Applied Soft Computing 27 (2015) 474–486

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Discernible visualization of high dimensional data using label information

Asef Pourmasoumi Hasan Kiyadeh, Amin Zamiri, Hadi Sadohgi Yazdi*, Hadi Ghaemi

Computer Department, Ferdowsi University of Mashhad, P.O. Box: 9177948974, Mashhad, Iran

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 23 July 2014
Accepted 18 September 2014
Available online 30 September 2014

Keywords:
Visualization
Star Coordinate
High dimensionality reduction
Fisher's discriminant form

ABSTRACT

Visualization methods could significantly improve the outcome of automated knowledge discovery systems by involving human judgment. Star coordinate is a visualization technique that maps k-dimensional data onto a circle using a set of axes sharing the same origin at the center of the circle. It provides the users with the ability to adjust this mapping, through scaling and rotating of the axes, until no mapped point-clouds (clusters) overlap one another. In this state, similar groups of data are easily detectable. However an effective adjustment could be a difficult or even an impossible task for the user in high dimensions. This is specially the case when the input space dimension is about 50 or more.

In this paper, we propose a novel method toward automatic axes adjustment for high dimensional data in Star Coordinate visualization method. This method finds the best two-dimensional view point that minimizes intra-cluster distances while keeping the inter-cluster distances as large as possible by using label information. We call this view point a discernible visualization, where clusters are easily detectable by human eye. The label information could be provided by the user or could be the result of performing a conventional clustering method over the input data. The proposed approach optimizes the Star Coordinate representation by formulating the problem as a maximization of a Fisher discriminant. Therefore the problem has a unique global solution and polynomial time complexity. We also prove that manipulating the scaling factor alone is effective enough for creating any given visualization mapping. Moreover it is showed that *k*-dimensional data visualization can be modeled as an eigenvalue problem. Using this approach, an optimal axes adjustment in the Star Coordinate method for high dimensional data can be achieved without any user intervention. The experimental results demonstrate the effectiveness of the proposed approach in terms of accuracy and performance.

1. Introduction

Our 3-dimensional perspective limits our conceptual experience of higher dimension space. Nevertheless our interaction with high dimension spaces is getting more and more inevitable. Increasing development of science and technology has led to substantial growth in data production beyond any human being conception capability. Billions of Web pages in cyberspace, huge geographical information, large amounts of biological data and numerous amounts of business databases are just small portions of the available data.

This has been the main motivation of obtaining geometric models (like graphs where there are 2 or 3 variables) of multivariate relationships arising in analyzing large sets of high dimensional data. Consequently numerous "visualization" approaches are proposed that try to achieve the best map from *k*-dimensions to 2 or 3 dimensions which is discernible for the human brain, effortlessly.

Unprecedented growth of data production and the limited ability of the human brain have made data visualization an interesting subject in computer science during recent years. As Card et al. described, visualization is "the use of computer-supported interactive, and visual representation of abstract data to amplify cognition." Visualization is considered as one of the most intuitive methods for cluster detection and validation, and especially is performing well for the representation of irregularly shaped clusters [27,32].

Other approaches of overcoming the problems of high dimensionality are dimension reduction [4,20] and feature selection [24]. Data sampling and data summarization could also help to cope with

* Corresponding author.
E-mail addresses: asef.pourmasoumi@stu-mail.um.ac.ir (A.P.H. Kiyadeh), amin.zamiri@stu.um.ac.ir (A. Zamiri), h-sadoghi@um.ac.ir (H.S. Yazdi), hadi.qaemi@stu-mail.um.ac.ir (H. Ghaemi).

http://dx.doi.org/10.1016/j.asoc.2014.09.026
1568-4946/

large amount of data records [17,28]. Scientists interested in these fields face a similar problem in exploratory analysis or visualization of multivariate data.

Star Coordinate is a visualization technique for mapping k-dimensional data into Cartesian coordinates, in which the coordinate axes are arranged on a circle of a two-dimensional plane with the origin at the center of the circle. It is proved that in this mapping technique, a cluster can always be preserved as a point-cloud (or cluster) in the visual space through linear mappings. But the main problem arises when these mapped point-clouds overlap one another, making their boundaries indistinguishable. Therefore the user is given the ability to push and pull or rotate the axes until the desired outcome is achieved. However, an advantageous adjustment is difficult or even impossible for the human agent to achieve, when visualizing high dimensional data. As a result, some researchers have proposed various dimension reduction methods, as pre-processing steps before applying the Star Coordinate visualization technique.

In this paper, we focus on the problem of automatic axes adjustment in Star Coordinate technique for improved visualization results. Our goal is to find the best projection possible that can represent the original data topology in k-dimensional data especially where k is greater than 50, effectively making manual axes adjustment impossible. The rest of the paper is organized as follows. Section 1.1 presents a discussion of related work. The main features of the Star Coordinate algorithm are briefly discussed in Section 1.2. Then, the proposed method is introduced in Section 2. In Section 3, we present the experimental results that validate the cost model. Section 4 presents a discussion of the experimental results. Finally, Section 5 concludes the presented approach.

1.1. Related work

Numerous approaches have been proposed for the visualization of multi-dimensional datasets. Scatterplot matrix [9], parallel coordinates [21] and dimensional stacking [31] have been developed to address this issue. Parallel coordinates (PC) [21] is a well-known method in which features are represented by parallel vertical axes linearly scaled within their data range. Each sample is represented by a polygonal line that intersects each axis at its respective attribute data value. Parallel coordinates can be used to study the correlations among various attributes by spotting the locations of the intersection points [44]. Also, they are useful for detecting the data distributions and functional dependencies. The main challenge of parallel coordinate approach is the limited space available for each parallel axis. There are several extended method for parallel coordinate, such as Circular Parallel Coordinates [19] and Hierarchical Parallel Coordinates [13].

Ester et al. [10] proposed DBSCAN to discover arbitrarily shaped clusters. It may not handle data sets that contain clusters with different densities. The OPTICS method, derived from the DBSCAN algorithm, uses visualization for visual cluster analysis [1] and is useful for finding density-based clusters in spatial data. Like most of the clustering algorithms, OPTICS is a parametric approach. Yang et al. [39] proposed a visual hierarchical dimension reduction technique, which groups dimensions and visualizes data by using the subset of dimensions obtained from each group. In [2] and [36], some features that affect the quality of visualization have been introduced and some of the above systems are compared based on listed features.

Another famous approach for data visualization is Star Coordinate [25] and its extensions, such as VISTA [6]. The proposed method is based on the Star Coordinate technique. Star Coordinates arranges coordinate axes on a two-dimensional surface, where each axis shares the same origin point. It uses a linear mapping to avoid the cluster breaking after k-dimensional to 2D space mapping. (This has been proven in [8] mathematically). So far, several extensions for VISTA have been introduced. iVIBRATE [7] is a framework for visualizing large datasets using data sampling and the Star Coordinate model. In [37], an Enhanced VISTA is proposed which improves visualization and eases the human computer interaction. The experiments have shown that visual cluster rendering can improve the understanding of clusters, and validate and refine the algorithmic clustering result effectively [25].

VISTA is a very good interactive approach for visualization of k-dimensional data where $K < 50$, and its efficiency has been proven by various articles. The main shortcoming of this method is that the dimension must be less than 50. Since, according to each dimension of data, a coordinate axis is drawn, when the number of dimensions is more than 50, working with VISTA tools would be very exhausting for humans and, practically, its interactivity property would be useless. This problem becomes more serious when the number of dimensions is much greater than 50. However, there are many datasets with a large amount of features in the world, e.g., textual data, image data, bioinformatics data, etc.

In this paper we propose a novel semi-supervised visualization method for high dimensional data, where a fraction of the data is labeled. The visualization result achieved by applying this method is optimal in terms of discernibility by the user. This work extends Star Coordinates capabilities in working with high-dimensional datasets.

1.2. Star Coordination

Star Coordinates is a visualization technique for mapping high-dimensional data into two dimensions. In this technique a 2D plane is divided into k equal sectors (θ_i, the angle of the sectors, is set to $2\pi i/k$ by default). Therefore there are k coordinate axes, with each axis representing one dimension of data and all axes sharing their origins at the center of a circle on the 2D space (Fig. 1) having the same length [25]. Data points are scaled to the length of the axis, in way that the smallest is mapped to the origin and the largest to the other end of the axis. Then unit vectors on each coordinate axis are calculated accordingly to allow scaling of data values to the length of the coordinate axes.

