Eurographics Conference on Visualization (EuroVis) 2017
J. Heer, T. Ropinski and J. van Wijk
(Guest Editors)

Volume 36 (2017), Number 3

Linear Discriminative Star Coordinates for Exploring Class and Cluster Separation of High Dimensional Data

Yunhai Wang[1] Jingting Li[1] Feiping Nie[2] Holger Theisel[3] Minglun Gong[4] and Dirk J. Lehmann[3]

[1]Shandong Univeristy, China
[2] Northwestern Polytechnical University, China
[3]University of Magdeburg, Germany
[4]Memorial Univeristy, Canada

Abstract

One main task for domain experts in analysing their nD data is to detect and interpret class/cluster separations and outliers. In fact, an important question is, which features/dimensions separate classes best or allow a cluster-based data classification. Common approaches rely on projections from nD to 2D, which comes with some challenges, such as: The space of projection contains an infinite number of items. How to find the right one? The projection approaches suffers from distortions and misleading effects. How to rely to the projected class/cluster separation? The projections involve the complete set of dimensions/features. How to identify irrelevant dimensions? Thus, to address these challenges, we introduce a visual analytics concept for the feature selection based on linear discriminative star coordinates (DSC), which generate optimal cluster separating views in a linear sense for both labeled and unlabeled data. This way the user is able to explore how each dimension contributes to clustering. To support to explore relations between clusters and data dimensions, we provide a set of cluster-aware interactions allowing to smartly iterate through subspaces of both records and features in a guided manner. We demonstrate our features selection approach for optimal cluster/class separation analysis with a couple of experiments on real-life benchmark high-dimensional data sets.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

High dimensional data occur in a number of domains, such as biology, physics, or economy. To support the user gain insights from such data, visual cluster exploration is an effective way [CL06, SMT13], which visualizes the clusters in low dimensional space. Regarding this, star coordinates [Kan01] (SC) project n dimensional data onto a 2D space. By interactively changing the SC's parameters, the cluster structures of high dimensional data may be revealed. However, finding good projections through this kind of user interactions is a tedious, time-consuming, error-prune, trial-and-error process. Thus, we formulate star coordinates as a special case of linear dimensionality reduction [DL07, JW02] to automatically separate the data's clusters/classes best and in order to facilitate a reliable and simple feature selection techniques for our domain experts for both: labeled and unlabeld nD data. By interacting with our discriminative star coordinates (DSC), a purposeful visual cluster analysis of high dimensional data is enabled. Moreover, since the projection properties deciphers the influence of dimensions to the projection, our approach allows an efficient feature selection by the domain experts.

Feature selection is in that regard one of the most relevant applications for domain experts that aim to identify a subset of relevant features for model construction. Regarding cluster exploration, our application partners are interested in finding subspaces, i.e. a set of features/dimensions, that allows linear separation between the clusters. Although projection pursue [FT74] can automatically find the best linear separation according to a particular quality measure, it does not allow users to interactively explore the relationship between features and cluster separation. On the other hand, interactively exploring the projection space to find good projections from the scratch is a tedious procedure.

One popular example for this is the cluster separation in the WDBC data set: In good projections, doctors figured out that two global clusters separate well, which allow to discriminate between malignant and benign breast cancer tumour cells. Figure 1(left) shows the separation. This way, the doctors were able to identify which features/dimensions allow such a separation of the cell types.

However, apparently, this interactive process may be overwhelmingly time consuming and it guarantees no success. Even worse, success is just coincidently possible. Nevertheless, the domain ex-

perts still rely on the interactive exploration approach. Why is it so? Due to a simple fact: our domain experts are experts for their domain, but neither for data analysis nor for,e.g., machine learning. Thus, we provide a visualization approach that the domain experts can use without any further effort or knowledge. Nonetheless, it enables an efficient feature selection for interesting subspaces to reliably allow the linear separation of cluster/classes.

In summary, the main contributions of this paper include:

- We introduce data-driven algorithms for labeled and unlabeled data to get the best separating projections;
- We judge why our projections successfully address the issue of reliability, misleading, and the infinite number of projections;
- We illustrate how a (quick) feature selection is enabled by our approach and define a visual analytics scheme that facilitates interactive feature analysis and cluster exploration; and
- We demonstrate our prototype for a set of standard real benchmark nD data sets.

In the following we motivate our work by discussing the cluster/class separation background and by deriving and introducing the infinite number issue, the reliability issue, and the misleading issue, which will be successfully addressed by our technique.

2. Motivation

One relevant information in order to analyze nD data is to know how the data are grouped and separated in between, denoted clustering and class separation, respectively.

In detail, if data – such as illustrated in Figure 1(a) – are linearly separable, one can assume a linear boundary that separates the data in class A on the left side and class B on the right. The (toy) example in Figure 1(b) illustrates this (blue line). The less distant a data record is to this boundary, the less likely is it to have classified this record correctly. In other words, the inverse distance to the classification boundary is a measure for the reliability and quality of the classification. Thus, one would suppose the quality of classification of Figure 1(c) to be better than that of Figure 1(b), because the class-based records are more distant to the boundary. Clearly, one is interested in maximizing the empty margin (red lines) around the boundary in order to get reliably cluster/classes of the data. Moreover, the smaller the average distance of a cluster record is to its average point (called centroid), the more likely is it that records belong to this cluster and does not potentially shape up a cluster by its own. See Figure 1(e-f). In fact, compact clusters that minimizing their spread are preferred, such these in Figure 1 (f). To sum up, for reasons of reliability one is interested in clusters that maximize the margin and minimizing their spread, which well fits with the objective of LDA [**?**]. These are the separation properties that are usually to be considered as to be relevant and interesting in the nD data.

However, the separation properties of high-dimensional data itself cannot be visually analysed, since the lack of dimensionality. Thus, an established approach is to project the data (in a multivariate sense) onto a 2D domain. Figure 1(a) illustrates how such a projection from nD to 2D may look like. Unfortunately, there is an infinite number or multivariate projections, often condense in the term curse of dimensionality: Some projections show the data separation reliably. Some projections show separation, that does not

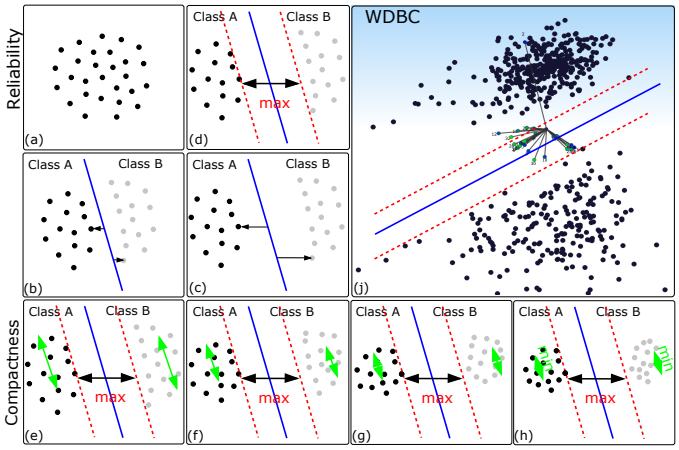

Figure 1: *Cluster/Class Separation Properties*

exist in the data, due to several reasons, such as distortion, overlapping, and so fourth in the projection process. These projection are not reliable at all. Moreover, the majority of projections show no separation properties at all.

