RESEARCH

Project-5

Sujan Darai (sujad96@zedat.fu-berlin.de), Matanat Mammadli (matanam94@zedat.fu-berlin.de), Samra Hamidovic (samrah96@zedat.fu-berlin.de)

Full list of author information is available at the end of the article

Abstract

Goal of the project: The project aims to explore interactive visualization techniques for high-dimensional datasets, including PCA, t-SNE, and UMAP, to uncover underlying structures and assess dimensionality reduction effectiveness. By implementing parallel coordinates plotting, linked scatter plots, and data inspection, we seek to understand dataset distributions, identify misclassifications or outliers, and demonstrate the practical utility of these visualization methods in facilitating data exploration and informed decision-making across diverse domains.

Main results of the project: The project found that the 2-D t-SNE and UMAP scatter plots for the breast cancer dataset unexpectedly lacked the benign class, while the digits dataset showed well-separated clusters for all 10 digits. Selected digit samples, mainly digits 8 and 9, were identified as outliers. The pairwise distance matrix effectively illustrated sample proximities, with darker colors indicating closer samples and lighter colors indicating farther apart samples.

Personal key learning:

1 Sujan Darai: Learned to use jscatter, understanding of the parallel coordinate plot.
2 Samra Hamidovic: Learned what brushing is, how to perform data-driven brushing technique on our dataset, and how to illustrate two different representations of embeddings/plot.
3 Matanat Mammadli: Learned about parallel coordinate plots, learned to use brush and linking techniques, and wrote results in a better way.

Estimated working hours:

1 Sujan Darai: 8 hours
2 Samra Hamidovic: 8 hours
3 Matanat Mammadli: 8 hours

Project evaluation: 1

Number of words: 2255

Keywords: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Brushing, Linking

1 Introduction

In the realm of data visualization, the evolution of interactive techniques has revolutionized the way data analysts interact with and derive insights from complex datasets. One such dynamic graphical method, brushing, allows for real-time interaction with data displays on computer graphics terminals, enabling analysts to select, emphasize, and manipulate specific points or regions of interest. The two main types of brushing operations, demand-driven brushing and data-driven brushing, offer distinct approaches to exploring multidimensional data, catering to both user-defined criteria and inherent data patterns.[1]

Moreover, linking in data visualization plays a crucial role in establishing connec-

tions between corresponding points or regions across multiple scatterplots or graphical displays. This technique enhances the exploration of relationships, patterns, and trends within the dataset, providing a comprehensive understanding of the data and uncovering valuable insights that may remain hidden in individual plots.[1]
This project at hand aims to delve into interactive visualization techniques tailored for high-dimensional datasets, such as PCA, t-SNE, and UMAP, to unveil underlying structures and evaluate the effectiveness of dimensionality reduction. Through the implementation of parallel coordinates plotting, linked scatter plots, and data inspection, our goal is to gain insights into dataset distributions, pinpoint misclassifications or outliers, and showcase the practical utility of these visualization methods in facilitating data exploration and supporting informed decision-making across diverse domains.
By leveraging these advanced visualization tools and techniques, we aspire to enhance the analytical capabilities of data scientists and domain experts, enabling them to extract meaningful insights, discover hidden patterns, and make data-driven decisions with confidence and clarity.

2 Goal of the project

The project sets out to explore interactive visualization techniques tailored for high-dimensional datasets, employing methods such as PCA and t-SNE. The primary aim is to unveil the underlying structures within the data and assess the effectiveness of dimensionality reduction techniques.
Through the implementation of parallel coordinates plotting, linked scatter plots, and data inspection, the project aims to achieve a comprehensive understanding of dataset distributions, identify potential misclassifications or outliers, and evaluate the performance of dimensionality reduction methods.
By using interactive visualization tools, we aim to help users explore complex datasets more easily. This will help people make better decisions in different areas. Ultimately, we want to show how these visualization techniques can be useful in analyzing data and understanding it better.