The mapping of a point from k-dimensional space to a point in the two dimensional Cartesian coordinates is determined by the sum of all unit vectors ($\vec{u}_{xi}$, $\vec{u}_{yi}$), on each coordinate multiplied by the value of the data element for that coordinate, as shown in Formula (1):

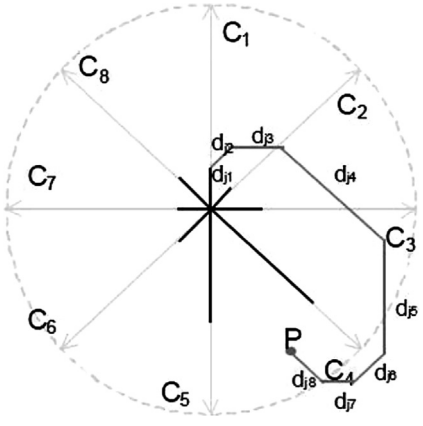

Fig. 1. The image of an 8-dimensional point in Cartesian coordinate [25].

$$P_j(x,y) = \left(\sum_{i=1}^{k} \vec{u}_{xi}(d_{ji} - \min_i), \sum_{i=1}^{k} \vec{u}_{yi}(d_{ji} - \min_i) \right)$$
$$D_j = (d_{j0}, d_{j1}, \ldots, d_{ji}, \ldots, d_{jk}), \vec{u}_i = \frac{\vec{C}i}{\max_i - \min_i} \quad (1)$$
$$\min_i = \min\left\{d_{ji}, 0 \le j \le \left|D\right|\right\}, \max_i = \max\left\{d_{ji}, 0 \le j \le \left|D\right|\right\}$$

where D_j is a k-dimensional data element and $P_j(x, y)$ is its two-dimensional projected point. The main idea of Star Coordinate is to arrange the coordinate axes on a two-dimensional plane, where the coordinate axes are not necessarily orthogonal to each other [25]. The projection of high dimensional data to 2D space inevitably introduces overlapping and ambiguities, and even bias. It means that multiple points in k-dimensional space may map into one point in Cartesian space. But it is shown that a cluster can always be preserved as a point-cloud in the visual space through linear mappings [8]. The only problem is that these point-clouds may overlap one another. To make sure one finds the best possible mapping, Star Coordinates and its extension VISTA [6] provide several visual adjustment mechanisms, such as axis scaling (α-*adjustment* in VISTA) and axis angle rotation. Both axis scaling and angle rotation are linear transformation. Since linear mapping does not break clusters, the clusters in the multi-dimensional space are still visualized as dense point-clouds (the "visual clusters") in two-dimensional space and the visible gaps between the visual clusters in two-dimensional visual space indicate the *real gaps* between point-clouds in the original high dimensional space [7]. In the following we mention these two mappings.

1.2.1. Rotation transformation

Rotating an axis modifies the direction of the axis's unit vector and changes the correlation of the corresponding feature (dimension) with other features. Axis rotation changes the direction of axes, thus making a particular data attribute more or less correlated with other attributes. This can resolve the overlapping problem substantially. It helps the user distinguish between clusters that may incorrectly overlap. This is possible by modeling the Star Coordinate using the Euler formula: $e^{ix} = \cos x + i \sin x$, where $z = x + iy$, and i is the imaginary unit. However, as experimental results have shown, adjusting the scaling transformation is enough in order to find a satisfactory visualization. Therefore we can leave θ_i to be constant as $\theta_i = 2\pi i/k$ [8].

1.2.2. Scaling (α-adjustment in VISTA) transformation

Scaling transformations allow users to change the length of an axis, thus increasing or decreasing the contribution of a particular data column (particular dimension or feature) on the resultant visualization [25]. Using axis scaling interactively, a user can observe that the data distribution changes dynamically. This is done by adding α to Formula (1) in iVIBRATE [7] as following:

$$P_j(x,y) = \left((c/k)\sum_{i=1}^{k} \alpha_i \vec{u}_{xi}(d_{ji} - \min_i), (c/k)\sum_{i=1}^{k} \alpha_i \vec{u}_{yi}(d_{ji} - \min_i) \right) \quad (2)$$

where $\alpha_i (i = 1 \cdots k, \alpha_i \in [-1, 1])$ provides the visually adjustable parameters. As mentioned in [7], $\alpha_i \in [-1, 1]$ covers a considerable range of mapping functions and this range combined with the scaling factor c, is effective enough for finding a satisfactory visualization. It is known that linear mapping does not break clusters, but may cause cluster overlaps [26]. In Fig. 2 initial data distribution of the Iris dataset from the UCI machine learning repository (available online from http://www.ics.uci.edu/~mlearn/databases/) is shown. Iris has a four-dimensional dataset with 150 records and 3 clusters. Fig. 3(A) depicts the original data distribution of Iris dataset together with the cluster indices achieved by applying the

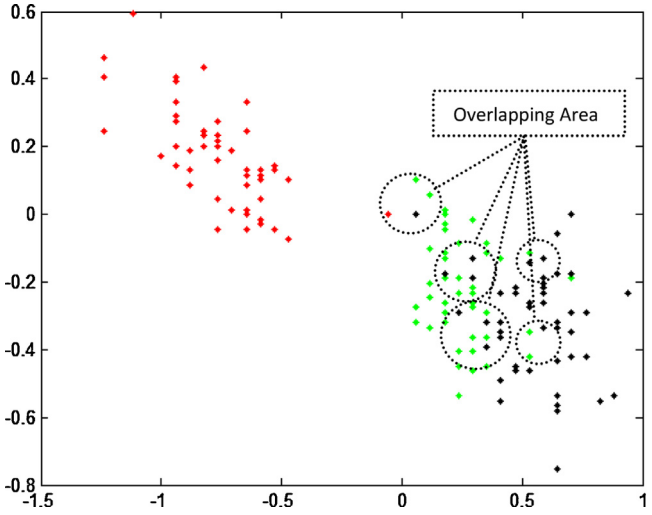

Fig. 2. The initial data distribution of clusters of Iris dataset with original labels.

K-means clustering algorithm in VISTA, in which clusters overlap. Fig. 3(B) shows a better separated cluster distribution of Iris using α-adjustment performed interactively by an expert user.

As shown in the above figure, the Star Coordinate approach can effectively visualize data using user interaction. However, as mentioned in [7], Star Coordinate or its variants such as VISTA [6] are limited to visualizing data with a maximum of 50 dimensions. When the number of dimensions is more than 50, visualization using user interaction is practically impossible. As a result, the cluster overlapping problem could not be resolved in high-dimensional data visualization using conventional approaches. In the proposed method, this problem is resolved, provided that a fraction of data (even if small) is labeled. Using this label information, k-dimensional data can be mapped onto the two-dimensional plane, with clusters of the data as recognizable as possible for the system user.

2. The proposed approach

In the proposed approach, label information of a fraction of the data is employed to enhance visualization results. In the field of pattern recognition, there is a similar subject named *semi-supervised clustering* [5] and in data mining it is known as *domain knowledge based clustering*. The effect of domain knowledge application in information visualization has been shown in the studies in these fields [6].

In order to find the best mapping for data visualization, the optimal configuration of axes in the Star Coordinate method should be determined first. The optimal state, called the Discernible Visualization, is defined by the condition where mapped clusters are as dense as possible while the sum of distances between their centroids (means) is maximized. In this configuration, visual perception of clusters is improved, since cluster boundaries are more recognizable to the human eye. To achieve this, we incorporate label information in the visualization process. The labels could be determined manually by the user, or could be the result of a conventional clustering method. In the latter there is no need for human intervention, thus making the visualization process fully automatic. However manual labeling by the user is usually more reliable. With more label information available, a better visualization result can be achieved, as the process becomes more similar to supervised visualization. However, as shown in the experimental results, this method achieves satisfactory results with even limited labels as well.

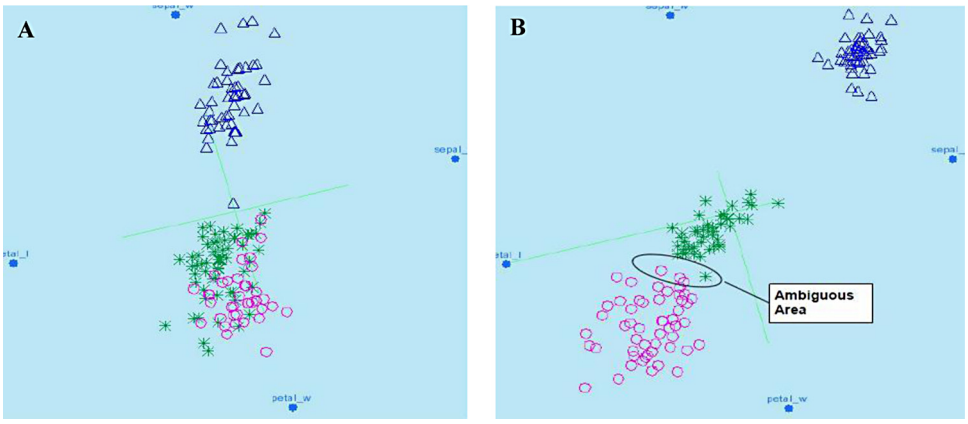

Fig. 3. (A) The initial data distribution of clusters of Iris produced by k-means in VISTA [6]. (B) Visualization of Iris dataset with k-means labels after α-adjustment [6].

One of the most interesting features of the Star Coordinate method is that the overall properties and relations of clusters are preserved in the mapped space. But this does not guarantee that clusters do not overlap. As mentioned earlier, to resolve the overlapping problem, two transformations were introduced: scaling (or α-adjustment in VISTA) and rotation. In [8], it has been shown empirically that adjusting α alone (without any rotation adjustment) is effective enough to achieve any desired visualization. In Proposition 1, we provide a proof for this statement.

Proposition 1. For every modified mapping constructed by performing a rotation with degree θ over the initial mapping in Star Coordinate method, there exists a scaling adjustment with value α that results in the same mapping.

