Eurographics / IEEE Symposium on Visualization 2011 (EuroVis 2011)
H. Hauser, H. Pfister, and J. J. van Wijk
(Guest Editors)

Volume 30 (2011), Number 3

Assisted Descriptor Selection Based on Visual Comparative Data Analysis

Sebastian Bremm[1] and Tatiana von Landesberger[1,2] and Jürgen Bernard[1] and Tobias Schreck[1]

[1]Technische Universität Darmstadt, Germany
[2]Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany

Abstract

Exploration and selection of data descriptors representing objects using a set of features are important components in many data analysis tasks. Usually, for a given dataset, an optimal data description does not exist, as the suitable data representation is strongly use case dependent. Many solutions for selecting a suitable data description have been proposed. In most instances, they require data labels and often are black box approaches. Non-expert users have difficulties to comprehend the coherency of input, parameters, and output of these algorithms. Alternative approaches, interactive systems for visual feature selection, overburden the user with an overwhelming set of options and data views. Therefore, it is essential to offer the users a guidance in this analytical process.
In this paper, we present a novel system for data description selection, which facilitates the user's access to the data analysis process. As finding of suitable data description consists of several steps, we support the user with guidance. Our system combines automatic data analysis with interactive visualizations. By this, the system provides a recommendation for suitable data descriptor selections. It supports the comparison of data descriptors with differing dimensionality for unlabeled data. We propose specialized scores and interactive views for descriptor comparison. The visualization techniques are scatterplot-based and grid-based. For the latter case, we apply Self-Organizing Maps as adaptive grids which are well suited for large multi-dimensional data sets. As an example, we demonstrate the usability of our system on a real-world biochemical application.

Categories and Subject Descriptors (according to ACM CCS): I.7 [Information Interfaces and Presentation]: — I.5.2 [Pattern Recognition]: Design Methodology—Feature evaluation and selection

1. Introduction

Exploration of and search in large data sets are important tasks in various application domains such as biology, finance, architecture, music, or emergency management. These applications handle objects of various types including molecules, music files, videos, images, 3D models, etc.

The analytical tasks in these areas are usually supported by efficient clustering and data retrieval algorithms relying on the calculation of object similarity. Although various methods for measuring data similarity exist, *descriptors of data elements* (i.e., multi-dimensional feature vectors, or feature sets) is commonly used in many applications. Data descriptors represent objects by a n-dimensional vector of numerical values (i.e., features). The similarity between objects is then calculated applying vector distance measures. The results are used as input to data analysis algorithms integrated in the application. The quality of the data descriptors has a major impact on the analytical results, therefore a lot of attention is given to finding suitable data description.

Determination of suitable data description is highly data and use case dependent. It should capture the relevant information from the input objects. Usually, the objects can be represented by descriptors in several ways, each capturing different data properties (see Figure 1 for an illustration). Which one is used in the analytical task, depends highly on the current task and semantics of the descriptor. For example, in biochemistry, the analyst may concentrate on aromatic properties of the molecules or on their fragment complexity. Finding an optimal data description (i.e., a multivariate feature vector) is not a trivial task, in particular for unlabeled data. On one hand, a descriptor with higher dimensionality can be calculated in order to capture as much avail-

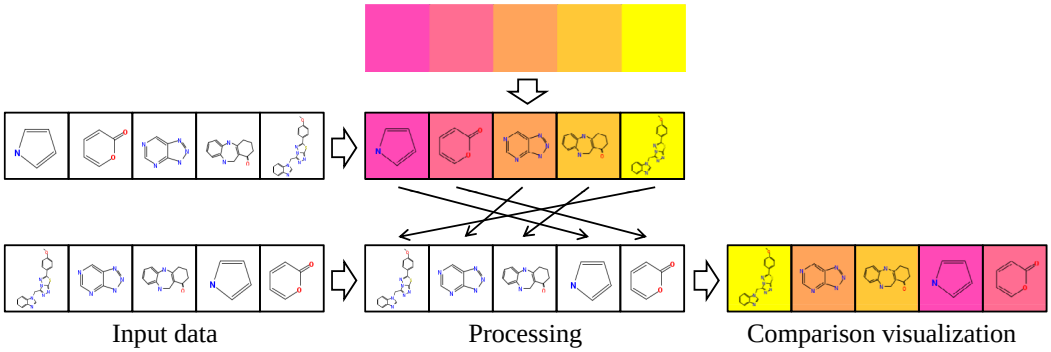


Figure 1: *Two meaningful data descriptors of biochemical data and their comparison. Each descriptor captures different data properties (atom resp. nitrogen count). Left: The input data is sorted according to each descriptor. Center: Color is mapped to the first ordering. The sorting is compared using connectors. Right: Compact comparison view. Color mapping based on object identity revealing descriptor correspondence.*

able information as possible. This can be done by extracting more features or by combining available descriptors. However, such larger descriptors increase calculation complexity, may include redundant information and can decrease significance of information about object similarities. In the latter case, distances between points become more equally distributed and therefore less informative (the so-called "curse of dimensionality") [AHK01]. On the other hand, if extracting only a small set of features (low-dimensional) necessary information can remain un-captured and different objects may not be discriminated from each other.

To tackle the *problem of determining a suitable dimensionality of data descriptors*, two main approaches exist. One possibility is to reduce dimensionality by projection or combination of similar features into one final dimension. A disadvantage of this approach is that the resulting dimensions are difficult to interpret, as they do not have a specific semantic meaning. The second approach is the selection of distinct important features from the original descriptor (i.e. multivariate feature vector). A crucial task here is to decide whether an object property should be disregarded or not, which is highly data and task dependent. These approaches often consider features individually disregarding groupings of features that should remain together.

