High/Low Dimensionality. A high and a low dimensionality of a cluster can both be considered useful in different applications. While a high dimensionality is often interpreted as more descriptiveness, we argue that a low dimensional cluster can also be of interest, especially if a higher dimensional subspace contains the same cluster structures. That means, fewer dimensions correspond to lower cluster complexity. However, clusters with a very low dimensionality ($\sim 1-3$ dimensions) are typically of no interest since no deeper knowledge can be extracted.

C3.4 High/Low Cluster Size. While most subspace clustering algorithms favor clusters with many members, we believe that in some applications clusters with a small cluster size are important, esp. when combined with *C3.1* and *C3.2*. Possible use case: a dataset contains many obvious structures, while smaller clusters may contain unexpected patterns.

3.2 Visual Design- and Interaction Requirements for Subspace Cluster Evaluation

In the following we summarize design requirements to assess the quality criteria as categorized above. In Section 4 we showcase one possible instantiation of the design requirements in our SUBEVAL framework.

Cluster vs. Clustering. Crucial for the design of an evaluation system is to distinguish between the evaluation of a single *cluster* and the evaluation of a *clustering* result. For a single cluster, the different cluster characteristics (*C3*) are of interest, independent of a potential redundancy (*C1*) or coverage (*C2*) aspect. Likewise, for a clustering result the overall quality information, such as redundancy (*C1*) and coverage (*C2*) is important, i.e., a high-quality result can still contain a few clusters with e.g., low compactness (*C3.1*).

Reasoning for a Good/Bad Quality. Another important aspect is to distinguish between a *cluster/clustering quality* and *explanations/reasons* for a good/bad quality. The first aspect primarily states whether a clustering is useful or not, while the second one requires a more fine-grained level for an in-depth understanding.

Interactive Visualizations. For many of the presented quality criteria it is not mandatory to develop complex visualizations. Simple visual encodings and well-established visualizations, such as bar- or line charts, allow to extract quickly useful meta-information (e.g., the redundancy of dimensions in subspaces or the number of not clustered data records). We show examples in Figures 5 and 6. Even simple visualizations become powerful analysis tools if interactivity and faceted-browsing is applied, i.e., an analyst interactively selects all subspace clusters containing a frequently occurring dimension and gets details on-demand, such as data

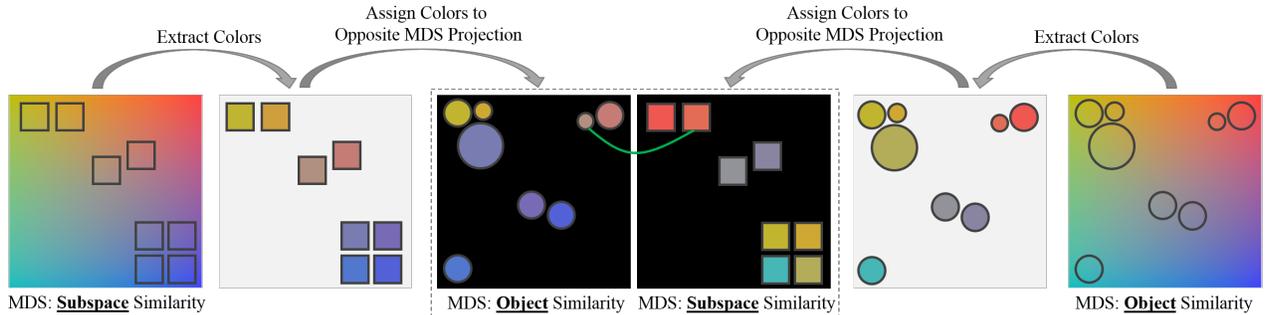


Figure 3: Schema for the two interlinked MDS projections: An MDS projection is computed for both, the subspace- and object similarity of the subspace clusters using an appropriate distance function. Afterwards, the projection is mapped on top of a perceptual linear 2D colormap where similar color correspond to a nearby location in MDS projection (similar objects). Finally, the colors of the subspace clusters of the *object similarity* projection are assigned to the clusters in the *subspace similarity* projection and vice versa. Interpretation: Nearby subspace clusters in a MDS projection with the same color are similar in both, the object and subspace space; nearby clusters with different colors in the object similarity projection are only in their cluster members, but not in the subspace.

distribution and commonalities of the selected clusters. This technique is known as *linking-and-brushing* [5].