Proof: In order to demonstrate the equivalency of performing a rotation with degree θ and scaling with value α, one must argue that the effect of any change in the original mapping by performing the rotation can be achieved by a proper α-vector scaling, while keeping θ unchanged. This means that any possible modification to the final visualization by rotating the axes θ degrees could be obtained by adjusting α in $[-1, 1]$. Suppose that $P_x(t)$ and $P_y(t)$ are the projection of a k-dimensional sample t on the x-axis and the y-axis in 2D-space, respectively. This projection can be formulated as below:

$$P_x(t) = \sum_{i=1}^{k} \alpha_i x_i \cos(\theta_i) \tag{3}$$

$$P_y(t) = \sum_{i=1}^{k} \alpha_i x_i \sin(\theta_i) \tag{4}$$

Now assume that the θ value of jth dimension has changed. It will be shown that this change is equivalent to performing axes scaling with value α' over the original mapping. This can be written as:

$$\sum_{i=1\&i\neq j}^{k} \alpha_i x_i \cos(\theta_i) + \alpha_j x_j \cos(\theta_j - \theta_0) = \sum_{i=1}^{k} \alpha'_i x_i \cos(\theta_i) \tag{5}$$

$$\sum_{i=1\&i\neq j}^{k} \alpha_i x_i \sin(\theta_i) + \alpha_j x_j \sin(\theta_j - \theta_0) = \sum_{i=1}^{k} \alpha'_i x_i \sin(\theta_i) \tag{6}$$

Expanding the cosine, Eq. (5) can be rewritten as:

$$\sum_{i=1\&i\neq j}^{k} \alpha_i x_i \cos(\theta_i) + \alpha_j x_j \cos(\theta_j)\cos(\theta_0) + \alpha_j x_j \sin(\theta_j)\sin(\theta_0) = \sum_{i=1}^{k} \alpha'_i x_i \cos(\theta_i) \tag{7}$$

With the left hand side of the above formula being constant (S_{total}) and defining $C_i = \alpha'_i x_i \cos(\theta_i)$, we have:

$$S_{total} = C_1 + C_2 + \ldots + C_k, -k \le S_{total} \le k \tag{8}$$

Eq. (6) can also be written as follows, if the sinus formula is expanded:

$$\sum_{i=1\&i\neq j}^{k} \alpha_i x_i \sin(\theta_i) + \alpha_j x_j \sin(\theta_j)\cos(\theta_0) - \alpha_j x_j \cos(\theta_j)\sin(\theta_0)$$

$$= \sum_{i=1}^{k} \alpha'_i x_i \sin(\theta_i) \tag{9}$$

Similarly we consider the left hand side constant and define $C'_i = \alpha'_i x_i \sin(\theta_i)$, therefore:

$$S'_{total} = C'_1 + C'_2 + \ldots + C'_k, -k \le S'_{total} \le k \tag{10}$$

The final system of equations is then defined as follows:

$$\begin{aligned} &S_{total} = C_1 + C_2 + \cdots + C_k, -k \le S_{total} \le k, \\ &S'_{total} = C'_1 + C'_2 + \ldots + C'_k, -k \le S'_{total} \le k \\ &C'_i = C_i \times \tan(\theta_i) \end{aligned} \tag{11}$$

Solving the above system results in the desired α-vector. In the system of equations in (11), there are two equations and k unknown variables. Since the number of unknown variables is larger than the number of equations, there are generally an infinite number of solutions. So, at least one α-vector for the problem can be found. This means that for any change in θ value of any given dimension, there exists at least one α vector which results in the same outcome. □

In the next section, we initially assume that the label information is only available about two clusters, for the sake of simplicity. We will then drop this assumption and discuss the scenarios where more than two clusters are available in Section 2.2.

2.1. Two-class case

Consider the simple case where there are two types of labels available for the labeled fraction of data. One set of k-dimensional samples $\{x_1^1, x_2^1, \cdots, x_{n_1}^1\}$ are labeled as w_1 and the other set $\{x_1^2, x_2^2, \cdots, x_{n_2}^2\}$ are labeled as w_2. Since label information is usually limited, n_1 and n_2 (the size of the first and the second set respectively), are much smaller than the total number of data points (N) in the dataset ($n_1 \ll N, n_2 \ll N$). Using labels, the proposed

method finds an optimal α-vector adjustment which projects k-dimensional data into 2D-space in such a way that mapped clusters are heterogeneous and homogeneous. For this purpose, we formulate our objective function (Formula (12)) as a Fisher discriminant analysis objective function [14]. The Fisher discriminant, used in linear classification, is proved to have a global solution which can be found by solving an eigenvalue problem. Hence, our solution is globally optimal.

$$J(\alpha) = \frac{F_1}{F_2} = \frac{\alpha^T(S_B)\alpha}{\alpha^T(S_W)\alpha} \tag{12}$$

In this formula F_1 and F_2 denote the inter-cluster and intra-cluster distance respectively, which could be written based on the axes scaling parameter (α) and the between and within-class scatter matrices (S_B and S_w) which emerge in the formulations provided in Lemma 1 and Lemma 2. An increase in inter-cluster distance implies more separation in cluster means, while a decrease in intra-cluster distance represents more dense clusters in the mapped space. Our objective function seeks an optimal axis scaling parameter (α) which minimizes the sum of Euclidian distance of each point to its cluster mean and maximizes distance between centroids (means) of clusters.

The above formula is in Fisher discriminant form for which an optimal global solution α could be found by calculating the eigenvector of $S_w^{-1}S_B$.

In Proposition 2 it is proven that the solution to the optimal stretching of Star Coordinate axes is obtained by maximizing the objective function shown in Eq. (12). The following lemmas are required before giving a proof.

Lemma 1. Let μ_1 and μ_2 denote the mean or centroid for clusters w_1 and w_2, respectively. Then, the distance between the means of clusters, μ_1 and μ_2 in 2D-spaces, can be shown to be a quadratic form:

$$F_1 = vdist(\mu_1, \mu_2) = \left\|\mu_2 - \mu_1\right\|^2 = \alpha^T(S_B)\alpha \tag{13}$$

Proof: μ_1 is the mean data in cluster w_1 and using Star Coordinate it can be calculated as:

$$\begin{aligned}
\mu_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{k}\alpha_j x_{ij}^1\cos\theta_j, \sum_{j=1}^{k}\alpha_j x_{ij}^1\sin\theta_j\right)\\
&= \frac{1}{n_1}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{k}\alpha_j x_{ij}^1\cos\theta_j, \sum_{i=1}^{n_1}\sum_{j=1}^{k}\alpha_j x_{ij}^1\sin\theta_j\right)\\
&= \left(\sum_{j=1}^{k}\alpha_j\cos\theta_j(1/n_1\sum_{i=1}^{n_1}x_{ij}^1), \sum_{j=1}^{k}\alpha_j\sin\theta_j(1/n_1\sum_{i=1}^{n_1}x_{ij}^1)\right)\\
&= \left(\sum_{j=1}^{k}\alpha_j\cos\theta_j X_j^1, \sum_{j=1}^{k}\alpha_j\sin\theta_j X_j^1\right)
\end{aligned} \tag{14}$$

where X_j^1 is the mean of jth dimension of w_1:

$$X_j^1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}^1 \tag{15}$$

Similarly μ_2 can be written as:

$$\mu_2 = \left(\sum_{j=1}^{k}\alpha_j\cos\theta_j X_j^2, \sum_{j=1}^{k}\alpha_j\sin\theta_j X_j^2\right) \tag{16}$$

Then, the distance between the means of clusters, μ_1 and μ_2 calculated as follow:

$$\begin{aligned}
F_1 &= vdist(\mu_1, \mu_2) = ||\mu_2 - \mu_1||^2\\
&= \left\|\sum_{j=1}^{k}\alpha_j\cos\theta_j X_j^2 - \sum_{j=1}^{k}\alpha_j\cos\theta_j X_j^1, \sum_{j=1}^{k}\alpha_j\sin\theta_j X_j^2 - \sum_{j=1}^{k}\alpha_j\sin\theta_j X_j^1\right\|^2\\
&= \left\|\sum_{j=1}^{k}\alpha_j\cos\theta_j(X_j^2 - X_j^1), \sum_{j=1}^{k}\alpha_j\sin\theta_j(X_j^2 - X_j^1)\right\|^2
\end{aligned} \tag{17}$$

Using the notation:

$$\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_k] \tag{18}$$

$$\begin{aligned}
Z &= [z_1, z_2, \cdots, z_k]\\
&= [\cos\theta_1(X_1^2 - X_1^1), \cos\theta_2(X_2^2 - X_2^1), \cdots, \cos\theta_k(X_k^2 - X_k^1)]
\end{aligned} \tag{19}$$

$$\begin{aligned}
Z' &= [z_1', z_2', \cdots, z_k']\\
&= [\sin\theta_1(X_1^2 - X_1^1), \sin\theta_2(X_2^2 - X_2^1), \cdots, \sin\theta_k(X_k^2 - X_k^1)]
\end{aligned} \tag{20}$$

We can then write Eq. (17) as:

$$\begin{aligned}
&vdist(\mu_1, \mu_2) = \left\|\alpha^T Z, \alpha^T Z'\right\|^2 = (\alpha^T Z, \alpha^T Z')^T(\alpha^T Z, \alpha^T Z')\\
&= \left((\alpha^T Z)^T(\alpha^T Z) + (\alpha^T Z')^T(\alpha^T Z')\right) = (Z^T\alpha)(\alpha^T Z) + (Z'^T\alpha)(\alpha^T Z')
\end{aligned} \tag{21}$$

It is clear that $Z^T\alpha = \alpha^T Z$ *and* $Z'^T\alpha = \alpha^T Z'$, therefore (21) can be written as:

$$vdist(\mu_1, \mu_2) = (\alpha^T Z)(Z^T\alpha) + (\alpha^T Z')(Z'^T\alpha) \tag{22}$$

which can be converted in the quadratic form:

$$\begin{aligned}
F_1 &= vdist(\mu_1, \mu_2) = \alpha^T(ZZ^T)\alpha + \alpha^T(Z'Z'^T)\alpha\\
&= \alpha^T(ZZ^T + Z'Z'^T)\alpha = \alpha^T S_B\alpha
\end{aligned} \tag{23}$$

In order to have the best visualization in 2d-space, clusters must be heterogeneous. This means the distance between the data points of two different clusters must be as large as possible. Maximizing Eq. (23) can lead to a satisfying separation between projected classes in 2-D space. Moreover, visualized clusters should be as dense as possible. This can be achieved by minimizing the trace of the sum of the covariance matrix of each class w_1 and w_2.