3 Data and preprocessings

3.1 Data

In this project, two datasets were used both of which were taken from the scikit-learn library. One dataset is the digit dataset, and another is the cancer dataset from Wisconsin. The digit dataset contains 1797 samples which store grayscale images of handwritten digits of 10 classes from 0 to 9. Each of the digits is represented by the 8x8 pixel array of elements where pixel values vary from 0 to 16, meaning pixel intensity. Each sample has 64 features in total in the digit dataset. The digit dataset is primarily useful for machine learning algorithms for the classification task for digit recognition.
Conversely, the breast cancer dataset comprises 569 samples of breast biopsies in which the 30 numerical features characterize each sample. The dataset contains information about the different cells contributing to the cancer diagnostics whether the tumor is malignant or benign. It has only 2 classes 0 and 1 where 0 represents benign and 1 represents malignant conditions. The breast cancer dataset is a classic

dataset utilized in binary class classification tasks for machine learning algorithms. We have not used the metabolomics dataset in this report because of the time limit. Since both of the datasets have only numerical columns or features, null values were checked and no null value was found. It was then loaded into the jupyter notebook. Both the digits and breast cancer datasets were ready to use for the dimensionality reduction procedures and other tasks specified in the goal section.

3.2 Preprocessing

In the preprocessing section, no data preprocessing was needed because the data was already stored as NumPy arrays and it doesn't contain any null values.

4 Methods

4.1 Brushing

Brushing in data visualization is a dynamic graphical method that enables data analysts to interact in real-time with data displays on computer graphics terminals. It involves the manipulation of a mouse-controlled brush over scatterplots or other visual representations of data to select, emphasize, or manipulate specific points or regions of interest. The primary objective of brushing is to facilitate interactive exploration and analysis of multidimensional data, allowing analysts to uncover patterns, relationships, outliers, and other insights within the dataset. There are two types of brushing: Demand-driven brushing and data-driven brushing.

Demand-driven brushing allows analysts to select specific data points or regions based on their explicit input or query. Analysts can specify the criteria for selecting data points, such as setting thresholds or defining ranges, and the brushing operation is carried out based on these predefined conditions. Demand-driven brushing empowers analysts to focus on specific subsets of data that meet their criteria, enabling targeted analysis and exploration of relevant data points.[1]

Data-driven brushing automatically identifies and highlights data points or regions based on inherent patterns, relationships, or outliers within the dataset. This type of brushing operation is driven by the characteristics of the data itself, such as clustering algorithms, trend detection, or outlier identification techniques. Data-driven brushing helps analysts uncover hidden insights and relationships within the dataset without the need for explicit user input, allowing for more exploratory and discovery-driven analysis.[1]

4.2 Linking

Linking in data visualization allows data analysts to establish connections between corresponding points or regions across multiple scatterplots or graphical displays. The primary goal of linking is to facilitate the exploration of relationships, patterns, and trends that may exist between variables represented in different views of the data. By linking related data points across visualizations, analysts can gain a comprehensive understanding of the dataset and uncover valuable insights that may not be apparent when examining individual plots in isolation.[1] In our project, we used linking (`jscatter.link()`) to combine 2-D t-SNE embedded scatter plots of the digits and the breast cancer datasets.

4.3 PCA & t-SNE

Principal Component Analysis (PCA) was utilized to reduce the dimensionality of the dataset while retaining its essential variability. The process involved mean-centering and scaling the data to unit variance. A covariance matrix was then computed to capture the relationships between variables. Eigenvalues and eigenvectors of this matrix were calculated, with the eigenvectors representing the principal components (directions of maximum variance) and eigenvalues indicating the variance amount. The data was projected onto these principal components, creating uncorrelated variables ordered by explained variance. This transformation simplified the dataset, making it more manageable for analysis and visualization.[2]

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used to visualize high-dimensional data in two or three dimensions. Developed by Laurens van der Maaten and Geoffrey Hinton, t-SNE preserves local data structures by mapping similar high-dimensional points to nearby points in a lower-dimensional space. It models pairwise similarities using Gaussian distributions in high dimensions and t-distributions in low dimensions. By minimizing the Kullback-Leibler divergence between these distributions through gradient descent, t-SNE iteratively refines the low-dimensional representation. Key parameters such as perplexity, learning rate, and the number of iterations are optimized for the best results. This method is particularly effective for identifying clusters and relationships in complex datasets, making it a valuable tool for data exploration and interpretation in fields like genomics and image processing.[3]

5 Results and discussion

We are going to discuss the results we obtained after applying all the aforementioned methods to our datasets.