From this, a set of questions arise, which motivates our work:

- **Infinity Projections Issue:** How can we find a finite set of projections (or a single) that separates the projected points best, from the number of infinite ones?
- **Reliability Issue:** How can we maximize the reliability of the seen clusters, i.e., how to maximize the margins and compactness of the clusters/classes?
- **Misleading Issue:** How can we be sure that the seen separation illustrated the the nD data separation correctly?

In the next section we revise work which is related to ours.

3. Related work

In this section, we review related work from the field of projection-based data visualization.

Multidimensional Data Visualization: By showing all pairwise combinations of scatterplots, a scatterplot matrix [BC87] reveals all pairwise correlations. Parallel coordinates [Ins85] represent the dimensions as a set of parallel axes and render each data tuple as a polyline. These methods are tailored to visualize correlation and trends, but are not effective in cluster analysis. Recently researchers have tried to enhance the cluster analysis capability of these methods [JLJC05, ZYQ*08], however, they can handle 20 dimensions at most due to their poor scalability. See Keim et al. [AMS02] for more details.

The most often used dimension reduction (DR) methods include PCA, LDA [JW02], projection pursuit (PP) [FT74], and many variants of MDS [BG05]. PCA is an unsupervised method that pursues a subspace preserving the maximal data variances, while LDA selects the best subspace to linearly separate different classes from a labeled data set. To combine the advantages of these two methods, Choo et al. [CBP09] propose a two-stage framework for vi-

sualization of labeled data. LDA and PCA both assume the data or class follows a Gaussian distribution, which might not be true for some data. In contrast, PP pursues the most interesting projection where the "interestingness" is defined as the non-Gaussianity of the projection [FT74]. All these methods belong to linear DR methods that find the subspace by seeking a projection matrix. In contrast, MDS [BG05] is a non-linear method that aims to preserve the distances between data pairs in a low dimensional space. To reduce the computation cost, part-linear multidimensional projection (PLMP) [PSN10] and local affine multidimensional projection (LAMP) [JPC*11] construct the embedding space through a subset of representatives.

It is challenging to learn how each dimension is related to the clustering result, because original data attributes are lost in the final visualization. To address this issue, LDA or its variants are integrated together with star coordinates [RSRDS16, VLL11], whose axes vectors are defined by the projection matrix. Our DSC further extends such work to unlabeled data and provides a set of interactive methods that facilitates the user to explore how each dimension contributes to the class/cluster structures.

Star Coordinates: The method of star coordinates is proposed by Kandogan et al. [Kan01]. It is defined by uniformly arranging n coordinate axes on a circle with the origin at the center. With the $2 \times n$ projection matrix defined by n axes, star coordinates represent a 2D linear embedding of the original data. Based on such low dimensional projections, Friedman et al. [FT74] find projection that robustly reveals structures in the data with the projection pursuit method. To make a complete view of the data, Asimov introduced the grand tour [Asi85] to visualize high dimensional data with a sequence of two-dimensional embedding. Due to its efficiency in visualizing high-dimensional clusters, Teoh and Ma [TM03] use star coordinates to facilitate interactive visual classification. Recently, Lehmann and Theisel [LT13] extend these approaches to an orthography-preserving star coordinates, they provide optimal and short data tours with them [LT15], and they generalize them in a concept of general projective maps [LT16]. To increase estimation accuracy, Sanchez and Sanchez [RSS14] suggest to combine data centering with the orthography-preserving star coordinates [LT13].

The affine multivariate projection techniques still suffer from the mentioned issues in Sec. 2. It remains unclear how to select a separating projection from the infinite projection spaces, and if a projection separates cluster/classes well, it is unclear how reliable/misleading these clusters are w.r.t. the original nD data, and thus feature selection remains challenging.

Feature Selection: The feature selection itself has been considered from the perspective of the general classifier description [GE03] or of the feature classifier in nD data [SP14]. Either way, there is still a need for visualization-based feature selection approaches, which provide intuitive ways to rank features, compare features among dimensions, and merge or combines features. Most work in this area focus on using statistical metrics to characterize the relationship between features [Guo03, SS05, TFH11]. Recently Krause et al. [KPB14] propose an interactive features selection framework, which interacts directly with the feature selection and classification algorithm. Having similar spirit with this work, our DSC aims to interactively find features that allow linear separation between clusters.

4. Technical Background for Affine Projections

Given a data set $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^n$, which has been centered and normalized, affine star coordinates project a nD point $\mathbf{x}_i$ to a 2D point $\mathbf{x}'_i$ by the matrix multiplication

$$\mathbf{x}'_i = \mathbf{G}^T \cdot \mathbf{x}_i, \tag{1}$$

with $\mathbf{G}^T = (\mathbf{g}_1, \cdots, \mathbf{g}_n)$ is the $2 \times n$ projection matrix. Here, the anchor point $\mathbf{g}_j \in \mathbb{R}^2$ deciphers the influence and weight of dimension j to the projection, which can be interactively changed by the user in order to visually explore the data. Figure 2 illustrates star coordinates for a 3D data set.

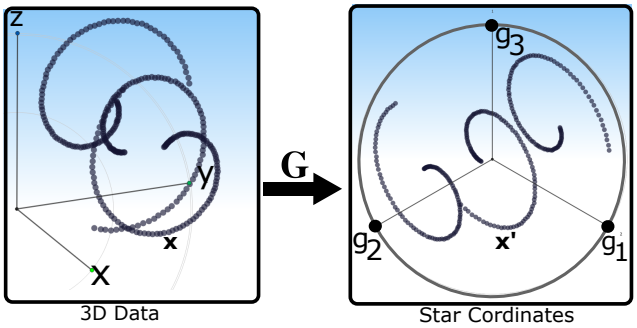

Figure 2: *3D data and a related Star Coordinates.*

Apparently, manually finding a proper $\mathbf{G}$ is challenging. Thus, automatic projection selection techniques have been more and more used in recent years, for instance, by minimizing a certain objective function $\text{argmin}_{\mathbf{G}} f_{\mathbf{X}}$. The objective $f_{\mathbf{X}}$ defines the features of interest to be captured in the data. We also rely on an optimization process in order to define the best class separating projection matrix $\mathbf{G}_{dsc}$, called **Discriminative Star Coordinates** (DSC). It will subsequently be introduced in detail.