Multi-Granularity Analysis. To get detailed information of the quality of subspace clustering result at different granularity levels, a multi-level analysis from overview to detail is required (see also the visual information seeking mantra by SHNEIDERMAN [26]: *Overview first, zoom and filter, then details-on-demand*). In the following, we describe a potential workflow with three granularity levels (*L1-L3*):

L1 Overview. The user needs to quickly develop an overview of the redundancy aspect (*C1*) for all detected clusters to decide whether a result is generally useful or not. Quality must be assessed separately, but also related in two spaces: cluster member- and dimension space. Redundancy is highly correlated with similarity as many highly similar cluster imply a notion of redundancy. Therefore, an appropriate visualization must be able to depict (relative) similarity between data objects, as well as between dimension combinations. One possible visualization technique to fulfill these visual properties is Multi-dimensional Scaling (MDS) [11], as depicted in Figure 1. MDS approximates the high-dimensional distances in a low (2D) dimensional space, thus making it suitable for depicting redundancy aspects (*C1*). Set-oriented distance functions such as the Jaccard Index or the Overlap Coefficient are a possible mean to intuitively compute the similarity between two clusters or subspaces:

$$Jaccard_Similarity(c_i, c_j) = 1 - \frac{|c_i \cap c_j|}{|c_i \cup c_j|}$$

A similarity value of 0 refers to two completely identical clusters. Likewise, the similarity can be computed between two subspaces. Based on the similarity notion of a specific application, a different distance function can be applied. Other subspace cluster properties, such as the cluster size or compactness, can be encoded with additional visual variables (e.g., color or size) into the points of the projection or by bar charts as presented, e.g., in Figures 5 and 6.

L2 Cluster comparison. At the cluster comparison level, the user needs to validate a potential object- and/or dimension redundancy identified in (*L1*). The analyst will also have to examine the coverage of the cluster members and di-

mensions, and particularly compare the coverage of multiple clusters. As one potential solution we propose one MDS projection per subspace cluster, illustrating the object similarity by location in the MDS projection and highlight the cluster members accordingly as further described in Section 4.2. Another approach to analyze common members/dimensions in different clusters are Parallel Set visualization [6].

L3 Data Instance. At the last analysis level, the user needs to investigate the properties of a single selected cluster. Only at this fine-grained detail level the analyst will understand why specific objects are clustered within a subspace, and, more importantly, to find potential reasons why a clustering fails to identify a valid object to cluster relationship. One possible approach to analyze the data distribution of high-dimensional data are Parallel Coordinates [18], which show the distribution of multiple data objects among a large set of dimensions. It might be useful to combine the Parallel Coordinates with a box plot or another density measure in order to compare the data objects with the underlying data distribution of the dataset. An example for such an enhanced parallel coordinates plot can be found in Figure 1.

4. SUBEVAL: INTERACTIVE EVALUATION OF SUBSPACE CLUSTERINGS

In the following section, we introduce SUBEVAL which is one instantiation of the previously described multi-granularity analysis. The overview level (*L1*) uses two inter-linked MDS projections to simultaneously analyze cluster member- and dimension redundancy (Section 4.1). Section 4.2 (*L2*) introduces CLUSTDNA for detailed redundancy analysis and Section 4.3 (*L3*) describes DIMENSIONDNA to explore the distribution on a data instance level of one selected cluster.

4.1 Interlinked MDS for Cluster Member and Dimension Space Exploration

At the overview level, redundancy aspects (*C1*) are focused by visualizing the relative pair-wise similarity relationships of all clusters with the help of a MDS projection. In SUBEVAL simultaneously two interlinked MDS projections are used: the left MDS plot illustrates the similarity of subspace clusters w.r.t. the cluster members, and the right MDS

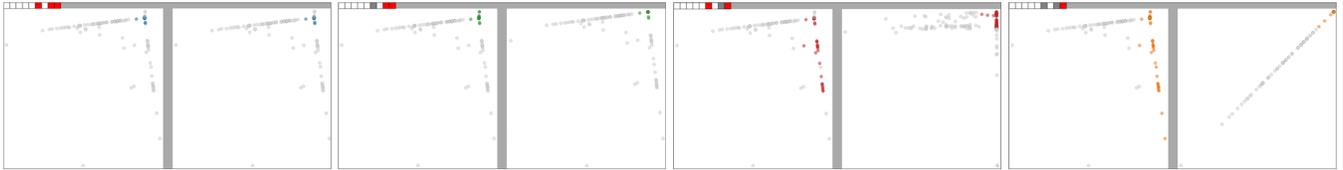


Figure 4: ClustDNA to compare the topology of 4 selected subspace clusters: each combined scatter plot represents an MDS projection of all data objects of the dataset in the subspace projection (right) and the SuperSpace (left; union of dimensions of all selected clusters). Cluster members are marked in color. The dimensions of the subspace are indicated by the top glyph (red = subspace -, grey = SuperSpace dimension).

plot depicts the similarity w.r.t. the dimension similarity. The user can change the similarity definitions in order to account for the different understanding of redundancy in the subspace analysis process. SUBEVAL supports multiple set-oriented similarity measures (e.g., Jaccard Index). Advanced measures as proposed in [28], are planned for future work.