Selection of the descriptors usually includes comparison of all possible sets of descriptors. Comparing the various multi-variate descriptors with differing dimensionality during and after the selection process is difficult [DB04]. Moreover, the evaluation of many methods is possible only in supervised way (objects having known labels). Many datasets however do not have labels, as they are costly to provide. In order to support this cumbersome procedure for unlabeled data, various algorithmic-based selection methods have been developed (see Section 2). They have common **problems**:

1. They often work in an automatic way without user involvement and they need a set of properly chosen input parameters. The *setting of these parameters* is difficult for domain experts that do not have expertize in data mining.
2. The algorithms do not take into consideration groupings of features (e.g., data descriptors composed of several features) that need to be *conserved together*.
3. The feature selection algorithm assigns global scores for decision on feature selections. These scores do not regard *local differences* in the data descriptions. Such differences occur when a subset of objects is well captured by the descriptor although the whole data set description is not satisfactory. The scoring results for data sets with specific local groupings often fail the scoring threshold although they may reveal interesting information.

In this paper, **we present** a novel *visual analysis approach for determining data descriptions suitable for the task at hand*. It addresses the problems stated above. We provide users with guidance in the data analysis process, as it has been shown useful for supporting dimension reduction tasks [IMI*10]. In contrast to previous work, our approach is based on comparative analysis suitable for multi-variate data descriptors with differing dimensionality also for unlabeled data. As description selection consists of several steps, we support the user in this incremental process. Our **contributions** are as follows:

1. We introduce a system for *comparative multi-variate data descriptor analysis*. It includes automatic descriptor recommendations and guidance highlighting interesting patterns such as borderline decisions of the automatic analysis. In this way, we support non-expert users.
2. We propose a specialized *score for comparing multi-variate descriptors* with varying dimensionality. The score is used for automatic recommendation.
3. We propose to use *color-coding for comparison of de-*

scriptors. The color coding provides data comparison in one single view (see Figure 1 for an illustration).

4. We develop dedicated *visualizations for comparison of multi-dimensional data descriptors*. These techniques are based on low-dimensional data presentation (scatterplot-based and grid-based) using color as comparison attribute. For large data sets, we employ adaptive grids with clustering properties – Self-Organizing Maps. These views allow for spotting overall similar descriptors and locally similar object groups in heterogeneous data sets.

We apply our techniques on real-world (biochemistry) and synthetic data sets demonstrating their usefulness.

The paper structure: Section 2 presents related work on algorithmic feature selection, data visualization and their interactive combination. It also introduces the Self-Organizing Map algorithm. Section 3 depicts our approach. It introduces the process as a whole, and then describes each part in more detail. Section 4 explains further aspects of our approach. Section 5 discusses color map choices. Section 6 shows applications of our approach on real data. Finally, Section 7 concludes and outlines future work.

2. Related Work

Finding an appropriate description for complex data types such as music [MM05], 3D objects [BKSS07], time series [Keo06], graphs [vLGS09] or biochemistry [BMGR04] data is a recent topic in various research areas. This description can be used for example in various data analysis, classification or search scenarios. Note that in this paper, we assume unclassified data in an exploratory analysis scenario.

The choice of relevant data descriptions (i.e., feature sets) is usually supported by feature selection algorithms (see Section 2.1). For exploration of the descriptors, interactive visual representations are used (see Section 2.2). Recent Visual Analysis tools combine both approaches in order to exploit their advantages (see Section 2.3).

2.1. Automatic Dimensionality Reduction

Data descriptors consist of a set of features (numeric values) representing complex data types. However, finding a proper descriptor is a challenge. Low dimensionality may lead to under-representation of the objects, and high-dimensional descriptors may suffer from problems such as "curse of dimensionality", where the distances between near and far objects converge [AHK01, BGRS99]. To tackle this problem, two main algorithmic approaches have been proposed: 1) dimension reduction and 2) feature selection. They often consider dimensions individually, disregarding possible semantic groupings of dimensions.

Dimension reduction techniques create an abstract reduced data description by projection or transformation of the original features into lower dimensional space. Examples are PCA [Jol02], MDS [BG97], Sammon's mapping [EC01] or spectral clustering [NJW01]. The resulting dimensions have no direct equivalence to the original dimensions. Therefore, the output dimensions are hard to interpret directly.

Feature selection approaches identify a set of meaningful features as subset of the input dimensionality. The final descriptor is built by interactive refinement of feature selections. These approaches can be classified according to the applied evaluation criterion into filter and wrapper approaches [KJ97]. Filter methods rely on an evaluation of the properties of every feature individually. If a certain criterion (e.g., entropy of distances [DCSL02]) is fulfilled, the feature is selected as relevant [KR92, Kon94, AD91]. Wrapper methods extend feature selection process with an additional step – e.g., clustering. The clustering results of the selected feature sets are evaluated to determine their relevance. However, the comparison of the results is difficult owing to variable number of resulting clusters and differences in the underlying feature space dimensionality [DB04].

The usage of algorithmic approaches solely is difficult for non-expert users as they need setting of, possibly extensive, number of input parameters and work in a black-box manner.

2.2. Visualization of High Dimensional Data

Finding appropriate parameters for various data analysis algorithms is crucial, but difficult for users who are not experts in data mining techniques (e.g., engineers or biologists). Using an interactive visualization, the user can steer the analysis process more intuitively. Often matrices of different feature representations are used for comparing high dimensional data. Wilkinson et al. [WAG06] show an overview of the descriptor space by visualizing scatterplotts for all pairs of input dimensions. Sips et al. [SNLH09] extend this approach by showing scatterplot matrices where each displayed axis is a combination of input features. Alternatively, an interactive visualization of a confusion matrix can be used to build combination models [TLKT09, KLTH10].

