Subspace Search and Visualization to Make Sense of Alternative Clusterings in High-Dimensional Data

Andrada Tatu[*]
University of Konstanz
Germany

Fabian Maaß[†]
University of Konstanz
Germany

Ines Färber[‡]
RWTH Aachen University
Germany

Enrico Bertini[§]
University of Konstanz
Germany

Tobias Schreck[¶]
University of Konstanz
Germany

Thomas Seidl[‖]
RWTH Aachen University
Germany

Daniel Keim[**]
University of Konstanz
Germany

Abstract

In explorative data analysis, the data under consideration often resides in a high-dimensional (HD) data space. Currently many methods are available to analyze this type of data. So far, proposed automatic approaches include dimensionality reduction and cluster analysis, whereby visual-interactive methods aim to provide effective visual mappings to show, relate, and navigate HD data. Furthermore, almost all of these methods conduct the analysis from a singular *perspective*, meaning that they consider the data in either the original HD data space, or a reduced version thereof. Additionally, HD data spaces often consist of combined features that measure different properties, in which case the particular relationships between the various properties may not be clear to the analysts a priori since it can only be revealed if appropriate feature combinations (subspaces) of the data are taken into consideration. Considering just a *single* subspace is, however, often not sufficient since different subspaces may show complementary, conjointly, or contradicting relations between data items. Useful information may consequently remain embedded in *sets of subspaces* of a given HD input data space.

Relying on the notion of subspaces, we propose a novel method for the visual analysis of HD data in which we employ an interestingness-guided subspace search algorithm to detect a candidate set of subspaces. Based on appropriately defined subspace similarity functions, we visualize the subspaces and provide navigation facilities to interactively explore large sets of subspaces. Our approach allows users to effectively compare and relate subspaces with respect to involved dimensions and clusters of objects. We apply our approach to synthetic and real data sets. We thereby demonstrate its support for understanding HD data from different perspectives, effectively yielding a more complete view on HD data.

Index Terms: H.2.8 [Database Applications]: Data mining; H.3.3 [Information Search and Retrieval]: Selection process; I.3.3 [Picture/Image Generation]: Display algorithms

1 Introduction

The analysis of high-dimensional (HD) data is an ubiquitously relevant, yet notoriously difficult problem. Problems exist both in automatic data analysis and in the visualization of this kind of data. On the visual-interactive side, a limited number of available visual variables and limited short-term memory of human analysts make it difficult to effectively visualize data in high numbers of dimensions. For automatic pattern detection, a typically employed paradigm is the one of clustering, which identifies groups of objects based on their mutual similarity. Unlike traditional clustering methods, for the mentioned HD data considering all features simultaneously is not effective anymore due to the so-called curse of dimensionality [3]. As dimensionality increases, the distances between any two objects become less discriminative. Moreover, the probability of many dimensions being irrelevant for the underlying cluster structure increases.

Global dimensionality reduction or feature selection methods do not solve this problem as clusters may be located in *different subspace projections* of the feature space, i.e., projections obtained by considering subsets of the original dimensions. For such scenarios the clustering structure tends to be obfuscated in the original feature space and traditional clustering algorithms as well as visual analysis based on the full-space may fail. For large feature spaces, interesting patterns may often be located only in subspace projections of the data. As insights may not be hidden in only one single subspace, relevant analysis should consider also multiple subspaces and their interrelations. Especially, for HD data we can expect to have different views on the same data [11, 21], i.e., the same objects might group differently given different subspace perspectives (see Figure 1 for an illustration). The existence of alternative relevant subspaces may stem from the data description process, when during preprocessing, features (dimensions) which describe different semantic properties of the data, are combined. For instance, in demographic analysis, households are often described by an array of many variables, combinations of which constitute different conceptual domains, such as wealth, mobility, or health. Likewise, it may be the combination of otherwise not semantically related dimensions, which by their combination give rise to interesting patterns. In the Data Mining community, a class of so-called *Subspace Analysis* algorithms has been proposed to cope with the problem of identifying interesting subspaces and clusters from a HD data set. To date, however, there has been a very limited focus on the presentation and interpretation of the generated output. Furthermore, subspace analysis often produces highly redundant results that need to be further manipulated in order to get meaningful results [19].

We propose an initial step towards the use of Visual Analytics as a way to explore alternative views generated by subspace analysis algorithms. We define an analytical pipeline made of algorithmic and visual components that permits to single out and explore alternative views in the data. After being analyzed by a subspace search algorithm, the data is structured and further processed in an interactive visualization environment to reduce redundancy.

The main contribution of this paper is the operative definition and implementation of this multistep pipeline which permits to sift through an exponential number of subspace candidates and to reduce the problem to a handful of relevant views. More specifically, we (1) introduce a mechanism to deal with subspace redundancy

[*] e-mail: tatu@inf.uni-konstanz.de
[†] e-mail:maass@inf.uni-konstanz.de
[‡] e-mail:faerber@cs.rwth-aachen.de
[§] e-mail:bertini@inf.uni-konstanz.de
[¶] e-mail:tobias.schreck@uni-konstanz.de
[‖] e-mail: seidl@cs.rwth-aachen.de
[**] e-mail: keim@inf.uni-konstanz.de

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

by defining topological and dimensional subspace similarity and by allowing flexible and interactive subspace aggregation; (2) we provide a well-reasoned interactive visualization environment that permits to compare and assess alternative views by visually comparing topological and dimensional similarities and strike a balance between visual complexity and level of detail.

We evaluate our method through two case studies. The first is based on synthetic data to check whether the tool does what it is supposed to do. The second is based on real-world data to demonstrate how the tool can help finding and interpreting alternative views in HD data. We believe these results show the potential of Visual Analytics in the context of automated mining algorithms. It furthermore shows how the use of Visual Analytics can enhance the understanding of the results of automated data analysis methods, and lead to new questions concerning more effective or more efficient algorithms.

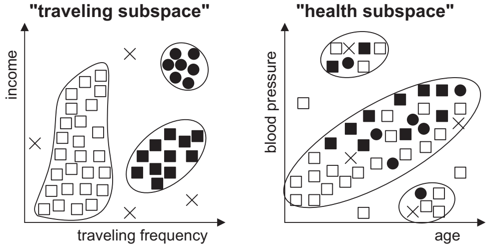

Figure 1: Alternative data distributions and groupings [20] in two different subspaces of a larger HD data space (domain here: demographic data analysis). Our proposed visual analysis method integrates the notion of alternative subspaces into the analysis process and links it to the task of comparative cluster analysis.

2 Subspace Analysis

In this section, we discuss the challenges for visual subspace analysis in more detail and explain how we tackle these with our new interactive, explorative framework supported by subspace search algorithms.

As is commonly known in subspace clustering, dealing with HD data in its subspace projections faces two main challenges. The first, serious challenge is a reasonable scalability w.r.t. the dimensionality of the data set. As for a d-dimensional data set the number of possible subspaces $S \subseteq \{1,\ldots,d\}$ is $\sum_{k=1}^{d}\binom{d}{k} = 2^d - 1$, many subspace clustering approaches do not scale well for very HD data. Every algorithm has to employ some strategy and heuristics to cope with such an exponential search space. The second, closely related challenge is dealing with high redundancy, that stems from the high similarity of the exponentially many subspaces. If two subspaces share a high proportion of dimensions, they are likely to exhibit a very similar clustering structure [11]. A large search result with high redundancy is, however, not beneficial for the user as it masks the complete information and is hard to interpret.