Visual Mapping for Redundancy Analysis.

In the MDS projection, each subspace cluster is represented by a single point/glyph. In order to compare the clusters with in the corresponding counter-MDS plot we use a 2-dimensional color schema [8, 27] that links position with color (similar position = similar color; see Figure 1 (1) and (2)). The basic intuition is that in the left MDS projection (object similarity) the cluster member similarity is encoded by the 2D coordinates (position), while the dimension similarity is encoded in color in the same projection. In other words, proximity corresponds to similar/redundant clusters w.r.t. objects and a similar color indicates similar/redundant clusters in dimension similarity. The same is true for the subspace similarity in the right projection: similarity is encoded by the position, while color is used to encode the similarity in cluster member aspect. The interpretation of our interlinked MDS representation is as follows: clusters being close to each other and share a similar color in one MDS projection are similar, hence redundant in both, the cluster member and subspace aspect (C1.1)+(C1.2). Subspace clusters, which are close in the cluster member projection, but different in their coloring are similar in their cluster members, but different in their subspace topology (C1.3).

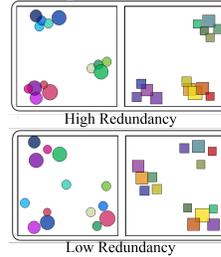
Computation of Coloring in MDS Projections.

The computation of our linked MDS projections is illustrated in Figure 3. First, the two MDS projections for the cluster member and subspace similarity are computed *independently* using a user-defined distance function. Afterwards, both projections are mapped *independently* on top of a perceptual linear 2D colormap [23]. A nearby location in the MDS projection (high similarity) is mapped to a similar color. Up to this point, the visual variables position and color are calculated independently and are not comparable between the two MDS plots. We can now apply the color information of the clusters in one MDS projection on top of the clusters in the opposite projection. By exchanging the semantic color mapping schemes between the two plots, the cluster member MDS can still indicate a (dis-)similarity in their cluster members (visually encoded by the point's location), but the coloring reflects the subspace similarity. Alike, the subspace similarity view reflects the dimension similarity by means of the points' locations, but allows perceiving the cluster membership similarities via the color encoding.

Interpretation of MDS Structures.

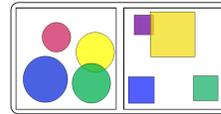
In the following, we give guidelines on how to interpret the visual appearance of the different MDS plots with respect to the presented quality criteria in Section 3.1.

High- and Low Redundancy (C1).



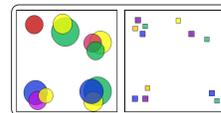
Similar objects have been clustered in similar subspaces: we can see groups of clusters in which colors are similar (top). Opposed to low redundancy (bottom), we can see groups of clusters, too, but either in different subspaces or with different objects. Thus, close clusters have dissimilar colors.

Big (Non-compact) Clusters (C3.1 + C3.4).



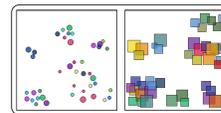
Clusters with many members or dimensions are illustrated by large glyphs in the MDS plots. Compactness can be additionally visualized by a more detailed glyph representation.

Too low-dimensional clusters (C3.3)



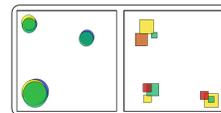
If the relevant subspace is too low dimensional the inferable insights are too trivial and no deeper conclusion about dependencies between the dimensions are possible. Too low-dimensional clusters can be seen by rather small glyphs in the subspace MDS projection. This is especially true for clusters with many cluster members (C3.4).

Small Splinter Clusters (C1.3 + C3.2)



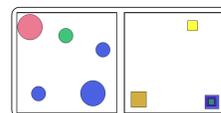
The result contains many small clusters indicated by small glyphs. These clusters do not provide a good generalization of the data; general conclusions cannot be extracted.

Cluster Splitup in Subspaces (C1.3)



Split of clusters in subspaces: nearly identical object sets are clustered in different subspaces, indicated by largely overlapping cluster circles. Although this does not imply redundancy (colors are different, thus each cluster contains new information), it provides reason to suspect that these objects actually form a cluster in a single high-dimensional subspace.