Visualization of high-dimensional data can be supported by a **Self-Organizing Map (SOM)**. It is a neural network algorithm that combines dimension reduction, clustering and layout of the data [Ves99, Koh01]. The dimension reduction is achieved by a projection of multivariate input data onto a low dimensional grid of prototypes (the map) in a way that it approximately preserves the topological properties of the data set (i.e., two data points that are close in original space are usually close in the lower dimensional space). The algorithm can handle large data sets and offers good clustering results. A SOM result can easily be visualized using the SOM grid by showing the reference prototypes (e.g., using multidimensional visualization techniques) or the nearest member to the prototype [Ves99].

2.3. Visual Analytics Approaches for Dimensionality Reduction

Integration of user feedback in the analysis process is crucial when use case dependent parameter adjustment and result evaluation are required. Many approaches interactively combining automatic calculations with visualization in the area of feature selection have been proposed. Choo *et al.* [CLKP10] presented a framework for data classification combining matrix with parallel coordinates. A scatterplot shows the resulting projection of a Linear Discriminant Analysis [Fuk90] which is iteratively refined during the classification process. Usually, dimension reduction methods focus on preserving structures of the high dimensional space. Johansson and Johansson enable the user to rank the importance of those structures by interactive steering of quality metrics [JJ09]. Tatu et al. [TAE*09] proposed analytical methods to find and filter important structures to reduce the complexity of the resulting visualization. The DimStiller [IMI*10] framework supports the whole process of feature selection. Additionally, the user gets guidance in every step of the pipeline, e.g., regarding parameter choice. It however focuses on dimension reduction only for individual features and does not consider local similarities.

Building upon these approaches, we propose a strategy where the user can decide on the level of process automation from a fully automatic up to a step by step assisted workflow.

3. Approach

Our approach aims at finding suitable descriptors of a given dataset. The input consists of several multivariate descriptors each potentially with a different dimensionality. The output is a subset of independent descriptors suitable for the use case at hand. In the workflow, the descriptors are compared pair wise for finding groups of similar ones and thereby to choose representatives among them. It supports the scalability of our approach w.r.t. the input dimensionality.

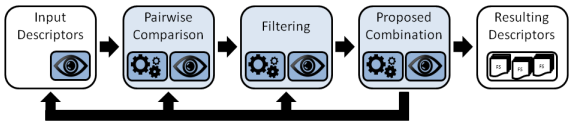


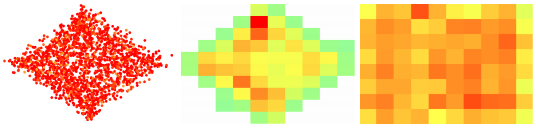
Figure 2: *Schema of the descriptor selection process. Every step (blue) encompasses automatic data processing and visualization part. The input consists of many descriptors for one data set. These are compared and filtered resulting in a proposed set of independent descriptors. This is an interactive, guided analysis process. Feedback loops allow the user to refine results on demand.*

In order to adapt the algorithms to a given use case and improve result specificity, the user is involved in the process. We offer guidance to ease the analysis for non-expert users. We have developed a dedicated pipeline supporting the data descriptor finding process (see Figure 2). This pipeline consists of several steps that are supported by both algorithmic and visual means (see Sections 3.1–3.4). Every step of the pipeline supports visualization and is interactively steerable. By combining the automatic and visualization functions, the results can be iteratively refined by the analyst.

First, the input data (a set of descriptors) can be explored using dedicated visualizations. They support the scalability w.r.t. the number of input data items. For large data sets, visual clustering using SOM is employed. Then, a pairwise comparison of the descriptors shows both global and local similarities between them. These results are used for filtering of similar descriptors. The final recommendation step shows an overview of results of automatic pre-processing, recommendations and offers the user the possibility to interactively refine the results. The analyst receives additional guidance by the highlighting of interesting or critical patterns. Feedback loops allow the user to interactively refine the results.

3.1. Basis Visualization of the Descriptors

At the beginning of the analysis, it is important to get an overview of the input data set. When choosing the visualization design, we focused on its re-usability in the whole workflow. It eases the correspondence of the representations and possibility to compare the data across displays.



(a) Scatterplot (b) Grid view (c) SOM view

Figure 3: *The visualization approaches. a) Scatterplot with overplotting for large data sets. b) Grid-based view showing inhomogeneous data distribution across display and empty space. The color coding denotes data density in each cell - green (low) to red (high)). c) SOM view with homogeneous data distribution and good usage of the display space.*

We have chosen to use a low-dimensional data display (in 2D), where each multi-dimensional descriptor space is mapped onto equally dimensionalised space. We employ both scatterplot-based and grid-based displays of the data (see Figure 3). The advantage of this display is that it may be applied to both, the initial overview of the data and the pairwise comparison of data descriptors (see Section 3.2).

Scatterplot visualization is often used to present the projected descriptor space in two dimensions. The objects are represented as points, and their size or color indicate their projection quality [SvLB10]. The advantage of this display is its familiarity and easy interpretability, however it suffers from overplotting (in particular for large data sets, see Figure 3). Although extensions of scatterplots overcoming most

of these shortcomings exist, this problem still prevails for large complex data sets.

In order to overcome the overplotting problem in scatterplots, we propose a **grid-based data visualization**. A simple approach would be to overlay a regular grid upon the scatterplot and color-code the data density in each grid cell. This approach is suitable, if the objects are equally distributed in space. However, in case of heterogeneous data distribution, important information might be lost (owing to high data density) and parts of the display might remain unused (empty cells). We address this issue by using an adaptive grid. We propose to use a Self-Organizing Map (SOM) (see Section 2.2 for more information). This combines dimensionality reduction with a grid-based visualization. It adapts to the density of objects in high dimensional space and therefore offers a more detailed overview of the data space (see Figure 3).

3.2. Pairwise Comparison of Descriptors

The first step of the pipeline is the comparison of the input descriptors for finding redundant information. Identifying correlated descriptors is a common technique in this respect. However, it often needs class labels or is restricted to equally dimensioned descriptors. Our approach is able to handle unlabeled heterogeneous multi-dimensional input descriptors. We propose a new score and dedicated views for comparing the descriptor similarity.

