VISA: Visual Subspace Clustering Analysis

Ira Assent Ralph Krieger Emmanuel Müller Thomas Seidl
Data management and data exploration group, RWTH Aachen University, Germany
{assent,krieger,mueller,seidl}@cs.rwth-aachen.de

ABSTRACT

To gain insight into today's large data resources, data mining extracts interesting patterns. To generate knowledge from patterns and benefit from human cognitive abilities, meaningful visualization of patterns are crucial. Clustering is a data mining technique that aims at grouping data to patterns based on mutual (dis-)similarity. For high dimensional data, subspace clustering searches patterns in any subspace of the attributes as patterns are typically obscured by many irrelevant attributes in the full space. For visual analysis of subspace clusters, their comparability has to be ensured. Existing subspace clustering approaches, however, lack interactive visualization and show bias with respect to the dimensionality of subspaces.

In this work, dimensionality unbiased subspace clustering and a novel distance function for subspace clusters are proposed. We suggest two visualization techniques that allow users to browse the entire subspace clustering, to zoom into individual objects, and to analyze subspace cluster characteristics in-depth. Bracketing of different parameter settings enable users to immediately see the effect of parameters on their data and hence to choose the best clustering result for further analysis. Usage of user analysis for feedback to the subspace clustering algorithm directly improves the subspace clustering. We demonstrate our visualization techniques on real world data and confirm results through additional accuracy measurements and comparison with existing subspace clustering algorithms.

1. INTRODUCTION

Increasingly large data resources in life sciences, mobile information and communication, e-commerce, and other application domains require computer-based techniques for gaining knowledge. More and more data are produced by sensor networks, technical or financial monitoring systems, scientific experiments, or telecommunication networks. Scientists developing new drugs, system administrators monitoring complex technical processes, and decision makers being responsible for complex social or technical systems require an overview and even a deeper understanding of their monitored systems. Moreover, yet unknown dependencies in the observed measurements are crucial for development of new models that detect and explain causalities.

While computers efficiently support statistical and associative analysis of large amounts of data, they do not reach the high cognitive abilities of human users. Human experts are able to quickly identify both correlations and irregularities if data or results are appropriately visualized [13]. As "visual animals", humans excel at expressing and analyzing visual entities [19]. Tasks that are inherently difficult for computers such as parametrization, or are subjective by their very nature such as redundancy or relevance to a given task, can be effectively supported by human computer interaction. Visualization plays a key role in interaction as it build an interface between an automated output and the human user. Visual data analysis allows to combine the efficiency of computers with the cognitive strength of human users [13]. For data analysis or mining tasks, visualization is the central step from patterns to knowledge in the knowledge discovery process [11].

Clustering is one of the major knowledge discovery tasks. It aims at summarizing data base objects such that similar objects are grouped together while dissimilar ones are separated [11]. In noisy data or data with many attributes, clusters are often hidden in subspaces of the attributes and do not show up when clustering over the full attribute space. A global reduction to relevant attributes is often infeasible, as relevance of attributes is not necessarily a globally uniform property in the data [1; 12; 3]. Subspace clustering thus aims at identifying clusters in any possible attribute combination. As the number of possible subspace is exponential in the number of dimensions, this is a challenging task both with respect to efficient runtimes of the algorithm as well as to the typically enormous number of (redundant) output clusters.

For full-space clustering, different visualization techniques exist [8; 25; 9; 16; 14]. In all these approaches the same fixed set of dimensions is encoded in the visual model. These methods are thus as limited as full-space clustering: clusters are detected only if visible in the chosen mappings. As typically clusters are hidden in different subspace dimensions they cannot be detected by globally defined projections and encodings.

Meaningful subspace cluster visualization requires aggregation of similar results and occlusion of redundant information. Users need compact representation of different subspace clusterings for visual analysis of relevant aspects and feedback to the algorithm. Visualization thus requires a measure of similarity for the output, as well as techniques to reduce the overwhelming number of (redundant) results common in subspace clustering. Exploiting user feedback, subspace clustering algorithms may be focused to relevant parts or subspaces of the data through explicit guidance.

In this paper, we propose a novel distance measure for subspace clusters that reflects their inherent connections or differences, respectively. Based on subspace clustering distance models, VISA (visual subspace clustering analysis) is able to produce expressive visual diagrams that allow for meaningful user interaction.

For traditional full-space clustering, density-based approaches have shown to successfully mine clusters even in the presence of noise [7]. The idea is to define clusters as dense areas separated by sparsely populated areas. Density-based clustering has been extended to subspace clustering in previous works [12; 17]. As dimensionality is ignored in these approaches, density in subspaces of different dimensionalities is not comparable. Existing approaches which do not take this effect into consideration hence check incomparable values against the same threshold. Our density-based subspace clustering approach DUSC (dimensionality unbiased subspace clustering) based on statistical foundations takes the dimensionality into account. This method eliminates dimensionality bias and leads to comparable clustering results between different subspaces. This allows for meaningful analysis of visualized subspace clusters following our novel VISA method.

In this work, we propose visualization techniques for subspace clustering. Our contributions include

- unbiased, comparable results based on a statistically sound subspace clustering model
- powerful yet compact visualization capable of dealing with clusters in many different projections
- meaningful ranking of the most "interesting" results to guide analysis of the output
- user interaction for parameter setting, exploration and feedback
- handling of redundant output

This paper is structured as follows: we review related work on both subspace clustering and visualization techniques in the following section. In Section 3 we formalize unbiased density-based subspace clustering. Novel visualization techniques for subspace clustering results, including discussion of how to rate (dis-)similarity and redundancy of the output, are presented in Section 4, before we conclude our paper.