Cluster Splitup in Objects (C3.2)



Split of clusters w.r.t. objects: a cluster might be divided into multiple clusters. We can discriminate between two cases: (1) a single cluster is partitioned in a single subspace (rare case) (c.f., blue circles), or (2) a cluster is partitioned and lives in differ-

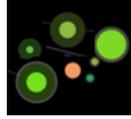
ent subspaces, which is a typical case for projected clustering algorithms like Proclus [1].

Visual Enhancement and User Interaction.

Further visual encodings can be mapped on top of the enhanced MDS representation to iteratively add more details to the clusters. The additional information adds another level of complexity to the visualization. Therefore, the user can optionally add them, if needed for an analysis purpose.

Glyph Representation: The size of the points in the MDS projection can be mapped to e.g., the cluster- or subspace size. This representation allows assessing the characteristics *C3.3* and *C3.4* of all clusters.

Furthermore, additional cluster characteristics can be added to the glyph representation. For example, the compactness can be illustrated by the size of an inner circle in the glyph. A combination of multiple criteria in a pie-chart like fashion is also imaginable. A mouse over provides additional information for a cluster (e.g., size or members).



Linking and Brushing. We implemented a linking and brushing functionality between the two MDS projections. Moving the mouse over one cluster in the left projection highlights the same cluster in the right projection and vice versa. The user is able to apply a lasso selection and highlight all selected clusters in the opposite plot (c.f. Figure 1).

Selection and Filtering. Selected subspace clusters can be further analyzed by (*L2*) CLUSTDNA (Section 4.2) and (*L3*) DIMENSIONDNA (Section 4.3). Additionally, the selected clusters can be reprojected into the MDS space to remove outlier-clusters which may distort the projection.

Ground Truth Comparison. Finally, SUBEVAL allows to add potential ground-truth clusters to the projections. Using this feature, external evaluation methods can be enhanced by (1) comparing the similarity of all detected clusters with the ground truth and see for example, that multiple clusters are similar to the benchmark, and (2) the multi-level analysis of SUBEVAL enables the user to visually analyze the structure of a ground truth cluster (c.f. DIMENSIONDNA) to decide whether a ground truth cluster is actually appropriate.

4.2 ClustDNA: Comparison of Cluster Topologies in Subspace Projections

At the second analysis level of SUBEVAL, a user is able to analyze and/or justify the redundancy of a small selection of subspace clusters (e.g., the four selected blue clusters in Figure 1 (1)). Our idea is to show for every selected cluster, both, all data objects and the cluster topology with a visualization, called CLUSTDNA. To understand the similarity between the different objects and the accordingly generated clustering structures, we rely again on a MDS projection. For every cluster we compute a projection in the respective subspace and assume that redundant subspace clusters result in similar MDS projections. Furthermore, we compare each subspace projection with a MDS projection containing the union of dimensions of all selected subspace clusters. We call the unified combination of dimensions SUPERSPACE. A comparison with these SUPERSPACE helps to decide whether a subspace of all dimensions results in a more profound cluster.

An example of CLUSTDNA can be found in Figure 4. Each selected subspace cluster is represented by a combination of two MDS projections: SUPERSPACE (left) and subspace of cluster (right). The cluster members are colored whereas

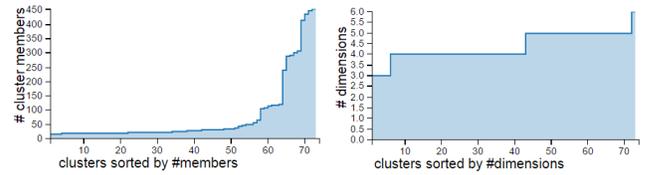


Figure 5: Distribution of the #of cluster members (left) and the #subspaces (right).

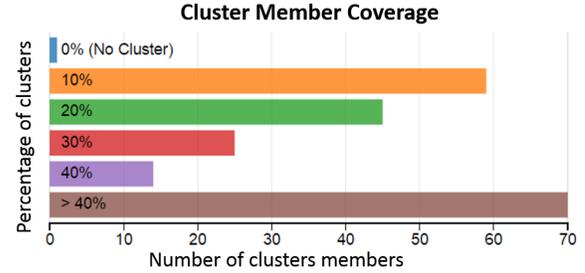


Figure 6: Bar charts to analyze the object coverage: a few objects are not clustered (blue), about 60 objects are a member in 1 – 10% of the clusters and more than 70 objects are a member in more than 40% the clusters.

non-cluster members are represented in grey. The small glyph at the top indicates the dimensions of each subspace (red = subspace -, grey = SUPERSPACE dimensions).