A core task in analysis of HD data is to apply a clustering method to reduce data complexity and identify groups of data for comparison. Different clustering algorithms follow different clustering notions, e.g., there exist density- (e.g., DBSCAN [9]) or compactness-based (e.g., k-Means) clustering methods, and their outcomes often depend crucially on non-intuitive parameter settings. Usually several clustering attempts are required until the user has a usable result. It is obvious that high runtimes of subspace clustering processes (see Section 6.3) are not tolerable for such a workflow. Consequently, we decided to start the visual data exploration one step *before* the actual clustering process and decouple subspace search and the actual clustering. Dedicated subspace search algorithms [2, 7, 15] have been designed to efficiently filter and rank the possible subspaces according to specific quality criteria (or interestingness measures, see also below). After subspace search has taken place, an arbitrary clustering approach can be used to cluster in the identified subspaces.

The use of subspace search for our purposes has several advantages: (1) It helps to effectively filter out those subspaces that based on low interestingness do not need to be considered by the user. (2) Subspace search approaches are designed to reduce the search space efficiently and they do not need to compute clusters. And (3) although, subspace search approaches themselves also rely on certain assumptions of what makes a subspace interesting, these assumptions do not necessarily lead to very different subspaces among different approaches. Therefore, the results are not as biased as they are for different clustering algorithms, which enables the user to already obtain valuable results with one subspace search approach. For example, the quality assessment based on the k-NN distance [2], favors neither the DBSCAN nor the k-Means clustering notion. And (4), integrating the subspace search into the HD analysis offers the user the opportunity to obtain a visual, intuitive overview of the clustering structure *before* even starting the actual clustering. Thus, the user can assess the potential of the data to deliver valuable clustering results at all; decide which subspaces are to be clustered; decide which clustering notion to follow in each subspace (since the notion does not need to be the same for all); more easily determine meaningful parameter settings for clustering approaches.

Subspace search methods guide their search process by specific interestingness scores that are defined heuristically. For example, the method proposed in [7] considers as interestingness score the variation of the density of objects across a regular cell-based partitioning of a given subspace. The underlying assumption is, that higher variation of density provides higher probability that the subspace shows meaningful structure. As another example, the SURFING method [2] relies on the histogram of the k-nearest neighbor distances for all objects in a given subspace. It considers subspaces with non-uniform distance distributions more interesting (as they are an indication of the presence of strong clusterings). The underlying assumption is that for subspaces that show meaningful structures (e.g., clusters), different k-NN distances will occur. These and other measures aim at identifying subspaces that show a high "contrast" with respect to the distribution of objects, allowing to spot meaningful structure in the subspaces.

Subspace search methods also typically contain heuristic approaches for early abandoning uninteresting subspaces, as exhaustive search would be prohibitively expensive. SURFING for example is based on a bottom-up strategy for searching subspaces by increasing dimensionality. It is based on testing additional dimensions for subspaces already known to be interesting. The list of currently interesting subspaces is continuously pruned to keep only the most interesting subspaces and speed up the search. SURFING has no dimensionality bias, assumes no specific clustering structure and in practice, it is parameter free. Due to these properties, we rely on this method in our proposed approach, using the implementation provided to us by the original authors, but other subspace search algorithms could be easily used as well.

Overall, using the results of a subspace search algorithm as a starting point for our visualization has many advantages. Subspace search methods such as SURFING employ efficient search strategies tackling the efficiency challenge of subspace analysis. However, they typically do not solve the challenge of high redundancy. This is exactly where our proposed visual analytical workflow introduced next, starts from.

3 Proposed Analytical Workflow

We propose a carefully designed visual-analytics workflow for subspace-based exploration of HD data, making use of algorithmic subspace search in combination with visual-interactive representations for user-based filtering and exploration. Our approach starts

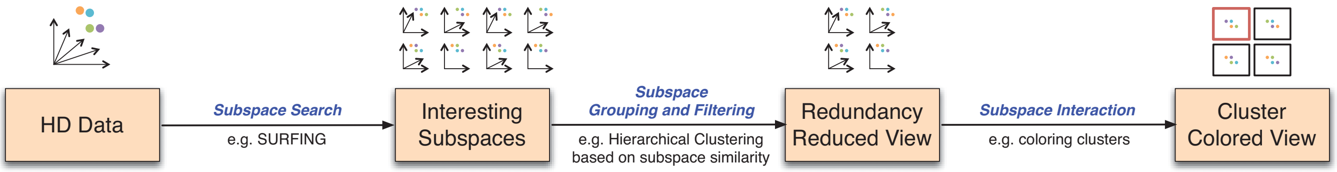

Figure 2: Our proposed analysis pipeline. A subspace selection algorithm is applied to automatically identify a candidate set of interesting subspaces. A filtering step reduces the potentially large and redundant set of automatically obtained subspaces to a user-selectable number of representing subspaces. Visual-interactive user exploration then proceeds on the subspace representations. Subspace analysis is also supported by comparative cluster views, allowing users to identify meaningful similar, complementary or even conflicting clustering structures in the set of subspaces.

(1) with an automatic *subspace search step*, where a large number of interesting subspaces is selected by a subspace search algorithm. Current subspace search methods provide an algorithmic handling of the problem of finding interesting subspaces, yet they often produce too many subspaces that may also be redundant and thereby overwhelm the interactive analysis (see also Section 2). We therefore employ *similarity-based grouping of subspaces* (2) and perform the interactive exploration of interesting subspaces based on a few group representatives. Appropriate visual representations and interactions support the *visual interactive analysis* (3) for better understanding the subspace search results, including the support for comparative cluster analysis.

Figure 2 depicts our proposed analytical workflow. We next detail the technical design decisions made for each of the analysis steps, including discussion of alternatives.

3.1 Generation of interesting subspace candidates

The advantages for choosing subspace search, and in particular SURFING, have been already discussed in detail in Section 2. We observe that typically subspace search algorithms output a huge number of subspaces. Since the examination of all subspaces is infeasible, a common approach is to filter the subspaces based on a certain threshold. This, however, ignores the fact, that the first ranked subspaces might be only slight variations (i.e., high overlap of dimension sets) of the same subspace and therefore are redundant to each other. Yet, interesting subspaces with substantially different dimension sets, as compared to the top ranked results, could be found at much later ranking positions, and run the risk to be neglected from the analysis. Therefore, we apply a grouping step based on an appropriately defined notion of subspace similarity, as described next.

3.2 Similarity-based subspace grouping and filtering

Given a large number of candidate subspaces, we apply hierarchical grouping and filtering to yield a smaller set of mutually sufficiently different, yet individually interesting groups of subspaces for interactive analysis. Our filtering and grouping operation is based on a custom similarity function defined on pairs of subspaces according to two main criteria: (1) overlap of the sets of dimensions that constitute the respective subspaces, and (2) resemblance in the data topology given in the respective subspaces.

(1) Similarity based on dimension overlap: Subspaces can be similar regarding their constituent dimensions. We use the *Tanimoto Similarity* [23] on bit vectors indicating the contained (active) dimensions in a respective subspace (1 denotes an active dimension, 0 the converse). The Tanimoto Similarity is then computed as the fraction of dimensions contained in both subspaces (AND-ing of the bit vectors), among the total number of different dimensions occurring in the subspaces (OR-ing of the bit vectors).