2. RELATED WORK

Clustering generates groups of similar objects while assigning dissimilar objects to different clusters [11]. In density-based clustering, clusters are dense areas separated by sparsely populated areas as in DBSCAN [7]. It has shown to be capable of detecting arbitrarily shaped clusters even in noisy settings. Traditional clustering does not scale to high-dimensional spaces. As clusters do not show across all attributes, they are hidden by irrelevant attributes [4]. Global dimensionality reduction techniques such as principal component analysis are typically not appropriate, as relevance is not globally uniform [5].

To detect locally relevant attributes, subspace clustering searches for clusters in any possible subset of the attributes [21]. As the number of subspaces is exponential in the number of dimensions, this is a challenging task. CLIQUE discretizes the data using grids to reduce computational complexity, yet misses clusters which spread across several grid cells [1]. SCHISM extends CLIQUE using a variable threshold to cope with different dimensionalities, yet relies on heuristics and grid-based discretization for pruning [23]. Consequently, completeness is lost as in all grid-based approaches. Specialized algorithms for categorical data or sequences [26; 2] require discretization as well. SUBCLU extends non-discretized density-based clustering to subspaces [12]. Its result, however, is biased with respect to dimensionality, i.e. while expected density changes with dimensionality, fixed density thresholds are used for pruning.

Visual data analysis techniques benefit from human cognitive abilities in data mining through user interaction. For traditional clustering in general, various visualization techniques have been proposed [8; 25; 9; 16; 14]. In all these cases, the same dimensions of the data space are encoded by the same projections or parameters in the visual model. The methods thus behave as limited as full-space clustering does: Clusters are only detected if they are visible in the respective chosen mappings. As typically clusters are hidden in different subspace dimensions they cannot be detected by globally defined projections and encodings.

Additionally, subspace clustering visualization has to deal with the exponential number of subspace projections and the typically enormous redundancy of the result. As different projections contain different clusters, visualization should provide an overall overview as well as means for in-depth analysis and interaction. Special cases like mosaic encodings in gene-expression analysis [6] order clusters by biological properties like position of a gene on the chromosomes. This underlying ordering does not extend to other application domains. As there is no inherent ordering in general subspace clustering applications, lack of (dis-)similarity measures and poor comparability of results are major hindrances for visualization.

3. SUBSPACE CLUSTERING

Our VISA (visual subspace clustering analysis) approach visualizes subspace clustering for user interaction. As mentioned above, this requires comparable clustering results in any subspace. Existing subspace clustering approaches do not take changing expected density values for different dimensionalities into account. Consequently, these approaches suffer from an effect we call *dimensionality bias* that hinders meaningful comparison of results.

We formalize density estimation and our notion of dimensionality bias, before proposing an unbiased density estimator that leads to comparable subspace clustering results.

Let $\mathbf{U} = [0, \mathbf{v}]$ be a universal domain for all dimensions, $\mathbf{D} = \{1, \ldots, d\}$ be an index set, and $\mathbf{DB} \subseteq \mathbf{U^D}$ a d-dimensional data base with n objects. A subspace $\mathbf{U^S}$ is the projection of $\mathbf{U^D}$ to the r dimensions specified by the index set $\mathbf{S} = \{s_1, \ldots, s_r\} \subseteq \mathbf{D}$. Analogously, let $\mathbf{DB^D}$ denote the original data base and $\mathbf{DB^S}$ its projection to the dimensions in $\mathbf{S}$. For ease of notation, we refer to a subspace $\mathbf{U^S}$ by its index set $\mathbf{S}$. The definition of density-based subspace clusters extends standard notions in density-based clustering [7]. Let $\|.\|^{\mathbf{S}}$ denote the restriction of norm $\|.\| : \mathbf{U^D} \to \mathcal{R}$ to the dimensions in subspace $\mathbf{S}$. The area of influence is the neighborhood in subspace $\mathbf{S}$:
$\mathcal{N}_\varepsilon^{\mathbf{S}}(o) = \{p | p \in \mathbf{DB}, \|p - o\|^{\mathbf{S}} \leq \varepsilon\}$. Typically, density of

an object o is determined by simply counting the number of objects in a fixed ε-range $\mathcal{N}_\varepsilon^{\mathbf{S}}(o)$. We generalize this idea by assigning weights to each object contained in $\mathcal{N}_\varepsilon^{\mathbf{S}}(o)$.

DEFINITION 1. ***Density Measure***
Let $\mathcal{W}$ be an arbitrary weighting function $\mathcal{W} : \mathcal{R} \to \mathcal{R}$. Based on $\mathcal{W}$, a generalized density measure $\varphi^{\mathbf{S}}(o)$ for an object o in subspace $\mathbf{S}$ is defined as:

$$\varphi^{\mathbf{S}}(o) = \sum_{p \in \mathcal{N}_\varepsilon^{\mathbf{S}}(o)} \mathcal{W}\left(\|o - p\|^{\mathbf{S}}\right)$$

Thus, an object o in subspace $\mathbf{S}$ is called *dense* if the weighted distances to objects in its area of influence sum up to more than a given density threshold $\varphi^{\mathbf{S}}(o) \geq \tau$.

