iVisClassifier: An Interactive Visual Analytics System for Classification Based on Supervised Dimension Reduction

Jaegul Choo* School of Computational Science and Engineering

Hanseung Lee† School of Electrical and Computer Engineering

Jaeyeon Kihm† School of Interactive Computing

Haesun Park* School of Computational Science and Engineering

Georgia Institute of Technology

Abstract

We present an interactive visual analytics system for classification, iVisClassifier, based on a supervised dimension reduction method, linear discriminant analysis (LDA). Given high-dimensional data and associated cluster labels, LDA gives their reduced dimensional representation, which provides a good overview about the cluster structure. Instead of a single two- or three-dimensional scatter plot, iVisClassifier fully interacts with all the reduced dimensions obtained by LDA through parallel coordinates and a scatter plot. Furthermore, it significantly improves the interactivity and interpretability of LDA. LDA enables users to understand each of the reduced dimensions and how they influence the data by reconstructing the basis vector into the original data domain. By using heat maps, iVisClassifier gives an overview about the cluster relationship in terms of pairwise distances between cluster centroids both in the original space and in the reduced dimensional space. Equipped with these functionalities, iVisClassifier supports users' classification tasks in an efficient way. Using several facial image data, we show how the above analysis is performed.

Index Terms: H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces—Theory and methods

1 Introduction

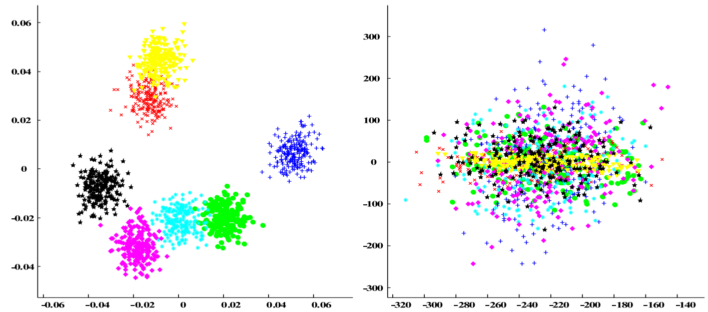
Classification is a widely-used data analysis technique across many areas such as computer vision, bioinformatics, text mining, etc. Given a set of data with known cluster labels, i.e., under a supervised setting, it builds a classifier (a training phase) to predict the label of new data (a test phase). Examples of classification tasks include facial recognition, document categorization, spam filtering, and disease detection.

Numerous classification algorithms such as an artificial neural network, decision trees, and support vector machines have been developed so far, and each method has advantages and disadvantages making it more suitable in certain domains. Even with its broad applicability, however, most of the classification algorithms are often performed in a fully automated manner that prevents users from not only understanding how the algorithm works on their data but also reflecting their domain knowledge into the classification process. Ironically, as classification algorithms become more sophisticated and advanced, they tend to be less interpretable to users due to their complicated internal procedure. These limitations may cause unsatisfactory classification results in real-world applications such as biometrics in which the reliability of the system is critical [27]. In some cases, there may be no option other than the manual classification process without being supported by automated techniques.

*e-mail: {jaegul.choo, hpark}@cc.gatech.edu

†e-mail: {hanseung.lee, jkhim3}@gatech.edu

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 29, Salt Lake City, Utah, USA
978-1-4244-9487-3/10/$26.00 ©2010 IEEE



(a) LDA (b) PCA

Figure 1: 2D Scatter plots obtained by two dimension reduction methods, LDA and PCA, for artificial Gaussian mixture data with 7 clusters and 1000 original dimensions. A different color corresponds to a different cluster.

This paper addresses how to visual analytics systems support automated classification for a real-world problem. As in other analytical tasks, the first step is to understand the data. From a classification perspective, users need to gain insight in terms of clusters such as how much the data within each cluster varies, which clusters are close to or distinct from each other, and which data are the most representative ones or outliers for each cluster. The next step is to understand both the characteristics of the chosen classifier itself and how they work on the data at hand. For instance, decision trees give a set of rules for classification, which are simple to interpret, and users can see which features in the data play an important role. In addition, analysis of misclassified data provides a better understanding of which types of clusters and/or data are difficult to classify. Such insight can then be fed back to the classification process in both the training and the test phases. In the training phase, users can refine the training data or modify the automated classification process for better performance in the long run. In the test phase, users can actively participate in determining the label of a new data by verifying each result that the automated process suggests and by performing further classification based on the interaction with a visual analytic system. The latter case ensures nearly perfect classification accuracy while maintaining much better efficiency than purely manual classification.

Not all classification algorithms are suitable for interactive visualization of how they work. Moreover, when the data is high dimensional such as image, text, and gene expression data, the problem becomes more challenging. To resolve this issue, we choose the classification method based on linear discriminant analysis (LDA) [9], one of the supervised dimension reduction methods. Unlike other unsupervised methods such as multidimensional scaling (MDS) and principal component analysis (PCA), which only use data, supervised ones also involve additional information such as cluster labels associated in the data. In case of LDA, it maximally discriminates different clusters while keeping the relationship among data within each cluster tight in the reduced dimensional space. This behavior of LDA has two advantages for interactive classification systems. The first one is that LDA is able to visualize the data so that their cluster structure can be well exposed. For example, as seen in Figure 1, LDA reveal the cluster structure better than PCA, and through LDA, users can easily find the cluster relationship and explore the data based on it. The other advantage is that the reduced dimensional representation of the data by LDA does not require a sophisticated classification algorithm in general since the data is already transformed to a well-clustered form, and such a transformation would map an unseen data item to a nearby area of its true cluster. Thus, after applying LDA, a simple classification algorithm such as k-nearest neighbors [7] can be performed, which has been successfully applied to many areas [4, 20]. Owing to this simplicity, users can get an idea about how the new data would be classified by looking at a nearby region based on visualization through LDA.