4.3 DimensionDNA: In-Depth Analysis

At the third analysis level, a user needs to be able to analyze one particular selected cluster to identify good/bad clustering decisions of an algorithm. SUBEVAL implements an enhanced Parallel Coordinates (PC) [18] visualization called DIMENSIONDNA. Classical PC are combined with a heatmap to illustrate the data density of the entire dataset in each dimension (Figure 1 (right)). Each vertical bar represents one dimension. The minimum value of the dimension is mapped to the bottom of the bar, linearly scaled to the top (maximum value). The white-to-black colormap encodes the number of objects falling into a specific range (dark = many objects; bright = few objects). Records of a selected cluster are visualized as a connected line (red) among all dimensions of the dataset. The subspace dimensions are highlighted.

Using DIMENSIONDNA, a user can analyze the compactness (*C3.1*) of the cluster members in the (subspace) dimensions in order to see whether a subspace cluster is valid. When selecting multiple clusters, the user is able to analyze the cluster’s redundancy (*C1*) and separability (*C3.2*). The underlying distribution of every dimension helps the analyst to inspect outliers or distortions that prevent an algorithm to identify clusters in particular dimensions.

4.4 Cluster Meta-Data Analysis

To provide additional information of detected subspace clusters (or a selection thereof), SUBEVAL comprises several visualizations to analyze the clusters’ meta-data:

Cluster- and Subspace Size Distributions: Figure 5 shows a line plot to assess the distributions of the *cluster size* (left) (c.f., *C3.3*) and *subspace size* (right) (c.f., *C3.4*). A user is able to see whether an algorithm produced rather small, large, or different sized subspace clusters.

Object Coverage Analysis: The bar-chart in Figure 6 is targeting *C2.1 -Object Coverage*, where we visualize the relationship between the number of (non-)clustered data objects. The non-clustered objects can be further investigated with the DIMENSIONDNA plot, while the redundancy aspects of the object-to-cluster assignment (*C1*) can be analyzed by interactions on the bar chart. It shows the number of objects (*x*-axis) which do not belong to any cluster (blue bar), and the number of members being part in $p\%$, of the clusters. The more this bar-chart is shifted to the bottom, the more often specific cluster members occur in multiple clusters.

Dimension Coverage Analysis *C2.2* is targeted with an interactive bar-chart showing how many subspaces a dimension is allocated. The user can subsequently investigate dimensions, which occur frequently or never in any subspace, with the DIMENSIONDNA plot.

Dimension Co-occurrence: Besides the coverage aspect, the user is able to analyze, which dimensions co-occur in the subspaces by choosing one or multiple dimensions. The chart is updated by filtering for subspaces containing the selected dimensions.

All charts can be interactively filtered. A selection of one e.g., dimension in the coverage analysis, or clusters of a specific size will update all other visualizations accordingly, thus allowing an analyst to concentrate on clusters of interest.

5. EXPERIMENTS

We describe two use cases to show the usefulness of SUBEVAL to visually evaluate the quality of subspace clusterings. SUBEVAL is implemented in Java/JavaScript in a server-client fashion using `d3.js`¹ for the visualizations. In the supplementary material² we provide a video and give the user the opportunity to explore the use cases with SUBEVAL.

Use Case 1: Redundancy Analysis.

In the first use case, we want to show the usage of SUBEVAL for the detection and analysis of redundancy. We apply the well-known CLIQUE [2] algorithm to the real-world GLASS dataset with 214 objects and 9 dimensions. CLIQUE is a grid-based algorithm which is known to detect many redundant clusters. For the GLASS dataset, 132 subspaces³ are detected.

In the first step, we analyze the cluster member coverage of our result (Figure 6). Except for one outlier (blue bar) we can quickly see that all data objects belong to at least one cluster. However, more than 70 data objects (30% of the dataset) are part of more than 40% of the clusters resulting in a noticeable degree of member overlap in the clusters.

