Improving Projection-based Data Analysis by Feature Space Transformations

Matthias Schaefer[a], Leishi Zhang[a], Tobias Schreck[a], Andrada Tatu[a], John A. Lee[b], Michel Verleysen[b] and Daniel A. Keim[a]

[a]University of Konstanz, Konstanz, Germany
[b]Université catholique de Louvain, Belgium

ABSTRACT

Generating effective visual embedding of high-dimensional data is difficult - the analyst expects to see the structure of the data in the visualization, as well as patterns and relations. Given the high dimensionality, noise and imperfect embedding techniques, it is hard to come up with a satisfactory embedding that preserves the data structure well, whilst highlighting patterns and avoiding visual clutters at the same time. In this paper, we introduce a generic framework for improving the quality of an existing embedding in terms of both structural preservation and class separation by feature space transformations. A compound quality measure based on structural preservation and visual clutter avoidance is proposed to access the quality of embeddings. We evaluate the effectiveness of our approach by applying it to several widely used embedding techniques using a set of benchmark data sets and the result looks promising.

Keywords: Visual Analytics, Projection-based Data Analysis, Feature Space Transformation, Quality Measures

1. INTRODUCTION

A good starting point for analyzing high-dimensional data is to map it to a 2D or 3D display as a scatter-plot-like visualization to see the structure and patterns in the data. We call such an approach *projection-based* analysis. A large number of Dimension Reduction (DR) techniques exist for performing such a task, however it is hard to come up with a satisfactory embedding (projection) that provides both good reflection of the data structure and clear indication of class boundaries at the same time. This is due to many factors. First of all, DR techniques only provide approximations of the distances between data items in the data space. It is nearly impossible to get a "perfect" embedding which shows the exact structure of the data in the visual space. Secondly, high-dimensional data often contains irrelevant attributes which obscure the real distance between data items, and the noise introduced by such irrelevant information may lead to a cluttered visual display where class boundaries are blurred. Such limitations hinder the human analyst from understanding relationships between data items and identifying patterns in the data. Although supervised DR techniques aim at providing better group separation based on class labels in the data, methods that maintain good comprise between structural preservation and visual cluttering avoidance (i.e., classes are well separated and easy to identify in the visual display) are lacking. Furthermore, there are no existing measures that can be directly applied to evaluate the quality of a given embedding based on all the above mentioned criteria.

We present a computational framework that tackles the problem of high-dimensional data embedding by improving the quality of an embedding by extending relevant features in the feature vector space. The framework first identifies relevant features in the data and then extends them in the original feature space using various transformation strategies. Given an initial embedding, users can interactively update the configuration by enhancing the influence of relevant features using different transformation strategies. A compound quality measure, which takes into consideration both structural

Further author information:
Matthias Schaefer: E-mail: Matthias.Schaefer@uni-konstanz.de, Telephone: (+49) 07531 884794
Leishi Zhang: Email: Leishi.Zhang@uni-konstanz.de, Telephone: (+49) 07531 883637
Tobias Schreck: Email: Tobias.Schreck@uni-konstanz.de, Telephone: (+49) 07531 883375
Andrada Tatu: Email: Andrada.Tatu@uni-konstanz.de, Telephone: (+49) 07531 884364
John A. Lee: Email: John.Lee@uclouvain.be, Telephone: (+32) 2764 9528
Michel Verleysen: Email:michel.verleysen@uclouvain.be, Telephone: (+32) 1047 2551
Daniel A. Keim: E-mail: Daniel.Keim@uni-konstanz.de, Telephone: (+49) 07531 883161

preservation and visual clutter avoidance is designed to assess the quality of the embeddings. The structural preservation is evaluated by existing structural quality measures. The visual clutter avoidance is evaluated by a density function that measures the overlap between classes and an area measure that calculates the size of overlap regions between classes. The combined measure provides a clear indication of the trustworthiness of an embedding in terms of both structural preservation and class separation. With an appropriate graphical user interface the user can achieve the best compromise between the two according to their preferences. Our proposed approach works in a supervised manner, that is, class labels are used to evaluate the visual quality of an embedding in terms of class separation. Unlike most of the existing supervised embedding techniques, which are based on specific DR algorithms, our approach is DR technique independent. It can be applied to improve an initial embedding generated by any DR algorithm.

The main contributions of this paper include: 1) an improved projection-based data analysis framework which transforms the feature vector space by extending the identified relevant features; 2) a new quality measure to automatically evaluate projection displays, integrating structure preservation and clutter avoidance; and 3) an evaluation of the effectiveness of different feature space transformations strategies, as a guideline for further development.

In Section 2, we introduce related work, including DR techniques that are commonly used for embedding high-dimensional data and various quality measures for evaluating the visual quality of embeddings. Section 3 details our framework for improving projection-based data analysis, and enumerates a number of possible feature vector transformation strategies. Section 4 discusses quality measures for evaluating a given embedding in terms of structural preservation and clutter avoidance. In Section 5, we illustrate the effectiveness of our approach by testing with synthetic and real data. In Section 6 we discuss the evaluation results before drawing conclusion and discussing possible extension of the work in Section 7.

2. RELATED WORK

Analyzing high-dimensional data is challenging, as the high dimensionality induces problems in both automatic analysis and visualization. On the automatic side, redundant and noisy dimensions may degrade the performance of the analysis algorithms.[1] On the visualization side, one has to cope with the conflict between the limited number of visual dimensions and the large number of data dimensions. To visualize the structure of high-dimensional data effectively, the data items have to be mapped in such a way that similar items are close to each other and dissimilar ones are far apart. This is usually achieved by a DR technique which tries to approximate the distances between data items in data space to the corresponding Euclidean distances in the visual display. In this section, we briefly review DR techniques that are commonly used for embedding high-dimensional data, as well as measures for evaluating the quality of an embedding.

2.1 Dimension Reduction for High-dimensional Data Visualization

DR is a well-studied topic by the machine learning community. The idea behind all DR techniques is that they should produce low-dimensional representations that preserve meaningful structural properties of data. In general, these properties formalize proximity relationships. They can be similarities (adjacencies, dot products) or dissimilarities (distances, angles). A large number of DR techniques exist, ranging from classical approaches such as *Principal Component Analysis (PCA)*,[2] *Classical Multidimensional Scaling (MDS)*,[3] to more recent extensions such as *Curvilinear Component Analysis (CCA)*,[4] *Isomap*,[5] *Generative Topographic Map (GTM)*,[6] and *Stochastic Neighbor Embedding (SNE)*.[7]

For generating a visual embedding, the training data can be either *labeled* or *unlabeled*, leading the development of *supervised* and *unsupervised* DR algorithms, which are often studied separately. While unsupervised DR aims at representing high-dimensional data in lower-dimensional spaces in a faithful way, supervised DR tends to emphasize features relevant for a given labeling of the data in the final embedding such that the visualization provides better class separation. However many DR techniques have variations for both supervised and unsupervised learning. The algorithm we propose in this paper works in a supervised manner, however, it can be applied to improve an embedding generated by any DR technique as long as trustworthy class labels are provided. Next, we focus on a few representative DR techniques as a basis for understanding the fundamental principle of DR. Comprehensive surveys of DR techniques can be found in.[8–12]

PCA and classical MDS are probably the two linear DR techniques for embedding high-dimensional data most widely used in data visualization. Also known as Karhunen-Loève transform, PCA reduces the dimensionality of the data to summarize the most important parts and simultaneously filters out noise. The algorithm de-correlates variables and selects those that bear most of the data variance for projection. PCA was later extended to Classical MDS which starts from either

a Gram matrix or a matrix of pairwise Euclidean distances instead of the sample covariance to compute the projections along the principle components.