Lemma 2. The trace of the sum of the covariance matrix of each class w_1 and w_2 can be show to be a quadratic form:

$$F_2 = trace(\Sigma_1 + \Sigma_2) = \alpha^T(S_w)\alpha \tag{24}$$

Proof: The trace of the sum of the covariance matrix of each class can be written as follows:

$$F_2 = trace(\Sigma_1 + \Sigma_2) = trace(\Sigma_1) + trace(\Sigma_2) \tag{25}$$

where Σ_1 is the covariance of class w_1 and Σ_2 is the covariance of class w_2. The covariance of the data in class w_1 in k-dimensional space is equal to:

$$\begin{aligned}\sum_1 &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T \\ &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-\mu_1)\cos\theta_j, \sum_{j=1}^{k}\alpha_j(x_{ij}-\mu_1)\sin\theta_j\right)\times \\ &\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-\mu_1)\cos\theta_j, \sum_{j=1}^{k}\alpha_j(x_{ij}-\mu_1)\sin\theta_j\right)^T\end{aligned} \tag{26}$$

Using Eq. (14), it can be written as:

$$\begin{aligned}\sum_1 &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j, \sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right)\times \\ &\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j, \sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right)^T \\ &\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left[\begin{array}{cccc}\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j\right)\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j\right) & \left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j\right)\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right) \\ \left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right)\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j\right) & \left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right)\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\sin\theta_j\right)\end{array}\right]\end{aligned} \tag{27}$$

The trace of the covariance matrix is calculated as follow:

$$trace\left(\sum_1\right) = \left(\sum_1^{11} + \sum_1^{22}\right) \tag{28}$$

where $\sum_1^{11}$ and $\sum_1^{22}$ are the elements of the diagonal of $\sum_1$ Firstly, $\sum_1^{11}$ is computed as follow:

$$\begin{aligned}\sum_1^{11} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_j^1)\cos\theta_j\right) \\ &\times\left(\sum_{j=1}^{k}\alpha_j(x_{ij}-X_i^1)\cos\theta_j\right)\end{aligned} \tag{29}$$

Introducing the vector T_i, defined as:

$$T_i = \left[(x_{i1}-X_1^1)\cos\theta_1, (x_{i2}-X_2^1)\cos\theta_2, \cdots, (x_{ik}-X_k^1)\cos\theta_k\right] \tag{30}$$

We can write $\sum_1^{11}$ as:

$$\begin{aligned}\sum_1^{11} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\alpha^T T_i)(\alpha^T T_i)^T = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\alpha^T T_i)(T_i^T\alpha) \\ &= \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i T_i^T)\right)\alpha\end{aligned} \tag{31}$$

Similarly introducing T_i' as:

$$T_i' = \left[(x_{i1}-X_1^1)\sin\theta_1, (x_{i2}-X_2^1)\sin\theta_2, \cdots, (x_{ik}-X_k^1)\sin\theta_k\right] \tag{32}$$

We can obtain $\sum_1^{22}$ as:

$$\begin{aligned}\sum_1^{22} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\alpha^T T_i')(\alpha^T T_i')^T = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\alpha^T T_i')(T_i'^T\alpha) \\ &= \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i' T_i'^T)\right)\alpha\end{aligned} \tag{33}$$

Using (33) and (31), Eq. (28) can be written as:

$$\begin{aligned}trace\left(\sum_1\right) &= \left(\sum_1^{11} + \sum_1^{22}\right) \\ &= \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i T_i^T)\right)\alpha + \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i' T_i'^T)\right)\alpha \\ trace\left(\sum_1\right. &= \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i T_i^T + T_i' T_i'^T)\right)\alpha\end{aligned} \tag{34}$$

Similarly, one can obtain trace of $\sum_2$ as below:

$$trace\left(\sum_2\right) = \alpha^T\left(\frac{1}{n_2-1}\sum_{i=1}^{n_2}(U_i U_i^T + U_i' U_i'^T)\right)\alpha \tag{35}$$

where U_i and U_i' can be defined similarly to T_i and T_i', and finally, Eq. (25) can be written as follow:

$$\begin{aligned}F_2 &= trace\left(\sum_1 + \sum_2\right) \\ &= \alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_i T_i^T + T_i' T_i'^T) + \frac{1}{n_2-1}\sum_{i=1}^{n_2}(U_i U_i^T + U_i' U_i'^T)\right)\alpha = \alpha^T S_w \alpha\end{aligned} \tag{36}$$

Using Lemma 1 and Lemma 2, Proposition 2 can be proved straightforwardly.

Proposition 2. Finding the optimal scaling adjustment parameter α for visualization of k-dimensional data can be modeled as an eigenvalue problem and can be written as:

$$\text{maximize} J(\alpha) = \frac{F_1}{F_2} = \frac{\alpha^T(S_B)\alpha}{\alpha^T(S_W)\alpha} \tag{37}$$

Proof: As mentioned earlier, it is desired to determine an optimal vector α (scaling adjustment parameter) such that the resulting mapped clusters are as dense as possible while having maximum distance with one another. This could be translated into maximizing the distance between cluster means μ_1 and μ_2, and minimizing

the trace of the covariance matrix, which is formulated as the following objective function:

$$J(\alpha) = \frac{F_1}{F_2} = \frac{\left\|\mu_2 - \mu_1\right\|^2}{trace(\sum_1 + \sum_2)} \tag{38}$$

Using Lemma 1 and Lemma 2, the objective function can be calculated simply as:

$$J(\alpha) = \frac{F_1}{F_2} = \frac{\alpha^T(ZZ^T + Z'Z'^T)\alpha}{\alpha^T\left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}(T_iT_i^T + T_i'T_i'^T) + \frac{1}{n_2-1}\sum_{i=1}^{n_2}(U_iU_i^T + U_i'U_i'^T)\right)\alpha} \tag{39}$$

The above may be rewritten as:

$$J(\alpha) = \frac{F_1}{F_2} = \frac{\alpha^T(S_B)\alpha}{\alpha^T(S_W)\alpha}\cdots\square \tag{40}$$

As can be seen in Eq. (40), our final objective function converted into *Fisher's discriminant* form, the maximization of which has therefore a *global solution* (although not necessarily unique). This globally optimal α maximizing (40) can be calculated by solving an eigenvalue problem. Differentiating equation (40) with respect to α yields:

$$\begin{aligned}&(\alpha^T S_B\alpha)S_W\alpha - (\alpha^T S_W\alpha)S_B\alpha = 0\\ &S_B\alpha = \lambda S_W\alpha\\ &{S_W}^{-1}S_B\alpha = \lambda\alpha\end{aligned} \tag{41}$$

showing that α is an eigenvector and λ is an eigenvalue of the $S_w^{-1}S_B$ matrix. Since $\lambda = J(\alpha)$, our objective function is maximized when the columns of α are the eigenvector associated with the highest eigenvalue of $S_w^{-1}S_B$.

So, in order to find the best visualization possible, one has to calculate the eigenvector associated with the highest eigenvalue of $S_w^{-1}S_B$. This is the value of the required axes adjustment (α) that results in the desired mapping. Using Star Coordinate mapping together with an axes adjustment with vector α (Eq. (2)), k-dimensional data can be projected into a 2D-plane in the best possible form. As mentioned, θ_i is kept constant and equal to $2\pi i/k$.

2.2. Multi-class

If there are more than 2 clusters ($c \geq 2$ clusters), a natural visualization of the proposed method can be used. Thus, the generalization of the intra-class scatter matrix is as shown below:

$$S_W = \sum_{i=1}^{c} S_i \tag{42}$$

where

$$S_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(V_jV_j^T + V''_jV_j'^T) \tag{43}$$

$$V_j = \left[(x_{j1} - X_1^1)\cos\theta_1, (x_{j2} - X_2^1)\cos\theta_2, \cdots, (x_{jk} - X_k^1)\cos\theta_k\right] \tag{44}$$

$$V_j' = \left[(x_{j1} - X_1^1)\sin\theta_1, (x_{j2} - X_2^1)\sin\theta_2, \cdots, (x_{jk} - X_k^1)\sin\theta_k\right] \tag{45}$$

In (42) the S_i formulation is similar to Eq. (35). For the generalization of S_B, we define it as the following *multiple Fisher discriminant* [14]:

$$S_B = \sum_{i=1}^{c}\left(n_i(\mu_i - \mu)(\mu_i - \mu)^T\right) \tag{46}$$

where μ_i is the mean of labeled data in each cluster which can be calculated as Eq. (14). We define a total mean vector μ by

$$\mu = \frac{1}{n}\sum_x x = \sum_{i=1}^{c} n_i\mu_i \tag{47}$$

and then replacing Eq. (23) by Eq. (46), gives:

$$S_B = \sum_{i=1}^{c} n_i(M_iM_i^T + M_i'M_i'^T) \tag{48}$$

$$M_i = \left[(\cos\theta_1(X_1^i - X_1^t), \cos\theta_2(X_2^i - X_2^t), \cdots, \cos\theta_k(X_k^i - X_k^t)\right] \tag{49}$$

$$M_i' = \left[\sin\theta_1(X_1^i - X_1^t), \sin\theta_2(X_2^i - X_2^t), \cdots, \sin\theta_k(X_k^i - X_k^t)\right] \tag{50}$$

where X_j^i is the mean of the jth dimension of the labeled data in the ith cluster and X_j^t is the mean of the jth dimension of all