3.2.1. Automatic Pairwise Descriptor Comparison

Our proposed similarity score relies on nearest neighbor relations of objects in the two descriptor spaces (see Equation 1). The score is a sum of normalized neighborhood distance distortions over all data objects and is an extension of the projection precision score presented in [SvLB10] and applies also to descriptors with variable dimensionality.

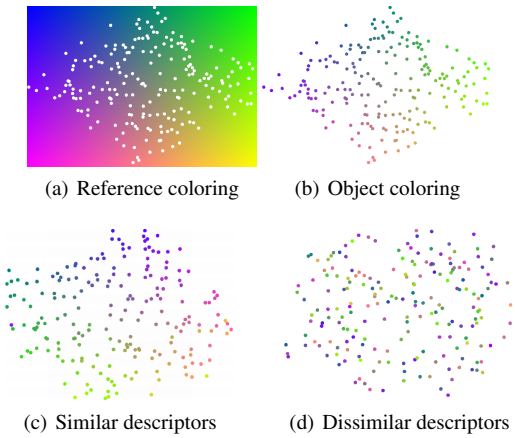
Let the l different input descriptors spanning input descriptor spaces D_1,...,D_l with the dimensionality $M^{D_1},..,M^{D_l}$. Let $O_1^{D_a},...,O_n^{D_a}$ be the n input objects described in the space D_a, $a \in 1,..,l$. Let further $d^{D_a}(O_x,O_y)$ be the distance of two objects O_x, O_y in D_a. Let $I_{O_x}^{D_a}$ be a sorted list of the k nearest neighbors of O_x in D_a. The similarity of two descriptors $s(D_a,D_b)$, $a,b \in 1,\ldots,l$ is defined as follows:

$$s(D_a,D_b) = 1 - \sum_{x=1}^{n} \sum_{y \in I_{O_x}^{D_a}} \left(\frac{d^{D_a}(O_x,O_y)}{\sqrt{M^{D_a}}} - \frac{d^{D_b}(O_x,O_y)}{\sqrt{M^{D_b}}} \right) \tag{1}$$

3.2.2. Visualization of the Pairwise Descriptor Comparison

The visualization of the descriptor comparison builds upon the visualizations of the data presented in Section 3.1. The low dimensional visualization relies on topology preservation property of dimension reduction algorithms. From a variety of approaches, a selection containing PCA, Kernel PCA, MDS, Sammons Mapping and SOM is considered.

For descriptor comparison, we extend the scatterplot and grid-based (SOM) visualization with similarity information. We propose a dedicated color visualization scheme for pairwise comparison of two descriptors in one single view. The data of the reference descriptor D_a are used for color coding of the data of the compared descriptor D_b. We apply a two-dimensional colormap for color coding as the individual views of the data are in 2D (see Section 5). The objects of the compared descriptors are then shown in the two-dimensional projected space. The color distribution of the objects in the compared space indicates the similarity of the two descriptors (see Figures 4 and 6).



(a) Reference coloring (b) Object coloring

(c) Similar descriptors (d) Dissimilar descriptors

Figure 4: *Scatterplot-based descriptor comparison visualization. Top: Object coloring. Bottom: The descriptor comparison. a) The reference color scheme mapped to the background and b) to the objects in the reference space. c) The homogeneous color gradient indicates a high similarity, d) the inhomogeneous gradient shows differing descriptors.*

The **scatterplot-based** comparison of two descriptors relies on the display of individual objects. For comparison, both positions of one object in spaces D_a and D_b have to be shown in one plot. We color code every object in the projection of D_a using a 2D colormap (see Figure 4a,b). This color is assigned to the corresponding objects in projection p(D_b), respectively. If objects have similar neighboring objects in both projections, their neighbors have similar colors in the visualization. In this way, local and global similarities of the two compared descriptors can be evaluated. In general, a homogeneous color distribution indicates a high similarity whereas a heterogeneous color distribution shows differences of the two descriptors (see Figure 4c,d).

The **grid-based visualization** of descriptor comparison is based on the result of a SOM projection. The SOM gives a good overview of the input space even for large data sets.

In analogy to the scatterplot view, we use a two-dimensional colormap to indicate the neighborhood coherency. The color mapping is based on a coloring of the reference grid using a two-dimensional colormap (see Section 5 for details) and object correspondence between grids. In the compared grid, the color of each cell C_x^{comp} is determined by the color of an corresponding cell C_{corr}^{ref} in the reference grid. However, in SOM view, objects from one compared cell may belong to several cells in the reference grid. In our approach, C_{corr}^{ref} is determined by the position of the majority of objects $OC_x^{comp} \in C_x{}^{comp}$ in C^{ref} (see Figure 5 for an illustration). In Section 4, we present further extension of this technique e.g., to visualize the cell distances.

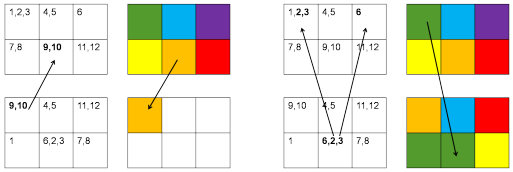
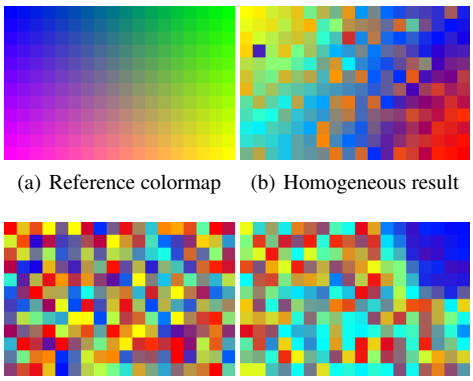