Inspired by the above ideas, we have developed a system called iVisClassifier, in which users can visually explore and classify data based on LDA. The first contribution of iVisClassifier lies in its emphasis on interpretation of and interaction with LDA for data understanding. Then, iVisClassifier features the ability to let users cooperate with the LDA visualization for the classification process. To show the usefulness of iVisClassifier, we present facial recognition examples, where LDA-based classification works well.

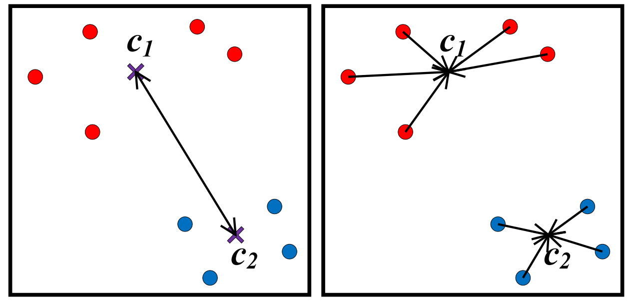
The rest of this paper is organized as follows. Section 2 discusses previous work related to interactive data mining systems and dimension reduction methods. Section 3 briefly introduces LDA and its use of the regularization in visualization, and Section 4 describes the details of iVisClassifier. Section 5 shows case studies, and Section 6 concludes our work.

2 Related Work

Supporting data mining tasks with interactive systems is an active area of study. As for clustering, an interactive system for hierarchical clustering was presented in [19], and a visualization-based clustering framework was proposed in [5], where users can analyze the clustering results and impose their domain knowledge into the next-stage clustering. In addition, various research has been conducted to make the dimension reduction process interactive. Yang et al. [28, 29] proposed a visual hierarchical dimension reduction method, which groups dimensions and visualizes data by using the subset of dimensions obtained from each group. Novel user-defined quality metrics was introduced for effective visualization of high-dimensional data in [14]. A user-driven visualization approach using MDS was proposed in [26].

However, in spite of the increasing demand from real-world applications, supporting classification tasks with an interactive visual system has not been studied extensively. Some studies [1, 2, 22] have tried to make a decision tree more interactive through visualization using circle segments [3] and star coordinates [15]. However, other classification methods have not been deeply integrated into interactive systems.

With respect to dimension reduction methods, a myriad of methods are still being proposed, and some of them claim their advantages on two or three-dimensional visualization. The recently proposed nonlinear manifold learning methods have shown the interesting ability to match the reduced dimensions to some semantic meanings such as the rotation of objects in image data [18, 21]. Another nonlinear method called t-SNE [24] has successfully revealed a hidden cluster structure in the reduced dimensional space for handwritten digit image and facial image data through computationally intensive iterations. While all the above-mentioned



(a) Maximization of distances between cluster centroids (b) Minimization of approximate cluster radii

Figure 2: Conceptual description of LDA. A different color corresponds to a different cluster, and c_1 and c_2 are the cluster centroids.

methods are unsupervised dimension reduction methods that do not consider cluster label information, supervised dimension reduction methods [9, 12], which explicitly utilize them in their computations, typically attempt to preserve the cluster structures by grouping the data with given labels.

Even with such technical advances, people still prefer traditional methods such as PCA, MDS, and self-organizing maps (SOM) because the state-of-the-art methods tend not to work universally for various types of data and they often lack interpretability. Motivated by this, a recently proposed system called iPCA [13] enables users to interact with PCA and its visualization results in the form of scatter plots and parallel coordinates. Our system shares a lot in common with iPCA in that users can play with LDA via scatter plots and parallel coordinates. Other than data understanding, however, our system aims further to support classification tasks utilizing the supervised dimension reduction.

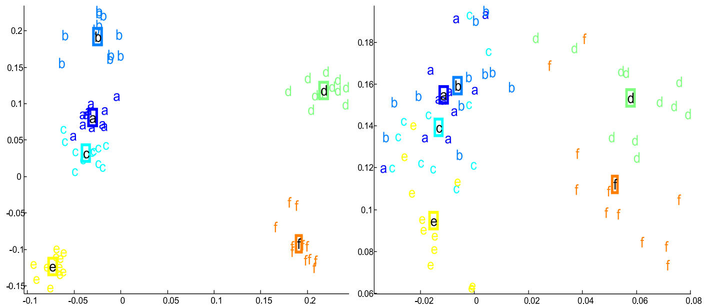
3 Linear Discriminant Analysis

In this section, we briefly introduce LDA and skip rigorous mathematical derivations due to a page limit. For more technical details about LDA and its use in visualization, refer to our previous work [6].

3.1 Concepts

LDA is a linear dimension reduction method that represents each of the reduced dimensions as a linear combination of the original dimensions. By projecting the data onto such a linear subspace, LDA puts cluster centroids as remote to each other as possible (by maximizing the weighted sum, B, of squared distances between cluster centroids, as shown in Figure 2(a)), while keeping each cluster as compact as possible (by minimizing the squared sum, W, of the disances between each data item in the cluster and its cluster centroid, as shown in Figure 2(b)), in the reduced dimensional space. Due to this characteristic, LDA can highlight the cluster relationship as shown in Figure 1(a), as opposed to other dimension reduction methods such as PCA. In LDA, this simultaneous optimization is formulated as a generalized eigenvalue problem that maximizes B while keeping its minimum value of W. Theoretically, the objective function value of LDA cannot exceed that in the original space, and such an upper bound is achieved as long as at least $k-1$ dimensions are allowed in LDA, where k is the number of clusters. Due to this characteristic, LDA usually reduces the data dimension to $k-1$.