One inherent limitation of linear approaches is that they cannot take into account nonlinear structures consisting of arbitrarily shaped clusters or curved manifolds. Among many nonlinear extensions of MDS which are designed to overcome such limitation,[4,13,14] Isomap is a interesting variation. Instead of using pairwise input-space distances as simple Euclidean distances, Isomap uses geodesic distances along the manifold of the data (technically, along a graph formed by connecting all k-nearest neighbors) to recover certain types of manifolds. SOM is one of the earliest nonlinear techniques which trains a discretized map representation of the input space of the training samples. A neighborhood function is used to preserve the topological properties of the input space. A notable recent extension of SOM is GTM, which is a generative version of SOM. GTM represents the probability density of high-dimensional data in a smaller number of latent or hidden variables using latent variable models.[15] The SNE method and its variant, t-SNE,[16] have recently received much attention. They are based on similarity preservation instead of distance preservation. They involve specific similarity definitions that show interesting invariance properties and makes them robust against the phenomenon of norm concentration.

Automatic DR techniques help to show underlying structure and relationships in high-dimensional data. However, with the increasing size and complexity of data, it becomes more and more difficult to generate meaningful transformations without interactive analysis based on integrating human knowledge and feedback during the learning process. This leads to the development of *interactive dimension reduction* techniques. Examples include the iPCA system[17] which supports the analysis of multivariate datasets through extensive interaction with the PCA output. The iVisClassifier system[18] facilitates the interpretability of the computational model applied to the data via interaction and multiple views projections. Furthermore, the DimStiller framework[19] defines an interactive workflow that guides users through the process of finding suitable dimension subsets. In[20] the importance of integrating interactions with statistic methods (in particular, DR techniques) to support explorative analysis of high-dimensional data is discussed. In[21] interactive selection of sets of features was proposed. The approach generates, for each candidate feature set, a projection. Using color-coding, a comparison matrix of the individual projections is provided which supports identification of similar and complementary feature sets. A related problem was addressed in.[22] There, Dendrogram structures were extracted from alternative feature sets, and applied for interactive comparison and selection of feature sets.

The above mentioned techniques show that a rich body of research exists on high-dimensional data visualization. However, how to appropriately compute and visualize structure and patterns in high-dimensional data remains a tough challenge due to the nature of the data and DR techniques.

2.2 Quality Measures

Despite the fact that DR research started over a hundred years ago and a large number of DR techniques have been developed, the question of quality assessment of a given embedding remains mostly unanswered until recent years. One possible way to measure the quality of a embedding is to use a so called *stress* or *strain* measure.[23–25] These measures often come with nonlinear DR methods, and are typically used as objective functions for measuring the quality of structural preservation in terms of how well do the Euclidean distances between pairwise data items in a low-dimensional embedding approximate the corresponding distances in high-dimensional data space. While strain and stress measures check the preservation of global structure of data with respect to distance/similarity preservation, in recent years, more and more research has been devoted to designing of new criteria for quality assessment with a broader applicability, taking into consideration also the small neighborhood preservation, examples include the trustworthiness and continuity measure,[26] the local continuity meta-criterion,[27] and the K-ary neighborhoods measure.[28] In the case of labeled data, the classification error is a typical choice, see for instance[29] and other references in.[30] The integration of classification error measures in the DR technique leads to better group separation in the final embedding.

As mentioned previously, apart from studying the structure of the data, analysts also expect to see patterns in the embedding. Such patterns include grouping information, such as classes (with labeled data) and clusters (with unlabeled data) and outliers. Although it is often intuitive to tell the visual quality of an embedding by seeing how cluttered it is and how badly groups overlap, a qualitative measure is still preferred by analysts for quality assessment. In recent years some research has been devoted to the evaluation of visual quality of graphical representations of high-dimensional data in a broader sense.[31–34] Such graphical representations include not only visual embedding of high-dimensional data generated by DR techniques, but also other form of visualizations such as parallel coordinates[35] or scatter plot projection of data values over a subset of dimensions (usually two or three dimensions, mapped to x-, y- and z- axes in the visual display).

The visual quality measures can be based on different user tasks, for example, outliers, clusters, correlation, and abstraction level. Two existing visual quality measures are closely related to our problem.[33] The first measure is the *Histogram Density Measure* which was designed for ranking scatter plot visualizations. The second one is the *Class Density Measure* which is based on an image processing algorithm that transforms each group in a continuous, smooth density function based on local neighborhoods and measures the mutual overlap between pairwise of data items in the scatter plot. An overview of approaches that use quality measures in high-dimensional data visualization and a systematization based on a literature review is presented in.[36]

None of these above mentioned measures however provides the facility to assess the quality of a given embedding in terms of both structural preservation and class separation, as well as visual cluttering avoidance. In our work, we try to design a measure which takes into account both structural preservation quality and visual quality to make sure the embedding reflects the original structural of the data as well as provides clear (low-cluttered) visualization for pattern analysis.

3. FEATURE SPACE TRANSFORMATION

Most projection-based analyses involve two steps: Given a high-dimensional data set, a distance or similarity matrix that records pairwise distances (similarities) between objects is first calculated using a preselected measure; a DR technique is then applied, aiming to approximate in the projection space the pairwise distance (similarity) measured in the data space. The visual embedding is meant to help analysts to understand the data structure as well as identify meaningful patterns in the visual display, in particular, arbitrary shaped clusters. However, high-dimensional data often contains irrelevant dimensions which obscure the real distance between objects and even with carefully chosen DR techniques, the grouping information may still be hidden in the visual representation due to the noise. To reduce noise and preserve grouping information as well as structure of data, we propose a feature vector transformation approach which first design transformation strategy and then transform the original feature space by extending corresponding feature vectors. Such transformation is expected to provide better group separation in the final embedding. The resulting embedding can be evaluated by a quality measure to make sure the final embedding shows clear separation of classes, whilst still preserving data structure well. This is achieved by a quality measure combining class-related overlap measures on the one hand, and stress-based measures on the other. In this section we introduce our method for feature vector space transformation and exemplify our approach with two simple transformation strategies.

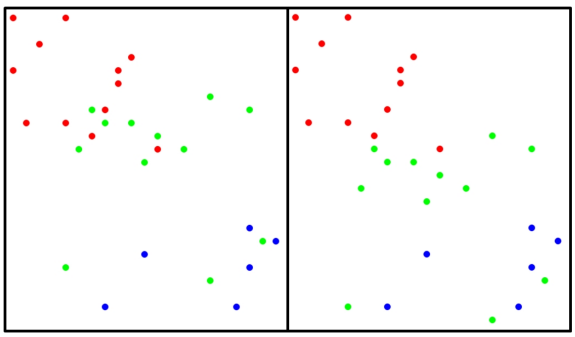

Figure 1. Example for a projection of two 2-dimensional data sets with three classes. Left: data set A, right: data set B. Classes 0, 1 and 2 are colored in red, green and blue.