(2) Similarity based on data topology: We also compare subspaces with regard to their data distribution. Specifically, we consider the similarity of k-NN relationships in the respective subspaces. For efficiency reasons, we compute the k-nearest neighborhood ($k = 20$) lists for a sample of 5% of the contained data points. The similarity between two subspaces is then evaluated as the average percentage of agreement of k-NN lists in the subspaces. This score measures the similarity of the k-NN topology of the data, where k is a parameter and can be adapted to the data sets at hand by the user. Note that also other similarity measures are in principle possible. For instance, the data could be clustered and the similarity between subspaces evaluated according to the resemblance of obtained clusterings by an appropriate measure such as the RandIndex [22].

These two distance functions are the basis for the subspace grouping step in our analytical workflow as follows:

(1) Subspace grouping: We apply hierarchical agglomerative grouping of subspaces based on the topologic distance function using Ward's minimum variance method [30]. Based on the dendrogram representation of the obtained hierarchical grouping, the user chooses the hierarchy depth level to select a number of groups. This way the user can easily decide how many clusters are desired for the analysis.

(2) Subspace filtering: Based on the previously achieved grouping of subspaces, we filter one subspace from each group as representative: For each group we consider the subspaces with the lowest dimensionality and choose the one which exhibits the highest interestingness score. We note that other rules for filtering representatives are possible, but find that this rule is robust and effective for users, as it tries to keep the dimensionality as low as possible.

These steps together with both distance functions take us further towards our goal of understanding the different kinds of relationships between subspaces. They can complement, confirm, or contradict each other and being aware of these relations can be crucial for further mining tasks.

		contained dimensions	
		similar	not similar
data topology	similar	truly redundant	confirmatory
	not similar	dominant dimensions	truly complementary

Figure 3: Filtering cases that can be supported by our two defined subspace similarity functions.

Four basic cases can be identified, each of which might be relevant for a given subspace analysis task: (1) Subspaces that are similar in both, their contained dimension sets and their data topology (truly redundant subspaces); (2) Subspaces that are dissimilar in both, their contained dimensions and their data topology (truly complementary subspaces); (3) Subspaces that are similar w.r.t. data topology but dissimilar regarding their contained dimensions (confirmatory subspaces: we confirm the same data relationships in different subspaces); and (4) Subspaces that are similar w.r.t. their contained dimensions, but dissimilar regarding topology (this is generally not expected but could indicate the existence of one or a few dimensions which are by their nature very dominant for the data topology). Figure 3 illustrates these four basic filtering cases.

3.3 Visual-interactive design

After hierarchical aggregation and/or filtering of the potentially redundant set of subspaces have taken place, we apply a set of analytical views for exploring and comparing the subspaces. Our displays are based on (1) scatterplot-oriented representations of individual subspaces or groups of subspaces, (2) similarity-based or linear list layouts for sets of subspaces, and (3) additional informative views (parallel coordinates and color-coding for comparison of groups in data).

The proposed design is the result of several iterations of alternative solutions in which we explored and compared several representations. Two design choices are worth discussing here: (1) the design of a visual representative for subspaces and (2) their layout. We decided to represent subspaces with scatter plots because they allow for the identification and comparison of groups in the data. More abstract representations (like simple colored marks) would require less space but would not allow the rich topological comparison provided by the scatter plots. In contrast, representations that are more complex like, e.g., parallel coordinates would provide a direct representation of the dimensions included in the subspace but would make their representation much more cluttered. As for the layout, we tried several tree and graph layouts to make the relationship between the subspaces and their shared dimensions explicit but we found that this rarely provides interesting insights and makes the visualization too cluttered to be of any use.

Scatter plots for subspaces can be generated by any appropriate projection technique such as PCA [14], MDS [8] or t-SNE [29], to name a few. We currently use MDS, but we experimented with others and any other technique could be used as an alternative. For a group of subspaces, one representative subspace is chosen (see below). To convey the involved dimensions, we also add an index glyph to the respective scatter plot (see Figure 4).

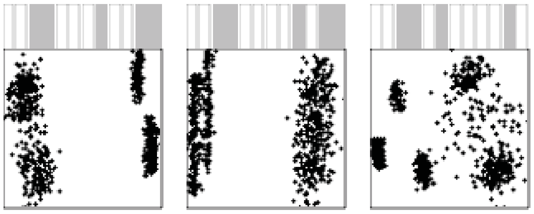

Figure 4: Subspace representation by 2D scatterplots with dimension glyph. We can see two 5D subspaces (left) and one 4D subspace (right) in the visual representations.

The analytical views are combined and linked in an application that consists of the following components:

Linearly sorted view of subspaces. To obtain a first overview of the output of the subspace search algorithm, we present all the subspaces in a linear view. The MDS scatter plots representing the individual subspaces are sorted left-to-right and top-down according to the interestingness index provided by the subspace search method. This view is exclusively used as a detail view for groups of topologically similar subspaces. Figure 5(1) illustrates the subspaces of the synthetic data set, which is described also later in Subsection 4.1.

Subspace group view. In this view, groups of subspaces that have been formed by hierarchical agglomerative grouping are shown. Each group is represented by one selected subspace from that group, using the filtering method as described in the previous Subsection.

The representative subspaces are each visualized by an MDS plot, and shown side-by-side (Figure 6(1) illustrates). A dimension histogram on top of it indicates the distribution of dimensions contained by the subspaces in the group, where the length of the bar encodes the frequency of the respective dimension. The last bar encodes the percentage of subspaces contained in this group. It is colored in orange to be easily distinguished from the others. Each group of subspaces from the preceding view can be expanded and its member subspaces can be seen and compared in detail (as Figure 6(5) illustrates). This allows a better understanding of the current similarity threshold, and allows to expand or further collapse the group structure based on visually perceived similarity between subspaces. The user can investigate how similar the distribution of dimensions is among different groups of subspaces. To this end, a click on the dimension histogram icon of one particular group will cross-highlight the dimensions of the selected group that are also contained by other clusters. In summary, the subspace group view allows a global comparison of non-redundant subspaces and their similarities concerning the contained data topology.

Dimension-based subspace similarity view. We also support the comparative analysis of all subspaces based on their similarity regarding the set of active dimensions. To this end, a global MDS layout, based on the Tanimoto distances between the subspaces, as described in Section 3.2, is generated. Figure 6(4) illustrates the subspace similarity view. For a high number of subspaces, this view can only provide an impression of the similarity relationships but by zooming more details become visible. The subspace group view (based on data topology distance) and dimension-similarity view (based on Tanimoto distance) are linked by color-coding (outer frame coloring). Thereby, we can compare the similarity of subspaces by their topological and dimension-overlap-based similarity.

Additional views and cluster comparison support. We also integrated details-on-demand for each subspace by a *parallel coordinates view* (Figures 5(3) and 6(3) illustrate). Highlighting contained dimensions helps to understand the difference of the subspaces in more detail. Furthermore, interactive exploration of the subspaces is enhanced by a *single subspace view*, providing an enlarged view of a selected subspace scatter plot (Figures 5(2) and 6(2) illustrate this). This view also allows to manually select clusters of objects by a lasso tool. Cross-coloring of the selected points among the other subspaces and within the parallel coordinates plot thus allows comparative exploration of grouping structures – a core problem in making effective use of alternative subspaces.