Although LDA can reduce the data dimension down to $k-1$ dimensions without compromising its maximum objective function value, it is often not enough to use for 2D or 3D visualization purposes. In this case, users can either select a few of the most significant dimensions or perform an additional dimension reduction step



(a) $\gamma = 10^2$ (b) $\gamma = 10^5$

Figure 3: Effects of a regularization parameter γ in $S_w + \gamma I$. It can control how scattered each cluster is in the visualization. The data is one of the facial image data called SCface, and we chose the first six persons' images.

to further reduce the dimension to two or three [6]. In iVisClassifier, we adopt the former strategy so that we can easily interpret the dimension reduction step while interacting with all the LDA reduced dimensions.

3.2 Regularization to Control the Cluster Radius

In regularized LDA, a scalar multiple of an identity matrix γI is added to the within-scatter matrix S_w, the trace of which represents W.[1] It was applied to LDA [8] in order to circumvent a singularity problem when the data matrix has more dimensions than the number of data items, i.e., an undersampled case. In addition, regularization also has an advantage against overfitting in the classification context.

On the other hand, a unified algorithmic framework of LDA using the generalized singular value decomposition (LDA/GSVD) was proposed [11], which broadens the applicability of LDA regardless of the singularity. For undersampled data, e.g., text and image data, LDA/GSVD can fully minimize the cluster radii, making them all equal to zero. However, making the cluster radii zero results in representing all the data points in each cluster as a single point. Although it makes sense in terms of the LDA criteria, it does not keep any information to visualize at an individual data level. Thus, we utilize regularization to control the radius or scatteredness of clusters in the visualization to either focus on the data relationship or the cluster relationship, as shown in Figure 3. In an extreme case, when we sufficiently increase the regularization parameter γ, S_w is almost ignored in the minimization term, i.e., $S_w + \gamma I \simeq \gamma I$, so that LDA focuses only on maximizing B without minimizing W. Mathematically, this case is equivalent to applying PCA on the cluster centroids [6].

3.3 Algorithms

To ensure real-time interactions, it is important to design an efficient algorithm for LDA. Therefore, we reduce the data matrix size by applying either QR decomposition for undersampled cases or Cholesky decomposition for the other cases before running LDA. The main idea here is to transform a rectangular data matrix of size $m \times n$ into a square matrix of size $\min(m,n) \times \min(m,n)$ without losing any information. Then, the GSVD-based LDA algorithm is performed on this reduced matrix much efficiently. For more details, refer to [16].

[1] Instead of W, the LDA formulation uses S_w, which is then replaced with $S_w + \gamma I$ by regularization. For more details, refer to [6].

4 System Description

4.1 Data Encoding

Given a data set along with its labels, iVisClassifier first encodes the data into high-dimensional vectors. In its current implementation, it takes text documents, images, and generic numerical vectors with comma-separated values. When dealing with image data, the pixel values in each image are rasterized to form a single column vector, and text data are encoded using the bag-of-words model. Such encoding schemes determine the dimensions of image and text data as the total number of pixels in a single image and the total number of different words, respectively, which can be up to the hundreds of thousands.

Along with numerical encoding, iVisClassifier has several optional pre-processing steps such as data centering and normalization that makes the norm of every vector equal. In addition, other domain-specific pre-processing steps are also provided, such as contrast limited adaptive histogram equalization [17] for image data and stemming and stop-word removal for text data.

4.2 Visualization Modules

Once the data matrix whose columns represent data items is obtained, LDA is performed on this matrix with its associated labels. Users can recompute LDA with different regularization parameter values γ through a horizontal slide bar interface until the data within each cluster are adequately scattered. As described in Section 3, LDA reduces the data dimension to $k-1$ where k is the number of clusters. Just as the reduced dimensions in PCA are in an order to preserve the most variance, those in LDA are also in an order for preserving the most value of the LDA criterion. That is, the first reduced dimension represents each cluster most compactly while keeping different clusters most distinctly. With this in mind, we visualize LDA results in four different ways: parallel coordinates (Figure 4A), the basis view (Figure 4B), heat maps (Figure 4C), and 2D scatter plots (Figure 4F).

Parallel coordinates

Parallel coordinates is a common way to visualize multi-dimensional data. In parallel coordinates, the dimension axes are placed side by side as a set of parallel lines, and the data item is represented as a polyline whose vertices on these axes indicate the values in the corresponding dimensions. The main problem of parallel coordinates is that it does not scale well in terms of both the number of data items and dimensions. However, LDA can deal with both problems effectively in the following ways. First, with a manageable number of clusters, k, LDA reduces the number of dimensions to $k-1$, without losing any information on the cluster structure based on the LDA criterion. In addition, in terms of the number of data items, LDA plays the role of data reduction for undersampled cases since it can represent all the data items within each cluster as a single point by setting $\gamma = 0$, which in turn visualizes the entire data as k items. The dimension-reduced data by LDA may suffer the same scalability problem when the number of clusters and/or the regularization parameter γ increases. Nonetheless, in most cases, LDA significantly alleviates the clutter in parallel coordinates in that dealing with a large number of clusters is not practical and that users can always start their analysis with $\gamma = 0$.