5. Discriminative Star Coordinates

In this section, we are looking for a projection matrix $\mathbf{G}_{dsc}$ that separates the classes/clusters best in the projection space. For this, we define our Discriminative Star Coordinates (DSC) $\mathbf{G}_{dsc}$ as:

$$\mathbf{G}_{dsc} = (\mathbf{e}_1(\mathbf{B})\ \mathbf{e}_2(\mathbf{B}))^T, \tag{2}$$

where $\mathbf{e}_i$ are the two eigenvectors to the two largest eigenvalues of an optimal separation matrix $\mathbf{B}$. The construction of $\mathbf{B}$ depends on whether the data are labeled or not. Thus, we explain in Section 5.1 the construction of $\mathbf{B}$ for labeled nD data; and in Section 5.2 for unlabeld nD data. Finally, we explain the visual design of our scheme and illustrate the discriminative star coordinates exemplarily.

5.1. Discriminative Star Coordinates for labeled Data

If the data comes with an a priory classification, i.e., the data are labeled, then both the number of classes and records per class are known. In this case, the optimal separation matrix $\mathbf{B}$, which maximizes the class separation in projection space, can be defined by discriminant analysis approaches from the field of machine learning by following [Fuk90].

For this, assume $\mathbf{X}$ consists of samples from k classes, the corresponding label is $\mathbf{y} = \{y_1, \cdots, y_n\}$, where $y_i \in \{1, \cdots, k\}$. Three scatter matrices need to be considered: total scatter $\mathbf{S}_t$, between-cluster scatter $\mathbf{S}_b$, and within-cluster scatter $\mathbf{S}_w$ are defined as follows [Fuk90]:

$$\mathbf{S}_t = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{S}_b + \mathbf{S}_w, \tag{3}$$

$$\mathbf{S}_b = \sum_{i=1}^{k} \frac{m_i}{m} (\mu_i - \mu)(\mu_i - \mu)^T, \tag{4}$$

$$\mathbf{S}_w = \sum_{i=1}^{k} \sum_{y_j = i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T, \tag{5}$$

where m is the number of data records, $\mathbf{x}_i \in \mathbb{R}^n$, $\mu = \sum_{i=1}^{m} \mathbf{x}_i / m$ is the mean of the data, m_i is the number of the records of the i^{th} class, and $\mu_i = \sum_{y_j = i} \mathbf{x}_j / m_i$ is mean of the i^{th} class. The within-cluster scatter of the projected $\mathbf{X}$ can be expressed as:

$$\begin{aligned} \mathbf{S}'_w &= \sum_{i=1}^{k} \frac{m_i}{m} (\mathbf{x}'_i - \mu'_{y_i})(\mathbf{x}'_i - \mu'_{y_i})^T \\ &= \sum_{i=1}^{k} \mathbf{B}^T (\mathbf{x}_i - \mu_{y_i})(\mathbf{x}_i - \mu_{y_i})^T \mathbf{B} \\ &= \mathbf{B}^T \mathbf{S}_w \mathbf{B}. \end{aligned}$$

For simillar reasons apply:

$$\mathbf{S}'_b = \mathbf{B}^T \mathbf{S}_b \mathbf{B}. \tag{6}$$

Classes/clusters are optimally separated if the between-class scatter $\mathbf{S}'_b$ is maximized and the within-class scatter $\mathbf{S}'_w$ is minimized, i.e., an optimal transformation $\mathbf{B}$ maximize trace $\mathrm{tr}(\mathbf{S}'_b)$ but minimize trace $\mathrm{tr}(\mathbf{S}'_w)$:

$$\max \frac{\mathrm{tr}(\mathbf{S}'_b)}{\mathrm{tr}(\mathbf{S}'_w)} = \max \frac{\mathrm{tr}(\mathbf{B}^T \mathbf{S}_b \mathbf{B})}{\mathrm{tr}(\mathbf{B}^T \mathbf{S}_w \mathbf{B})}, \tag{7}$$

approximated by

$$\max \mathrm{tr} \frac{\mathbf{B}^T \mathbf{S}_b \mathbf{B}}{\mathbf{B}^T \mathbf{S}_w \mathbf{B}}. \tag{8}$$

This gives $\mathbf{B}$ as the eigenvectors to

$$\text{the largest } k-1 \text{ eigenvalues of } \frac{\mathbf{S}_b}{\mathbf{S}_w} \tag{9}$$

for data $\mathbf{X}$ with k classes. Since it cannot find a discriminative 2D subspace to characterize the data with two classes, we use the state-of-the-art To address this issue, we solve Eq. 8 with the state-of-the-art method [JNZ09], which can efficiently find the global optimum.

5.2. Discriminative Star Coordinates for unlabeled Data

Unlabeled data does not have a known data classification. Thus, there is a need to integrate an initial classification or clustering process of the data into the process of maximizing the separation in the projection space, in order to reveal our discriminative star coordinates. For this, we subsequently following the *LDA-km* algorithm [DL07] to address this issue: Since so-called irrelevant dimensions may confuse clustering algorithms [PHL04], a K-Means clustering is applied in a certain and relevant subspace. This leads to the slightly modified optimization issue, as

$$\max_{\mathbf{B},\mathbf{y}} \frac{\mathrm{tr}(\mathbf{B}^T \mathbf{S}_b \mathbf{B})}{\mathrm{tr}(\mathbf{B}^T \mathbf{S}_w \mathbf{B})}. \tag{10}$$

A closed form solution is not available anymore, but an iterative pin-pong-esque algorithm does the trick, by alternatively fixing $\mathbf{B}$ and $\mathbf{y}$ [DL07]. For this, our approach initiated an initial $\mathbf{B}^T$ by a PCA [VDMPVdH09] of the data $\mathbf{X}$. Then, until it converges, our approach does:

- Fixing $\mathbf{B}$:
  $\mathbf{y}$ is obtained by performing K-means in the space $\mathbf{B}^T\mathbf{X}$ (to minimize the influence of the initial random centers, we run K-means several times and choose the one with the smallest within-cluster variation).
- Fixing $\mathbf{y}$:
  $\mathbf{B}$ is given by Eq. (9).

From experiences, the approaches converges usually in less than 10 iterations.

Finally, we are interested in finding the cluster number k which yield the optimal quality of separation. Thus, in order to quantitatively evaluate the quality of separation, we measure its quality with the *silhouette coefficient* [KR09], which penalizes class overlap:

$$Silh = \frac{1}{d} \sum_{i}^{d} \frac{b(\mathbf{x}'_i, \mathbf{x}'_j) - a(\mathbf{x}'_i, \mathbf{x}'_j)}{\max(a(\mathbf{x}'_i, \mathbf{x}'_j), b(\mathbf{x}'_i, \mathbf{x}'_j))} \in [-1, 1], \tag{11}$$

where a is the average and b is the minimal distances between member $\mathbf{x}'_i$ and the remaining members $\mathbf{x}'_j$ of the same class. Then, from a set of different cluster numbers $k_i, i = 1, ..., f$, the final number of clusters k is this that maximizes the average silhouette $\overline{Silh}$ [KR09] over all classes c in projection space.