The results of the inter-linked MDS projection can be found in Figure 1. We can see a large group of bigger clusters in the top left corner of the cluster member similarity projection. The clusters of the group share a common clustering topology, but have a different color encoding. This corresponds to similar clusters occurring in subspaces of different dimensions. Besides the smaller splinter clusters that occur in different (larger-dimensional) subspaces, the user is faced with four larger clusters (blue shaded on the left side). These clusters seem to have similar cluster members in similar subspaces and thus can be suspected redundant. We analyze this potential redundancy further with CLUSTDNA as shown

in Figure 4. In the dimension glyph on the top, we can see, that all four clusters share most of their dimensions. Another interesting observation is that the first and second clustering have an almost identical cluster topology which is visible through a similar MDS projection. The cluster on the right comprise only a single dimension in which all cluster members are almost identical. A user can conclude that the selected clusters are truly redundant. It would be sufficient to only report the first cluster without losing much knowledge about the data.

Finally, we select one of the redundant clusters and investigate the dataset distribution with the DIMENSIONDNA, as shown in Figure 1 (3). We can see that the cluster members are compact in the subspace dimensions *dim1,4,5*, but also in non-subspace dimensions *dim0,2,3,5*, and *dim7*. Hence, an analyst may question, why the aforementioned dimensions are not part of a subspace. In summary, a user can quickly see that the clustering result contains a few larger subspace clusters, but also many smaller splinter clusters and a few redundant clusters as described above. The shown results can be attributed to the bottom-up strategy of CLIQUE, which is known to produce a large number of redundant clusters. An analyst may either change the parameter settings or apply a different subspace clustering algorithm.

Use Case 2: Splinter Cluster Analysis.

In the second use case, we analyze a good performing subspace clustering algorithm (INSCY [4]) on the VOWEL dataset as experimentally identified in [24]. The dataset contains 990 object, described by 10 dimensions⁴. INSCY is an algorithm with a redundancy elimination strategy.

According to the experiments in [24], the algorithm performs well on the dataset with good external performance measures (compared to a ground truth). When analyzing the clustering result with SUBEVAL, we made the following observations: The size of the subspaces is homogeneous with a dimensionality between three and six dimensions. However, the number of cluster members varies significantly. Many clusters contain less than 30 members and only a few clusters have more than 300 members as shown in Figure 5. When encoding this information into the inter-linked MDS projection (c.f. Figure 7), we can see that the clustering contains a large number of small splinter clusters with a variety of different colors. This means that in a large number of subspaces, the algorithm detected small, less expressive clusters. The group of bigger clusters on the bottom left is apart from the splinter clusters and contains significantly more cluster members, hence a more general representation of the data. As visible from the similar coloring, there are many redundant clusters, which can be verified in the detail analysis. We select one of the clusters, as shown in Figure 7 (1), and analyze the data distribution with the DIMENSIONDNA (shown in Figure 7 (3)). The subspace contains three dimensions. *dim3*, however, does not seem to be compact and an analyst may question why this dimension is part of the subspace. It is therefore interesting that the algorithm performed well on the dataset according to the experiments in [24]. Based on our findings, an algorithm expert could improve the clustering results by a careful adjustments of the parameters.

¹<https://d3js.org/>

²<http://www.subspace.dbvis.de/idea2016>

³Parameter of CLIQUE for use case 1: -XI 10 -TAU 0.2

⁴Parameter of INSCY for use case 2: -gS 10 -mS 16 -de 10.0 -m 2.0 -e 8.0 -R 0.0 -K 1

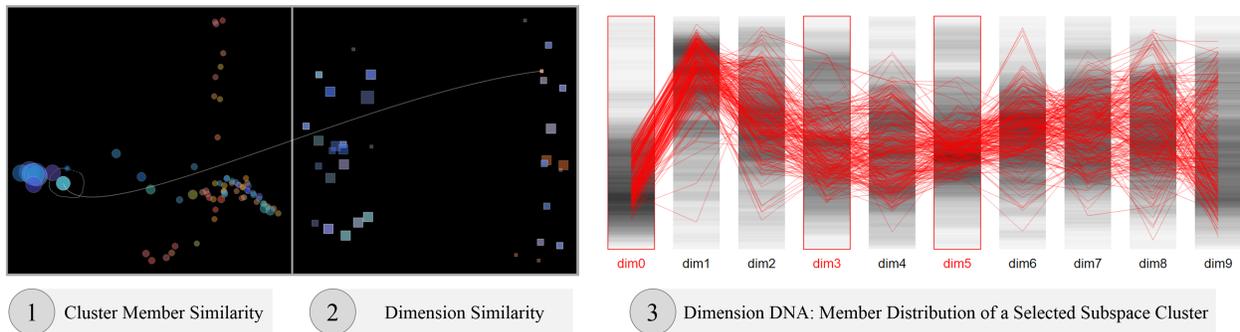


Figure 7: Use Case 2: (1) + (2) Group of large clusters with similar subspaces (blue group left) and many small splinter clusters with different colors (=different subspaces) (left). One cluster is selected for detailed analysis. (3) DimensionDNS: Visualizing the distribution of cluster members of the selected cluster. An analyst may wonder why the outliers in *dim3* and *dim5* are part of the cluster.