Our implementation of parallel coordinates has several interactions including a basic zoom-in/out function. First, users can control the transparency of the polylines to see how densely the lines go through a particular region. To this end, users can switch all the colors indicating cluster labels to a single one, e.g. black. In addition, iVisClassifier has several shifting and scaling options. One is to align the minimum value of each dimension at the bottom horizontal line in the view, and the other is to align both the minimum and the maximum values at the top and bottom line, respectively. iVisClassifier is also able to filter the data by selecting particular

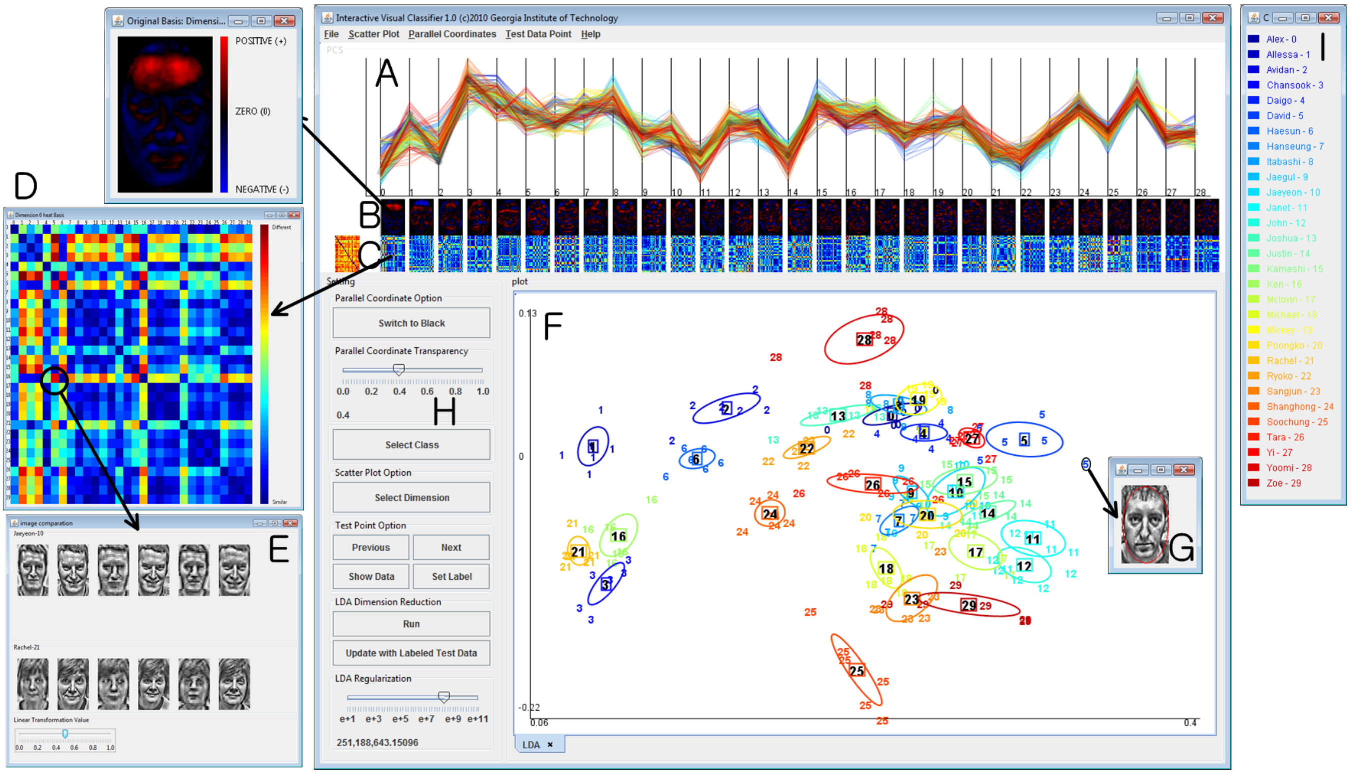


Figure 4: The overview of the system. SCface data with randomly chosen 30 persons' images were used, and different colors correspond to different clusters, e.g., persons. The arrow indicates a clicking operation. (A) Parallel coordinates view. The LDA results in 29 dimensions are represented. (B) Basis view. The LDA basis vectors are reconstructed in the original data domain, e.g., images in this case. (C) Heat map view. The pairwise distances between cluster centroids are visualized. The leftmost one is computed from the original space, and the rest from each of the LDA dimensions. Upon clicking, the full-size of a heat map is shown (D), and clicking each square shows the existing data in the corresponding pair of clusters (E). (F) Scatter plot view. A 2D scatter plot is visualized using two user-selected dimensions. When clicking a particular data point, its original data item is shown (G). (H) Control interfaces. Users can change the transparency and the colors in parallel coordinates. Data can be filtered at the data level as well as at the cluster level. The interfaces for unseen data visualize them one by one, interactively classify them, and finally updates the LDA model. A horizontal slide bar for the regularization parameter value in LDA controls how scattered each cluster is visualized. (I) shows the legend about cluster labels in terms of their assigned colors and enumerations.

clusters and/or data points in a certain range specified by a mouse pointer, and brushing and linking is implemented between parallel coordinates and scatter plots.

Basis view

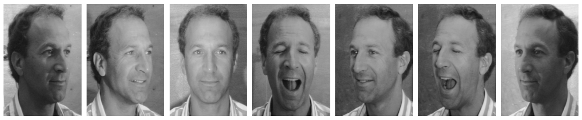
When data go through any kind of computational algorithms, it is crucial to have a better understanding of what happens in the process. For instance, even though the dimension reduction result is given by LDA, users may need to know the meaning behind each dimension and the reasons why those dimensions maximize the LDA criterion. Without such information, users cannot readily understand why certain data points look like outliers or certain clusters are prominent in the LDA result. Following this motivation, we provide users with the meaning of each reduced dimension of LDA in the following way.

First of all, LDA is a linear method where each reduced dimension is represented as a linear combination of those in the original space. Thus, we have a linear combination coefficient for each reduced dimension, which we call a basis vector, and the dimension of this basis vector is the same as the original dimension. For image data in which the original dimension is the number of pixels in the image, each coefficient value in this basis vector corresponds to each of the pixels. Based on this idea, we reconstruct the LDA basis in the original data domain, e.g., an image in our case. However, it is not always straightforward to convert the basis back to the original data domain. For example, pixel values in an image have a certain specifications that they have to be all integers between 0 and 255 while the LDA basis is real-valued with positive and negative signs mixed. In the past, several heuristics to handle this issue were used in the context of PCA by mapping basis vectors to grayscale images [23, 25] by taking either its absolute value or adding the minimum value. However, these heuristic methods lose or distort the information contained in the basis vectors. Therefore, we map positive and negative numbers in the basis vector into two color channels, red and blue, respectively. In this way, we obtain the reconstructed images of LDA basis vectors as shown in Figure 4B.