5.3. Visual Design of Discriminative Star Coordinates

In recent years, some standard design metaphors for multivariate projections have been established – amongst others – by [Kan00, LT13, FIB*14, LT15, LT16], on which our approach also rely on, which are:

- Each anchor point $\mathbf{g}_i$ of $\mathbf{G}_{dsc}$ is given as vector $\mathbf{g}_i - \mathbf{o}$ to the origin $\mathbf{o}$, placed in the center of the visualization space.
- Per anchor point, a circle with the radius of its magnitude is drawn to improve the user's accuracy in perceiving data.
- The dimension index i is given at each anchor point $\mathbf{g}_i$.
- Each projected point is colored w.r.t. it class/cluster id.

Here, we represent the contribution of each dimension by the length of its corresponding axis, where shorter axes indicate that the related dimensions will likely have less contribution in separating clusters [RSRDS16]. Beside these standard metaphors, we add two ovational metaphors to ease the inspection of the separation quality:

- All elements of a class are enveloped by a closed class-based colored contour. For this, our approach draws a smoothed version of the class-wise convex hull.

- The uncertainty for a projected record belonging to a certain class is the larger the more distant it is to its centroid. Thus, to visualize the confidence, we set its opacity based on its distance to its cluster centroid.

5.4. Outcome for Labeled/Unlabeled Data

We illustrated the DSC outcome for labeled/unlabeld data.

Labeled Data: For the case of labeled data, Figure 3 shows radial Star Coordinates (SC) [Kan00], orthographic star coordinates (OSC) [LT13], and our discriminative Star Coordinates (DSC) in comparison for two labeled data sets. On the top, a synthetic 3D data set with 2 classes based on two Gaussians can be seen, and below the 4D *Iris* data set with 3 classes is presented. Our DSC gives the best class separation.

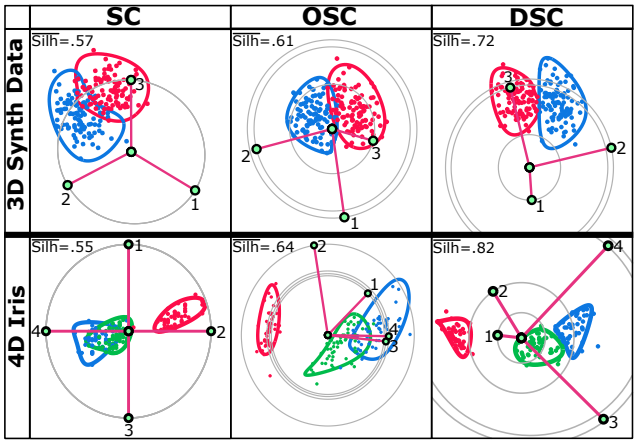

Figure 3: *Discriminative Star Coordinates for labeled data compared with traditional Star Coordinates for (top) a 3D synthetic data set with 2 classes and (bottom) the 4D Iris data set with 3 classes.*

Unlabeled Data: For the case of unlabeld data, Figure 4 shows a set of different iterations of the DSC approach from Section 5.2 for the unlabeled 16D *Pendigits* data set. Since we found that $k = 3$ clusters describe the inherent structure well, the figure illustrates the outcome for this and different iteration numbers (a-c). For reasons of completeness, we illustrate in (d) the case that K-means will apply on the nD dataspace, instead of ignoring irrelevant dimensions (cf. Sec. 5.2). In total, we see that the cluster separation increases quickly with the iteration number in our unlabeled DSC.

Interpretation: Note that in traditional Star Coordinates it is coincident if there will be a good cluster/class separation be seen, or not. In fact, it is completely random if any chosen configuration meet a well separating view or not, since the data were not considered. Thus, if no clusters/classes can be seen, it does not mean that the classes could not be well separated and no clusters exist, respectively. It just means that potentially an inappropriate configuration is used. In contrast, our data-driven DSC approach enforced class/cluster separation reliably, i.e., if no separation is seen than one can rely on the fact that no linear separation is possible; but if the data are linear separable then they will be reliably separated in the projection spaces.

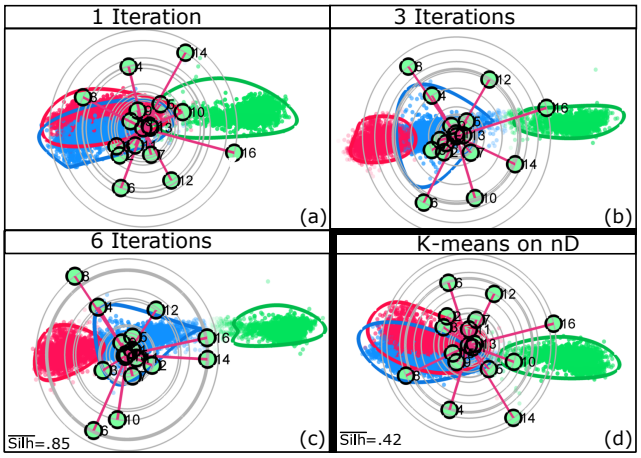

Figure 4: *Discriminative Star Coordinates for unlabeled Pendigits data.*

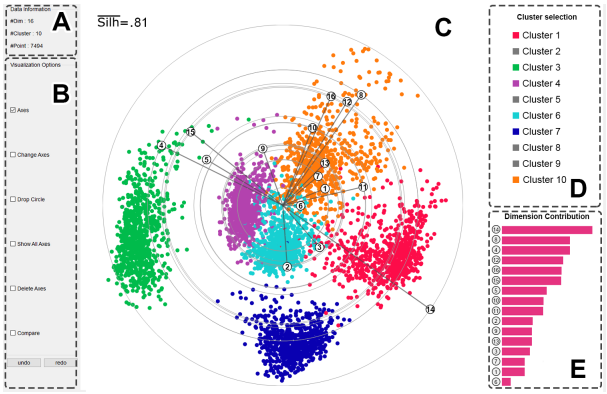

Figure 5: *The interface for visual cluster exploration and feature selection: (a) the command view; (b) the DSC view; (c) the cluster selection view; and (d) the bar view. In this example, the six selected classes are well separated in the DSC view, where the un-selected class is indicated by the gray icon in the cluster selection view.*

6. Visual Cluster Exploration and Feature Selection

In this section, we present our prototype for the DSC-based visual cluster exploration and feature selection. Our interface consists of five coordinated views, illustrated in Figure 5: (a) the command view; (b) the DSC view; (c) the cluster selection view; and (d) the bar view. In the command view, the visual parameters can be set, while the clusters of interest can be selected in the cluster selection view. The bar view provides a sorted overview of the feature contribution to the seen clustering, which gives a technique to detect in a simple way features that contribute in a majority sense to the seen projection. The related DSC projection itself is seen in the DSC view. In addition, in order to intuitively reveal the relationship between clusters and data dimensions, we provide a set of cluster-aware interactions.