3.1 Dimensionality Bias

Incomparable density values pose the following problem: the high discrepancy in density scales of low-dimensional or high-dimensional subspaces makes it impossible to find a suitable parameter for a fixed density threshold τ. If on the one hand τ is parametrized such that high-dimensional clusters with low expected density are detected then numerous excess pseudo-clusters are generated in low-dimensional spaces where expected density is high. On the other hand, a parametrization of τ which separates clusters from noise in low-dimensional spaces loses clusters in high-dimensional spaces. We assume that τ is fixed as dimensionality dependent thresholds can also be incorporated into the density measure.

To obtain comparable density values, unbiased density measures have to be independent of the dimensionality of the subspace. Statistically speaking, this corresponds to the same expected density value regardless of the dimensionality of the subspace.

DEFINITION 2. ***Dimensionality Unbiased Density Measure***
A density measure $\varphi^{\mathbf{S}}$ is dimensionality unbiased if its expected density is the same for any two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2 \subseteq \mathbf{D}$:

$$\forall\ \mathbf{S}_1, \mathbf{S}_2 : \quad E\left[\varphi^{\mathbf{S}_1}\right] = E\left[\varphi^{\mathbf{S}_2}\right]$$

We now show how dimensionality bias can be eliminated for any density estimator. As the expected density should be the same for any two subspaces, we normalize density estimators with their expected density. For any density measure $\varphi^{\mathbf{S}}$, the normalized measure $\frac{1}{E[\varphi^{\mathbf{S}}]}\varphi^{\mathbf{S}}$ is dimensionality unbiased. With linearity property of the expectation, this is straightforward: $E[\frac{1}{E[\varphi^{\mathbf{S}}]}\varphi^{\mathbf{S}}] = \frac{1}{E[\varphi^{\mathbf{S}}]}E[\varphi^{\mathbf{S}}] = 1$ for all subspaces. Thus, for any two subspaces, normalizing the density measure by the expected value of the subspace yields comparable density values for any two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$. Comparable density values yields consistent results. This is important both for the mining process as well as for visualizing coherent views.

The computation of an unbiased density measure is given in [3] for the Epanechnikov kernel as density measure [24]. Similarly, normalization of any other density measure with respect to dimensionality can be performed.

3.2 The unbiased DUSC approach

Intuitive density threshold. The density threshold is a core parameter since it sets the dividing line between dense objects and noise. As this parameter has to be set by the user it is important for users to have an intuitive understanding of this parameter. Commonly, users do not know density distribution apriori, which makes the choice of a density value difficult. We exploit the fact that in our approach density is measured with respect to the expected density as discussed before. Consequently, users do not need to specify absolute density thresholds, but only a factor by which the expected density has to be exceeded. Following the definition in the previous section, an object o is dense in subspace $\mathbf{S}$ according to the expected density $\alpha(\mathbf{S})$ iff:

$$\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o) \geq \mathbf{F}$$

where $\mathbf{F}$ denotes the density threshold. As the density factor $\mathbf{F}$ is independent of the dimensionality and data set size, it is much easier to specify and its setting can additionally be supported by an overall visualization of the clustering result.

Empty space problem. With increasing dimensionality the expected density and hence the expected number of objects contained in an area of influence drops exponentially [4]. This effect is termed "empty space problem" in statistics [24]. Compared to the expected density, an object may be determined as dense even if the area of influence is nearly devoid of observations, resulting in pseudo-dense single objects. To remove pseudo-dense objects, we introduce a specific density constraint on the expected density of η objects in the area of influence. The expected density value of an object o which contains η objects in the area of influence $E_\eta\left[\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o)\right]$ can be computed from the expected density of the density measure. Details for the Epanechnikov kernel can be found in [3].

To guarantee that objects are not considered dense if the ε sphere is virtually empty, a very small value for η is sufficient (generally two or three). Users typically do not need to change this value. Our new density-based subspace clustering model below combines the density constraints α and ω [3]. α and ω combined ensure an unbiased density notion without defining objects in nearly empty regions as dense.

Redundancy. Since the number of possible subspace projections is exponential in the number of dimensions, subspace clustering algorithms often produce numerous redundant subspace clusters. To avoid excessive cluster outputs which contain essentially the same information repeated in different dimensionalities, we check if a cluster $\mathbf{C}$ in subspace $\mathbf{S}$ is redundant. We define a cluster as redundant if (most of) the objects contained in the cluster are also contained in another cluster in a higher dimensional subspace $\mathbf{S}' \supset \mathbf{S}$. We use a parameter R to specify the degree of redundancy acceptable to the user. To restrict the output to a reasonable size a strict redundancy parameter is often appropriate and can be chosen from suitable visual representations.

So far, we have studied the density of individual objects. Subspace clusters, following density-based clustering paradigm, are connected sets of dense objects. To ensure that clusters reflect the inherent structure of the data, they should contain a certain minimum number of objects. This constraint *minSize* is typically about 1% of the data base size.