The main idea of the feature space transformation is to extend the relevant features (e.g., mean values of selected dimensions) for better group separation in the final embedding. This can be achieved by adding additional feature vectors to the original feature space to leverage the noise introduced by irrelevant dimensions. For selecting dimensions to extend, there are a number of intuitive guidelines, for example, one can check the *range* of data values over a dimension - typically the smaller the range, the less likely the class is going to overlap with others in the dimension therefore dimensions that have high range may be less relevant to the class separation. Another choice is the *spread* - if all the class members share similar values in one dimension, it is likely this dimension is discriminative to the class label. In other words, generally speaking dimensions that have high spread may be selected for extension. Next we use these two measures for relevant feature selection.

To illustrate the feature space transformation approach, we generate two simple 2-dimensional data sets, A and B, each contains 30 objects that belong to three different classes. Figure 1 shows the 2D projection of both data sets. Colors are used to indicate class labels.

One simple feature space transformation strategy is the addition of mean values. Although a feature space can be transformed in many different ways, besides this simple strategy. For example, *median* or *mode* can be applied instead of means value (depending on the nature of data) etc. Also, the number of extensions of can vary. The maximum number

of extensions can be the total number of dimensions n, in which case we extend all dimensions. Our experimental results show that this maximum extension leads to a good group separation but loss of similarity preservation between groups objects 6. Next, we illustrate the mean-value extension strategy using the two data sets. First of all, the mean values mv_{dc} of each dimension $d \in \{1,\ldots,n\}$ for each class c are calculated (see Table 1). We found a simple heuristic approach based on mean spread of dimensions may already be suitable for selecting which dimensions to extend. Ideally the selected dimensions should be discriminative to the class labels. In our case, we take the *range* and *spread* (see below for details).

Table 1. Mean values mv_{dc} for each dimension $d \in \{1,2\}$ for each class $c \in \{0,1,2\}$ of data set A and B.

Class	mv_{1c} data A	mv_{2c} data A	mv_{1c} data B	mv_{2c} data B
c=0	10	10	10	10
c=1	16	16	16	19
c=2	20	24	20	24

We start from calculating the range r_d for each dimension $d \in \{1,\ldots,n\}$ with class labels c (here $c \in \{0,1,2\}$):

$$r_d = max(mv_{dc}) - min(mv_{dc}) \tag{1}$$

Result in range values for data sets A and B as shown in Table 2. Next we calculate the spread_d measure for each dimension

Table 2. Range values r_d for each dimension $d \in \{1,2\}$ of data set A and B.

	d=1 data A	d=2 data A	d=1 data B	d=2 data B
min	10	10	10	10
max	20	24	20	24
r	10	14	10	14

which is defined as:

$$spread_d = \begin{cases} r_d^2/sd_d & if\ sd_d \neq 0 \\ r_d^2 & else \end{cases} \tag{2}$$

where sd_d is the standard deviation of the differences between the ordered mean values of the classes within dimension d. With this heuristic we can reward equally spread mean values, the higher the spread measure the better. Table 3 shows the spread values for the data sets A and B. In the next section, we define quality measures that can be applied to evaluate the

Table 3. Spread-measure spread_d for each dimension $d \in \{1,2\}$ of data set A and B.

	d=1 data A	d=2 data A	d=1 data B	d=2 data B
difference between c0 and c1	6	6	6	9
difference between c1 and c2	4	8	4	5
standard deviation	1.4	1.4	1.4	2.8
range	10	14	10	14
spread	70.7	138.6	70.7	69,3

effectiveness of the above mentioned heuristic approaches.

4. QUALITY MEASURES

As mentioned previously, to visualize high-dimensional data a good embedding is expected to approximate the data structure well and highlight patterns. There are a number of quality measures for evaluating these properties (see Subsection 2.2), however not many of them take both aspects into consideration and allows the user the define the best compromise. We propose a new quality measure which combines three score functions for evaluating structural and visual aspects of an visual embedding respectively and gives users the freedom of adjusting the weight for each score. First of all, for measuring the structural preservation, we use two alternative quality measures which assess how well the structure of data is preserved in the embedding. Secondly, for measuring overlapping between groups, we combine an area-based overlapping measure and a density-based overlapping measure. The former calculates the size of the overlapping regions

between groups. The latter measures how objects between different groups overlap in a particular region. The combination gives a good indication of how well groups are separated in the projection.

Note that a wide range of measures exist for calculating the structure preservation, the area of overlap between two regions and the density of overlapping objects inside a region. The measures we use in this paper are used to illustrate the basic principle of our approach and can be replaced by other measures of similar nature.

4.1 Stress measure

To evaluate structuring preservation we apply *Sammon's stress*[25] and *K-ary neighborhoods measure*.[37] While Sammon's stress focuses on the quality of global pairwise similarity preservation,[25] k-ary neighborhood measure also shows the quality of local neighborhood preservation of an embedding. The k-ary neighborhood measure records two types of neighborhood preservation errors in the embedding, neighborhood intrusion error and neighborhood extrusion error.[28]

Given a data with N objects, Sammon's stress computes an error E, which represents how well the present configuration of N points in the lower dimensional space fits the N points in the high-dimensional space:

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{3}$$

The distance between data item i and data item j in the original high-dimensional space is defined by d^*_{ij}, and the distance between their lower dimensional projections by d_{ij}. The stress value is calculated from the data of the projection resulting from the transformed features, against the initial feature vector data. With this procedural method it is ensured that the distance preservation is measured to the initial feature space.

The *K-ary neighborhoods measure* is defined as

$$Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \frac{|n^*_i(K) \cup n_i(K)|}{KN} \quad , \tag{4}$$

where $n^*_i(K)$ is the set of indices of the K nearest neighbors of the i-th datum in the HD space whereas $n_i(K)$ corresponds to the set of indices of the K nearest neighbors in the LD space.

4.2 Overlap measures

For pattern search, it is important to avoid visual cluttering in the projection such that patterns can be easily identified and groups can be easily perceived. Visual cluttering in an embedding can be caused by either overlaying of objects in the display or overlap between group boundaries. E.g., Figure 2 shows 3 different projections of the same data set. While the left projection has a cluttered region on the top left corner where objects are plotted on top of each other and the middle one has a big overlap region between purple and orange group, the right projection shows much clearer view of the groups. To achieve a less cluttered embedding, we designed two overlap measures to evaluate the visual cluttering level of an embedding. The first measure calculates the size of overlapping region between groups and the second measure sums up the density of objects in overlapping regions.