We used the data-driven brushing technique because we didn't know much about our datasets. We searched for the outliers in the 2-dimensional t-SNE embedded scatter plots of the digits and the breast cancer datasets (created with the `jscatter` package from python), marked them with the lasso selection box tool, and got the index with `scatter.selection()`. We could also plot some of these selected digits, as seen in the figure 5. But overall, we selected 5 outliers, 5 digits with indices (1658, 1582, 794, 1553, 129).

We performed Principal Component Analysis (PCA), like in the previous project, on the digits and the breast cancer datasets and transformed them to reduce their dimensions to 10. We then converted our data to DataFrame and convert PCA results and labels to a format suitable for `Plotly`. Then plotted our results as the parallel coordinates plots. However, one notable observation was that our scatter plots for the 2-D t-SNE embedded breast cancer dataset did not include the benign class (class 1, diseased), which is quite unusual. We were unable to determine the reason for this discrepancy.

Afterward, we defined the function to analyze these selected samples, which calculated pairwise distances between selected samples using Euclidean distance and created pairwise distance matrix plot. This function could also identify wrongly classified samples/outliers, but in our cases there were none.

We created 10-dimensional embeddings of the digits and breast cancer datasets

using PCA. These were plotted as the parallel coordinates plots and are illustrated in figure 3. The parallel coordinates plot for the handwritten digits dataset illustrates the relationships between the principal components (PCs) derived from the PCA transformation. Each line in the plot represents a digit from 0 to 9, with the PCs plotted along the axes. The color of each line corresponds to the digit it represents, facilitating the differentiation of digits within the plot. This visualization allows for the examination of patterns and clusters formed by the digits in the reduced-dimensional space. All lines representing different colors/classes were well-distributed, indicating well-separated clusters of the digits classes.

Similarly, the parallel coordinates plot for the breast cancer dataset showcases the relationships between the principal components generated by PCA. Each line in the plot corresponds to a sample from the dataset, with the PCs plotted along the axes. The color of each line indicates the class label, distinguishing between benign and malignant tumors. By observing the patterns formed by the lines in the plot, insights into the distribution and separability of the classes in the reduced-dimensional space can be gleaned. While playing with this interactive graph, we could observe the patterns formed by the lines in the plot, we noticed more yellow lines than violet, which could indicate that the yellow lines were drawn last, violet lines were not drawn on top of the yellow ones. This suggests that the data are ordered according to the classes.

Two-dimensional t-SNE and UMAP embedded scatter plots of the breast cancer and the digits datasets are shown in figure 1 and 2. As mentioned earlier, somehow the breast cancer data does not showcase class 1 or benign/diseased cases. But on the digits dataset, we can clearly see all 10 digits building well-separated and dense clusters. The distance between different classes (clusters) is large and the distance between components from the same class is small, creating compact clusters and indicating well-performed t-SNE dimensionality reduction. We chose our selected samples (from the digits dataset) via the brushing method from this scatter plot as well. The UMAP creates clearer and more distinguishable patterns in figures than the t-SNE (which could be because of parameter choices of t-SNE), which helps in maintaining the structure and separability of clusters more effectively as compared to t-SNE scatter plots.

We visualized all of our selected digits from the famous handwritten digits dataset as well, as seen in figure 6 but due to the tight layout adjustment being applied after the subplots have been created, there are some unexpected boxes underneath our images. Here we see clearly, that the 5 outliers we selected from the digits scatter plot happen to be digit 9 (two times) and digit 8 (three times). We selected them from two different clusters.

The pairwise distance matrix for the selected samples is displayed in figure 4. It provides a comprehensive overview of the distances between individual samples within the dataset. Each cell in the matrix represents the Euclidean distance between a pair of selected samples, with darker colors indicating greater distances.

Regions of the matrix with lighter colors indicate samples that are closer together in the high-dimensional feature space. Conversely, darker regions signify samples that are farther apart.