Heat maps

With heat maps, we visualize the pairwise distances between cluster centroids, where each heat map has $k \times k$ elements. The leftmost heat map in Figure 4C represents such information in the original high-dimensional space, and the following ones on the right side are computed within each reduced dimension of LDA. Through this visualization, we can get the information about which particular cluster is distinct from the other clusters and which cluster pairs are close or remote in each dimension. Furthermore, comparisons between heat maps of the original space and each of the reduced dimension show which cluster distances are preserved or ignored.

By clicking the (i, j)-th square in the enlarged heat map (Figure



(a) Weizmann

(b) SCface

Figure 5: A single person's image samples in two data sets.

4D), users can compare the data items in the i-th and j-th clusters as shown in Figure 4E. In addition, the slide bar at the bottom in Figure 4E enables users to overlap the data image with its corresponding basis image, which tells us how the pixels in these images are weighted in its corresponding dimension and why the data of the selected two clusters are closely or remotely related in this dimension, as shown in Figure 9.

Scatter plots

The scatter plot visualizes data points in the two user-selected reduced dimensions of LDA with a zoom-in/out functionality. In this view, a data item is represented as a point with an initial letter and a different color of its corresponding cluster label. Additionally, the first and the second order statistics per cluster, which are the mean and the covariance ellipse, give the effective information about clusters.

Our scatter plot view given by LDA allows users to interactively explore the data in view of the overall cluster structure in the following senses: 1. which data points are outliers or representative points in their corresponding clusters, 2. which data points are outliers or representative points in their corresponding clusters, 3. how widely the data points within a cluster are distributed and accordingly, which clusters have potential subclusters, and 4. which data points overlap between different clusters.

In addition, brushing and linking with parallel coordinates overcome the limitation that the scatter plot can only show two or three dimensions at a time. In this way, users can see how the selected data or clusters in the scatter plot behave in the other dimensions.

4.3 Classification Modules

After obtaining insight from exploring the data with known cluster labels, users can now interactively perform classification on the new data whose labels are to be determined. This process works as follows. First, a new data item is mapped onto the reduced dimensional space formed by the previous data. It is then visualized in parallel coordinates and in the scatter plot view. Such visualization significantly increases the efficiency of users' classification tasks by visually reducing the search space. Within this reduced visual search space, users can easily compare the new data item with the existing data or clusters nearby. When the new data point falls into a cluttered region where many different clusters overlap, users can select or filter out some data or clusters and recompute LDA with this subset of data including the new point, which we call a computational zoom-in process. In other words, LDA takes into account the selected clusters and/or those corresponding to the selected data, which requires a much smaller number of dimensions than $k-1$ for LDA to fully discriminate the selected clusters. Based on the new visualization generated in this way, uses can better identify which clusters the new point belongs to.

On completing the visually-supported classification process, users can assign a label to the new data item and optionally include the newly labeled data in future LDA computations, which is initiated only when users want to recompute them. The reason we do not force users to include every new data in LDA computations is that users' confidence level of the assigned label may not be high enough for some reason such as noise.

5 Case Studies

In this section, we present an interactive analysis using two sets of facial image data, Weizmann database[2] and SCface database [10], for facial recognition. Weizmann is composed of 28 persons' frontal images in a constant background, in which each person has 52 images. The variations within each person's images exist regarding viewing angles, illuminations, and facial expressions. We resized the original 512×352 pixel images to 64×44 pixel images, resulting in 2816 dimensional vectors. SCface is an image collection taken in an uncontrolled indoor environment using multiple video surveillance cameras with various image qualities. It is composed of 4160 static images of 130 subjects, of which we randomly selected 30 persons' images for our study, where each person has 32 images. Since the images in SCface generally contain parts other than a face, such as the upper body of a person and a different background, we have cropped a facial part using an affine transformation that aligns the images based on the eye coordinates. The image samples of two data sets are shown in Figure 5.

In the following, we present an exploratory analysis towards better understanding of both the data and the computational method we have used, i.e., LDA. Next, we describe how users interactively perform classification supported by iVisClassifier.

5.1 Exploratory Data Analysis

In general, understanding the data at the cluster level is essential to deriving an initial idea about the overall structure in a large-scale data set. In this sense, we can begin with the heat map view of the pairwise distances in the original space to look at how the clusters are related. From the heat maps shown in Figure 6(a) and 7(a), we can see that pairwise cluster distances vary more in Weizmann than in SCface. This view also reveals the clusters that look distinct from the other clusters, e.g., person 14 in Weizmann and person 7 in SCface. Element-wise comparisons reveal that persons 11 and 14 look quite distinct, which makes sense due to baldness and shirt colors, but persons 2 and 10 look similar in Figure 6(a). Similarly, persons 1 and 7 look different while persons 2 and 26 are indistinguishable in Figure 7(a).