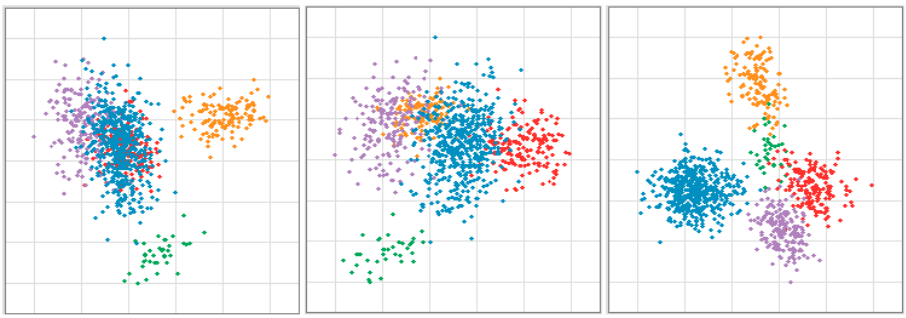

Figure 2. 2D projections of high-dimensional data with 5 classes, color of objects represent class labels.

Overlap area measure. To calculate the size of overlapping regions between groups of labeled points, we first need to define the region of each of the groups. The basic idea is to describe the group boundary by drawing a representative, enclosing hull of the objects that belong to the same group. A number of methods exist for computing boundaries for point sets, e.g., various convex hull generation methods[38–40] and the isocontour approach as proposed in.[41] In[42] we computed overlap between convex hulls for the visual comparison of the discrimination capabilities in alternative feature spaces. Here, we apply a different and less known hull formation method proposed by Moreira and Santos[43] which computes the region of a set of points as a concave hull based on a *k-nearest neighbors* approach. The advantage of this approach is that the generated region is usually more compact and better reflects arbitrary shaped groups in the embedding (see Figure 3). For each point that has to be connected to the next point, the algorithm first searches the best connection among its k nearest neighbors to make sure the result hull is as compact as possible. Compared to the isocontour approach, the parameter setting of Moreira and Santos' method is much simpler and can be automated. Only one parameter needs to be defined, which is the k parameter. It controls the smoothness of the computed hull. When k is set to 3, the algorithm automatically looks for the smoothest possible envelop, i.e., the most compact concave hull. In the worst case, when k is equal to *number of points-1*, the algorithm will output a convex hull.

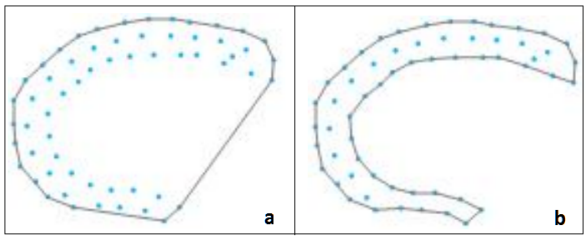

Figure 3. Regions of a set of points generated by: a) convex hull approach, b) concave hull approach.

Once the region of each group is defined, we calculate the overlap region $intersect(i,j)$ for each pair of groups i and j. The overlap area measure sums up the area of all the overlap regions between pairwise groups for the set g of groups:

$$ov_{reg} = \sum_{i=1}^{|g|-1} \sum_{j=i+1}^{|g|} intersect(i,j) \tag{5}$$

Figure 4 shows an example of the overlap area measure. Given three classes colored in red, green and blue, our method first computes the boundary of each class using the concave hull approach. The size of overlap region(s) between pairwise concave hulls is then calculated and summed up as the area overlap measure (black surrounded areas in Figure 4).

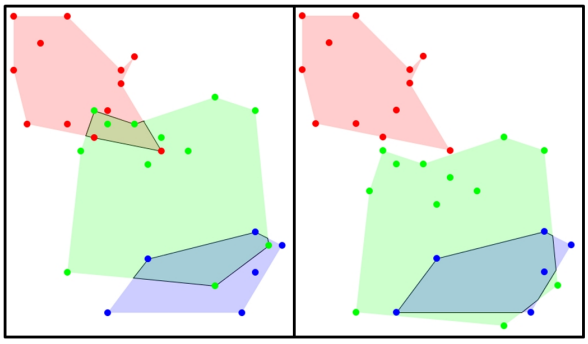
Figure 4. The black surrounded areas show the overlap of the concave hulls for data set A(left) and data set B(right).

Overlap density measure. The hull-based approach does not consider the possibly, non-uniform density of points. Therefore, we complement it by an overlap density measure that evaluates how strongly points are over-plotted in the visual display. The display area is divided into grid units (where the resolution of the grid can be adjusted). A Gaussian function G is used to determine whether a grid unit is occupied by a particular class depending on the density of objects inside the grid square. Once the Gaussian model for each class is computed, our approach checks pairwise classes to see how many grid units are occupied by more than two classes. The count will be summed up as overlap density measure. Equation (6) defines the overlap measure for a dataset with K classes and an image with P pixels. The Gaussian function is defined in Equation (7) where f is an indicator function which gives 1 in case when a grid square k is occupied, according to the Gaussian model, by a pair of classes, and 0 otherwise. In the remaining examples in this paper, we set grid resolution to 3 pixels and the σ value of the Gaussian model to 12 pixels. However the grid resolution can be adjusted and normalized to fit in the size of the display and the σ value can also be changed.

$$ov_{density} = \sum_{i=1}^{|K|-1} \sum_{j=i+1}^{|K|} \sum_{p=1}^{|P|} f(G_{ip}, G_{jp}) \tag{6}$$

$$f(G_{ip}, G_{jp}) = \begin{cases} 1 & if\ G_{ip} > 0\ and\ G_{jp} > 0 \\ 0 & else \end{cases} \quad (7)$$

Figure 5 shows two examples of overlap regions between classes. The overlap grid squares are shaded in gray, and the scale of the gray color indicates the density level.

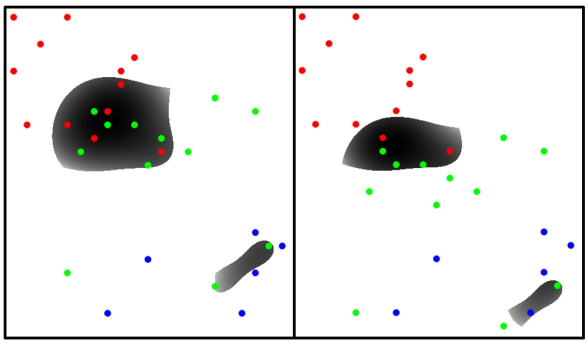

Figure 5. The gray shaded regions show the density overlap for data set A(left) and data set B(right).

Figure 4 and 5 show the visualization of area and density overlap of the projections of the two simple data sets A an B (Figure 1).

5. EXPERIMENTS

In this section we propose three different simple transformation strategies and demonstrate their effectiveness by applying each of them to transform eight different data sets. For each setting and each data set, we apply four different DR techniques to both the original and the transformed data and compare the quality of generated embeddings. We use two types of existing structural preservation measures, as well as the two overlap measures proposed in the previous section to evaluate the quality of each embedding. Next we detail the transformation strategies, the data, the embedding methods, the quality measures, and the result of the experiments.