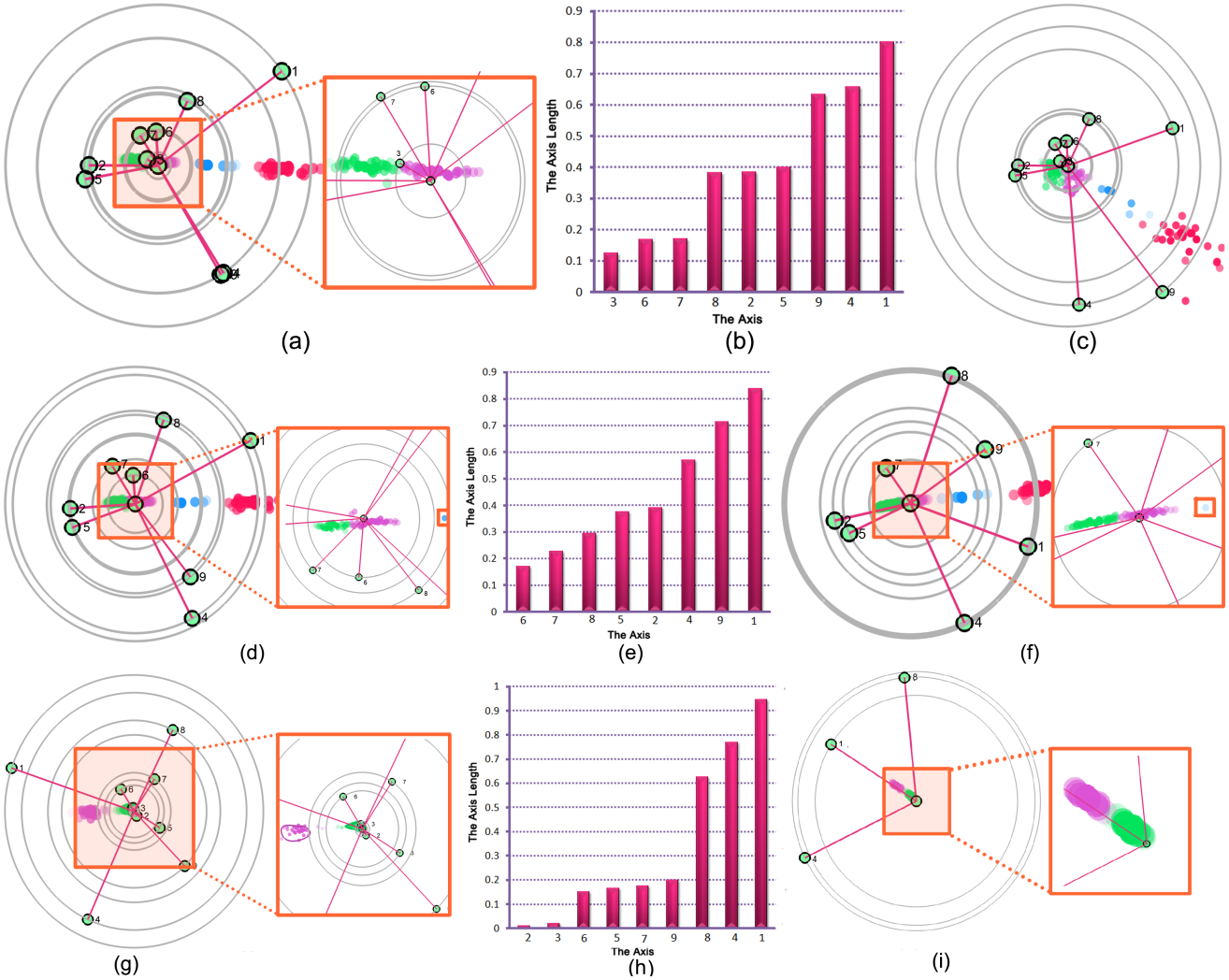

Figure 6: *DSC visualizations for the BreastTissue data set, where the number of clusters is set to 4. (a,b) The default DSC view and the bar chart view; (c) The DSC view generated by transforming the 1^{st} and 9^{th} axes; (d,e) The DSC view and the bar chart view generated by removing the 3^{rd} axis in (a); (f) The DSC view generated by removing the 6^{th} axis in (d), where a blue point highlighted in a red box shows the difference in cluster separation between (d) and (f); (g,h) The DSC view and the bar chart view generated by zooming in the green and magenta clusters; (i) The DSC view generated by using the dimensions corresponding to the right-most 3 axes (8^{th}, 4^{th} and 1^{st}) in (g).*

6.1. Anchor Point Interaction

Although the DSC configurations provided by LDA and ULDA maximize the separation of all clusters, the separation among clusters of specific interest to users may not be maximized. Thus, we allow interaction of the anchor points $\mathbf{g}_i$ of our DSC: Since the default DSC configuration visualizes the clusters with the maximal overall separation, an anchor point interaction cannot give a better configuration. However, this is not the goal here, but such interactions support to study how the dimension/features influences the separation degree of existing clusters. For example, shortening the 1^{st} axis in the example of Figure 6(a) reduces the separation between red and blue clusters, suggesting that the two clusters are separated along the first data dimension. Lengthening the 9^{th} axis enlarges the shapes of the green and magenta clusters, indicating that both clusters have intra-cluster variation along the 9^{th} dimension. Thus, such interaction is useful within the feature selection analysis process.

6.2. Iterative Feature/Dimension Selection

Since the length of each axis reveals the contribution of the corresponding dimension/feature to the clustering, the user can hypothesize which dimension is uninformative. For this, we introduce the following iterative process: By observing the bar chart of the sorted axes lengths (Figure 6(b)), the user can remove dimensions/features whose corresponding contributions are below a threshold. With the remaining subset of dimensions/features, the DSC is re-computed

and our coordinates view are updated. Through visually inspecting the change of the cluster separation, the user may determine whether the removed dimensions are indeed irrelevant or redundant. This visual analytics related iteration process could also be done automatically. But finding a proper threshold is not easy. Thus, our user-based approach integrate the domain knowledge of the user and the user's ability to judge the results into the parameter space exploration process. This could just hardly mimicked automatically and is useful when the axes lengths bring ambiguity.

Figure 6(d,e,f) show two iterations of axes filtering. From the bar chart view in Figure 6(b), the user can see that the 3^{rd} axis has the smallest length. Thus, the user may assume this axis does not affect the clustering a lot and may filter this axis and get a new result as shown in Figure 6(d,e). Comparing the zoomed view of Figure 6(d) with the one in Figure 6(a), the separations among different clusters have not been change much. This is consistent with the result from the qualitative measure. While removing the next smallest (the 6^{th}) axis in Figure 6(d), the separation between the blue and the magenta clusters becomes smaller; see the blue point highlighted by red boxes in Figure 6(d,f) is an example. This indicates that the 6^{th} dimension of the input data helps to separate the blue and the magenta clusters.

6.3. Cluster-based Zooming

DSC shows that all clusters are separated with the maximal separation, but it does not imply that the subset of clusters are maximally separated. To help the user to learn the separation of a subset of clusters and how the data dimensions contribute to such clusters, we allow the user to perform cluster-based zooming where DSC takes the points belonging to the selected clusters as the input data and visualize them to the user.