The resulting subspace cluster model taking these conclu-

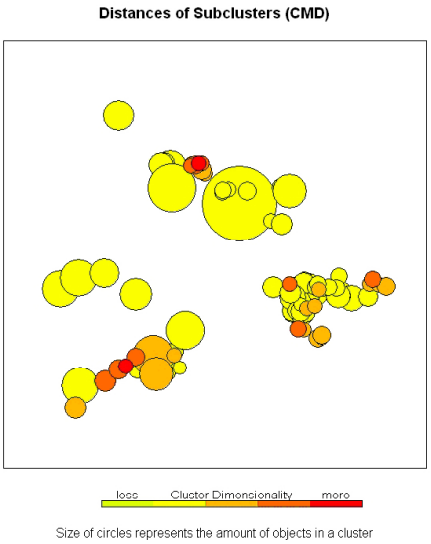

Figure 1: Subspace clustering overview: the diameter depicts the number of objects in the cluster, the color denotes the dimensionality (the darker, the more dimensions)

sions into account is formalized in the following.

DEFINITION 3. ***DUSC Subspace Cluster***

A set of objects $\mathbf{C} \subseteq \mathbf{DB}$ *in subspace* $\mathbf{S} \subseteq \mathbf{D}$ *is a subspace cluster if:*

- *objects in* $\mathbf{C}$ *are* $\mathbf{S}$***-connected:***
  $\forall o, p \in \mathbf{C}: \exists k: \forall i = 1, \ldots, k-1: \exists q_i \in \mathbf{C}:$
  $\|q_i - q_{i+1}\|^{\mathbf{S}} \leq \varepsilon \wedge q_1 = o, q_k = p$
- ***more dense than expected and not pseudo-dense:***
  $\forall o \in \mathbf{C}$: $\varphi^{\mathbf{S}}(o) \geq max\{\mathbf{F} \cdot \alpha(\mathbf{S}), \eta \cdot \omega(\mathbf{S})\}$
- $\mathbf{C}$ *is* ***maximal****, i.e. contains all* $\mathbf{S}$*-connected objects*
  $\forall o, p \in \mathbf{DB}$, o, p *S-connected* $\Rightarrow$ $(o \in \mathbf{C} \Leftrightarrow p \in \mathbf{C})$
- ***minimum cluster size:*** $|\mathbf{C}| \geq minSize$
- ***not redundant:*** $\neg\exists(\mathbf{C}', \mathbf{S}')$ *subspace cluster with*
  $\mathbf{C}' \subseteq \mathbf{C} \wedge \mathbf{S} \subset \mathbf{S}' \wedge |\mathbf{C}'| \geq R \cdot |\mathbf{C}|$

The DUSC subspace clustering model extends existing density-based notions of maximality and connectedness with statistically sound density computation via normalized Epanechnikov kernel and expected density. Clusters contain a significant part of the data, and are not redundant.
Evaluating the cluster model for all possible subspaces is infeasible as their number is exponential with respect to the dimensionality. In our subspace clustering algorithm we thus avoid excess subspace cluster evaluation through (1) a filter-and-refine architecture using the weakest density threshold of all subspaces as filter pruning, (2) a depth first computation on a specialized index structure, and (3) redundancy pruning.

4. VISA

The DUSC algorithm from the previous section is used to mine subspace clusters. The resulting patterns have to be analyzed by users for knowledge generation. This analysis is still challenging as there are many different subspace projections and possibly redundant subspace clusters. We define the set of subspace clusters that have to be analyzed:

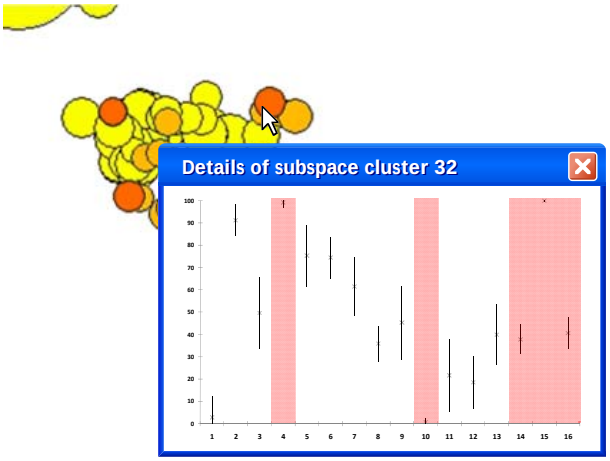

Figure 2: Detailed view for one subspace cluster: mean and variance for each dimension of the subspace cluster

DEFINITION 4. ***Subspace Clustering***

A subspace clustering is a set of clusters in their corresponding subspaces $\{(\mathbf{C}_1, \mathbf{S}_1), \ldots, (\mathbf{C}_n, \mathbf{S}_n)\}$*, where* $\mathbf{C}_i$ *is a subspace cluster in subspace* $\mathbf{S}_i$ *as specified in Definition 3.*

For user benefit, it is necessary to visualize subspace clusterings such that the entire output can be browsed even for clusters in different subspace projections, and that detailed views into individual subspace clusters are possible. Moreover, feedback for subsequent guiding of subspace clustering runs should be provided. This requires visual support for analysis of parameter settings as well as a focus on the most "interesting" results.
We therefore define a set of criteria that should be included in visualization:

DEFINITION 5. ***Visual analysis criteria***

A subspace clustering visualization should represent the following subspace clustering properties:

- ***space overlap****, i.e. the number of common subspaces for any two cluster subspaces* $\mathbf{S}_i$ *and* $\mathbf{S}_j$*:* $\mathcal{S} = |\mathbf{S}_i \cap \mathbf{S}_j|$
- ***object overlap****, i.e. the number of common objects in any two subspace clusters* $\mathbf{C}_i$ *and* $\mathbf{C}_j$*:* $\mathcal{O} = |\mathbf{C}_i \cap \mathbf{C}_j|$
- ***interestingness****, i.e. the factor* $\mathcal{I}$ *by which the expected density is exceeded on average for any subspace cluster* $\mathbf{C}_i$ *in subspace* $\mathbf{S}_i$*:*

$$\mathcal{I}(\mathbf{C}_i, \mathbf{S}_i) = \frac{\sum_{o \in \mathbf{C}_i} \varphi^{\mathbf{S_i}}(o)}{|\mathbf{C}_i| \cdot \alpha(\mathbf{S}_i)}$$

The above aspects are crucial for our VISA approach, since they are necessary for in-depth analysis of subspace clusterings, where large result sizes easily occlude the most interesting, novel patterns. Next, we show how compact representation and detail information on subspace clusters can be combined for browsing. After this, we focus on bracketing and in-depth analysis.

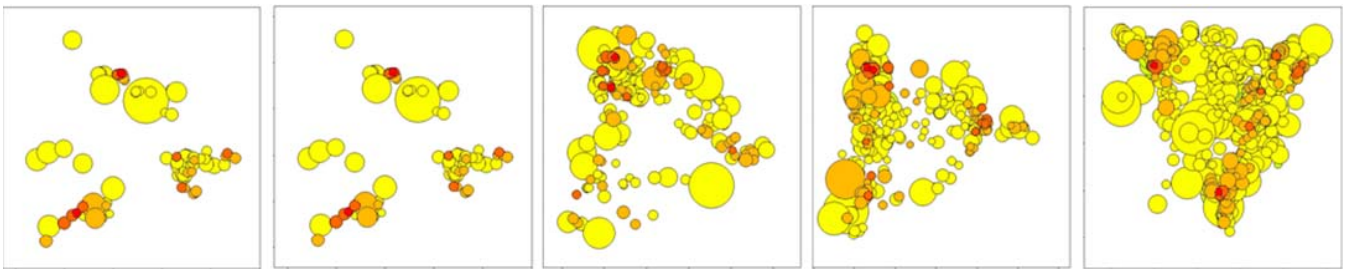

Figure 3: Bracketing Redundancy in MDS images (from left to right $R = 0\%, 2\%, 5\%, 7\%, 10\%$): for each parameter setting, the output is illustrated, allowing users to pick appropriate settings from the series of images easily

4.1 Browsing subspace clusterings

To give a general overview over the entire subspace clustering, basic structural information of subspace clusters should be compactly represented. Interactively, details should be provided for individual clusters during browsing.

The major challenge for any overview visualization of subspace clusterings lies in the comparability of subspace clusters. As subspace clustering algorithms may identify patterns in completely different or overlapping subspaces, we have to define (dis-)similarity of subspace clusters on a general scale. We propose a distance function for comparing any two subspace clusters that takes both the subspace overlap and the object overlap into account. It is defined as a convex sum of the difference in subspaces and the difference in cluster objects:

DEFINITION 6. ***Subspace Cluster Distance***

The distance between two subspace clusters $\mathbf{C}_i, \mathbf{C}_j$ in subspaces $\mathbf{S}_i, \mathbf{S}_j$, respectively, is defined as the convex sum of subspace distance and object distance:

$$\beta \left(1 - \frac{|\mathbf{S}_i \cap \mathbf{S}_j|}{|\mathbf{S}_i \cup \mathbf{S}_j|}\right) + (1-\beta)\left(1 - \frac{|\mathbf{C}_i \cap \mathbf{C}_j|}{\min\{|\mathbf{C}_i|, |\mathbf{C}_j|\}}\right)$$

This distance function thus allows comparing subspace clusters for a general overview. Independent of any application specific properties of the data, similarity can be measured by incorporating the criteria presented in Definition 5.

Normalization of both subspace and object distance is to a range of zero to one, respectively. We use two different normalizations for object and subspace distances.

1.) Object distance: a fully redundant subspace cluster is a lower dimensional projection of an interesting subspace cluster (object overlap). For visualization, interesting subspace clusters should be stacked upon their redundant counterparts, i.e. the distance should be zero. This is achieved by normalizing object distance by minimum cluster size. For redundant subspace clusters, the minimum in the denominator is the same as the intersection in the numerator, and object distance is indeed zero.

2.) For subspace distances, clusters in subsets of dimensions may very well differ in all objects. They share subspace projections, yet do not actually overlap. Their subspace distance should therefore still reflect the difference in the remaining dimensions, and not be zero. To this end, normalization is by the union of subspaces. For visualization, this means that sub-projections are located close to each other.

Object overlap can be further favored by choosing β smaller than 0.5 to give more weight to object distance.

An overview over the distances of all subspace clusters can be achieved by multidimensional scaling (MDS) [18]. Simply put, it allows for nonlinear projection of objects from their original distance space to a 2D or 3D visualization space, preserving mutual distances as much as possible. In our MDS image of the subspace clustering result, the distance information is enriched by information on the size and the dimensionality of a subspace clusters. For a general overview size and dimensionality of subspace clusters can be visualized by the radius of the circles and their color, respectively (see Figure 1). Consequently, browsing the entire output is possible. However, this enriched MDS image by itself does not provide sufficient information for user analysis. To get a better understanding of why subspace clusters are similar or different and for browsing the actual objects in subspace clusters, detailed information should be available upon click on subspace cluster representations. As individual objects in subspace clusters are typically more than two-dimensional, we represent the distribution of objects in a subspace cluster by a mean and variance plot for all dimensions. Additionally, we highlight those dimensions that are relevant for the subspace cluster (see Figure 2).