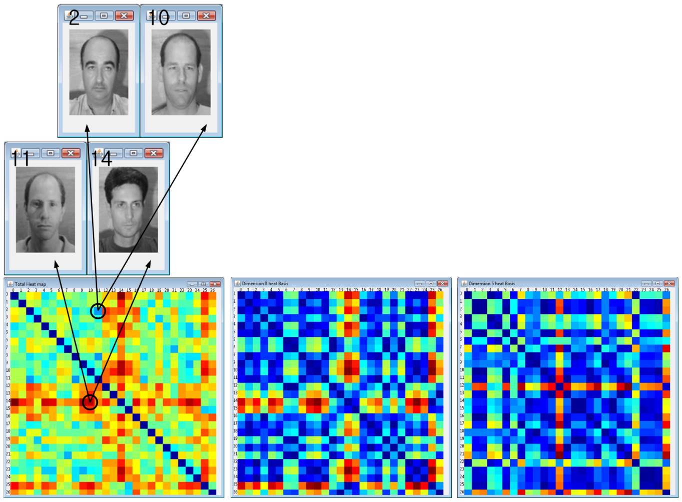
Next, let us look at the heat maps of the LDA dimensions shown in Figures 6-7. The first dimension turns out to reflect the most distinct clusters in the original space. In addition, the heat maps in the LDA dimensions have mostly blue-colored elements, i.e., almost zero, except for a few rows and columns, which indicates that each of the LDA dimensions tends to discriminate only a few clusters.

Next, Figure 8 shows the image reconstruction of the first six LDA bases for both data sets. It is interesting to see that in both cases, the forehead part is heavily weighted in the first dimension,[3] and then in the second dimension, the forehead part is differentiated into upper and lower parts. This indicates that the forehead part is the most prominent factor for facial recognition based on LDA in our data.

Basis images can be overlapped with the original images to highlight the region in the images that is heavily weighted in a specific reduced dimension. The example shown in Figure 9 was obtained by selecting one of the most remote cluster pairs (red-colored one in Figure 6(b)) in the first dimension. In the region covered by a blue color, we can see that the pixel values are quite different, i.e.,

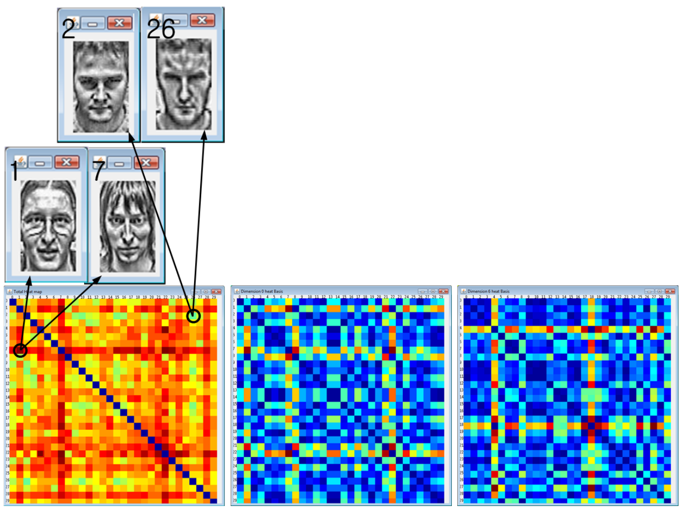
[2] http://www.wisdom.weizmann.ac.il/˜vision/databases.html

[3] Negative weighting coefficients represented as blue colors are equivalent to positive ones by negating the basis and the corresponding coordinate values of the data.



(a) The original space (b) The first dimension (c) The fifth dimension

Figure 6: Heat map view of the pairwise cluster distances of the Weizmann data set.



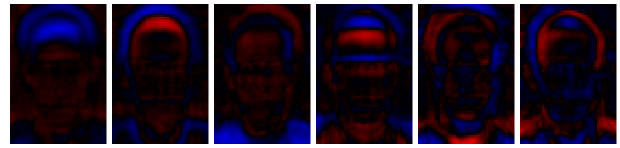
(a) The original space (b) The first dimension (c) The sixth dimension

Figure 7: Heat map view of the pairwise cluster distances of the SC-face data set.

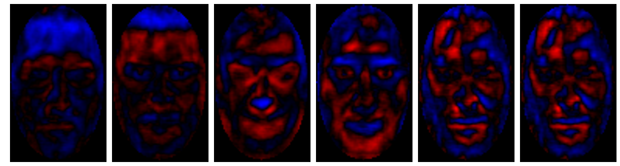
light in the first cluster and dark in the second cluster, which puts them far apart in the corresponding reduced dimension.

5.2 Interactive Classification

As described in Section 4.3, the main benefit of iVisClassifier for classification is that it visually guides users to the correct clusters for unseen data while allowing users to have control over the classification process. In general, most of the new data would be closely placed to their corresponding clusters in the scatter plot. If only a few clusters are found nearby, e.g., when a point to classify is placed near the cluster 7, which is almost isolated from the other clusters at the leftmost part in Figure 10(a), then by checking some of the nearby data in the cluster 7, users can quickly classify them into their corresponding clusters. However, a problem arises when the new point is visualized near a cluttered region as shown in Figure 10(a). With this visualization, we have a less clear idea as to which clusters to look at because numerous clusters exist near the point of interest. In this case, we can select a subset of data points around it and then recompute the dimension reduction only with



(a) Weizmann



(b) SCface

Figure 8: Reconstructed images of the first six LDA bases.

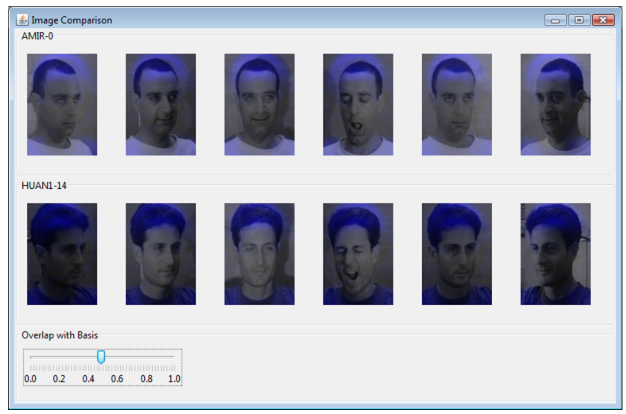


Figure 9: The effect of overlapping a basis image over the original data. Users can see which part of images are weighted by a basis vector.