6. DISCUSSION AND FUTURE WORK

While our technique has proven useful for an efficient and effective visual comparison of subspace clusters regarding certain quality aspects, we identify areas for further research.

Alternative Visual Design. The inter-linked MDS projection between the cluster member and dimension similarity of subspace clusters may be difficult to read and requires some training for unfamiliar users. The same is true for the CLUSTDNA visualization. Furthermore, MDS projections face generally the problem of overlapping points and might not show the actual similarity between all combinations of points as discussed below. Therefore, we are planning to improve the MDS projection and also work on different visual representations for the overview of subspace clusterings. Node-link diagrams as introduced in [30] may be an interesting starting point to this end.

MDS projects data points into a 2D space by preserving the pair-wise distances between all data points as well as possible. Depending on the distance distributions, the 2D projection may not reflect the actual relationships correctly. Then, objects appearing close in the projection might be dissimilar in their original space, and far apart objects may be similar. Independent of the quality, a MDS projection is typically interpreted by a user as it is, without considering a possible error which lead to wrong analysis results. SUBEVAL already provides methods for drill-down to justify presumptions in a different view. Later, we also want to address the quality of the MDS projection by visualizing the difference between the similarity in the MDS projection and the real data characteristics, or rely on further techniques for visualization of projection quality [25].

SUBEVAL is designed to analyze one subspace clustering result at a time. A comparative evaluation of several clustering results would be beneficial to compare the influence of minor changes in the parameter settings. We plan to extend SUBEVAL for a comparative analysis of multiple clusterings.

Application to Related Approaches. The analysis goal of subspace clustering differs significantly from other analysis techniques like *subspace outlier detection* (SOD) [33] and *subspace nearest neighbor search* (SNNS) [19]. While SOD tries to identify subspaces in which outliers exist, SNNS identifies nearest neighbor sets to a given query in different subspaces. Although the analysis goal differs, both techniques share the same evaluation challenges like subspace clustering, i.e., redundant subspaces and results (outliers or nearest

neighbors). In the future, we want to extend SUBEVAL for the quality assessment of SOD and SNNS. For the inter-linked MDS projection we need to develop quality measures for the redundancy definition. DIMENSIONDNA can be applied to both techniques. Also, we need to develop visualizations to access the meta-data of the respective analysis.

SUBEVAL is designed for the quality assessment of subspace clusterings, however, it can also be used for the evaluation of full-space clusterings, particularly with partially overlapping clusters. For the MDS projection, an appropriate measure is needed to compute the similarity between clusters. One option is to compute the distance between cluster centroids or the pair-wise distances between all cluster members. DIMENSIONDNA and CLUSTDNA can also be applied to investigate cluster topologies and member distributions.

Open Source Framework. SUBEVAL is part of SUBVA (Subspace Visual Analytics), a novel open-source framework for visual analysis of different subspace analysis techniques. Besides providing implementations of recently developed visualizations, such as SUBVIS [20], SUBVA integrates the well-known OpenSubspace framework [24] as a module, allowing analysts to apply the most commonly used subspace clustering algorithm to a given dataset. We will distribute the framework on our website⁵ and provide the source code in the supplementary material.

7. CONCLUSION

This paper presented SUBEVAL, a subspace evaluation framework for the simultaneous assessment of several quality characteristics of one subspace clustering result. SUBEVAL combines expressive visualizations with interactive analysis and domain knowledge, and complements, potentially advancing standard evaluation procedures with a more comprehensive, multi-faceted approach. We summarized state-of-the-art evaluation methods for subspace clustering algorithms and showed that, besides classical measures, visualizations can be an insightful approach to the evaluation and understanding of subspace clustering results. We also outlined ideas for extensions of our approach.

Acknowledgments

The authors are grateful for the valuable discussion and work that contributed to the underlying framework of J. Kosti, F.

⁵<http://www.subva.dbvis.de>

Dennig, M. Delz, and S. Wollwage. We wish to thank the German Research Foundation (DFG) for financial support within the projects A03 of SFB/Transregio 161 “Quantitative Methods for Visual Computing” and DFG-664/11 “SteerSciVA: Steerable Subspace Clustering for Visual Analytics”.