Figure 5: *Schema of the SOM comparison coloring. Left: An unambiguous color assignment, where all cell members from the compared SOM are grouped in one cell of the reference SOM. Right: The color assignment using majority principle – the cell color is used where the most elements are situated.*



(a) Reference colormap (b) Homogeneous result

(c) Inhomogeneous result (d) Locally homog. result

Figure 6: *Grid-based descriptor comparison visualization using SOM. a) The reference color scheme, b) homogeneous color gradient indicating a high similarity, c) the inhomogeneous coloring for differing descriptors, d) locally homogeneous coloring showing descriptors well discriminating subgroups of objects.*

3.3. Filtering of Redundant Descriptors

A high similarity of two descriptors implies that both carry the same information regarding the neighborhood distribution of the described objects. The task is, to remove this redundant information from the descriptor set. The filtering of redundant descriptors is based on the similarity scores calculated for all descriptor pairs. The result is visualized and serves as starting point for the interactive analysis process (see Section 3.4). Note that the pre-filtered descriptors can be interactively viewed and the pre-filtering can be rejected by the user in the next step of the pipeline. This is in particular important for borderline decision cases or in cases where user knowledge contributes to the decision making.

3.3.1. Automatic Descriptor Filtering

The automatic descriptor filtering is based on the pairwise similarity scores. If the similarity of two descriptors is high, they contain redundant information which should be included only once within the final descriptor. The filtering relies on the similarity threshold h, which specifies the maximal distance up to which two descriptors are considered similar. h affects the number and size of the resulting groups of descriptors, and is interactively set to specify the target number of groups desired. Let S be an ordered list of scores $S = \{s(D_a, D_b)\}, 1 \leq a, b, \leq l$ starting with the best one (the highest). All $s(D_a, D_b) >= h$, are regarded as similar. If a pair of descriptors $\{D_a, D_b\}$ satisfying the threshold exists, the descriptor with the higher average similarity to all other similar group members remains in S (see below).

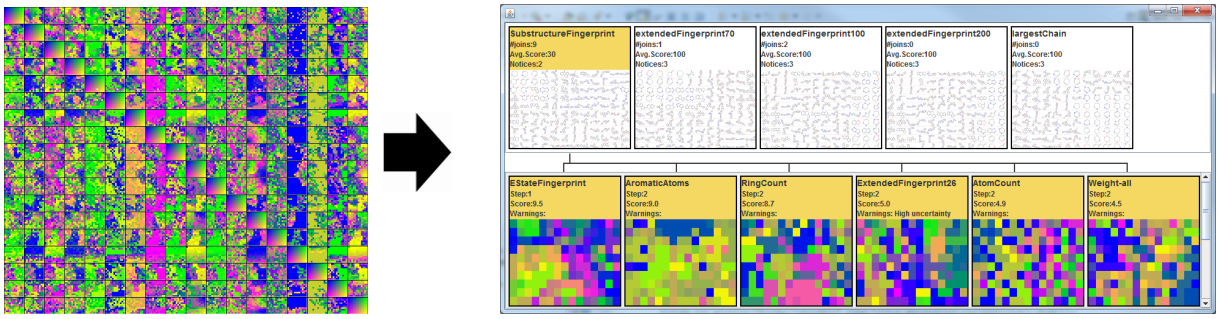
1: **for all** $s(D_a, D_b) \in S$ **do**
2: **if** $s(D_a, D_b) \geq h$ **then**
3: **if** $average(s(D_a, D_x)) \geq average(s(D_y, D_b))$: $s(D_a, D_x) \geq h, s(D_y, D_b) \geq h$ **then**
4: Remove D_b: remove $\forall s(D_b, D_y) \in S$
5: **else**
6: Remove D_a: remove $\forall s(D_x, D_a) \in S$
7: **end if**
8: **end if**
9: **end for**

3.3.2. Visualization of All Comparison Results

To enable the user to get an overview of multiple pairwise comparisons, we propose a matrix and an ordered list view of the comparison visualizations. It shows visualizations of all pairwise descriptor comparisons (see Figure 7a). These views can be filtered and sorted by a comparison score. This provides a better overview of the comparisons in particular for data with many descriptors.

3.4. Recommendation Visualization and Exploration

The automatically calculated proposal for descriptor selection is presented to the user in the last step of the pipeline. The result inspection is supported by interactive visual exploration of descriptor similarity. The user can choose from just applying the proposed combination or to inspect and adjust the steps of the process. User involvement in the process is advantageous especially in borderline cases where the automatic filtering decision is close to the decision criteria (similarity threshold). Interactive data space exploration



(a) Descriptor comparison matrix (b) Result view

Figure 7: *Visual descriptor comparison. Left: Initial overview of pairwise descriptor correspondence. Right: The result view after descriptor filtering. The top row shows the selected descriptors with the data views. The bottom row shows the comparisons of one descriptor with related descriptors (in yellow). This supports understanding of the filtering decisions.*

can reveal new information helping the user to make better decisions on descriptor filtering. For example, the scoring function cannot reveal local similarities between objects in two descriptor spaces. However these can be highly relevant for the usage of descriptors for certain object classes (e.g., people in 3D objects). The inspection is supported also in algorithmic way, highlighting such borderline decision cases. The algorithmic and visual support of the adjustment process is described in the following.

3.4.1. Automatic Support for the Result Exploration

During the filtering process, we automatically detect cases recommended for further examination by the analysts. These so-called *examination markers* are either cases where the score $s(D_a, D_b)$ was very close to the threshold h. A close score may indicate low confidence of the filtering decision or may indicate that the comparison result shows local abnormalities. These local structures are not considered in the calculation of the pairwise comparison score, but might be interesting in search scenarios where the local neighborhood of the input object is more important than the rest (see Section 6 for an illustration).