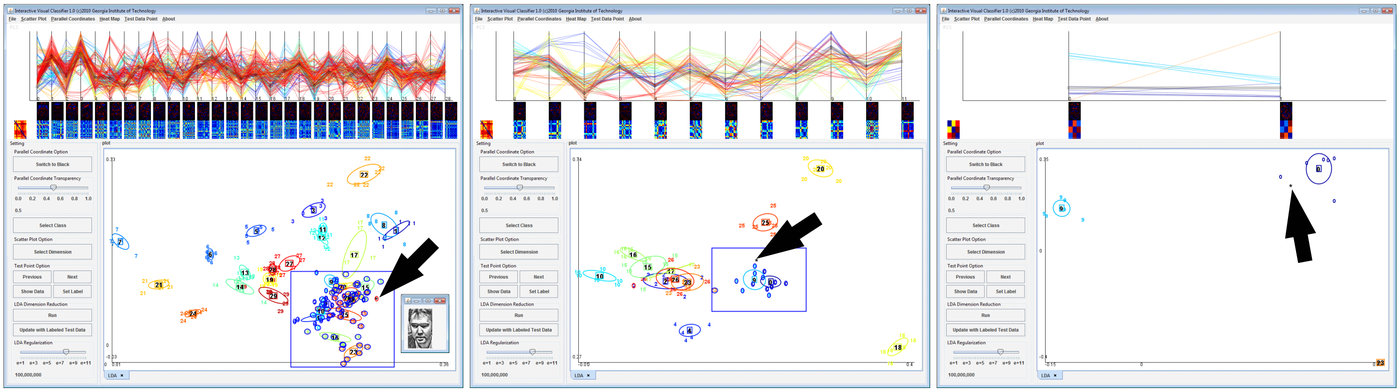
this subset. Figure 10 shows that this process guides the new point to its true cluster.

Another scenario for interactive classification in iVisClassifier is cooperative filtering between parallel coordinates and the scatter plot. Figure 11(a) shows a case where the new point is placed in an ambiguous region to classify. As we find that the new point (shown in a gray color in parallel coordinates) goes through the top region in dimension 7, we can filter the data in this dimension, and accordingly, the selected data are also highlighted in the scatter plot with a black circle, as shown in Figure 11(b). Additional filtering in the scatter plot by selecting either nearby clusters or data items ends up with only one possible cluster, as shown in Figure 11(c).

Once some of the new data are assigned their labels, users can recompute LDA by taking into account the newly labeled data. Figure 12 shows the distributions of the new data whose label is '0' before and after LDA recomputation with a newly labeled data item. As we can see, the rest of the unseen data in the cluster 0 becomes closer to its centroid after LDA recomputation, which indicates that the updated LDA dimensions potentially better discriminates the unseen data.

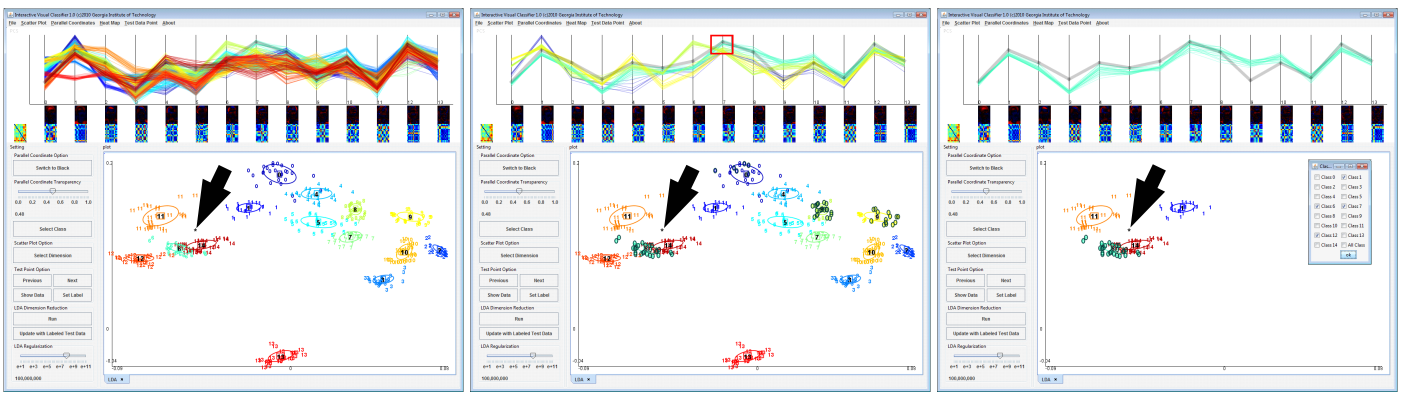
6 Conclusions and Future Work

In this study, we have presented iVisClassifier, a visual analytics system for clustered data and classification. Our system enables users to explore high-dimensional data through LDA, which is a supervised dimension reduction method. We interpret the effect of regularization in visualization and provide an effective user-interface in which users can control the cluster radii depending on whether they focus on the cluster- or the data-level relationships. In



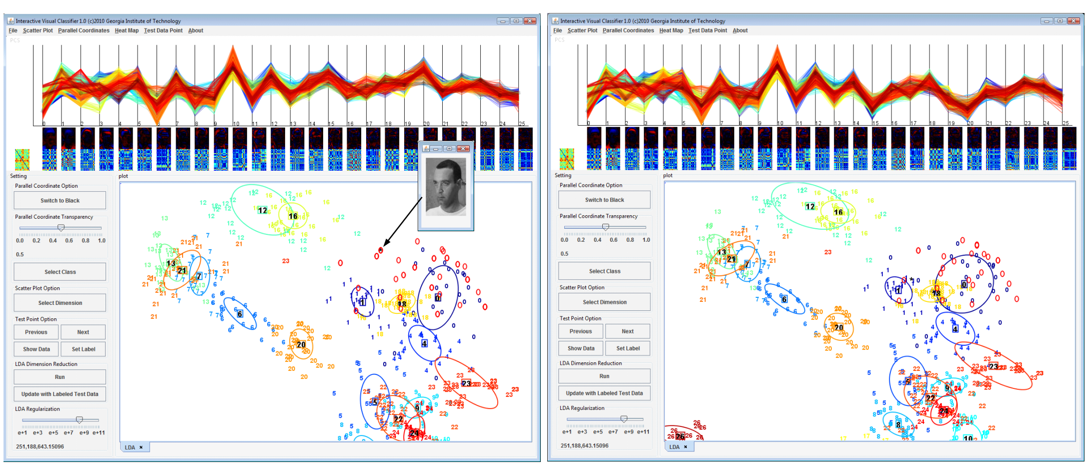
(a) The initial filtering (b) The second filtering (c) The final visualization result