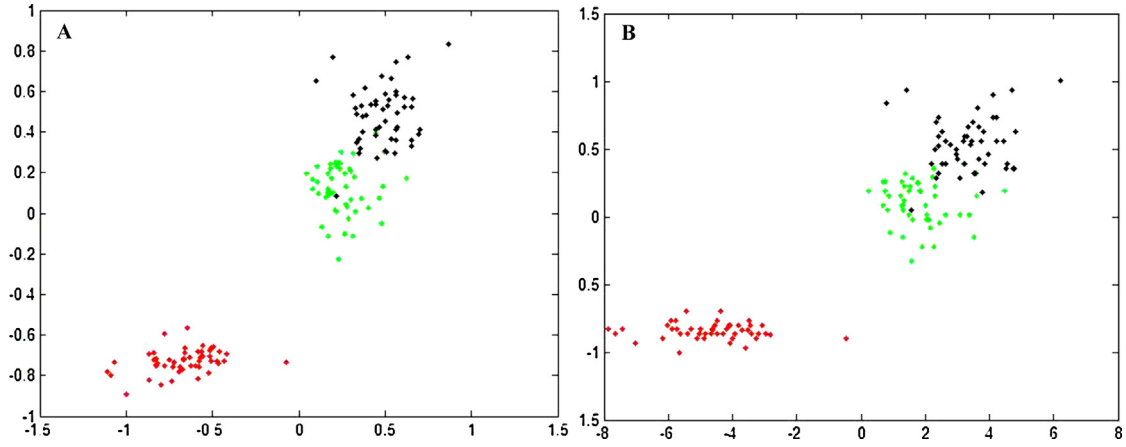

Fig. 4. Visualization of the Iris data set using the proposed two-class method (A) ($w_1 = 6$, $w_2 = 9$) (B) ($w_2 = 7$, $w_3 = 8$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

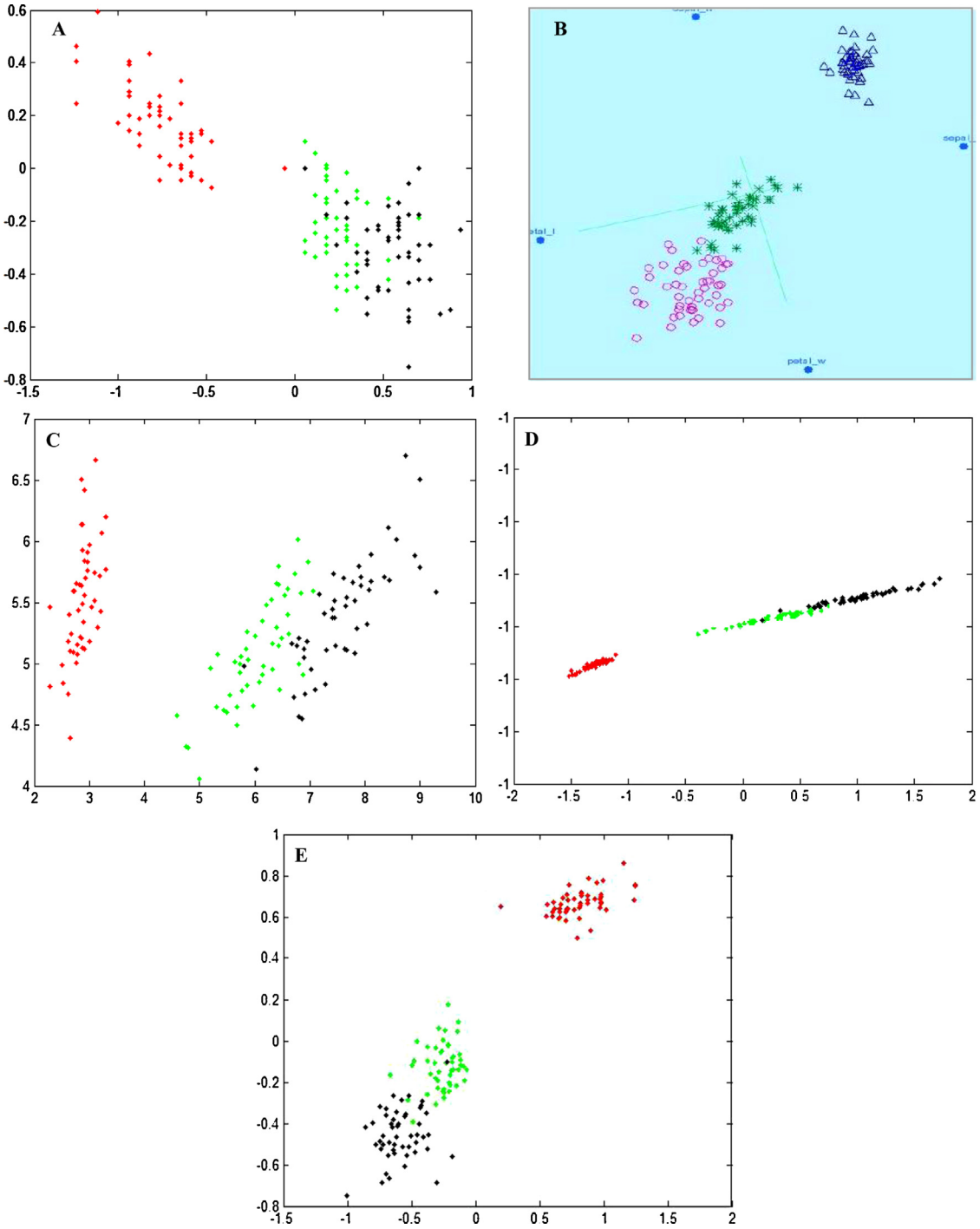

Fig. 5. (A) Initial mapping of the Iris dataset into the 2D plane (B)Visualization of the Iris data set after editing by a human expert [6], (C) using the PCA (D) using the Locally Liner Embedding approach (with 50 for number of neighbors) (E) using fully automatic multi-class proposed method ($w_1 = 4$, $w_2 = 3$, $w_3 = 3$).

labeled data. Finally, similarly to Eq. (40), we can write the objective function as:

$$J_{multi}(\alpha) = \frac{\alpha^T(S_B)\alpha}{\alpha^T(S_W)\alpha} = \frac{\alpha^T\left(\sum_{i=1}^{c} n_i(M_i M_i{}^T + M_i' M_i'{}^T)\right)\alpha}{\alpha^T\left(\sum_{i=1}^{c}\left(\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(V_j V_j{}^T + V_j' V_j'{}^T)\right)\right)\alpha} \tag{51}$$

One can find the optimal vector α that leads to dense and separated visualization of data clusters by maximizing equation (51). Again, similar to the previous section, with the calculated α-vector and Star Coordinate mapping, the best projection of k-dimensional data into 2D-space can be performed.

2.3. Computational complexity analysis

The proposed approach is straightforward to implement, has polynomial time complexity, and converges to a global solution.

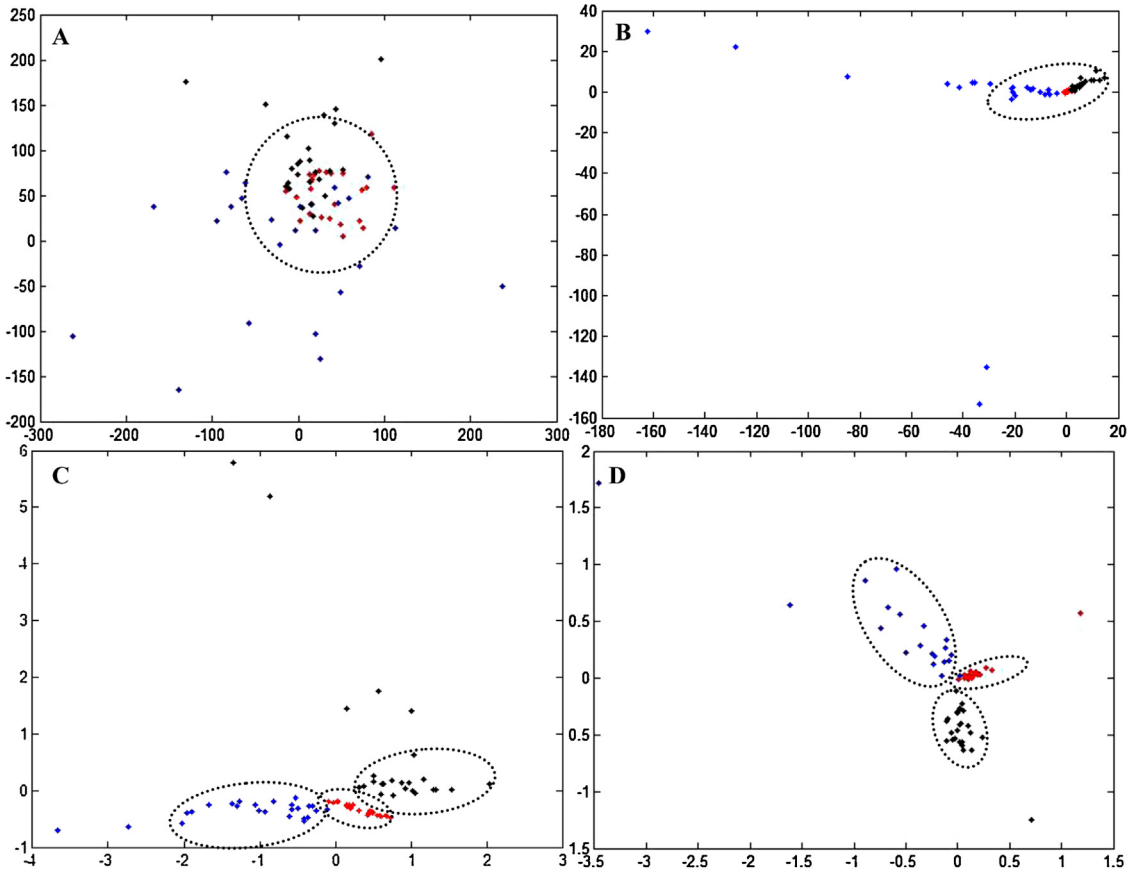

Fig. 6. Visualization on the DUC 2007 data set (A) using Star Coordinate method with no adjustment, (B) using PCA (C) using Locally Liner Embedding approach (with 25 for the number of neighbors) (D) using the proposed approach.