4 Application

We now demonstrate the analytical capabilities of our proposed approach. First, we use synthetic data as a proof of concept and exemplify the suggested workflow. We show how that relevant subspaces can conveniently be identified. Then, we describe an explorative setting in which interesting findings in alternative subspaces of a real world data set are obtained.

4.1 Application Scenario 1: Synthetic Data

We used a 750 record sample of the first 12D synthetic data set presented in [10] (data set No. 2). This data set consists of four 3D Gaussian clusters and two 6D Gaussian clusters. The remaining dimensions contain uniformly distributed random noise. The first step of our approach is to determine the interesting subspaces of the high-dimensional data set, by running automatic subspace search using SURFING (see Section 3). This subspace search returns a

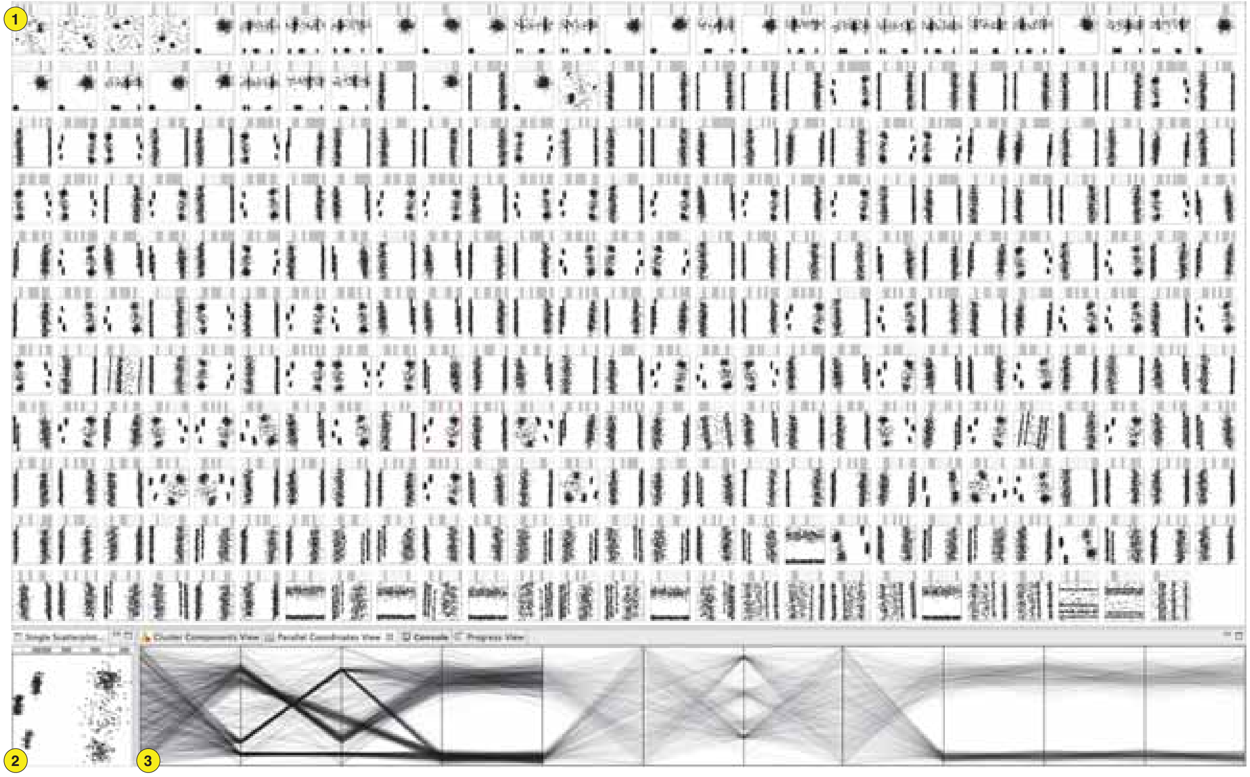

Figure 5: (1) *Linearly sorted view* of subspaces for the 12D synthetical data set from [10] showing the full result of SURFING, consisting of 296 subspaces. The selected subspace in this view is shown in a (2) *single subspace view* to enable interaction and in (3) a *parallel coordinates view* with the subspace dimensions as the first axes (highlighted), and all the other data dimension as the last axes.

total of 296 subspaces identified as interesting, out of the 4095 possible subspaces. To get a first impression of these subspaces, we use the **linearly sorted view of subspaces** shown in Figure 5, relying on MDS representations of the data in the subspaces, and sorted by the interestingness score in decreasing order.

The view shows the diversity of subspaces identified during the automatic step. The first elements in the first row of the view are very similar in terms of the point distribution (showing mostly scattered and spherical point distributions). However, at later positions, we also see other varieties of point distributions, including parallel stripe patterns, and stripes mixed with spherical patterns. In a normal (non-visual) analysis case, relying just on the subspaces ranked top by the interestingness score, the analyst might miss some of these different characteristics of the subspaces.

The overview also confirms that the subspace search did return a lot of redundant subspaces, judging by the shape of the MDS projection representations. The next step is therefore to group the subspaces according to their similarity, allowing the user to abstract to a smaller number of relevant subspaces to compare them in detail. We used our similarity function based on the data topology, creating a hierarchal agglomerative clustering. Figure 6(1) shows that the number of subspaces can be reduced considerably in a meaningful way by the user. The navigation buttons, as shown in Figure 6(6), allow the user to move through each dendrogram level and to find the desired level of redundancy. Here the dendrogram was cut at 0.73, very close to the root. As a result, six groups are found and visualized by their representatives. The number of groups can be variated, and the user can also investigate different levels in the dendrogram hierarchy. In this data we quickly found that six groups is the right level of detail for our further investigation.

We investigate the components of each group of subspaces in more detail. Figure 6(5) shows the group detail view of the orange, green, and purple subspace groups as framed in Figure 6(1). Topologically similar subspaces are grouped together. In this way, the analyst is given an overview of the existing groups and, if needed, can further compare individual group components.

On top of the scatterplots a dimension histogram is indicating the distribution of dimensions for each group. The last bar of the histogram is marked in orange and represents the percentage of subspaces contained in this group. It is scaled logarithmically, so that this bar is also visible for groups with few elements. A click on the dimension histogram of one group representative highlights its dimensions in all the other representatives. In Figure 6(1) the green group was clicked. To understand why the green- and gray-framed groups are split, we can consult the additional view in Figure 6(4). It shows an MDS layout of all interesting subspaces based on the dimension overlap (Tanimoto) similarity. In this view closeness of two subspaces corresponds to dimension similarity. We see that the green- and gray-framed cluster groups are located on the far left side in the plot. This shows us that the subspaces are similar in terms of dimensions, but being in different groups, they must show different topological similarity according to our similarity measure. This can be explained as all the subspaces of the gray-framed group contain dimension $d12$, while none of the subspaces in the green-framed group contain this dimension. This is visible by the bars in the dimension histogram of the gray-framed group. As it is not highlighted, it is not contained in the marked green-framed group. This dimension is obviously responsible for a different data distribution.