Transformation Settings Given a labeled data set with n dimensions and m items, the feature space can be transformed in many ways. In this section we apply three very simple transformation strategies to demonstrate the effectiveness and generalizabililty of our approach. The basic idea is to extend the mean values of irrelevant dimensions to leverage noise in the data and thus avoid visual cluttering in the embedding. In our first and second strategy, we select the dimension that has the highest range / spread to represent the irrelevant dimension in the data. We extend the feature vector space by adding an additional dimension with mean value of the irrelevant dimension assigned to all the data items to reduce the noise introduced by the irrelevant dimension. To further analysis the compromise between structural preservation and visual clutter avoidance, our third strategy extends all n dimensions in the data by adding an additional dimension for each of them to the original feature space, with mean values of the dimension assigned to all data items. More specifically, our experiment extends a given feature space in the following ways:

1. Extend the initial feature vector with the mean value of the dimension that has the highest range.
2. Extend the initial feature vector with the mean value of the dimension that has the highest spread.
3. Extend the initial feature vector with mean values over all dimensions.

Data To evaluate the effectiveness of the different transformation settings, we test each approach with eight datasets (see Table 4 for more details). The data we chose consists of four synthetic datasets, created by Gaussian functions with a grid, and four benchmark datasets that have been used by various recent visualization publications, including the *ecoliProteins* data which encodes amino acid proteins sequences from the *E.colie bacteria*,[44] the *yeast* dataset which denotes cellular localization sites of proteins citeuci, the *tse300* dataset which records the weekly price history of 300 TSE index stocks in the year 2002,[44] and the *bbdm13* which is a subset from a hospital based case-control study designed to examine the epidemiology of fibrocystic breast disease.[45]

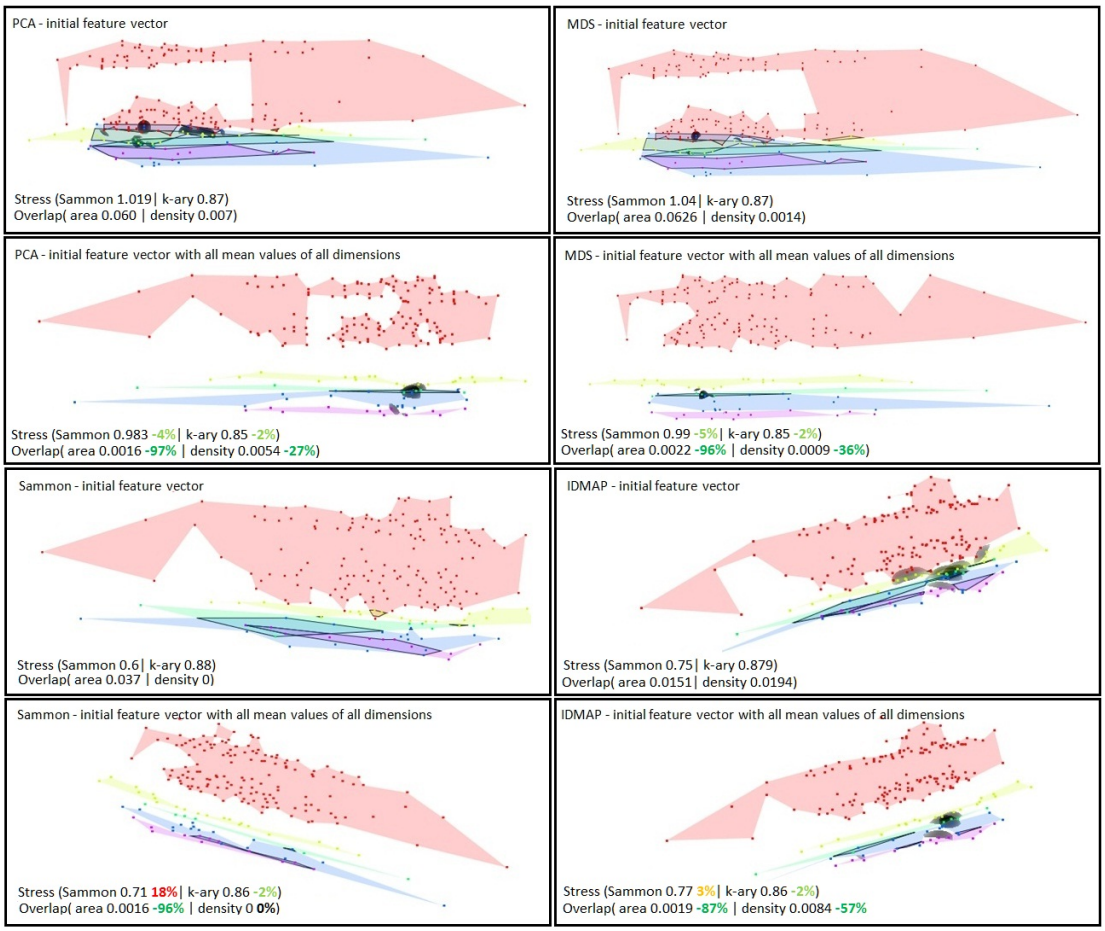

Figure 6. Embeddings of original and transformed 'bbdm13' data using 4 different DR technique (PCA, Classical Scaling, Sammon's Mapping, and IDMAP)

Table 4. List of data sets, ordered by number of dimensions.

Name	Type	Points	Dimensions	Classes	Provenance
twoSquare	synth.	968	3	4	[46]
gauss-d5-3c	synth.	500	5	3	[46]
gauss-d5-5c	synth.	500	5	5	[46]
ecoliProteins	real	332	7	8	visumap [44]
yeast	real	1452	8	9	uci [47]
tse300	real	244	9	8	visumap [44]
gauss-d10-5c	synth.	500	10	5	[46]
bbdm13	real	200	13	5	umass [45]

Embedding Methods We generated 2D embeddings of both the original data and the transformed data using four DR techniques (PCA, Classical Scaling, Sammon's mapping and IDMAP) implemented in the PEx (Projection Explorer) System[12] using the default parameter settings. PEx is a widely visualization tool for creating and exploring visual representations of high-dimensional data. PCA, Classical Scaling (MDS) are two of the most widely used linear DR techniques by existing visualization systems. Sammon's mapping is one of the most known nonlinear multidimensional scaling methods.[25] IDMAP is another nonlinear extension of MDS. It is an improved embedding technique for supporting visual exploration of multidimensional datasets.[48]

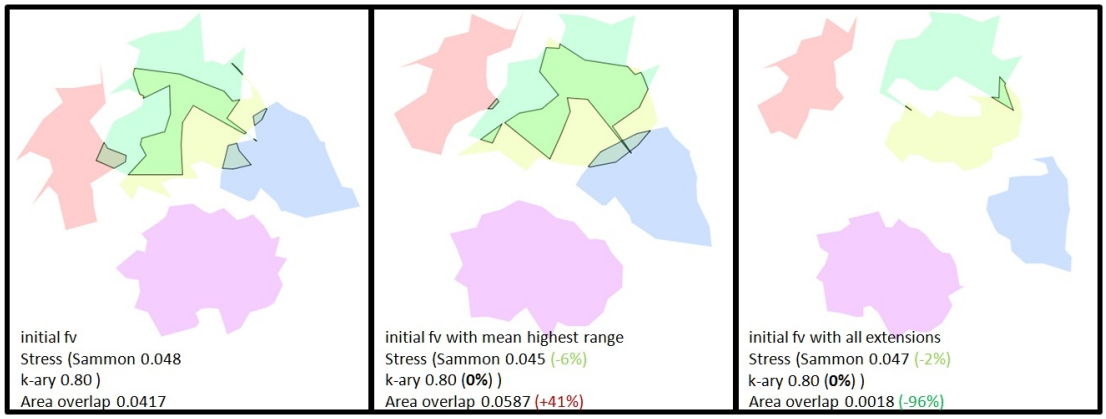

Figure 7. Overlapping areas in Sammon's Mapping of 'gauss-d10-5c' data set. Left: with the initial feature vector; middle: with the extension of the feature vector with the mean value of the dimension that has the highest range; right: with the extension of the feature vector with all mean values of all dimensions.