Figure 6(g,h) shows the DSC view and the bar chart view by zooming in the selected green and magenta clusters. Comparing the axes lengths shown in Figure 6(g) with Figure 6(a), we can see the 2^{nd} and 5^{th} axes become less important for separating the green and magenta clusters while the 7^{th} and 8^{th} axes become more important. By carefully observing the bar chart view (Figure 6(h)), we can see that the first three axes (8^{th}, 4^{th} and 1^{st}) contributes 83% and thus we hypothesize whether these three dimensions are sufficient to separate the two classes. Hence, the user further picks these three axes and gets the DSC view (Figure 6(i)), where two clusters are close to each other but still can be separated. As a result, the user can learn that the corresponding three dimensions are sufficient to discriminate these two clusters.

7. Case Study: DSC-based Feature Selection in Practice

We have implemented and tested our prototype visualization system on a PC with an Intel Xeon E5540 2.53 GHz CPU and 4.0 GB RAM using C++. Our system can achieve interactive visualization for the data sets used in this paper. Since both labeled and unlabeled data can be visualized with our system, we demonstrate its effectiveness with two data sets, where we analyze the data with the following pipeline: (1) run DSC to get the cluster visualization; (2) remove irrelevant axes by observing the axes lengths; (3) rotate and scale the axes. As linear DR methods, LDA and LDA-km both are very fast where the computation of the DSC for all tested data sets can be finished in less than 1 second.

7.1. Unlabeled USDA Food Data

First we present a case-study on the USDA food composition data set (http://www.ars.usda.gov/), which was organized by Tatu et al. [TMF*12]. After preprocessing, this data contains 722 records and each record consists of 18 dimensions where each dimension represents one type of nutrients.

Although many approaches [Jai10] can automatically find the number of clusters, it is not easy to incorporate user's experience into them. Instead, our ULDA configured DSC provides an interactive way to find the proper number of clusters. Figure 7(a,b,c,d) show the clustering results by setting number of clusters to 3, 4, 5, and 6, respectively. We can see almost all clusters in Figure 7(a,b,c) are well separated, while the new cluster in cyan shown in Figure 7(d) only contains one point which overlaps with the green yellow cluster. Hence, the proper number of clusters appears to be 5. This is consistent with the result generated by using Bayes Information Criterion (BIC) [Jai10].

Since the cluster structures in Figure 7(d) are consistent with the ones shown in Figure 7(a,b,c), one of co-authors, a machine learning expert, wants to analyze how different features contribute to such structures. Figure 7(e) shows the contribution of each dimension/feature to the clustering result of Figure (d), where we can see that the lengths of the left-most four axes : 1^{st} (Beta_Carot), 8^{th} (Magnesium), 17^{th} (Vit_E), and 11^{th} (Sodium), are noticeably shorter than the rest. In contrast, the right-most 4 axes 18^{th} (Water), 3^{rd} (Carbohydrt), 4^{th} (Energ_Kcal), and 10^{th} (Protein) take major roles and the sum of the contributions of these dimensions is 68.3%. Considering this, we remove the first four axes and get a new DSC visualization (Figure 7(f)).

Compared Figure 7(f) with Figure 7(a), we can see that five clusters in Figure 7(f) are still separated although the red cluster has a little overlapping with the green yellow and blue clusters. Adjusting the axes in Figure 7(f), 5 clusters are shown more clearly in Figure 7(h). Figure 7(g) illustrates the contributions of the rest axes to the clustering in Figure 7(f), where the most important 4 axes are the same with the ones in Figure 7(e) and contribute 73.4%. From this, we can conclude that there is a high level of redundancy in the original data, and the main nutrients in discriminating different foods are Water, Carbohydrt, Energy and Protein. By exploring 216 subspaces, Tatu et al. found that Protein is a dominant dimension in clustering by exploring 216 subspaces, which is consistent with our conclusion.

7.2. Labeled PENDIGITS Data

The PENDIGITS data set [AA97] contains 7494 training samples, where each sample is a digit and by 8 $(x;y)$ coordinates leading to a 16-dimensional feature vector. Note that the coordinates are both resampled along the pen's original trajectory and are normalized. As a result, the reconstructed samples may look distorted and may not look like the represented digits. For example, Figure 8(f,g,i,j) show four digits 6, 0, 9 and 4.

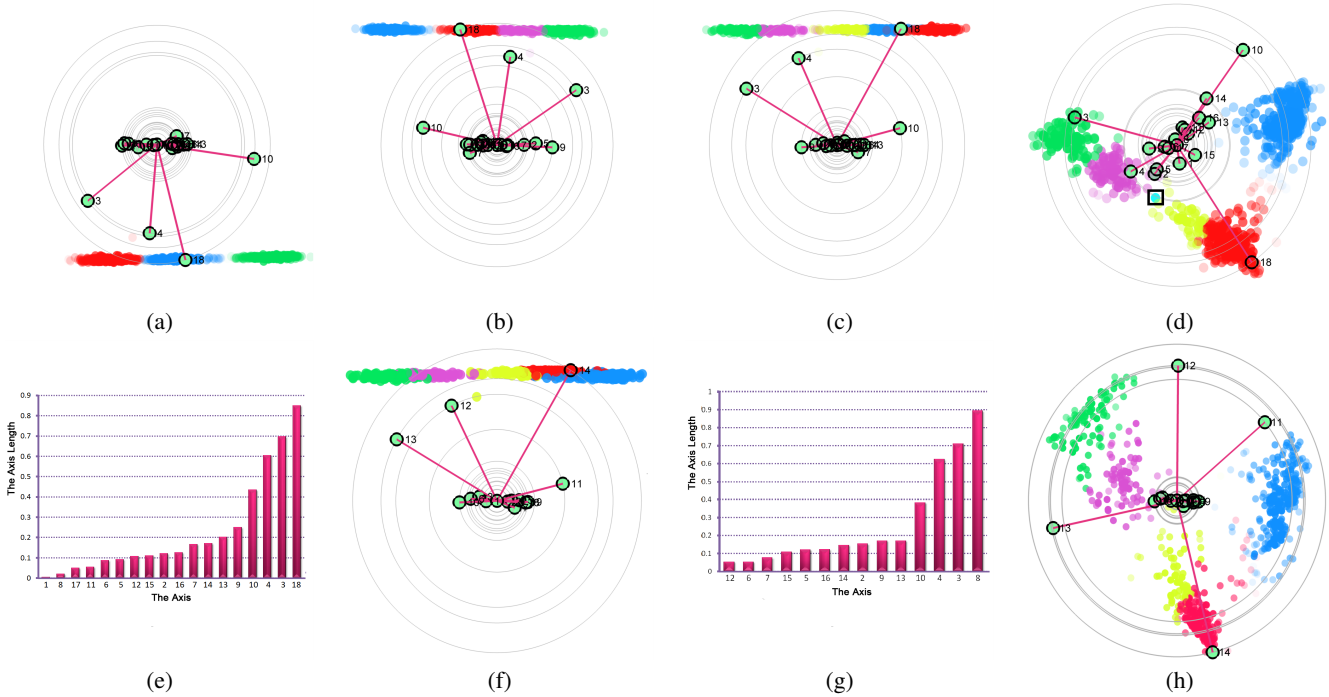

Figure 7: *The star coordinates visualization for the unlabeled USDA data set. (a-d) show the star coordinates where the numbers of cluster are set to 3, 4, 5, 6, respectively; (e) the bar chart view of the DSC result in (c); (f,g) the DSC view and the bar chart view generated by removing first four axes whose length are noticeably shorter than the rest shown in (e); (h) The DSC view generated by manipulating the most important 4 axes, allowing the five clusters to be shown more clearly.*