We can also go one step further in detailed comparison of subspaces by cross-color-coding clusters of points in the MDS representation. Our lasso tool allows the user to manually mark clusters of points in the MDS subspace representation, which allows to cross-compare the groupings among different subspaces. For example, we manually marked six separate clusters of points in the pink-framed subspace group (group number two in Figure 6(1))

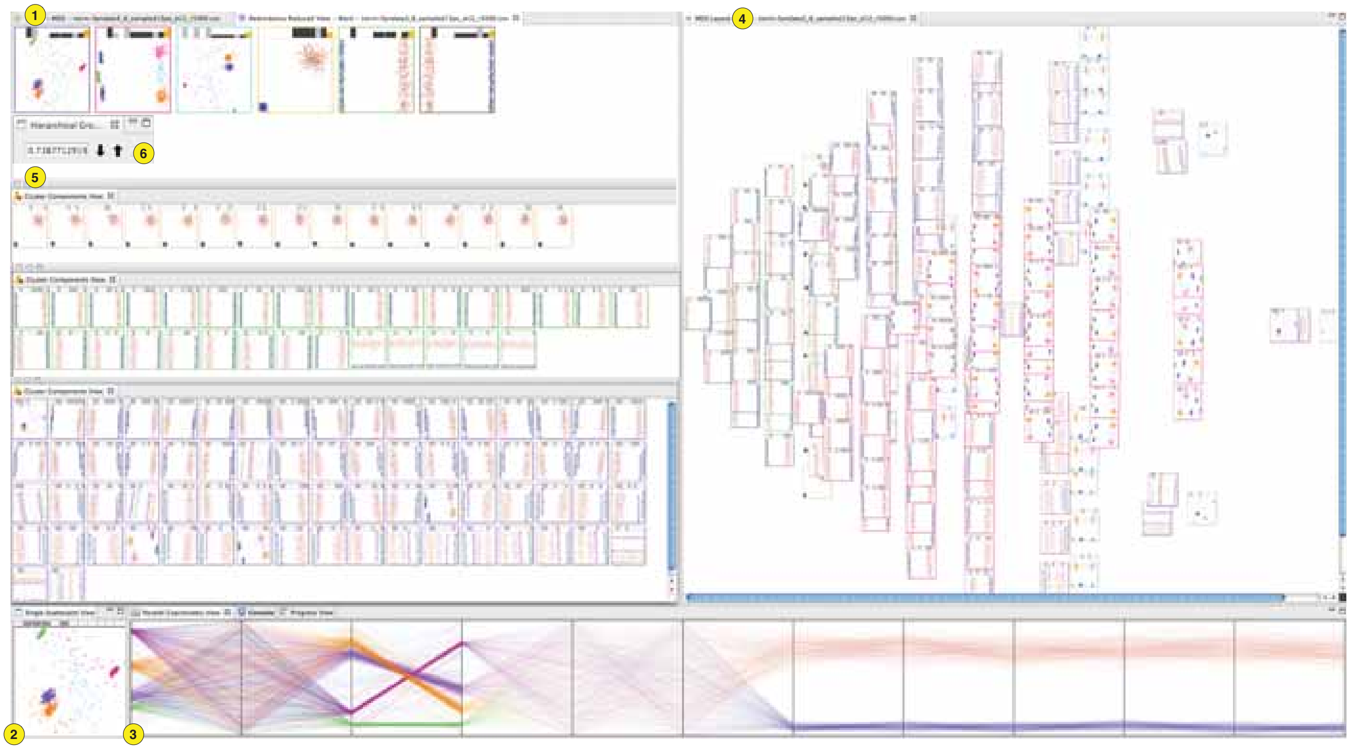

Figure 6: (1) *Subspace group view* for the $12D$ synthetic data set with six subspace groups. (2) *Single subspace view* showing the representative subspace for the first group. (3) Details-on-demand in the *parallel coordinates view* for the selected subspace. (4) The MDS layout of the subspace search results based on their dimension similarity. (5) *Group detail view* for the three (orange, green, purple) subspace groups. (6) Hierarchical navigation buttons.

and assigned distinct colors. By analyzing the distribution of colors among subspace group representatives, we see that other subspaces merge some of these clusters and spread others. This is also true for the purple framed group representative. The dark blue and pink point cluster (the upper most in the original colored subspace) are clustered in the purple subspace but some of their points also became noise in this subspace.

Summing up, we can see how our visual analytics workflow helps to deal with the extensive number of possibly interesting subspaces in a natural overview-first based visual analytics workflow. In a first step, the SURFING approach reduced the number of subspaces of the 12 dimensional data set from 4095 to 296 interesting ones. Since this set of subspaces still showed a high redundancy, in our next step we grouped them using our topological similarity measure. Based on the grouped subspaces, further investigations coud take place for comparing the relations and distributions among points of data within the subspaces.

4.2 Application Scenario 2: Exploration/discovery

We will now demonstrate the exploratory functionalities of our proposed approach based on a real data set. We analyze the USDA Food Composition Data Set (http://www.ars.usda.gov/), a full collection of raw and processed foods characterized by their composition in terms of nutrients. The database contains more than 7000 records and 44 dimensions. After removing missing values and outliers, as well as normalization 722 records (foods) remained for which we selected 18 dimensions of the data set that where interpretable to us.

From this input data set, application of the SURFING algorithm returned 216 interesting subspaces for further exploration. To get a first impression of this data, we investigated the *linearly sorted view* (see Figure 8 for a cut-out). Many subspaces, in particular those ranked with a high interestingness index, show a rather skewed distribution of points in our projection representation, concentrating along the edges of the diagrams. Only later in the ranking, we start to see the projections forming out more structure, that could be meaningful. The red color framed subspace in Figure 8 seems to be very interesting, forming long, clear stripes. With the help of the *single subspace view*, we further investigated this subspace $(Iron, Maganase, Vit_D)$ by coloring each stripe with a different color and compared the formation of these clusters across the other subspaces. Most of them seemed to be overspread by the cyan class (see Figure 8 right).

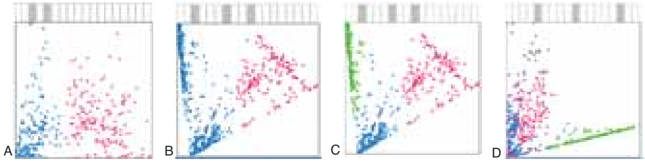

Figure 7: (A) Interesting spotted subspace $(Carbohydrat, Fibre)$ presenting two clusters. (B) Subspace $(Carbohydarte, Lipid, Protein)$ in the same cluster group of (A) where the cluster structure changes. (C) Green marked third cluster in subspace from (B). (D) Subspace $(Fiber, Protein, Vit_D)$ of orange color-framed subspace group, where the alternative clustering of points is visible.

At the same time, it is clear that a high level of redundancy is still present, and a further grouping is deemed necessary. Therefore, we continued with our next analytical step, the subspace grouping by agglomerative hierarchical clustering. We obtained different groups of subspaces and found out that these clearly striped clusters only appear in subspaces containing Vit_D.