In this section, the time complexity of the proposed method a for two-class visualization is determined.

1. The run time for calculating Eqs. (14) and (16) has complexity of $O(kn_1 + kn_2)$ where k is the number of dimensions.
2. The time complexity for calculating F_1 in Eq. (23) is $O(k^2)$.
3. The time complexity for calculating F_2 in Eq. (36) is $O(k^2n_1 + k^2n_2)$.
4. In Eq. (41), the run time is dominated by the solution for eigenvectors. Since S_w is a $k \times k$ matrix, it has time complexity of $O(k^3)$.

The values of n_1 and n_2 (number of labeled data points) in high dimensional data sets can be ignored. Thus, the time complexity of our proposed method for a two-class visualization in high dimension data sets is $O(k^3)$. However, there are many techniques available for speeding up the eigenvector problems. For example, specialized methods for sparse, symmetric eigenvalue problems can be used to reduce the complexity [11].

One can easily calculate the time complexity of the proposed method for c-cluster data visualization. Similar to the two-class visualization, if the number of dimensions is greater than n_i and c, the time complexity will be equal to $O(k^3)$

3. Experimental result

In this section, we present several experimental results that illustrate the effectiveness of the proposed method. We test our approach on four data sets. Some properties of the data sets are shown in Table 1. The proposed method was implemented in MATLAB running under Windows XP. The results of the experiments have been compared to those of VISTA and also with a dataset visualized manually by an expert user. Moreover, since there is no comparable technique for extending Star Coordination to high dimensional spaces, we have compared our approach with the Principle Component Analysis (PCA) [22] and Locally Linear Embedding (LLE) [34] methods. PCA is a *linear* feature reduction technique which has been used for many visualization applications [15]. In this technique the most important features of the data are extracted to decrease the dimension of the input space. The Locally Linear Embedding (LLE) algorithm addresses the *nonlinear* dimensionality reduction problem. LLE emphasizes the local linearity of the manifold and assumes that the local relations in the original data space are also preserved in the projected low-dimensional space [34]. We choose the LLE method for our comparison because of its widespread application in dimensionality reduction problems.

Table 1
Data sets description.

Data set name	Data set specifications		
	#Dimensions	#Clusters	#Samples
Iris	4	3	150
DNA	180	3	1400
DUC2007	3846	3	76
DUC2006	4176	5	100

3.1. Iris

The Iris dataset is a famous data set in the pattern recognition literature and can be obtained from the UCI machine learning website[1]. Iris has 150 instances, 4 features and 3 clusters.

[1] http://www.ics.uci.edu/~mlearn/Machine-Learning.html.

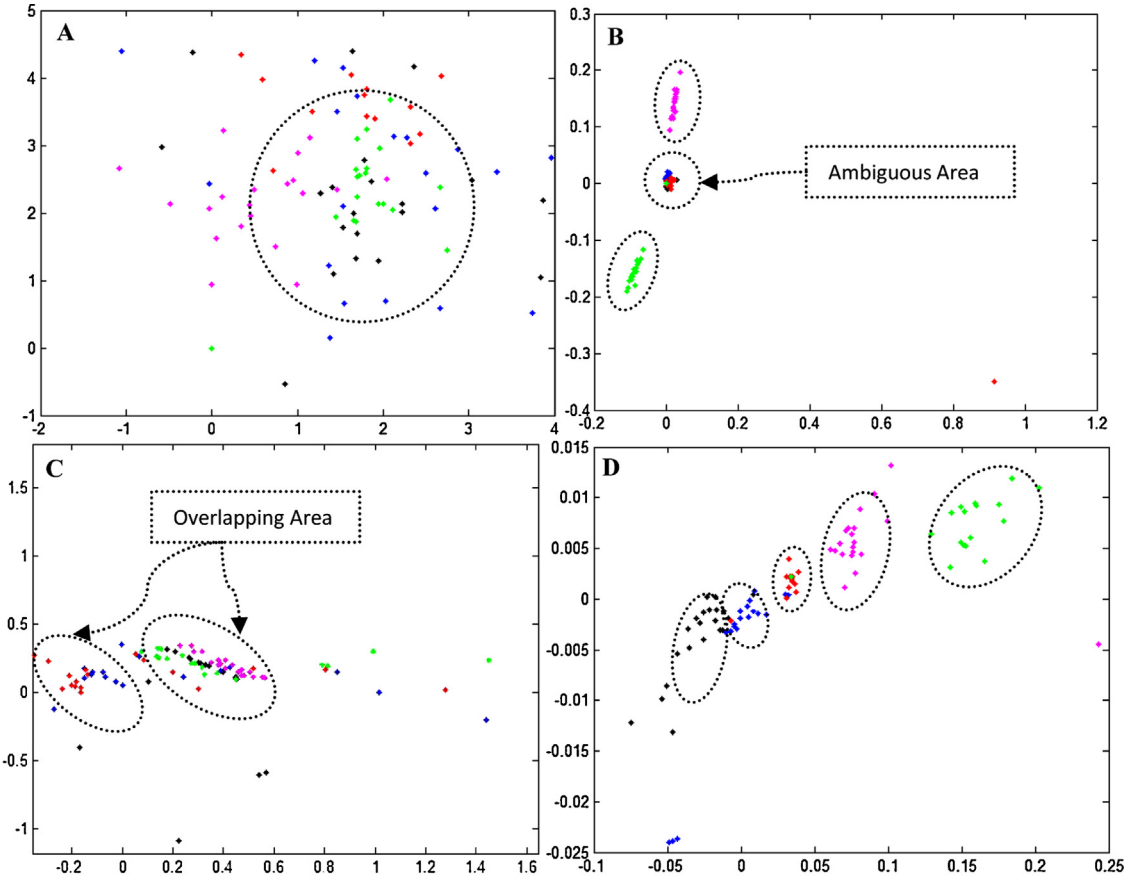

Fig. 7. Visualization of the DUC 2006 data set (A) using Star Coordinate method with no adjustment, (B) using PCA, (C) using Locally Liner Embedding approach (with 20 for the number of neighbors), and (D) using the proposed approach.

In Fig. 4, the visualization of the Iris data set, using the proposed two-class method is shown. In Fig. 4A, the visualization of the Iris data using 6 labeled samples from class1 and 9 labeled samples from class 2 are shown. In Fig. 4B, the visualization using 7 labeled samples from class 1 and 8 labeled samples from class 2 is shown. These data samples can be selected by a random sampling [38] of the original dataset or by using common methods in active learning.

In comparison to the initial visualization of the Iris data in Fig. 2, it is clear that the proposed method has automatically solved the problem of overlapping clusters. In Fig. 4, class 1 (red points) is linearly separated from the two other classes, and class 2 (green points) and class 3 (black points), which are not linearly independent, have less overlap. Also in Fig. 4, all clusters are denser than the clusters visualized in Fig. 2.

In Fig. 5, the visualization of the Iris data set, using the proposed multi-class method is depicted. In this case, 4 data samples from class w_1, 3 samples from class w_2,and 3 samples from class w_3 are selected randomly as labeled data input.

Fig. 5A depicts the initial mapping of the Iris dataset using Star Coordinate without any user adjustment. Fig. 5B illustrates the result of this mapping after several adjustments by an expert user [6]. As shown in this figure, three clusters are detected after manual adjustment. The expert used the α-transformation and θ-rotation tools of VISTA [6] to detect these 3 clusters. Despite the fact that the pink and green clusters are partially overlapping, the total distribution of data clusters are very clear and well-separated. Fig. 5C shows the result of PCA mapping for visualization of the Iris dataset. Three classes are distinguishable but the cluster boundaries are somehow vague. Fig. 5D shows the result of LLE dimensionality reduction method applied to the Iris dataset. As shown in this figure, although the clusters are concentrated, two of them (the green and black clusters) are not easily distinguishable. In Fig. 5E, the visualization of Iris using our multi-class proposed method is shown. In comparison to Fig. 5B, it can be seen that our approach has achieved almost the same amount of overlapping. The minimized cluster overlapping area demonstrates the effectiveness of our proposed method and the results are very close to those of human inference. In this experiment, a small fraction of the data samples are labeled. The effect of the number of labeled samples over the visualization results is studied in the next section.

In the next set of experiments the method is evaluated against three high-dimensional datasets: DUC2007, DUC2006 and DNA[2]. DNA is a dataset from *Statlog*. DUC2007 and DUC2006 are a series of question answering data that have been conducted by the National Institute of Standards and Technology (NIST).