In this case study, we focus on explore how to separate different clusters (digits) and detect the outliers from the clustering results. Since this data has been widely used to test dimension reduction methods, we first compare the projections generated by three DR methods: projection pursuit (PP), the method used by Van Long and Linsen [VLL11] (VL), and LDA, as shown in Figure 8(a,b,c), respectively. We can see that none of them can separate all ten classes, but they produce the class structures with different levels of separation, where the silhouette coefficients of these three results are 0.14, -0.05 and 0.26, respectively. In particular, LDA results the classes with the minimal overlapping (see Figure 8(c)) while VL produces classes with the largest overlapping. Although both methods maximize the separation between classes, LDA simultaneously minimizes the spread so that the classes in Figure 8(c) are more compact than the ones in Figure 8(b). Although PP is not targeted at class separation, its resulted class structures have better separation than the ones in Figure 8(a). This indicates that some classes do not follow Gaussian distributions so that LDA cannot clearly separate all of them. On the other hand, integrating LDA with our DSC can reveal how each dimension contributes to the class structures. As shown in Figure 8(a), the last $(x;y)$ (15^{th} and 16^{th} dimensions) plays the most important role in differentiating different digits. This suggests that where the writing ends provides the strongest clue about which digit is written.

Moreover, the presented cluster-based zooming allows the user to further explore clusters of interest and see whether and how they can be separated. For example, the red and blue clusters (digits 0 and 6, respectively) overlap with each other in Figure 8(a,b). Zooming into these two clusters makes them separated as shown in Figure 8(c), except for an outlier highlighted in a black box. To investigate why this outlier appears in the red cluster, we found its corresponding sample and compare it with representative samples from both clusters. Figure 8(e,f,g) show these three samples, where we can see the outlier sample looks different from representative samples from both clusters, which explains that why it is projected to the far corner of the cluster for digit 0. Figure 8(d) shows another example, where yellow, orange, and light blue clusters (represent digits 4,9,1, respectively) are separated after cluster-based zooming, except one orange point appears in the yellow cluster. To inspect the corresponding sample for the outlier (Figure 8(h)), we compare it with two representative samples of the yellow and orange clusters. From Figure 8(h,i,j), we can see that the shape of this outlier sample indeed looks ambiguous.

8. Discussion

Our visual analytics approach facilitates an interactive and iterative feature selection approach in order to find features that separate classes/clusters efficiently, successfully, and quickly. Some real-life examples were illustrated in Sec. 7. Moreover, our approach decouples domain experts from the need to have broad knowledge in

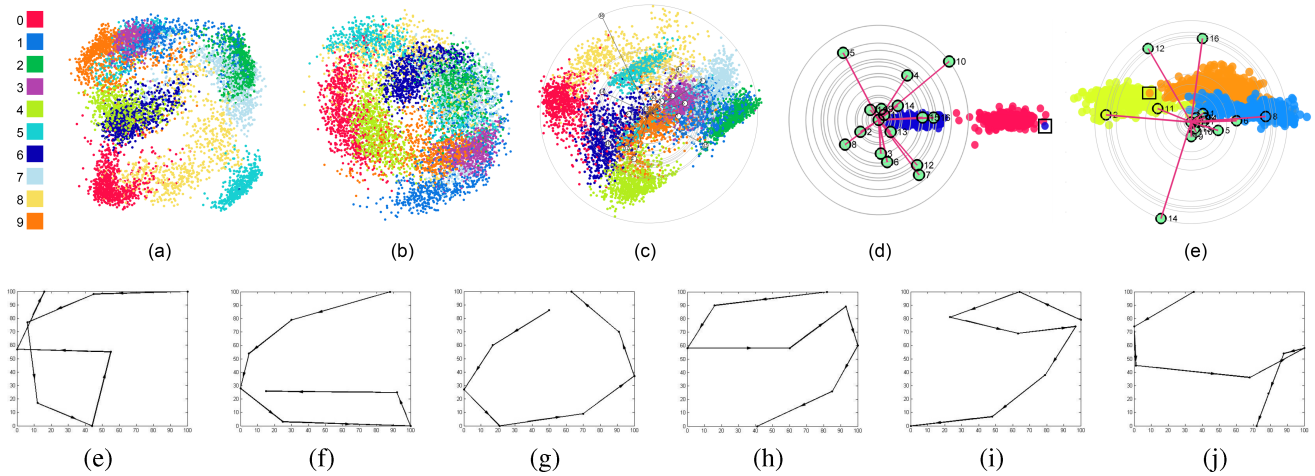

Figure 8: *The star coordinates visualization for the labeled PENDIGITS data set, which contains 10 different clusters and each cluster corresponds to a digits. (a) The scatterplot visualization generated by projection pursuit result; (b) The scatterplot visualization generated by the result of the method [VLL11]; (c) The default DSC view; (c) The DSC view generated by zooming in the red and blue clusters (represent 0 and 6, respectively), where the outlier is highlighted with a black box; (d) The DSC view generated by zooming in the yellow, orange and light blue clusters (represent 4, 9, and 1, respectively), where the outlier is highlighted with a black box; (e) The digit 6 corresponding to the highlighted blue outlier in (c); (f,g) The representative samples from the clusters for digits 0 and 6, respectively; (h) The digit 9 corresponding to the highlighted orange outlier in (d); (i,j) The representative samples 9 and 4 from the two clusters, respectively.*

computer science or machine learning to conduct exploration tasks, and allows them to solely focus on their specific domain.

For this, our DSC approach manages labeled and unlabeld data, and it addresses the three introduced issues that domain experts (and users in general) have: It detects data-driven a projection that separates classes/clusters well and provides it to the user. This way, the **Infinity Projection Issue** is addressed and a trial-and-error-based interactive search (for separation exploration) in the unlimited projection spaces is not required anymore. The computational time is negligible compared to the time consuming interactive search. In that regard, the DSC is a subset but superior to the available visual affine projection approaches, such as SC or OSC, where it is random to meet a separating view. Due to the data-driven nature, our DSC maximizes margins and the compactness in a linear separable manner, meaning that it gives reliable the view with the best separation quality (cf. **Reliability Issue**). Again, this fact makes it superior to SC and OSC in general: these approaches are not data-driven and even a well-separating view could be based on misleading distortion effects [LT13] and is thus not reliable. Since even the compactness is maximized w.r.t. structures in the data, the DSC shows cluster without the misleading effect (cf. **Misleading Issue**): a seen non-compact cluster is non-compact because its compactness cannot be enlarged anymore. Thus, the user can rely on that the cluster is really non-compact in the data. The same applies also for compact clusters in the data.