4.2 Bracketing

Bracketing refers to a technique originally from photography. Several different camera settings are used to take a series of pictures of the same subject. Photographers then pick the best setting among the resulting pictures. It has been discussed in human computer interaction in [22]. We propose bracketing for subspace clustering visualization as a useful technique that demonstrates the effects of parametrization. It provides not just a single subspace clustering output, but a series to choose from. This allows users to analyze the effect of different parameters at a single view.

As we have seen already in Figure 1, the clustering shows groups of subspace clusters. In their center we have the desired high-dimensional (red) subspace clusters containing few objects (small circle). These high-dimensional clusters in lower projections again form clusters, which are depicted as large yellow circles. This effect can be reduced by our redundancy parameter R, as we can see in Figure 3. The overwhelming result for $R = 10\%$, i.e. redundancy of 10% is permitted, gives us evidence of poor cluster quality as in large low-dimensional clusters not only hidden clusters are found but also noise. As we will see later in accuracy analysis, removing such redundant clusters leads to a significant quality improvement of the overall clustering result.

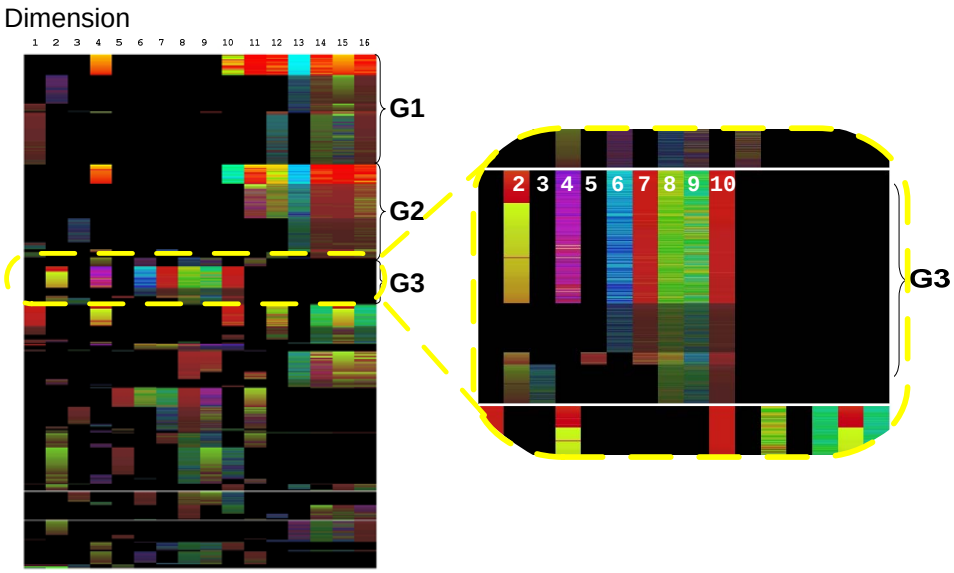

Figure 4: Matrix of subspace clusters groups. Rows represent the cluster and columns the dimensions, groups are separated by white lines (see right part). Color map HSV: hues represent the value in each dimension of the subspace cluster; saturation and value represent the factor $\mathcal{I}$ of interestingness.

4.3 In-depth analysis

As illustrated in Section 4.1, subspace clusterings visualized in MDS images typically form groups of similar subspace clusters. For in-depth analysis, visualizing characteristics of each of these groups of subspace clusters in a compact way is important. Characteristics are joint and distinct properties of groups of subspace clusters. Properties of importance are the interestingness $\mathcal{I}$ of a subspace cluster as well as the subspace and object overlap. Visualizing these properties is necessary for in-depth analysis of the overall subspace clustering.

Subspace clusters are grouped if they share dimensions or if they have objects in common, as defined by our subspace cluster distance function. To visualize these groupings such that similarities show up as clear visual patterns, we propose to plot all groups of subspace clusters and the objects contained in each group. Each group is defined as the set of all subspace clusters similar to a subspace cluster of high interest (see Def. 5). We call the most interesting subspace cluster of a group the *anchor* of the group. A *group* is then defined as all subspace clusters of at most ϑ distance to the anchor.

To visualize a group of subspace clusters we use a matrix representation. We start with the subspace cluster having the highest interest. Hence this cluster is the anchor of the first group. Starting with the anchor each subspace cluster of a group is represented by its objects. Each object is depicted by one row whereas the columns illustrate the dimensions of the object. To allow an in-depth analysis we use different color codes to visualize all characteristics of an object, that is its interestingness as well as the values in each dimension. If a dimension is not part of a cluster the column is black.