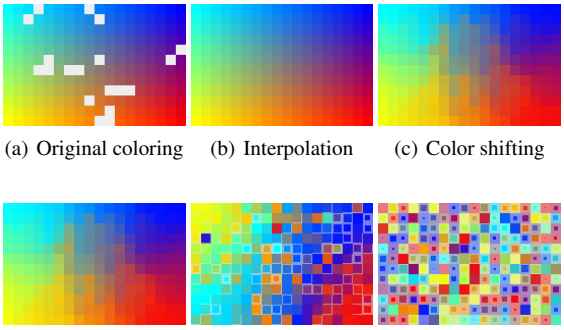
3.4.2. Visualization of the Proposed Descriptors

The recommended set of descriptors is visualized on the basis of their two-dimensional projections (see Figure 7b top row). The view includes additional important descriptor information (e.g., the number of similar descriptors or examination markers). This summary overview is an entry point to a deeper examination of the decision space. For example, the comparison of one selected descriptor with similar filtered descriptors is shown on demand (see Figure 7b bottom row). In this way, the understanding and adjustments of the filtering results are supported. This exploration may lead to adjustments in the process – feedback to the previous interactive steps of the pipeline.

4. Extensions

For better visual quality of the SOM comparison view, we have implemented the following additional data representations: color interpolation, color shifting for reference SOMs, and visualization of color unreliability. All of this options are interactively steerable.

Depending on the input data structure, the result of the SOM algorithm can include few empty cells. They represent a area of the featurespace without data samples. We enable the user to visually compare this areas in different feature spaces by *interpolation* of the colors of neighboring cells (see Figure 8b).



(a) Original coloring (b) Interpolation (c) Color shifting

(d) Coloring with reliability

Figure 8: *Visual extensions. Top: Illustration of improvement of visual display. Bottom: Display of coloring quality in SOM comparison. The columns show examples of SOMs. Left: a reference SOM, center: a homogeneous SOM, right: an inhomogeneous SOM.*

The SOM forms an adaptive grid, so the distances between neighboring cells are not constant. Therefore, we shift the colors of the two-dimensional colormap *according to the*

distance of the SOM cell centers (see Figure 8c). This function resembles the so-called U-Matrix which helps to identify the structure of the SOM clustering [VA00].

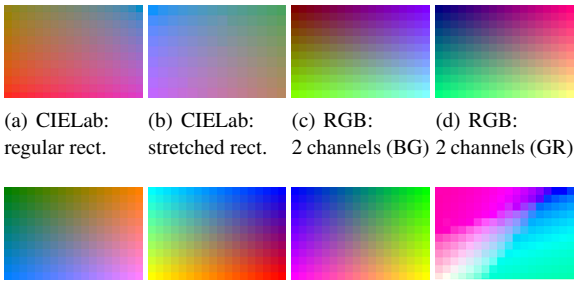
The SOM coloring uses matching of cell elements between the reference SOM and the compared SOM. In the easiest case, all objects O_i of the compared cell $|C_i|$ are in one cell of the reference SOM, so the *unreliability of the cell coloring* is zero (see Section 3.2 for details on SOM cell coloring). If the objects of a cell in the compared SOM (descriptor D_b) are distributed over several cells in the reference SOM (descriptor D_a), a higher unreliability is expected. The unreliability of a cell C_i in the compared SOM is measured by a score u_{C_i}, which takes into consideration the distance of the cells in the reference SOM to which the elements of the compared SOM cell C_i are matched and the selected majority cell R_i. The distance is calculated as the distance of the descriptors in the cell centers $d({R_i}^{D_a}, {R_k}^{D_b})$ using Euclidean distance measure. If the cells are located close, the object distribution can be handled as similar.

Visualization of the reliability can affect the cell color (via alpha channel or one of the axes of the color space) or cell size (reduced corresponding to the unreliability score). Cell size encoding has turned out to be very effective and intuitive (see Figure 8d). The background color of the reduced cell is colored in the cell color with a higher, user steerable transparency. In this way, the impression of the SOM coloring remains stable, so the color gradient is still visible and on the other hand, the cell reliability is easy to evaluate.

5. Two-Dimensional Color Maps

Coloring the data in a two-dimensional space, such as in SOM grid, is a challenging task. It is difficult to balance at the same time the following beneficial properties: a perceptual linearity of the color space, a high color resolution and the preservation of all pairwise prototype distances [KVK99]. For two-dimensional coloring, in particular SOM coloring, a number of color-based visualization techniques were proposed [KK98, Him98, KVK99, KVK00, Him00]. The idea is to apply high-contrast color space to illustrate the SOM grids distance relations as good as possible. These approaches use extraction of subspaces from, for example, the RGB or the CIELab color space. Compared to RGB, the CIELab color space is perceptually linear, which is beneficial for expressing distance relations with color. In return, the RGB color space is a regular cube and therefore quite easy to implement, whereas the CIELab color space has an irregular 3D shape and suffers from an additional projection needed to access CIELab.

Our decision for a two dimensional color map is the result of a comprehensive comparison of current colormap techniques. Figure 9 contains some of the most promising color map approaches for the grid of 18x12 cells. Note that the grid resolution can be adjusted. In our opinion, the CIELabs, benefiting from a perceptually linear color scheme, cannot be adequately exploited, because we can either use only a little linear subspace with a low color resolution, or need to apply an additional nonlinear algorithm to project the SOM grid to the CIELab color space. A demonstration of the poor color resolution can be seen in Figure 9 a) and b). They show that the resulting color contrast is so low that adjacent grid coordinates can not be distinguished clearly. After extensive experiments with the RGB color space, considering two channel and three channel approaches, we made the decision to follow Himbergs approach [Him00] to use a linear section of the RGB cube with maximized color resolution. Our goal was to increase the perception of color differences in SOM comparison. Thus, our colormap is spanned with the four corner colors cyan, yellow, blue and red (see Figure 9g).