3.2. DUC2007

The DUC2007 dataset consists of some topics and a set of 25 relevant documents per topic. We consider three random topics as clusters where each cluster contains 25 document instances. After performing text mining pre-processing steps, such as tokenizing, stop words removing and stemming, we established the feature vectors of every document. As shown in Table 1, each vector has 3846 features. Fig. 6A illustrates the result of the basic Star Coordinate method on the DUC 2007 data set. As can be seen, the clusters overlap is high, making them difficult to distinguish. In Fig. 6D, the visualization of the high-dimensional data of the DUC 2007 using our proposed multi-class approach is shown, where the three

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html.

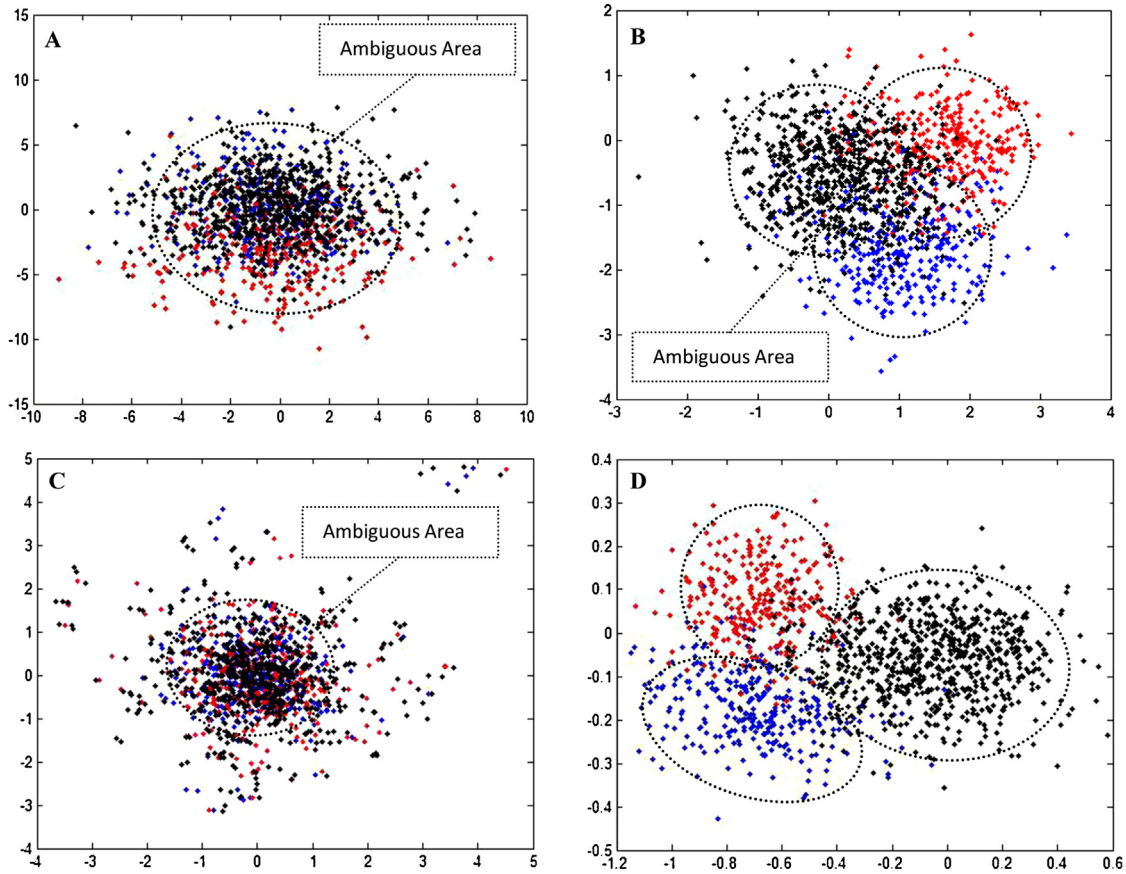

Fig. 8. Visualization on the DNA data set (A) using Star Coordinate with no adjustment method, (B) using PCA (C) using Locally Liner Embedding approach (with 300 for the number of neighbors), and (D) using the proposed approach.

clusters can be detected easily. In this experiment, we randomly label 9 instances (three from each class). In Fig. 6B, PCA results are shown where the clusters do not overlap but the cluster boundaries are only recognizable with color information (i.e. if cluster colors are removed from the results, only one cluster is recognizable), unlike Fig. 6C and 6D where the clusters are separated completely.

In many applications, such as image processing or text mining, the number of data dimensions are much larger than the number of instances, therefore the intra-class scatter matrix S_w can be singular. This is known as the *singularity* or *under sampled* problem [30]. In such cases, a constant *regularization term* can be added to the diagonal elements of the scatter matrix. So for the DUC2007 data set, Eq. (41) can be calculated with a *regularization parameter* γ as shown below [12]:

$$(S_W^{-1} + \gamma I)S_B\alpha = \lambda\alpha \tag{52}$$

The *Regularization parameter* can have a value between 0 and 1. A large γ may significantly disturb the information in the matrix, while a small λ may not be effective enough to solve the singularity problem. In the present experiments, we used $\gamma = 10^{-5}$.

3.3. DUC2006

Similar to DUC2007, the DUC2006 dataset also consists of some topics but with a set of 20 relevant documents per topics. In this case five random topics where selected. After performing the pre-processing steps mentioned in the previous experiment, we established the feature vectors of every document. As shown in Table 1, each vector has 4176 features. Two instances of each class were labeled (total of 10 labeled data). The visualization results of the initial Star Coordinate method outcome, PCA algorithm, LLE and the proposed approach applied to this dataset are shown in Fig. 7.

In Fig. 7 A (Star Coordinate initial result with no manual adjustment) the clusters are complelety overlapping and no single cluster can be detected. Two clusters in the PCA visualization result (Fig. 7B) are recognizable, however the other three clusters overlap. In this case, the user misguidedly recognizes only three clusters. In Fig. 7C the output of the LLE method is shown. Here the clusters are not fully separated and there are many noisy data points. Fig. 7D demonstrates the visualization results of the proposed approach, where all five clusters are separated and easily recognizable by the user, despite the fact that some noisy points can appear in the final result.

3.4. DNA

The DNA training data contains three classes with a total of 1400 instances, each of which having 180 features. In this case we labeled 6 instances from each class. Fig. 8A illustrates the result of the Star Coordinate method without any adjustment on the DNA dataset, and Fig. 8D shows the visualization of the high dimensional data of the DNA dataset using our proposed multi-class approach. In Fig. 8A, the clusters are almost complelety overlapped. Since the number of dimensions is greater than 50, VISTA [6] or iVIBRATE [7] cannot be used. As shown in Fig. 8D, our proposed method, without any direct user involvement, automatically reveals clusters clearly. By contrast, the PCA results (Fig. 8B) suffer from partial overlapping of the clusters. The LLE results (Fig. 8C) are not satisfying either and the clusters are not distinguishable.