Since we rely on the first largest eigenvalues of a well-separating transformation, there is still a loss of information in the projection process. This is an issue for projection techniques in general, since a low-dimensional embedding is not bijective. For reasons like that and to broaden the analytic possibilities in general, we allow in our visual analytics concept to interact with the anchor points in order to evaluate the dimension-wise influence for the cluster-separation, we provide an iterative feature selection concept to evaluate the separation of interesting subspaces, and an interactive cluster-zooming and inspection for individual cluster evaluation. In fact, we allow dimension-wise inspections and inspections of subsets of both features and records.

However, our approach comes with some limitations. First, we use the axis length to measure the feature contribution, which assumes the features are independent to each other. This might not be true if the features are high correlated with each other. Second, we consider only linear separation and we leave non-linear separation issues to future work. Thus, if no linearly well-separating view can be found by our technique, it does not mean that the data may be non-linear separable. Last, only continuous numeric data can be handled up to now. In general, it is challenging to deal with categorical data visually or even worse to mix up metric and categorical data. This restriction applies also to our approach but we are interested in figure this issue out in the future.

9. Conclusion

Our linear discriminative star coordinates shows clusters with the maximal linear separation and allows a quick feature selection application within a visual analytics scheme. This enables a set of interactions to study how each dimension is related to the clustering. This way, the user can analyze how clusters are formed in the high dimensional space, as illustrated for our benchmark data. In the future, we are going to extend this approach to non-linear separation schemes and we want to allow also the use of categorical data.

References

[AA97] ALIMOGLU F., ALPAYDIN E.: Combining multiple representations and classifiers for pen-based handwritten digit recognition. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition* (1997), vol. 2, IEEE, pp. 637–640. 7

[AMS02] A. K. D., MÜLLER W., SCHUMANN H.: Visual data mining. In *Star-Report, Eurographics* (2002). 2

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143. 3

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing scatterplots. *Technometrics 29*, 2 (1987), 127–142. 2

[BG05] BORG I., GROENEN P. J.: *Modern multidimensional scaling: Theory and applications*. Springer, 2005. 2, 3

[CBP09] CHOO J., BOHN S., PARK H.: Two-stage framework for visualization of clustered high dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (2009), pp. 67–74. 2

[CL06] CHEN K., LIU L.: ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Transactions on Information Systems 24*, 2 (2006), 245–294. 1

[DL07] DING C., LI T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the International Conference on Machine Learning* (2007), pp. 521–528. 1, 4

[FIB*14] FUCHS J., ISENBERG P., BEZERIANOS A., FISCHER F., BERTINI E.: The influence of contour on similarity perception of star glyphs. 2251–2260. 4

[FT74] FRIEDMAN J., TUKEY J.: A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers 100*, 9 (1974), 881–890. 1, 2, 3

[Fuk90] FUKUNAGA K.: *Introduction to statistical pattern recognition*. Academic press, 1990. 3, 4

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *J. Mach. Learn. Res. 3* (Mar. 2003), 1157–1182. 3

[Guo03] GUO D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*, 4 (2003), 232–246. 3

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91. 2

[Jai10] JAIN A. K.: Data clustering: 50 years beyond k-means. *Pattern recognition letters 31*, 8 (2010), 651–666. 7

[JLJC05] JOHANSSON J., LJUNG P., JERN M., COOPER M.: Revealing structure within clustered parallel coordinates displays. In *Proceedings of the IEEE Information Visualization Symposium* (2005), pp. 125–132. 2

[JNZ09] JIA Y., NIE F., ZHANG C.: Trace ratio problem revisited. *Neural Networks, IEEE Transactions on 20*, 4 (2009), 729–735. 4

[JPC*11] JOIA P., PAULOVICH F. V., COIMBRA D., CUMINATO J. A., NONATO L. G.: Local affine multidimensional projection. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2563–2571. 3

[JW02] JOHNSON R. A., WICHERN D. W.: *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002. 1, 2

[Kan00] KANDOGAN E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium* (2000), vol. 650, p. 9–12. 4, 5

[Kan01] KANDOGAN E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), pp. 107–116. 1, 3

[KPB14] KRAUSE J., PERER A., BERTINI E.: Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Trans. Vis. & Comp. Graphics 20*, 12 (2014), 1614–1623. 3

[KR09] KAUFMAN L., ROUSSEEUW P. J.: *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009. 4

[LT13] LEHMANN D. J., THEISEL H.: Orthographic star coordinates. *IEEE Trans. Vis. & Comp. Graphics 19*, 12 (2013), 2615–2624. 3, 4, 5, 9

[LT15] LEHMANN D. J., THEISEL H.: Optimal sets of projections of high-dimensional data. *IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE Information Visualization)* (2015). 3, 4

[LT16] LEHMANN D. J., THEISEL H.: General projective maps for multidimensional data projection. *Computer Graphics Forum (Proc. Eurographics) 35*, 2 (2016). 3, 4

[PHL04] PARSONS L., HAQUE E., LIU H.: Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter 6*, 1 (2004), 90–105. 4

[PSN10] PAULOVICH F. V., SILVA C. T., NONATO L. G.: Two-phase mapping for projecting massive data sets. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1281–1290. 3

[RSRDS16] RUBIO-SÁNCHEZ M., RAYA L., DIAZ F., SANCHEZ A.: A comparative study between radviz and star coordinates. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 619–628. 3, 4

[RSS14] RUBIO-SANCHEZ M., SANCHEZ A.: Axis calibration for improving data attribute estimation in star coordinates plots. *IEEE Trans. Vis. & Comp. Graphics 20*, 12 (2014), 2013–2022. 3

[SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Vis. & Comp. Graphics 19*, 12 (2013), 2634–2643. 1

[SP14] SINGH V., PATHAK S.: Feature selection using classifier in high dimensional data. *CoRR abs/1401.0898* (2014). 3

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4*, 2 (2005), 96–113. 3

[TFH11] TURKAY C., FILZMOSER P., HAUSER H.: Brushing dimensions-a dual visual analysis model for high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2591–2599. 3

[TM03] TEOH S. T., MA K.-L.: Starclass: Interactive visual classification using star coordinates. In *Proceedings of the Third SIAM International Conference on Data Mining* (2003). 3

[TMF*12] TATU A., MAAS F., FARBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (2012), pp. 63–72. 7

[VDMPVdH09] VAN DER MAATEN L., POSTMA E., VAN DEN HERIK J.: Dimensionality reduction: a comparative review. *J Mach Learn Res 10* (2009), 66–71. 4

[VLL11] VAN LONG T., LINSEN L.: Visualizing high density clusters in multidimensional data using optimized star coordinates. *Computational Statistics 26*, 4 (2011), 655. 3, 8, 9

[ZYQ*08] ZHOU H., YUAN X., QU H., CUI W., CHEN B.: Visual clustering in parallel coordinates. *Computer Graphics Forum 27*, 3 (2008), 1047–1054. 2