We therefore reset the coloring and started a new interactive analysis step, beginning with this stage of our workflow. After testing different filtering thresholds and comparing the topological- and the

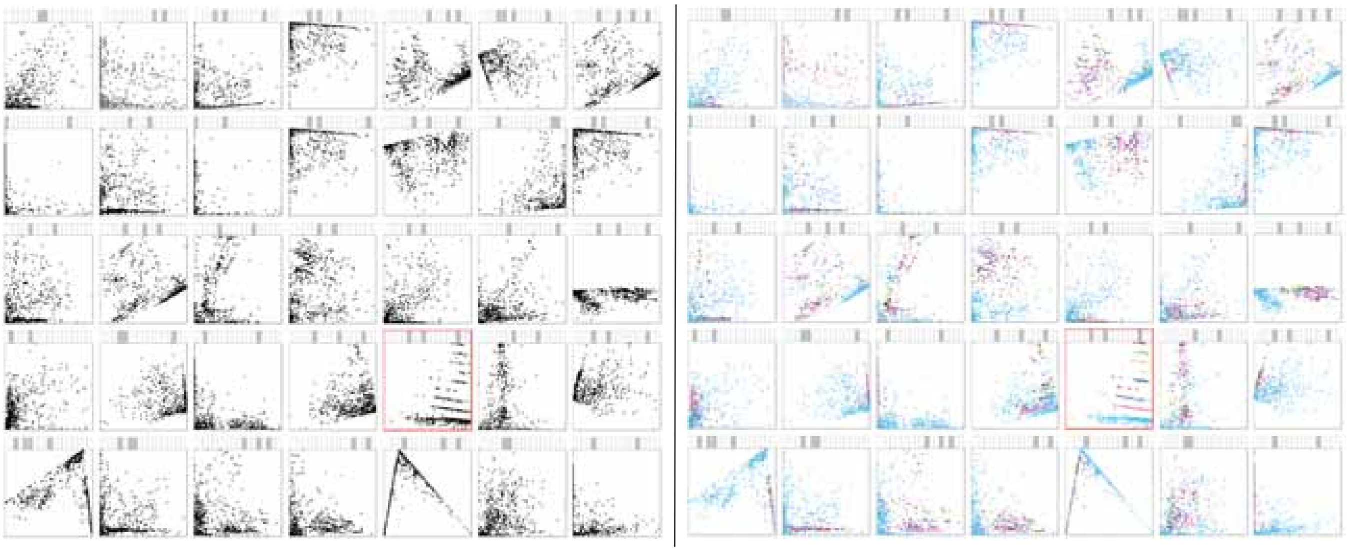

Figure 8: (1) *Linearly sorted view cut-out* of subspaces for the 18D USDA Food Composition Data Set. The full result of SURFING, consisting of 216 subspaces. We see a rather high level of redundancy. Subspaces exhibiting more structure are found in particular at the mid and end positions in the ranking. Relying only on the numerically top ranked results, we would have omitted such interesting cases from the analysis.

dimension-based similarity relations, we obtained a number of 12 groups, and considered this suitable for subsequent analysis.

From the reduced number of representative subspaces, one particular subspace stood out to us (see Fig 9(1) for the group representatives and Fig. 7(A) for the interesting spotted one). This subspace shows the most structure and allows to discern two point clusters (pink and blue). We selected this specific subspace group (framed brown in Figure 9) for further analysis. Cross-coloring is used to highlight its group components, that are shown at the bottom of the figure. It is visible that the group of subspaces are topologically similar, consequently this subspace is a valid representative.

In addition, we observe that there are some subspaces in this group where the clustering is changing. One example is shown in Figure 7(B). We assigned the green color to the outstanding points on the left side, as they seem to form a different structure. In the group view (see Fig. 9(1)) we can see that this green cluster overspreads on five of the 12 subspace group representatives. After a closer look to the components of the orange subspace group, we spotted a sharply defined green cluster (see Fig. 7(D) and highlighted in Fig. 9(2)). By highlighting the dimensions of the orange group, we can see that the brown group has a dominant dimension (*Protein*) that is not contained by any subspace of the orange group. We can therefore assume that this dimension is decisive for the clustering of the points. In the dimension-based similarity view (MDS Layout in Fig. 9(3)) the subspaces of the brown and orange groups are far apart from each other, which supports our finding that the groups contain different dimensions. Likewise we can see that the group components of the brown group are scattered across the MDS layout. This is due to the fact that the group subspaces are dissimilar in terms of their dimensions, but their topological similarity is dominated by the shared dimension (*Protein*).

Summing up, we demonstrated how our interactive, exploratory workflow can be applied to real data. Compared to the previous scenario, the information about the clusters is not known in real data sets, meaning that several interactive attempts are needed to investigate the vast number of interesting subspaces provided by the subspace search algorithm. With the help of the topological similarity functionalities, we could group the redundant clusters and have a closer look in their topological change. Using the different linked views of our approach helped us to identify different subspaces that present alternative clusterings.

5 Discussion and Possible Extensions

We will now summarize the main goal of our system, and discuss limitations and possible extensions next.

5.1 Summarizing the Main Goals of our Approach

Our presented approach supports visual-interactive analysis of HD data from *multiple perspectives* based on the notion of automatic subspace search. The core assumption for our approach is that useful information could be extracted in a comparative way from several different subspaces residing in a larger HD data space. This assumption is the key driving force behind subspace search and subspace clustering algorithms developed in the Data Mining community over the past few years. We exploit algorithmic subspace search in an encompassing visual-interactive system. Our approach is designed around Shneiderman's Visual Information-Seeking Mantra [27], applied to the problem of analyzing potentially large sets of subspaces. Modern subspace search methods such as SURFING efficiently identify candidate subspaces that are expected to exhibit informative structure without restricting on a specific nature of the structure. Specifically, interactively detecting and understanding relevant structures in subspaces is an explicit goal of our system. Our interactive support allows users to condense and compare subspaces, and even groups in data. Thereby, we close the analytical loop from algorithmic search of subspaces to sense-making by the user. Subspace search algorithms are very useful as a starting point. Since the identification based on interestingness is done heuristically, the search methods alone cannot solve the analytical problems at hand. To this end, capable visual-analytic systems need to be designed based on the output of the subspace search algorithm. We therefore designed, implemented, and applied an encompassing system design based on a subspace search method (exemplarily we used SURFING). It allows to explore HD data taking into account the curse of dimensionality and the possibility to find alternative clusters in different subspaces.

5.2 Limitations and Possible Extensions

We identify the following limitations and improvement opportunities for our approach.

Computational scalability. We designed and tested our system around data sets of moderate high-dimensionality of tens of dimensions. For higher-dimensional data, we will have to deal with