(a) CIELab: regular rect. (b) CIELab: stretched rect. (c) RGB: 2 channels (BG) (d) RGB: 2 channels (GR)

(e) RGB: 2 channels (RB) (f) RGB: 3 channels (g) RGB: 3 channels (h) XYZ

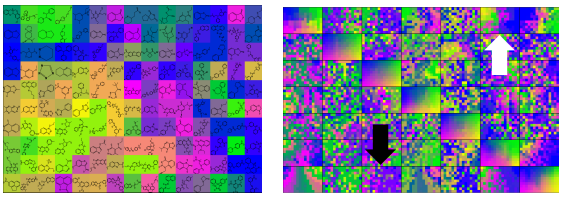
Figure 9: *Comparsion of colormaps for a 18x12 SOM grid. a) Rectangular cut out of the CIELab color space at L=55. b) Skewed rectangle cut (CIELab, L=55). c), d) and e) Two channels of RGB mapped to x and y axis, leaving the third constant. f) and g) Three channel RGB color scheme, diagonally cutting the RGB color cube [Him00]. h) Color scheme in XYZ color space.*

6. Application

In this section, we demonstrate our approach on a biochemical dataset following the introduced workflow. Researchers in biological and pharmacological sciences analyze large sets of molecules, e.g., as output of High Throughput Screenings (HTS). In HTS, many molecules are tested for reactivity with one specific molecule of interest. The resulting datasets contain several hundreds or thousands of molecules with high reactivity. The task of the analysts is to find few, promising compounds for further examination. The selection criteria are use case depended. Not only the data structure, but also user expertise and further factors such as costs have to be considered. Moreover, as shown in Figure 1, often there is more than one valid description of a given dataset. Therefore user interaction in the analysis process is needed.

The dataset contains 9989 molecules, described by 18 standard pharmacophore descriptors, divided into two groups. The first group consists of 11 basic, 1-D Quantitative Structure-Activity Relationship (QSAR) descriptors for, e.g., fragment complexity or the number of hydrogen-bond donors [BST04]. The second is a set of so called *fingerprints*, binary descriptors classifying whether the described molecule fulfills certain conditions or not. The group consists of 7 fingerprints with a dimensionality from 26 to 400.

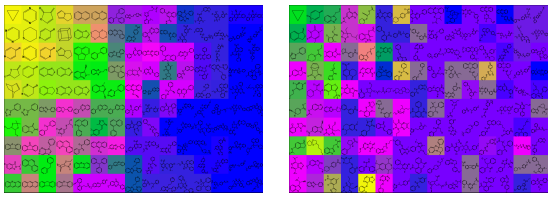
The analysis task combines two intentions: 1) Finding relevant groups of compounds and structures in the data, and 2) describing them as compactly as possible. As outcome of the automatic analysis process, five different groups of descriptors were proposed. Their overview is presented in Fig. 7. Each group is composed of one or more similar descriptors. Four of them were represented by fingerprints of varying dimensionality (307, 200, 100, 70). Details and iconic comparison views are provided on demand to analyze the groups in more detail. The first set, represented by the 307-dimensional substructure-based fingerprint contains 11 other descriptors, ordered by similarity (see Fig. 10a). It points out that the most similar descriptor, the 79-D EState fingerprint [HK95] exposes very similar local neighborhood relations to the 307-D substructure fingerprint and therefore can be used as its low dimensional replacement (see Fig. 10a).



(a) (b)

Figure 10: *a) Comparison of the 79-D EState fingerprint with the 307-D substructure fingerprint, showing their similarity. b) Matrix view on the group of 7 similar descriptors.*

To analyze the group in more detail, we switch to the matrix view and filter out SOM clustering results with a poor object distribution (Fig. 10b). One of the SOM comparisons shows a very homogeneous color gradient represents the descriptors for wight and number of atoms of the molecules (Fig. 10b white arrow & Fig. 11a). This validates an expectation of the coherence between weight and size. Looking at the comparison of the ExtendendFingerprint with the WienerNumber descriptor, we see that many cells are homogeneously colored (Fig. 10b black arrow & Fig. 11b). All of the purple molecules in the WienerNumber SOM are located in one cell of the ExtendendFingerprint SOM. If the pharmacologist is interested in these molecules, the WienerNumber descriptor is preferable. It leads to a higher diversity of the concerned molecules at a lower dimensionality (1 vs. 26).



(a) (b)

Figure 11: *a) Comparison of the weight to an atom count descriptor. The homogeneous color gradient validates the expected correlation of the descriptors. b) The 1-D WienerNumber descriptor shows a high separability for molecules which are all in one cell in the SOM of the 26-D ExtendendFingerprint.*

7. Conclusions and Future Work

In this paper, we have presented a novel system guiding analysts in the process of selecting suitable data descriptors. Our approach is based on a novel score for descriptor comparison applicable also for data descriptors with differing dimensionality for unlabeled data. We presented specialized visualizations for gaining an overview of both the descriptor space and descriptor comparison. We developed techniques for spotting high-quality local data descriptions in globally suboptimal data descriptions. The resulting comparison data space can be interactively explored.

The presented approach can be applied in various areas dealing with search and exploration of large data sets. For example, large video, image, 3D model or graph data sets can be easily analyzed. In order to demonstrate the usability of our system, we have used a scenario of selecting descriptors for biochemical data.

In the future, we would like to implement further algorithms for selecting interesting views and work on the estimation of initial parameters and their interactive steering. In particular, we would like to compare several scoring functions for their expressiveness and extend the pairwise comparison to simultaneous comparison of multiple descriptors. Combining of descriptors can be improved by additional heuristic algorithms including user feedback on the proposed elements. We would like to test our system with users in various application domains.