Quality Measures For each embedding, we evaluate the quality of its structural preservation by two measures, namely *Sammon's stress* and *k-ary neighborhood measure*. We also evaluate the quality of clutter avoidance by the two overlap measures, *overlap area* and *overlap density* as defined in the previous section.

Result The evaluation result are detailed in the table of Figure 9. Note that the Sammon's stress and k-ary neighborhood measure are always based on the distance matrix computed from the original data. First, we compare the difference between the stress measure and overlap measure before and after each transformation. The colored columns show the difference. Red indicates the transformation worsened the quality of embedding significantly (the value is more than 10 percent higher after transformation), orange indicates slightly worse performance (the value is higher, but less than 10% higher after transformation). Similarly, light green shows slight improvement ($< 10\%$) after transformation and dark green indicates significant improvement ($> 10\%$). Empty fields in the table indicate the choice of range value and spread value do not make a difference to the result.

From the result, it is not surprising to see that most the studied feature vector transformations lead to higher Sammon's stress value. This is not necessarily a bad sign, because the Sammon's stress records how well the lower dimensional embedding approximates the pairwise distances between data items in the high-dimensional data space. As mentioned previously, in high-dimensional data the real distances between data items are often obscured in the projection by the irrelevant dimensions and noise in the data. The feature space transformations aim to leverage noise in the data by extending relevant features, which naturally change the distances between pairwise data items in the transformed data space. With respect to neighborhood preservation, which is measured by the k-ary neighborhood measure, the result is significantly better than Sammon's stress. In majority of the cases the measure improves after transformation and there is no negative sign in the result. As a matter of fact, even with Sammon's stress, there are some positive cases. For example, with both *tse300* and *bbdm13* data sets, the quality of structural preservation actually improved after transformation in most of the cases. Given the fact that nearly all the overlap area and density measures decreased significantly after the transformation (as indicated by the dark green units in the table), this shows that transformations exist which can improve class separation without scarifying structure preservation.

The top two quarters in Figure 6 show the embeddings of the *bbdm13* data on 2D display. As we can see, there is no big change in the structure in terms of positioning of classes and data items in the display, and on the other hand the classes are better separated after the transformation. Even with the third transformation setting which simply adds all mean values over all dimensions.

6. DISCUSSION

The result of the evaluation is very encouraging and there is no DR technique or data set for which the transformation would not work at all. Even the trade-off between the structural preservation and the class separation is not as critical as

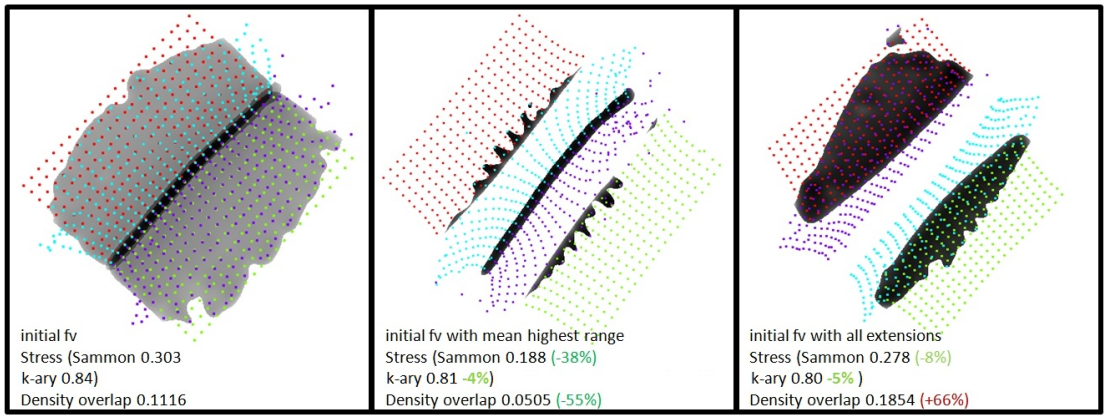

Figure 8. Area and density overlap of Sammon's Mapping of data set 'twosquare'. Left: with the initial feature vector; middle: extension of the feature vector with the mean value of the dimension that has the highest range; right: extension of the feature vector with all mean values of all dimensions.

we anticipated. A big advantage of the proposed approach is its flexibility. The framework is DR technique independent - it can be applied to improve the quality of an embedding generated by any DR techniques. Also it is easy to extend the framework by integrating user interactions and feedbacks. We next assess in more detail the experimental results, before discussing limitations and possible extensions of our approach.

6.1 Detailed Assessment of the Experimental Results

We see that for higher-dimensional data sets 9 Sammon's stress values increase. This is obvious, e.g., from comparing (*twoSquare*) (3 dimensions) to (*bbdm*) (13 dimensions in Figure 9, We note from our results, that the higher the dimensionality, the less the stress value increases. In other words, the transformation performs better for datasets with higher dimensionality. The same is true for the number of classes the dataset contains: The more classes the dataset contains, the less increase of the stress value after transformation, but the trend is not very obvious.

In terms of class separation we see the opposite effect by looking at the overlap measures: The overlap measures decrease more for data with fewer dimensions. There is no obvious correlation between the number of classes and the increasing overlap measures. Nearly all overlap measures in the table decreased after transformation, indicated by the predominant green numbers in Figure 9. That means the transformations perform well on avoiding over-plotting and visual cluttering. A decrease can also be seen in the example in Figure 6: The areas highlighted in black contours (overlap areas) and the gray shades (overlap density) are much smaller after the transformation. The area overlap measures decrease around 87% to 97%, and overlap density measures decreased around 27% to 57%.

We also studied the relationship between the number of extended dimensions and the quality of class separation in terms of both overlap area and overlap density. First we applied transformation setting 1, and 3 as discussed in Section 5 to the *gauss-10d-5c* data, and generated embeddings from both the original dataset and transformed datasets using the same DR technique. The result is shown in Figure 7 where overlapping regions are highlighted in black regions. From the 3 figures we can see that the 3 embeddings have nearly the same Sammon's stress value, however, the overlap area measure is worsened by the 1st transformation strategy (+41%), and improved substantially after applying the 3rd transformation strategy (-96%). We assume that extending more dimensions reduces the overall overlap area. Next we performed the same transformation settings to the *twosquare* dataset. Here we see a different correlation between the number of extended dimensions and overlap density. As we can see from Figure 8, the 1st transformation strategy gives the smallest overlap density measure (-55%) and at the same time reduced Sammon's stress measure substantially (-38%). The 3rd transformation strategy on the other hand worsened the over-plotting problem to a large extend (+66%) although it slightly improved the structural preservation (Sammon's stress -8%).It appears that adding too many dimensions brings negative effects to the overlap density measures.