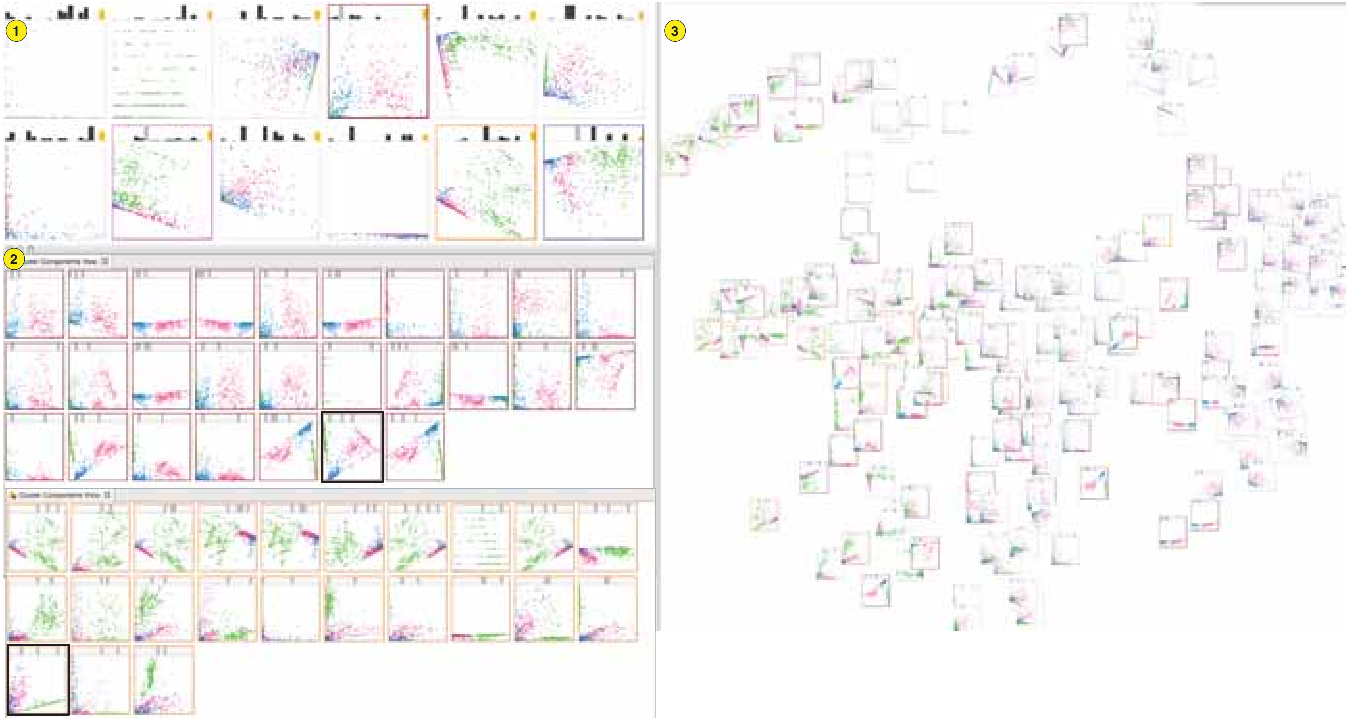

Figure 9: (1) *Grouped view* of subspaces for the 18D USDA Food Composition Data Set with 12 group representatives. (2) The brown and orange group components are shown in the *components view*. (3) MDS Layout of the total number of subspaces with cross-colored group representatives.

scalability issues in (1) computational complexity of the subspace search and (2) scalability of the visual representation of subspaces. Regarding (1), the search space increases exponentially with dimensionality. Subspace search algorithms probably need more aggressive filtering mechanisms to keep the number of searched subspaces tractable. A dynamically adjustable threshold could be useful here. However, we still need to ensure that no relevant results are excluded. To this end, sensitivity analysis is needed.

Visual scalability. Regarding (2), also scalable visual representations are needed for higher-dimensional data. We need to scale with the number of subspaces and the representation of each subspace. Hierarchical grouping of subspaces is already included in our system to scale with the number of subspaces. The linearly sorted view per se does not scale with many subspaces, yet it can be restricted to the representative subspaces obtained from hierarchical grouping. Visual representation of subspaces takes place by projection to show the data points and an index view to show contained dimensions. In particular, the latter will only scale for a limited number of dimensions. How to design set-oriented views to compare many sets of dimensions is a challenging problem that if solved, would improve our tool.

Projection-based subspace representation. We currently represent the subspaces by MDS projections of the data residing in respective subspaces. However, projection typically induces loss in information, that could be incorporated in our visualization, e.g., by showing the stress values in an overlay visualization [24]. In our experiments, MDS performed very well compared to using PCA. Yet, it would be interesting to test other projections. Also, other subspace representations besides scatterplots could be thought of, in essence similar to Value-and-Relation displays [33]. Likewise, many different, useful similarity notions to group and compare subspaces, such as notions based on stress measures, implicit clustering structures, relations to outliers, Scagnostics features [32], etc. could be employed. Testing them in different application domains is considered valuable future work. We note that our analytical approach can easily accommodate alternative subspace search algorithms, representations, and filtering options.

Interpretable Dimensions. To relate subspaces and data groups in subspaces, it is important for the analyst to be aware of the meaning of the dimensions of the respective subspace. Our index-based glyph does not convey information about the type of dimension. More semantically meaningful dimension representations would be useful. Detail-on-demand functions could be added to help the user interpret the involved dimensions and properties of the data points more efficiently.

Definition of interestingness and sensitivity to noise. Subspace search algorithms heuristically identify subspaces as interesting based on certain properties of object relations. Based on the user and application, additional interestingness formulations are possible and should be supported. Following best practices in data analysis, we have applied a data cleaning step (outlier and missing value removal) to our tested data before we fed it into our system. The SURFING algorithm is not robust with respect to missing values, whereas it seems to be robust with respect to outliers. The original paper does not discuss this aspect and we did not further investigate it. The projections used to represent data distributions in subspaces are sensitive to outliers and may generate clamped distributions if not pre-processed. We postpone the analysis of this problem to future work.

Automatic support for cluster comparison. Adding automatic clustering of data points in subspaces would be useful as a post-processing step. Equipped with automatic clustering, we can color-code the found clusters. This could lead to new visual-oriented interestingness measures useful for selecting interesting subspaces in the future. User interaction with the subspace search output could be a useful analytical feature for refinement. Allowing expert users to split or merge subspaces, or construct new subspaces by adding or removing dimensions, would be one option.

Usability and user adoption. Our current system design targets users with expertise in data mining. End-user applications, e.g., in Market Segment analysis, could benefit from subspace analysis. Yet we recognize that for end-users, the interface of our system would need to be customized, possibly. Our experience in collaborating with data mining experts showed that the tool can be useful not only for data exploration but also as an evaluation tool to assess the output generated by subspace analysis algorithms.

6 Related Work

6.1 Visualization and Clustering of HD Data

Visualization of HD data is a long-standing research topic. Classic approaches include parallel coordinates, scatter plot matrices, glyph-based and pixel-oriented techniques [31]. By an appropriate sorting of dimensions and mapping them to visual variables, these methods allow to overview and relate high-dimensional input data, however we may run into scalability problems for large numbers of dimensions or records. Dimension reduction methods such as PCA [14] or MDS [8] can be used to reduce the data to a smaller number of dimensions for subsequent visualization.

Identification and relation of groups of data is a key explorative data analysis task. Often, user interaction is needed to identify and revise the number and characteristics of data clusters found by automatic search methods. To this end, visual-interactive approaches are useful. Although, many methods have been proposed, we can only highlight few of them in an exemplary manner. In [25], interactive exploration of hierarchically clustered data along a dendrogram data structure is proposed to help users find the right level of clusters for their tasks. In [34], the parallel coordinates approach serves as a basic display to show data clustering results allowing to compare clusters along their high-dimensional data space. Also, 2D projections, possibly in conjunction with glyph-based representation of clusters, are widely employed, a recent example is [6].

These approaches to visualization and clustering in HD data spaces all have in common that they are based on a given full (or reduced) dimensionality of the input data set. Thereby, they show only a *singular* perspective of the usually multi-faceted HD data, that might not be the most relevant one. As we show in this paper, it is also useful to explore HD data for patterns in subsets of its full HD input space to increase potential data insight.