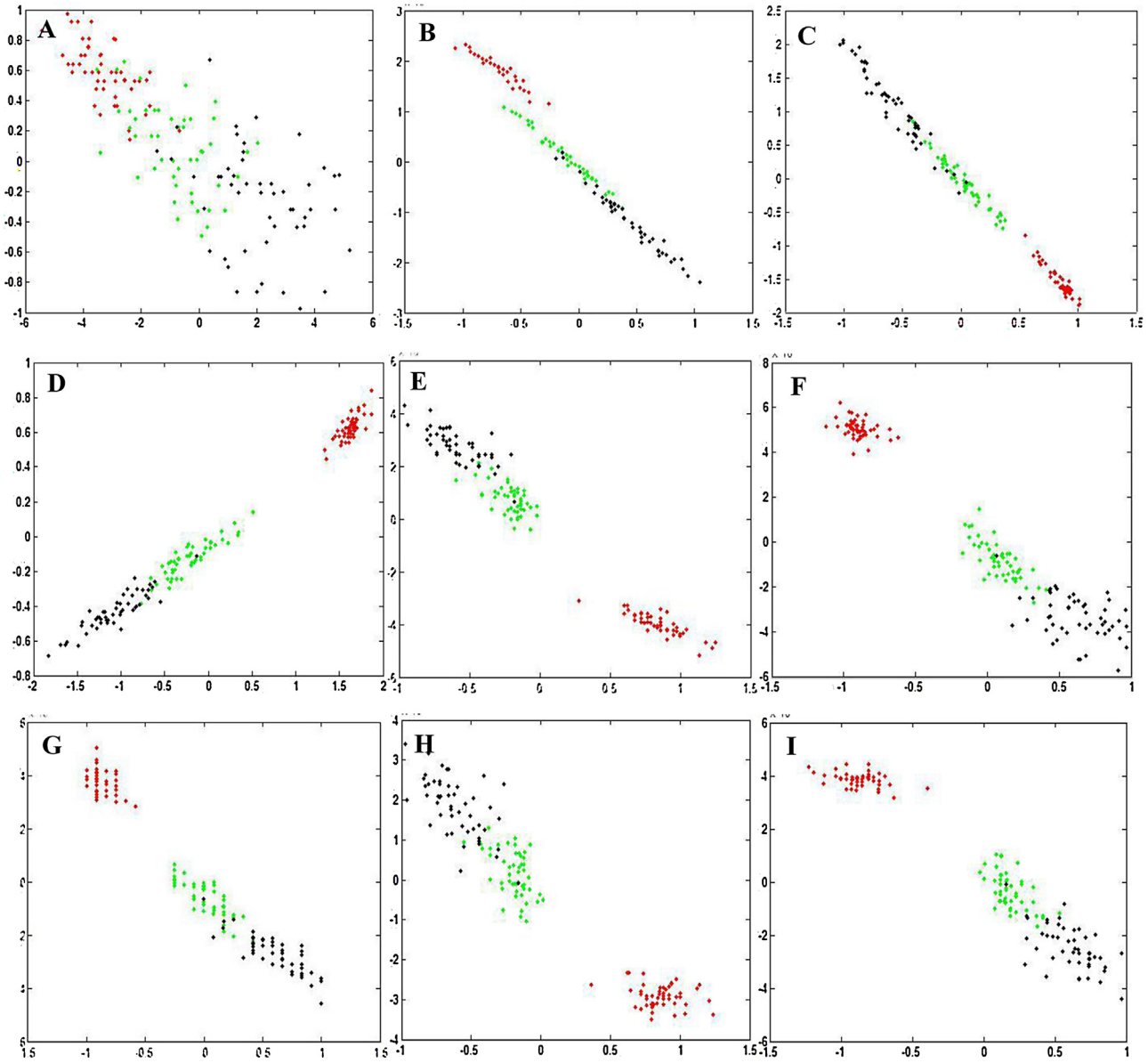

Fig. 9. Effect of the number of labeled samples on the visualization of the Iris dataset. In each picture three numbers of labeled samples a, b and c are used, for clusters w_1, w_2 and w_3 respectively. The number labeled samples in each picture are: A (1,1,1), B(1,1,1), C(2,1,2), D(2,2,1), E(2,2,2), F(5,3,3), G(6,9,7), H(32,21,29), I(30,33,29).

4. Discussion

4.1. Effect of the number of labeled samples

In this section, the effect of the number of labeled samples on the proposed method is studied. As shown in Fig. 9, the results are satisfactory over a wide range of values but, as the number of labeled samples decreases, the overlapping between clusters increases.

As shown in Fig. 9A and B, when the number of labeled samples is too small (1 per cluster), the degree of cluster overlap is high. In Fig. 9F, 11 samples (less than 0.08% of all 150 data points) are used but cluster separation is still satisfactory. As can be seen in Fig. 9E–I, a large increase in label information does not affect visualization significantly. Therefore, using even a small number of labeled samples can lead to satisfactory visualization of multi-dimensional data. As mentioned earlier, the samples to be labeled are selected randomly in the present study. The effect of other methods of selection is the subject of the Active Learning field and is not considered in this study.

4.2. Runtime comparison

In the following table the runtime of the proposed method, together with Star Coordinate and the PCA method, are compared on a 2.5 dual core CPU running Microsoft Windows XP. The runtime for the low-dimension data (DNA and Iris datasets) is almost equal but in higher dimensions the PCA method runtime increases substantially, so that for the DUC2006 and DUC2007 datasets the

Table 2
Runtime comparison.

Data set name	Data visualization algorithms		
	Star Coordinate	Proposed method	PCA method
Iris	0.0007	0.002	0.0003
DNA	0.03	0.11	0.11
DUC2007	0.03	98.4	404
DUC2006	0.06	156.2	565

PCA takes almost four times longer to complete compared to our method (Table 2).

5. Conclusion and future work

In this paper, we presented an extension to the Star Coordinate method that enables the application of this method to the visualization of high-dimensional data and requires no manual axes adjustment by the user. Our approach addresses the main problem with the Star Coordinate approach, namely that when the number of data dimensions is large (about 50 and more) manual modification to visualization parameters is almost impossible to achieve. We showed that the best data visualization is achieved where the axes are adjusted until mapped point clouds (clusters) in the mapped plane are as dense and as separated as possible. Our new approach is designed to achieve this optimal mapping, where visualization results are easily discernible by the user, using a set of labeled samples, even if the number of labeled samples is limited. By incorporating label information we modeled the objective function of the axes adjustment problem in Star Coordinate as *Fisher's discriminant* form. Therefore it is proven that the problem has a *global solution* and has *polynomial* time complexity. Experiments on different data sets have shown the potential of the proposed method in visualizing high-dimensional data.

The application of the proposed method in feature space is yet to be studied, since we modeled Star Coordinate visualization as a *Fisher's discriminant* form, which is only suitable for linearly separable data. Therefore it cannot visualize non-linearly separable clusters effectively. We are interested in extending our method to non-linear separable data by mapping the input space onto the feature space. Moreover, the effect of the number and selection method of labeled data (which is the focus of research activities in Active Learning) should be further analyzed.

References

[1] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering points to identify the clustering structure, in: Proceedings of the ACM SIGMOD Conference, 49-6, 1999.

[2] E. Bertini, A. Tatu, D. Keim, Quality metrics in high-dimensional data visualization: an overview and systematization, in: InfoVis – IEEE Information Visualization Conference, Providence, USA, 2011.

[4] K. Chakrabarti, S. Mehrotra, Local dimensionality reduction: a new approach to indexing high dimensional spaces, in: Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), 2000.

[5] O. Chapelle, B. Scholkopf, A. Zien (Eds.), Semi-supervised Learning, MIT Press, Cambridge, 2006.

[6] K. Chen, L. Liu, VISTA: Validating and refining clusters via visualization, J. Inform. Vis. 3 (4) (2004) 257–270.

[7] K. Chen, L. Liu, iVIBRATE: Interactive visualization-based framework for clustering large datasets, ACM Trans. Inform. Syst. (TOIS) 24 (2) (2006) 245–294.

[8] K. Chen, H. Xu, F. Tian, S. Guo, CloudVista: Visual cluster exploration for extreme scale data in the cloud, in: Scientific and Statistical Database Management Conference, Portland, OR, 2011.

[9] W.S. Cleveland, Visualizing Data, AT&T Bell Laboratories, Murray Hill, NJ, Hobart Press, Summit, NJ, 1993.

[10] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.

[11] D.R. Fokkema, G.L.G. Speijpen, H.A. Vandervorst, Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils, SIAM J. Sci. Comput. 20 (1) (1998) 94–125.

[12] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (1989) 165–175.

[13] Y.-H. Fua, M.O. Ward, E.A. Rundensteiner, Hierarchical parallel coordinates for exploration of large datasets, in: Proceedings of the IEEE Conference on Visualization'99, 1999, pp. 43–50.

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Inc., Boston, 1990.

[15] A.N. Gorban, B. Kegl, D.C. Wunsch, A. Zinovyev, Principle Manifolds for Data Visualization and Dimension Reduction, Springer, 2007.

[17] V. Gupta, G.S. Lehal, A survey of text summarization extractive techniques, J. Emerg. Technol. Web Intell. 2 (August (3)) (2010).

[19] P.E. Hoffman, Table Visualizations: A Formal Model and Its Applications (Doctoral Dissertation), Computer Science Department, University of Massachusetts, Lowell, 1999.

[20] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: Proceeding of 30th STOC, 1998.

[21] A. Inselberg, Multidimensional detective, in: Proceedings of the IEEE Symposium on Information Visualization, 1997, pp. 100–107.

[22] J.E. Jackson, A user's guide to principle components, in: Wiley Series in Probability and Mathematical Statistics, Wiley-Interscience, 1991, Reprinted 2003.

[24] H. Liu, H. Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic Publishers, 1998.

[25] E. Kandogan, Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, in: Proc. of ACM SIGKDD Conference, ACM Press, New York, 2001, pp. 107–116.

[26] D. Keim, Visual exploration of large data sets, ACM Commun. 44 (8) (2001) 38–44.

[27] D.A. Keim, et al., Visual analytics: scope and challenges, in: S. Simoff, M.H. Boehlen, A. Mazeika (Eds.), Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, Springer, 2008.

[28] G. Kollios, D. Gunopulos, N. Koudas, S. Berchtold, Efficient biased sampling for approximate clustering and outlier detection in large data sets, IEEE Trans. Knowl. Data Eng. 15 (5) (2003) 1170–1187.

[30] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, M.R. Thomas, Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data, Appl. Stat. 44 (1995) 101–115.

[31] R.S. Michalski, A planar geometric model for representing multidimensional discrete spaces and multiple-valued logic functions, in: Technical Report UIUCDCSR-78-897, University of Illinois, Urbana-Champaign, 1978.

[32] M.C.F. Oliveira, H. Levkowitz, From visualization to visual data mining: a survey, IEEE Trans. Vis. Comput. Graph. 9 (3) (2003) 378–394.

[34] S.T. Rowis, L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding, Science, 290.

[36] A. Tatu, et al., Visual quality metrics and human perception: an initial study on 2D projections of large multi-dimensional data, in: Proc. International Conf. on Advanced Visual Interfaces (AVI), ACM, 2010.

[37] K. Thangavel, P. Alagambigai, EVISTA – interactive visual clustering system, Int. J. Recent Trends Eng. 2 (November (1)) (2009).

[38] J. Vitter, Random sampling with a reservoir, ACM Trans. Math. Softw. 11 (1) (1985) 37–57.

[39] J. Yang, W. Peng, M. Ward, E. Rundensteiner, Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets, in: IEEE Symposium on Information Visualization, 2003. INFOVIS 2003, 21–21, 2003, 2003, pp. 105–112.

[44] P.C. Wong, R.D. Bergeron, 30 Years of multidimensional multivariate visualization, in: Scientific Visualization Overviews, Methodologies, and Techniques, IEEE Computer Society Press, 1997, pp. 3–33.