Figure 10: Interactive classification by computational zoom-in. Recursive visualization by recomputing LDA for interactively selected subsets of data guides a new point into its corresponding cluster. The thick arrow indicates the new point position.



(a) The initial visualization (b) The filtering in parallel coordinates (c) The filtering in the scatter plot

Figure 11: Interactive classification by mutual filtering. Filtering both in parallel coordinates and the scatter plot leads to a single cluster. The thick arrow indicates the new point position.



(a) Before labelling the test point (b) After labelling the new point

Figure 12: Effects of LDA recomputation with including a newly labeled point in the existing data. The arrow indicates the newly labeled point, and the red circles represent the distribution of the remaining unseen data in the cluster 0.

addition, iVisClassifier facilitates the interpretability of the computational model applied to their data. Various views such as parallel coordinates, the scatter plot, and heat maps interactively show rich aspects of the data. Finally, we showed that iVisClassifier can efficiently support a user-driven classification process by reducing humans' search space, e.g., recomputing LDA with a user-selected subset of data and mutual filtering in parallel coordinates and the scatter plot.

As our future work, we plan to improve our system to better handle other types of high-dimensional data and their classification tasks. Although our system can currently load and visualize other types of high-dimensional data such as text data, how we accommodate the basis view and blend the data item with the basis in the original data domain, as shown in Figure 9, would be the main issues.

In addition, although our tool works well when there is a reasonable number of clusters, it may not scale well when we have many clusters, e.g., hundreds of people in facial recognition. To handle this problem, we are considering the hierarchical approaches that group the clusters based on their relative similarities to keep the number of clusters manageable in an initial analysis.

Finally, the computation of LDA can be burdensome for user interactions when we have a large-scale data. Novel interactions with LDA provided by iVisClassifier motivate the new types of dynamic updating algorithms based on the previous LDA results in various situations. For instance, updating the LDA results when changing the regularization parameter value has not been studied before. Thus, we are currently exploring for various situations and their corresponding updating algorithms when computational algorithms are integrated into user-interactive systems.

7 Acknowledgements

The work of these authors was supported in part by the National Science Foundation grants CCF-0728812 and CCF-0808863. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: an interactive approach to decision tree construction. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–396, New York, NY, USA, 1999. ACM.

[2] M. Ankerst, M. Ester, and H.-P. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 179–188, New York, NY, USA, 2000. ACM.

[3] M. Ankerst, D. Keim, and H. Kriegel. 'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets. In *Proc. Visualization*, Hot Topic Session,, 1996.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[5] K. Chen and L. Liu. ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Trans. Inf. Syst.*, 24(2):245–294, 2006.

[6] J. Choo, S. Bohn, and H. Park. Two-stage framework for visualization of clustered high dimensional data. In *IEEE Symposiu on Visual Analytics Science and Technology, 2009. VAST 2009.*, pages 67 –74, oct. 2009.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-interscience, New York, 2001.

[8] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition, second edition*. Academic Press, Boston, 1990.

[10] M. Grgic, K. Delac, and S. Grgic. SCface - surveillance cameras face database. *Multimedia Tools and Applications Journal*, 2009.

[11] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995 –1006, aug. 2004.

[12] M. Jeon, H. Park, and J. B. Rosen. Dimensional reduction based on centroids and least squares for efficient processing of text data. In *Proceedings of the First SIAM International Workshop on Text Mining*. Chiago, IL, 2001.

[13] D. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. iPCA: An Interactive System for PCA-based Visual Analytics . *Computer Graphics Forum*, 28(3):767–774, 2009.

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

[15] E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 107–116, New York, NY, USA, 2001. ACM.

[16] H. Park, B. Drake, S. Lee, and C. Park. Fast Linear Discriminant Analysis using QR Decomposition and Regularization. *Technical Report GT-CSE-07-21*, 2007.

[17] E. Pisano, S. Zong, B. Hemminger, M. DeLuca, R. Johnston, K. Muller, M. Braeuning, and S. Pizer. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, 11(4):193–200, 1998.

[18] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

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

[20] K. Y. Tam and M. Y. Kiang. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *MANAGEMENT SCIENCE*, 38(7):926–947, 1992.

[21] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[22] S. Teoh and K. Ma. StarClass: Interactive visual classification using star coordinates. In *Proceedings of the 2003 SIAM International Conference on Data Mining (SDM03)*, 2003.

[23] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

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

[25] M. Vasilescu and D. Terzopoulos. Multilinear Analysis of Image Ensembles: TensorFaces. *Necture Notes in Computer Science*, pages 447–460, 2002.

[26] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 57 –64, 0-0 2004.

[27] R. Willing. Airport anti-terror systems flub tests; Face-recognition technology fails to flag "suspects". *USA TODAY, September 2, 2003. Available at http://www.usatoday.com/travel/news/2003/09/02-air-secur.htm.*

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

[29] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *VISSYM '03: Proceedings of the symposium on Data visualisation 2003*, pages 19–28, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.