To highlight subspace clusters of high interestingness we use the factor $\mathcal{I}$ as saturation and value in HSV color space [10]. Hence interesting subspace clusters can be easily identified as the active dimensions are represented by bright and intensive colors. To represent the value for each dimension we use the hue value according to the HSV space. By using all possible hue values the complete color range is used to encode the value of an object but other mappings can be used as well. Further on, objects are ordered with descending interest. Redundant objects that are already visualized

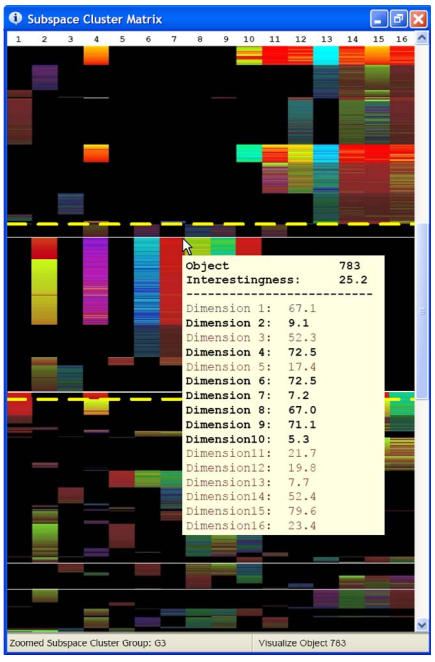

Figure 5: User Interface for a Subspace Clusters Matrix: By pointing on individual objects additional information about the cluster is visualized

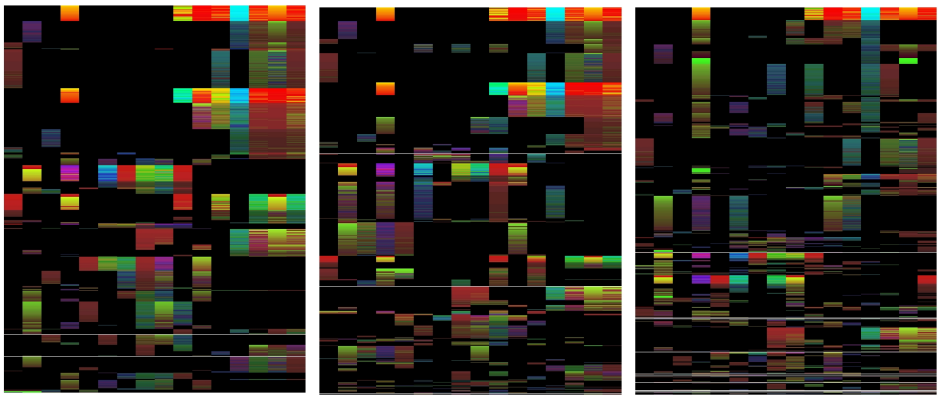

Figure 6: Bracketing Matrix for $R = 0\%, 5\%, 10\%$: for each parameter setting, the output is illustrated, allowing users to pick appropriate settings from the series of images easily

in other clusters are omitted in order to obtain a compact and descriptive representation of the complete clustering. Once all objects of a subspace cluster have been presented the next subspace cluster with respect to the next anchor is displayed.

Figure 4 illustrates the subspace cluster groups for the Pendigits data set using $R = 0\%$. As discussed, a group of subspace clusters typically has some dimensions in common (the *core dimension* of a group). Our proposed subspace clustering matrix allows for visual recognition of core dimensions. Consider the group $G3$ zoomed in on the right side of Figure 4, dimensions 2, 4 and 6 to 10 can be easily identified as core dimensions. By comparing the different groups illustrated in Figure 4 we can see that many subspace clusters identified correlations in dimension 12 to 16 (e.g. $G1$, $G2$, etc.), while some other groups also have a core containing the dimensions 6 to 10 (e.g. $G3$). Further on, a density connected fullspace cluster (a cluster considering all dimensions) is not found (no completely colored row contained in Figure 4). Additionally, detailed information about an object can be obtained by moving the mouse over the object. Figure 5 illustrates a user interface for visualizing a subspace cluster matrix. Information about the intersestingness and the values of an object can be presented in a tool tip. Hence, details about the *center of a group* (i.e. the objects of highest interest) can be easily obtained by the user by moving the mouse over the first objects of a group. For precise selection of individual objects the scroll bar can be used for an interactive zoom of specific areas of the subspace clustering matrix (illustrated in Figure 5 by the area surrounded by the yellow lines). The area above and below the yellow lines shows the original compact representation as depicted in Figure 4. In-between the lines, rows are enlarged by a specified factor. Scrolling up or down, the zoom area can be varied.

Bracketing as shown in Figure 6 illustrates the subspace cluster groups for different redundancy setting (left part $R = 0\%$, middle $R = 5\%$ and right part $R = 10\%$). As can be easily seen, a clear clustering structure is obtained if redundancy is removed (leftmost matrix). Small subspace groups are removed if less redundancy is allowed. Hence, removing redundancy improves the clustering structure. We validate this visual impression of improved quality by additional measurements on the accuracy and quality of DUSC, comparing it to two recent subspace clustering algorithms, SUBCLU [12] and SCHISM [23]. In addition to the previously used Pendigits data, we used Glass and Vowel from [20] and Shapes from [15]. Quality is determined using the entropy, i.e. $H(\mathbf{C}) = -\sum_{i=1}^{k} p(i|\mathbf{C}) \cdot \log(p(i|\mathbf{C}))$ for k class labels in cluster $\mathbf{C}$. For a set of clusters we take the average entropy weighted by the number of objects per cluster. For readability, inverse entropy is normalized to a range of 0% to 100%.Coverage (C) is the percentage of objects in any subspace cluster. It indicates the ratio of clustered objects to noise. The amount of noise in a data set is typically not known apriori, but noise is present in most real world data sets. As sparsely populated regions often exhibit a weak correlation to the class label, quality may increase if coverage decreases.