6.2 Automatic and Visual-Interactive Feature Selection

In Machine Learning, feature selection is the problem of selecting from a large space of input features (or dimensions) a smaller number of features that optimize a measurable criterion, e.g., the accuracy of a classifier [18]. Most automatic feature selection methods rely on supervised information (e.g., labeled data) to perform the selection. Therefore, they are not directly applicable to the explorative analysis problem. In existing works involving visual-interactive selections or comparison of features, the Rank-by-Feature Framework [26] provides a sorted visual overview of the correlation among pairs of features. In [13], the selection of input features was supported by a measure of the interestingness of the visual view provided by candidate features. An interactive dimensionality reduction workflow was presented in [12], relying on visual approaches to guide users in selecting features.

In [4] and [5], interactive visual comparison was proposed to relate data described in different given feature spaces based on 2D mappings and tree structures extracted from the different data spaces. Furthermore, in [17] a visual design based on network and heat map visualization was proposed to relate clusterings in different subsets of dimensions. In [34], dimensions are hierarchically clustered based on a simple value-oriented similarity measure. Based on this structure, user navigation can take place to identify interesting subspaces. In a recent work [35], the output of this simple search method was visualized by tree- and matrix-based views, where each dimension combination was represented by a single MDS plot.

In summary, many of these methods are applicable to compare data regarding different criteria. However, most of them assume the feature selection to be performed globally and do not take the subspace search problem directly into account.

6.3 Subspace cluster analysis and visualization

As traditional full-space clustering is often not effective for revealing a meaningful clustering structure for HD data, in the emerging research field of subspace clustering [16] several approaches aim at discovering meaningful clusters in locally relevant subspaces. The problem of finding clusters in HD data can be divided into two subproblems: *subspace search* and *cluster search*. The first one aims at finding the subspaces where clusters exist, the second one at finding the actual clusters. The large majority of existing algorithms considers the two problems simultaneously and produces a set of clusters, where each cluster is typically represented by a pair (O,D) with O being the set of clustered objects (rows of the original data table) and D being the subset of relevant dimensions (columns of the original data table). Several methods have been proposed, that differ to the clustering search strategy and constraints with respect to the overlap of clusters and dimensions [7, 15, 21].

Only few works to date have considered visualization support for subspace clustering. The VISA [1] system uses visualization to help in interpreting the subspace clustering result. A global view shows the similarity between clusters in terms of the number of records and dimensions, and a detail view shows properties of individual clusters. A disadvantage of this approach is that no visualization or comparison for the data distribution in respective subspaces is supported. Heidi Matrix [28] uses a complex arrangement of subspaces on a matrix representation based on the computation of the *kNN* in each subspace. The complex visual mapping scheme may not be easy to use and its effectiveness to the best of our knowledge has not been evaluated yet. [10] proposes an approach for finding and visualizing interesting subspaces in astronomical data. Candidate subspaces are found from the data and ranked by a quality metric based on density estimation and morphological operators.

We note that if we apply one of these subspace clustering visualizations, we immediately inherit two main challenges of this paradigm that is still considered an open research issues, namely: the efficiency challenge (relating to subspace cluster search) and the redundancy challenge (relating to the typical redundancy of the outputs generated).

7 Conclusions

We presented an encompassing visual-interactive system for subspace-based analysis in HD data. Subspace-based analysis can constitute a new paradigm for HD data analysis since informative structures in the data can be found and compared in different subspaces of a larger HD input space. We defined, implemented, and demonstrated an analytical workflow based on automatic subspace search. A larger set of automatically identified interesting subspaces is grouped for interactive exploration by the user. A custom subspace similarity function allows for comparing subspaces. Our approach is able to effectively pin down several interesting views and helps to come up with specific findings regarding similarities of groups in data. We discussed a set of possible extensions of the system, which could be addressed as future work.

Acknowledgements

The research leading to these results has received funding from the ”SteerSCiVA: Steerable Subspace Clustering for Visual Analytics” DFG-664/11 project. We thank our colleagues Miloš Krstajić, Svenja Simon, and Leishi Zhang for their helpful comments on this project. We are thankful to Bilkis Ferdosi and Peer Kröger for sharing their subspace search implementations with us. The reviewers' comments were most helpful in improving this paper.

References

[1] I. Assent, R. Krieger, E. Müller, and T. Seidl. Visa: visual subspace clustering analysis. *SIGKDD Explor. Newsl.*, 9(2):5–12, 2007.

[2] C. Baumgartner, C. Plant, K. Kailing, H.-P. Kriegel, and P. Kröger. Subspace selection for clustering high-dimensional data. In *Proceedings of the Fourth IEEE Conference on Data Mining (ICDM)*, pages 11–18. IEEE CS Press, 2004.

[3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ”nearest neighbor” meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, pages 217–235, 1999.

[4] S. Bremm, T. v. Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum*, 30(3):891–900, 2011.

[5] S. Bremm, T. v. Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher. Interactive visual comparison of multiple trees. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 31–40. IEEE CS Press, 2011.

[6] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17:2581–2590, 2011.

[7] C.-H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 84–93. ACM, 1999.

[8] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, 1994.

[9] M. Ester, H.-P. Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI Press, 1996.

[10] B. J. Ferdosi, H. Buddelmeijer, S. Trager, M. H. F. Wilkinson, and J. B. T. M. Roerdink. Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 35–42. IEEE CS Press, 2010.

[11] S. Günnemann, E. Müller, I. Färber, and T. Seidl. Detection of orthogonal concepts in subspaces of high dimensional data. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM)*, pages 1317–1326, 2009.

[12] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 3–10. IEEE CS Press, 2010.

[13] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 15:993–1000, 2009.

[14] I. Jolliffe. *Principal Components Analysis*. Springer, 3rd edition, 2002.

[15] K. Kailing, H.-P. Kriegel, P. Kröger, and S. Wanka. Ranking interesting subspaces for clustering high dimensional data. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 241–252, 2003.

[16] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1–58, 2009.

[17] A. Lex, M. Streit, C. Partl, and D. Schmalstieg. Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 16(6):1027–1035, 2010.

[18] H. Liu and H. Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.

[19] E. Müller, I. Assent, S. Günnemann, R. Krieger, and T. Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 377–386, 2009.

[20] E. Müller, S. Günnemann, I. Färber, and T. Seidl. Discovering multiple clustering solutions: Grouping objects in different views of the data. In *Proceedings of the 10th IEEE Conference on Data Mining (ICDM)*, page 1220, 2010.

[21] D. Niu, J. G. Dy, and M. I. Jordan. Multiple non-redundant spectral clustering views. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 831–838. Omnipress, 2010.

[22] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[23] D. J. Rogers and T. T. Tanimoto. A Computer Program for Classifying Plants. *Science*, 132(3434):1115–1118, 1960.

[24] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Palgrave Macmillan Information Visualization*, 9(3):181–193, 2010.

[25] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86, 2002.

[26] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis)*, pages 65–72, 2004.

[27] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages (VL)*, pages 336–343. IEEE CS Press, 1996.

[28] S. Vadapalli and K. Karlapalem. Heidi matrix: nearest neighbor driven high-dimensional data visualization. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery (VAKD)*, pages 83–92. ACM, 2009.

[29] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[30] J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[31] M. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Taylor & Francis, 2010.

[32] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pages 157–164. IEEE CS Press, 2005.

[33] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis)*, pages 73–80. IEEE CS Press, 2004.

[34] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the Symposium on Data Visualization (VISSYM)*, pages 19–28. Eurographics Association, 2003.

[35] X. Yuan, Z. Wang, and C. Guo. Mds-tree and mds-matrix for high dimensional data visualization. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis)*, 2011. Poster abstract.