6.2 Limitations and Possible Extensions

The contradictory effect on the stress and clutter measures brings up the challenge in finding the optimal number of extended dimensions as well as transformation strategy. In general, the space of possible feature transformations is huge.

In this paper, we were able to study just a limited set of simple transformations. The comparison of the proposed transformations shows that the range heuristic mostly performs better than the spread heuristic in our experiment. In terms of decreasing overlap measures the extension of the feature vector with all mean values of all dimensions often performs better. However, the conclusions we draw here are based on a limited number of transformation setting, DR techniques and relatively small datasets.

The quality measures cover important aspects of projection (stress and overlap). Defining overlap is subtle, and requires definition of a hull model. Concave hulls form intuitive and compact shapes, but other shapes are possible. The quality measures could serve as objective criteria for an in-depth search over the space of possible transformations, identifying the best result. A related extension possibility is to integrate it in an interactive analysis environment where the user can explicitly change the extensions method, and select dimensions for extension, and interact with the embedding results. The user could be allowed to set a trade-off to weight the different quality measures, arriving at an application- and user-dependent best choice.

Going one step further, users could be allowed to mark specific subsets of the data, e.g., specific classes, which are considered with priority in the quality criterion and a respective automatic search. An interactive user interface could let the user select relevant features and adjust transformation settings.

Note that our approach is based on availability of class labels. The feature extensions proposed should be compared against supervised projection techniques, which employ the class information by the definition of the projection. Such a comparison will be interesting to do. Our approach considers inter-class separation. However, also intra-class clutter should be avoided. To this end, measures based on the uniformity of the spatial distribution of points in the visualization, could be thought of. Finally, how to measure and improve the quality of projections in case where no class labels are available, is a challenging problem.

7. CONCLUSION & FUTURE WORK

We proposed a framework for improving the quality of a low-dimensional embedding of high-dimensional data by means of simple feature space transformation. The methods are simple and based on the idea of reducing the influence of irrelevant features and noise in the data, and extending by average features within classes of data points. A series of quality measures for assessing the structural preservation quality and the visual quality of a given embedding was proposed. The analyst can apply different transformation strategies to improve the quality of an initial embedding generated by any embedding techniques. An encompassing experimental evaluating assessed the effectiveness of the approach. The result of the evaluation are promising and showed that nearly all transformations benefited by better class separation, and in most of the cases, the structural preservation is not much affected. Our approach is simple to implement and generically applicable to any projection technique which operated on features.

We also identified a number of interesting extensions of our approach. The quality measures and extensions could be interactively controlled by the user. Automatic search in the space of possible transformations could bring up a number of candidate views for the user to inspect. Finally, our approach should be experimentally compared against supervised projection methods, which implicitly leverage class information in doing the projection.

ACKNOWLEDGMENTS

The authors wish to thank Halldor Janetzkó for his support in some parts of the implementation and Geoff Ellis for his feedback.

REFERENCES

1. A. Hinneburg, C. Aggarwal, and D. Keim, "What is the nearest neighbor in high dimensional spaces?," in *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pp. 506–515, 2000.
2. I. Jolliffe, *Principal Components Analysis*, Springer, 3rd ed., 2002.
3. W. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika* **17**, pp. 401–419, December 1952.
4. P. Demartines and J.Hérault, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of datasets," *IEEE Transactions on Neural Networks* **8**(1), pp. 148–154, 1997.

5. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science (New York, N.Y.)* **290**, pp. 2319–2323, Dec. 2000.
6. C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: The generative topographic mapping," *Neural Computation* **10**, pp. 215–234, 1998.
7. G. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems* **15**, pp. 833–840, 2003.
8. J. Lee and M. Verleysen, *Nonlinear dimensionality reduction*, Springer, 2007.
9. A. Wismüller, M. Verleysen, M. Aupetit, and J. A. Lee, "Recent advances in nonlinear dimensionality reduction, manifold and topological learning," in *ESANN*, 2010.
10. L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *Tilburg University Technical Report* , 2009.
11. J. Venna and S. Kaski, "Comparison of visualization methods for an atlas of gene expression data sets," *Information Visualization* **6**, pp. 139–154, May 2007.
12. F. V. Paulovich, M. C. F. Oliveira, and R. Minghim, "The projection explorer: A flexible tool for projection-based multidimensional visualization," in *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing*, *SIBGRAPI '07*, pp. 27–36, IEEE Computer Society, (Washington, DC, USA), 2007.
13. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**, pp. 2323–2326, 2000.
14. M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*, pp. 585–591, MIT Press, 2001.
15. C. L. John, *Latent variable models: an introduction to factor, path, and structural analysis*, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1986.
16. L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research* **9**, pp. 2579–2605, 2008.
17. D. H. Jeong, C. Ziemkiewicz, B. D. Fisher, W. Ribarsky, and R. Chang, "iPCA: An interactive system for PCA-based visual analytics," *Comput. Graph. Forum* **28**(3), pp. 767–774, 2009.
18. J. Choo, H. Lee, J. Kihm, and H. Park, "iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction," in *IEEE VAST*, pp. 27–34, 2010.
19. S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller, "DimStiller: Workflows for dimensional analysis and reduction.," in *IEEE VAST*, pp. 3–10, IEEE, 2010.
20. A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North, "Observation-level interaction with statistical models for visual analytics," in *IEEE VAST*, pp. 121–130, IEEE, 2011.
21. S. Bremm, T. v. Landesberger, J. Bernard, and T. Schreck, "Assisted descriptor selection based on visual comparative data analysis," *Wiley-Blackwell Computer Graphics Forum* **30**(3), pp. 891–900, 2011. (Proceedings of Eurographics / IEEE-VGTC Symposium on Visualization 2011).
22. S. Bremm, T. v. Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher, "Interactive comparison of multiple trees," in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 31–40, IEEE Computer Society, 2011.
23. J. de Leeuw and W. Heiser, "Theory of multidimensional scaling," in *Handbook of Statistics*, ch. 13, pp. 285–316, North-Holland Publishing Company, Amsterdam, 1982.
24. J. Kruskal, "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new index of condensation," in *Statistical Computation*, R. Milton and J. Nelder, eds., pp. 427–440, Academic Press, New York, 1969.
25. J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Trans. Comput.* **18**, pp. 401–409, May 1969.
26. J. Venna and S. Kaski, "Neighborhood preservation in nonlinear projection methods: An experimental study," in *Proceedings of ICANN 2001*, G. Dorffner, H. Bischof, and K. Hornik, eds., pp. 485–491, Springer, Berlin, 2001.
27. L. Chen, *Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis*. PhD thesis, University of Pennsylviana, July 2006.
28. J. Lee and M. Verleysen, "Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods," in *JMLR Workshop and Conference Proceedings (New challenges for feature selection in data mining and knowledge discovery)*, Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. Van de Peer, eds., **4**, pp. 21–35, Sept. 2008.