The first column in Figure 7 shows the best quality (Q) results of DUSC with $R = 0\%$ as also seen from bracketing redundancy. Allowing more redundancy, coverage increases and quality goes down slightly. However, even for $R = 10\%$ DUSC shows better quality than the competing algorithms. The fact that coverage is not 100% indicates that DUSC can distinguish between noise and clusters in subspaces of varying dimensionalities as also illustrated by Figure 6. The pendigits data set, for example, contains handwritten numbers, some of which are clearly different from the rest of the data set. Biased algorithms like SCHISM and SUBCLU do not detect noise, but assign all objects to clusters. The last data set SHAPE contains rotated versions of 9 different shapes, but only 3 of the shapes clearly form clusters. Thus most of the objects have to be considered noise. DUSC detects the given clusters correctly while SCHISM detects only a small part of the clusters and SUBCLU mixes up clusters with noise (less than 100% quality).

	DUSC 0%		DUSC 5%		DUSC 10%		SUBCLU		SCHISM	
	Q	C	Q	C	Q	C	Q	C	Q	C
Pendigits	86	74	83	87	81	92	58	100	77	100
Glass	60	87	51	90	50	93	44	100	44	99
Vowel	82	70	79	100	74	100	10	100	42	100
Shape	100	31	100	31	100	31	98	82	100	1

Figure 7: Accuracy for real world data

Thus the visual impression of our subspace clustering visualizations (see Figure 6) can be confirmed by these measurements.

5. CONCLUSION

We introduced the first subspace clustering visualization to the best of our knowledge. Based on comparable results, i.e. unbiased subspace clustering, browsing of the result is possible through a novel distance function that reflects the subspace and the object overlap, respectively. Subspace clustering interestingness is incorporated to show the most relevant results. Interaction is possible through zooming in to objects in subspace clusters and through choice of adequate subspace clustering settings and feedback.

6. REFERENCES

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM Intl. Conf. on Management of Data*, pages 94–105, Seattle, Washington,USA, June 2-4 1998.

[2] I. Assent, R. Krieger, B. Glavic, and T. Seidl. Spatial multidimensional sequence clustering. In *Proc. 6th IEEE Intl. Conf. on Data Mining - Workshops*, pages 343–348, 2006.

[3] I. Assent, R. Krieger, E. Müller, and T. Seidl. Dusc: Dimensionality unbiased subspace clustering. In *Proc. IEEE Intl. Conf. on Data Mining*, Omaha, Nebraska, USA, 2007.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *Proc. 7th Intl. Conf. on Database Theory*, pages 217–235, Jerusalem,Israel, 1999.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2001.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proc. National Academy of Science of the USA*, volume 95, pages 14863–14868, 1998.

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proc. 2nd Intl. Conf. on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, USA, 1996.

[8] U. Fayyad, G. G. Grinstein, and A. Wierse, editors. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[9] M. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[10] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics (2nd ed. in C): Principles and Practice*. Addison-Wesley, Boston, MA, USA, 1996.

[11] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, USA, 2001.

[12] K. Kailing, Kriegel, H.-P., and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *Proc. 4th SIAM Intl. Conf. on Data Mining*, pages 246–257, 2004.

[13] D. Keim. Information visualization and visual data minig. *IEEE Transactions on Visualization and Computer Graphics*, 8:1–8, 2002.

[14] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proc. IEEE Intl. Conf. on Information Visualization*, pages 9–16, 2006.

[15] E. Keogh, L. Wei, X. Xi, S.-H. Lee, and M. Vlachos. LB_Keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *Proc. 32nd Intl. Conf. on Very Large Data Bases*, pages 882–893, 2006.

[16] B. Kovalerchuk and J. Schwing. *Visual and Spatial Analysis - Advances in Data Mining, Reasoning, and Problem Solving*. Springer, 2004.

[17] H.-P. Kriegel, P. Kröger, M. Renz, and S. Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In *Proc. 5th IEEE Intl. Conf. on Data Mining*, pages 250–257. IEEE Computer Society, 2005.

[18] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In *Psychometrika*, volume 29, pages 1–27. Springer New York, 1964.

[19] M. Lucente. *Diffraction-Specific Fringe Computation for Electro-Holography*. PhD thesis, Massachusetts Institute of Technology, 1994.

[20] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 2006.

[21] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Exploration Newsletter*, 6(1):90–105, June 2004.

[22] J. C. Roberts. Exploratory visualization using bracketing. In *Proc. ACM Working Conference on Advanced Visual Interfaces*, pages 188–192, New York, NY, USA, 2004. ACM.

[23] K. Sequeira and M. Zaki. SCHISM: A new approach for interesting subspace mining. In *Proc. 4th IEEE Intl. Conf. on Data Mining*, pages 186–193, Washington, DC, USA, 2004. IEEE Computer Society.

[24] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[25] T. Soukup and I. Davidson. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. Wiley, 2002.

[26] M. Zaki, M. Peters, I. Assent, and T. Seidl. CLICKS: An effective algorithm for mining subspace clusters in categorical datasets. *Data and Knowledge Engineering*, 60:51–70, January 2007.