Acknowledgments

This work was partially supported by the German Research Foundation (DFG) within the project Visual Feature Space Analysis as part of the Priority Program on Scalable Visual Analytics (SPP 1335).

References

[AD91] ALMUALLIM H., DIETTERICH T.: Efficient algorithms for identifying relevant features. In *Canadian Conf. on Artificial Intelligence* (1991), pp. 38–45.

[AHK01] AGGARWAL C., HINNEBURG A., KEIM D.: On the surprising behavior of distance metrics in high dimensional space. *Database Theory* (2001), 420–434.

[BG97] BORG I., GROENEN P.: *Modern multidimensional scaling: Theory and applications*. Springer, 1997.

[BGRS99] BEYER K., GOLDSTEIN J., RAMAKRISHNAN R., SHAFT U.: When is nearest neighbor meaningful? *Database Theory* (1999), 217–235.

[BKSS07] BUSTOS B., KEIM D., SAUPE D., SCHRECK T.: Content-based 3D object retrieval. *IEEE Computer Graphics and Applications 27 (4)* (2007), 22–27.

[BMGR04] BENDER A., MUSSA H., GLEN R., REILING S.: Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci 44*, 1 (2004), 170–178.

[BST04] BÖCKER A., SCHNEIDER G., TECKENTRUP A.: Status of HTS data mining approaches. *QSAR & combinatorial science 23*, 4 (2004), 207–213.

[CLKP10] CHOO J., LEE H., KIHM J., PARK H.: iVisClassifier: An Interactive Visual Analytics System for Classification Based on Supervised Dimension Reduction. In *IEEE Symposium on Visual Analytics Science and Technology* (2010), pp. 27–34.

[DB04] DY J., BRODLEY C.: Feature selection for unsupervised learning. *J. of Mach. Learning Research 5* (2004), 845–889.

[DCSL02] DASH M., CHOI K., SCHEUERMANN P., LIU H.: Feature selection for clustering-a filter solution. In *IEEE Int. Conf. on Data Mining* (2002), p. 115.

[EC01] EWING R. M., CHERRY J. M.: Visualization of expression clusters using Sammons non-linear mapping. *Bioinformatics 17*, 7 (2001), 658–659.

[Fuk90] FUKUNAGA K.: *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[Him98] HIMBERG J.: Enhancing SOM-based data visualization by linking different data projections. In *Int. Symp. on Intelligent Data Engineering and Learning* (1998), Eureka, p. 427.

[Him00] HIMBERG J.: A SOM based cluster visualization and its application for false coloring. In *IEEE Int. Joint Conf. on Neural Networks* (2000), vol. 3, p. 3587.

[HK95] HALL L., KIER L.: Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. of Chemical Information and Computer Sciences 35*, 6 (1995), 1039–1045.

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: DimStiller: Workflows for dimensional analysis and reduction. In *IEEE Conference on Visual Analytics Software and Technologies* (2010), pp. 3–10.

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 993–1000.

[Jol02] JOLLIFFE I.: *Principal component analysis*. Springer, 2002.

[Keo06] KEOGH E.: A decade of progress in indexing and mining large time series databases. In *Int. Conf. on Very Large Data Bases* (2006). Tutorial.

[KJ97] KOHAVI R., JOHN G.: Wrappers for feature subset selection. *Artificial intelligence 97*, 1-2 (1997), 273–324.

[KK98] KASKI S., KOHONEN T.: *Visual Explorations in Finance*. Springer, 1998, ch. Tips for processing and color-coding of Self-Organizing Maps, pp. 195–202.

[KLTH10] KAPOOR A., LEE B., TAN D., HORVITZ E.: Interactive Optimization for Steering Machine Classification. In *Conference on Human Factors in Computing Systems* (2010).

[Koh01] KOHONEN T.: *Self-Organizing Maps*. Springer, 2001.

[Kon94] KONONENKO I.: Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94* (1994), Springer, pp. 171–182.

[KR92] KIRA K., RENDELL L.: The feature selection problem: Traditional methods and a new algorithm. In *National Conf. on Artificial Intelligence* (1992), pp. 129–129.

[KVK99] KASKI S., VENNA J., KOHONEN T.: Coloring that reveals high-dimensional structures in data. In *Int. Conf. on Neural Information Processing* (1999), vol. 2, pp. 729 – 734.

[KVK00] KASKI S., VENNA J., KOHONEN T.: Coloring that reveals cluster structures in multivariate data. *Australian J. of Intelligent Information Processing Systems 6*, 2 (2000), 82–88.

[MM05] MIERSWA I., MORIK K.: Automatic feature extraction for classifying audio data. *Machine Learning 58*, 2–3 (2005), 658–659.

[NJW01] NG A. Y., JORDAN M. I., WEISS Y.: On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (2001), MIT Press, pp. 849–856.

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

[SvLB10] SCHRECK T., VON LANDESBERGER T., BREMM S.: Techniques for precision-based visual analysis of projected data. *Information Visualization 9*, 3 (2010), 181–193.

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

[TLKT09] TALBOT J., LEE B., KAPOOR A., TAN D.: EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In *Int. Conf. on Human Factors in Computing Systems* (2009), pp. 1283–1292.

[VA00] VESANTO J., ALHONIEMI E.: Clustering of the self-organizing map. *Trans. on Neural Networks 11*, 3 (2000), 586–600.

[Ves99] VESANTO J.: SOM-based data visualization methods. *Intelligent Data Analysis 3*, 2 (1999), 111–126.

[vLGS09] VON LANDESBERGER T., GÖRNER M., SCHRECK T.: Visual analysis of graphs with multiple connected components. In *IEEE Symp. on Visual Analytics Science and Technology* (2009), pp. 155–162.

[WAG06] WILKINSON L., ANAND A., GROSSMAN R.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *Trans. on Visualization and Computer Graphics 12* (November 2006), 1363–1372.