8. REFERENCES

- [1] C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park. Fast Algorithms for Projected Clustering. *SIGMOD Rec.*, 28(2):61–72, 1999.
- [2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2):94–105, 1998.
- [3] I. Assent, R. Krieger, E. Müller, and T. Seidl. VISA: visual subspace clustering analysis. *SIGKDD Explor. Newsl.*, 9(2):5–12, 2007.
- [4] I. Assent, R. Krieger, E. Müller, and T. Seidl. INSCY: Indexing subspace clusters with in-process-removal of redundancy. In *Proc. of ICDM*, pages 719–724, 2008.
- [5] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.
- [6] F. Bendi, R. Kosara, and H. Hauser. Parallel sets: visual analysis of categorical data. In *Proc. of InfoVis*, pages 133–140, 2005.
- [7] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is “nearest neighbor” meaningful? In *Database theory — ICDT’99*, pages 217–235, 1999.
- [8] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on visual comparative data analysis. *CGF*, 30(3):891–900, 2011.
- [9] S. Bremm, T. von Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher. Interactive Comparison of Multiple Trees. In *Proc. of VAST*, 2011.
- [10] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *TVCG*, 17(12):2581–2590, 2011.
- [11] T. F. Cox and M. A. Cox. *Multidimensional scaling*. CRC press, 2000.
- [12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of SIGKDD*, pages 226–231, 1996.
- [13] I. Färber, S. Günemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek. On Using Class-Labels in Evaluation of Clusterings. In *Workshop at SIGKDD*, 2010.
- [14] S. Günemann, I. Färber, E. Müller, I. Assent, and T. Seidl. External Evaluation Measures for Subspace Clustering. In *Proc. of CIKM*, pages 1363–1372, 2011.
- [15] S. Günemann, I. Färber, H. Kremer, and T. Seidl. CoDA: Interactive cluster based concept discovery. *Proc. of VLDB Endowment*, 3(1-2):1633–1636, 2010.
- [16] S. Günemann, H. Kremer, I. Färber, and T. Seidl. MCEplorer: Interactive Exploration of Multiple (Subspace) Clustering Solutions. In *Data Mining Workshops at ICDMW*, pages 1387–1390, 2010.
- [17] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition.
- [18] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. 2013:95–116, 2013.
- [19] M. Hund, M. Behrisch, I. Färber, M. Sedlmair, T. Schreck, T. Seidl, and D. Keim. Subspace Nearest Neighbor Search - Problem Statement, Approaches, and Discussion. In *Proc. of SISAP*, pages 307–313. 2015.
- [20] M. Hund, D. Böhm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D. A. Keim, L. Majnarić, and A. Holzinger. Visual analytics for concept exploration in subspaces of patient groups. *Brain Inf.*, pages 1–15, 2016.
- [21] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering pattern-based clustering, and correlation clustering. *ACM TKDD*, 3(1), 2009.
- [22] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *Proc. of ICDM*, pages 911–916, 2010.
- [23] S. Mittelstädt, J. Bernard, T. Schreck, M. Steiger, J. Kohlhammer, and D. A. Keim. Revisiting Perceptually Optimized Color Mapping for High-Dimensional Data Analysis. In *In Proc. of EuroVis*, pages 91–95, 2014.
- [24] E. Müller, S. Günemann, I. Assent, and T. Seidl. Evaluating Clustering in Subspace Projections of High Dimensional Data. *Proc. of VLDB Endowment*, 2(1):1270–1281, 2009.
- [25] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, 2010.
- [26] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. of Visual Languages*, pages 336–343. IEEE, 1996.
- [27] M. Steiger, J. Bernard, S. Mittelstädt, S. Thum, M. Hutter, D. A. Keim, and J. Kohlhammer. Explorative Analysis of 2D Color Maps. In *Proc. of Computer Graphics, Visualization and Computer Vision*, volume 23, pages 151–160, 2015.
- [28] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. A. Keim. Subspace Search and Visualization to Make Sense of Alternative Clusterings in High-Dimensional Data. In *Proc. of VAST*, pages 63–72, 2012.
- [29] A. Tatu, L. Zhang, E. Bertini, T. Schreck, D. A. Keim, S. Bremm, and T. von Landesberger. ClustNails: Visual Analysis of Subspace Clusters. *Tsinghua Science and Technology*, 17(4):419–428, 2012.
- [30] C. Vehlow, F. Beck, P. Auwärter, and D. Weiskopf. Visualizing the evolution of communities in dynamic graphs. *Comput. Graph. Forum*, 34(1):277–288, Feb. 2015.
- [31] M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210, 2002.
- [32] M. O. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., 2010.
- [33] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012.