29. L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of nonlinear manifolds," *Journal of Machine Learning Research* **4**, pp. 119–155, June 2003.
30. J. Venna and S. Kaski, "Comparison of visualization methods for an atlas of gene expression data sets," *Information Visualization* **6**, pp. 139–154, May 2007.
31. E. Bertini and G. Santucci, "Visual quality metrics," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, *BELIV '06*, pp. 1–5, ACM, 2006.
32. J. Johansson and M. D. Cooper, "A screen space quality method for data abstraction.," *Comput. Graph. Forum* **27**(3), pp. 1039–1046, 2008.
33. A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim, "Combining automated analysis and visualization techniques for effective exploration of high-dimensional data," in *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pp. 59–66, 2009.
34. M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan, "Selecting good views of high-dimensional data using class consistency," *Comput. Graph. Forum* **28**(3), pp. 831–838, 2009.
35. A. Inselberg, "The plane with parallel coordinates," *The Visual Computer* **1**(2), pp. 69–91, 1985.
36. E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *Proceedings of the IEEE Symposium on IEEE Information Visualization (InfoVis)* **17**, pp. 2203–2212, 2011.
37. J. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing* **72**(7–9), pp. 1431–1443, 2009.
38. R. Graham, "An efficient algorithm for determining the convex hull of a finite planar set," *Information Processing Letters* **1**(4), pp. 132–133, 1972.
39. R. Jarvis, "On the identification of the convex hull of a finite set of points in the plane," *Information Processing Letters* **2**(1), pp. 18–21, 1973.
40. J. Koplowitz and D. Jouppi, "A more efficient convex hull algorithm," *Information Processing Letters* **7**(1), pp. 56–57, 1978.
41. C. Collins, G. Penn, and M. S. T. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Trans. Vis. Comput. Graph.* **15**(6), pp. 1009–1016, 2009.
42. T. Schreck and C. Panse, "A new metaphor for projection-based visual analysis and data exploration," in *IS&T/SPIE Conference on Visualization and Data Analysis*, pp. 64950L.1–64950L.12, SPIE Press, 2007.
43. A. J. C. Moreira and M. Y. Santos, "Concave hull : a k-nearest neighbours approach for the computation of the region occupied by a set of points," pp. 61–68, 2007.
44. VisuMap Technologies Inc., "VisuMap Data Repository," 2011. http://www.visumap.net/, last accessed Nov. 2011.
45. University of Massachusetts, "Statistical Data and Software Help," 2011. http://www.umass.edu/statdata/statdata/, last accessed Nov. 2011.
46. M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Computer Graphics Forum* **31**(3), pp. 1335–1344, 2012. (Proceedings of Eurographics / IEEE-VGTC Symposium on Visualization 2012).
47. A. Frank and A. Asuncion, "University of California Irvine (UCI) Machine Learning Repository," 2010.
48. R. Minghim, F. V. Paulovich, and A. d. A. Lopes, "Content-based text mapping using multi-dimensional projections for exploration of document collections," in *Vis. and Data Analysis 2006*, *Proc. SPIE-IS&T Electronic Imaging*, pp. 259–270, SPIE, (San Jose, California, USA), 2006.

	feature vector with highest spread extension				feature vector with highest range extension				feature vector with all extension			
PCA	Sammon	k-ary	overlap area	overlap density	Sammon	k-ary	overlap area	overlap density	Sammon	k-ary	overlap area	overlap density
twoSquare	26%	-3%	0%	-72%					26%	-3%	0%	-72%
gauss-d5-3c	15%	-1%	-60%	-20%	0%	0%	-53%	-17%	78%	-2%	-100%	-100%
gauss-d5-5c	13%	-2%	-36%	-31%	7%	0%	-18%	-12%	39%	-2%	-94%	-76%
ecoliProteins	6%	-1%	-49%	-8%					73%	-3%	-89%	-42%
yeast	13%	0%	-26%	-16%	8%	0%	-15%	-11%	36%	-2%	-38%	-31%
tse300	-5%	0%	8%	-7%	-4%	0%	-79%	-5%	767%	-9%	-100%	-100%
gauss-d10-5c	0%	-1%	-33%	-8%					12%	-1%	-92%	-78%
bbdm13	-4%	-2%	-97%	-19%					-4%	-2%	-97%	-27%
MDS												
twoSquare	26%	-3%	0%	-31%			-47%	-5%	26%	-3%	0%	-31%
gauss-d5-3c	15%	-1%	-49%	-10%	0%	0%	-25%	-12%	78%	-2%	-100%	-100%
gauss-d5-5c	13%	-2%	-31%	-44%	7%	0%			39%	-2%	-90%	-91%
ecoliProteins	7%	-1%	-43%	2%			-12%	-10%	74%	-3%	-71%	-49%
yeast	14%	0%	-12%	-10%	8%	0%	-2%	-7%	36%	-2%	-39%	-31%
tse300	-5%	0%	-12%	-7%	-5%	0%			746%	-9%	-100%	19%
gauss-d10-5c	3%	-1%	-17%	38%					15%	-1%	-92%	-81%
bbdm13	-6%	-2%	-96%	-29%					-5%	-2%	-96%	-36%
Sammon												
twoSquare	-38%	-4%	-19%	-55%					-8%	-5%	-13%	66%
gauss-d5-3c	6%	0%	-75%	-13%	-3%	0%	-58%	0%	31%	0%	-100%	-100%
gauss-d5-5c	5%	-2%	-40%	-28%	2%	0%	-11%	-15%	28%	-2%	-99%	-95%
ecoliProteins	15%	-1%	-51%	33%					52%	-3%	-90%	-75%
yeast	-30%	-1%	-7%	8%	-28%	0%	4%	3%	-29%	-3%	-28%	-36%
tse300	-21%	0%	-90%	-91%	-22%	0%	-69%	-57%	913%	-6%	-100%	-100%
gauss-d10-5c	-6%	1%	41%	141%					-2%	1%	-96%	-100%
bbdm13	12%	-2%	-92%	0%					18%	-2%	-96%	0%
IDMAP												
twoSquare	77%	-4%	-88%	-59%					119%	-4%	-86%	-65%
gauss-d5-3c	11%	0%	-85%	-13%	0%	0%	-67%	-16%	53%	-1%	-100%	-100%
gauss-d5-5c	9%	-2%	-26%	-43%	4%	0%	-21%	-17%	38%	-3%	-100%	-84%
ecoliProteins	8%	-1%	-41%	-27%					50%	-4%	-87%	-70%
yeast	-6%	-1%	0%	-12%	-9%	0%	-11%	-4%	-3%	-3%	4%	-23%
tse300	-3%	0%	-92%	-9%	-2%	0%	15%	-5%	789%	-7%	-100%	-54%
gauss-d10-5c	-2%	0%	14%	-2%					7%	0%	-93%	-93%
bbdm13	1%	-2%	-87%	-52%					3%	-2%	-87%	-57%

Figure 9. Stress, area and density overlap for initial feature vectors and their transformations for 4 synthetic and 4 real data sets (ordered by number of dimensions) for projection techniques PCA, MDS, Sammon's Mapping and IDMAP. The changes are colored in (red) orange, if the transformations perform (very) badly with respect to the initial feature vector. If the transformations perform better or much better, the percentages for the changes are colored in light or dark green.