In our case, because we selected 5 samples or digits from our 2-D t-SNE embedded

digits scatter plot (with indices 1658 (digit 9), 1582 (digit 9), 794 (digit 8), 1553 (digit 8), 129 (digit 8)), in this matrix we see 5 indices (from 0 to 4) on the axes as well, that were assigned to our digits by the matrix, 0 representing our 1658, 1 representing 1582, 2 representing 794, 3 representing 1553 and 4 representing 129. Clusters or groups of samples with similar characteristics are manifested as blocks of darker colors, indicating their proximity to one another. Conversely, outliers or samples belonging to distinct groups are represented by lighter regions, indicating greater dissimilarity. Higher values (lighter colors) represent larger distances and lower values (darker colors) represent smaller distances. That's why on the diagonal line, where each cell corresponds to the distance between a sample and itself, we see the darkest blocks. Since the distance between an object and itself is always zero, the values on the diagonal are all zeros. The distance between samples 0 and 1 is represented as light green, suggesting that these digits were farther apart in the scatter plot. The distance between 3 and 4 is represented as dark green, indicating that these outliers were closer to each other and more similar. The distance between 0 and 2 is displayed as light yellow, which shows the dissimilarity and larger distance between these samples.

6 Contributions

Sujan Darai: Running and implementing subtask 0,1,2 and writing Data & Preprocessing

Matanat Mammadli: Running and implementing subtask 3 and writing Methods and Results & Discussion

Samra Hamidovic: Running subtasks 0,1,2,3 and writing Abstract, Introduction, Goal of the project and Methods

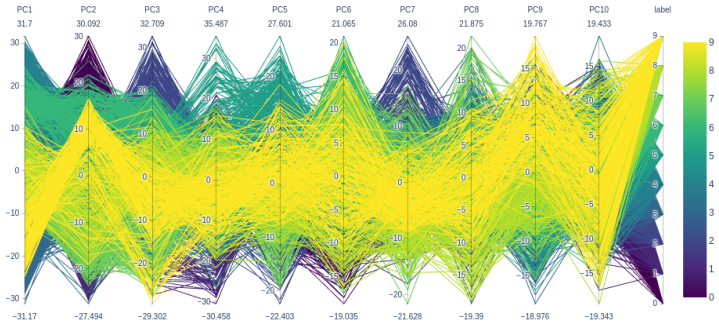
7 Appendix

(a) breast cancer data (b) digits data

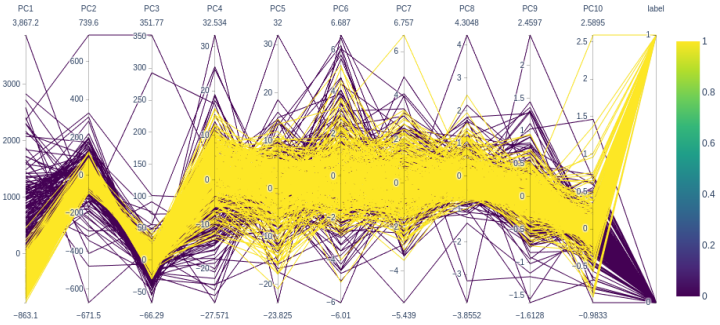
Figure 1: t-SNE scatter plots

(a) breast cancer data (b) digits data

Figure 2: UMAP scatter plots



(a) digits data



(b) breast cancer data

Figure 3: Parallel coordinates plot (10D PCA transformed)

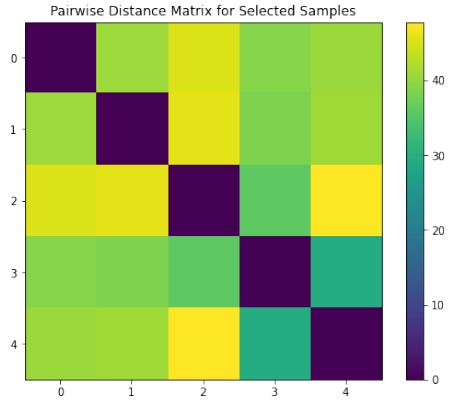
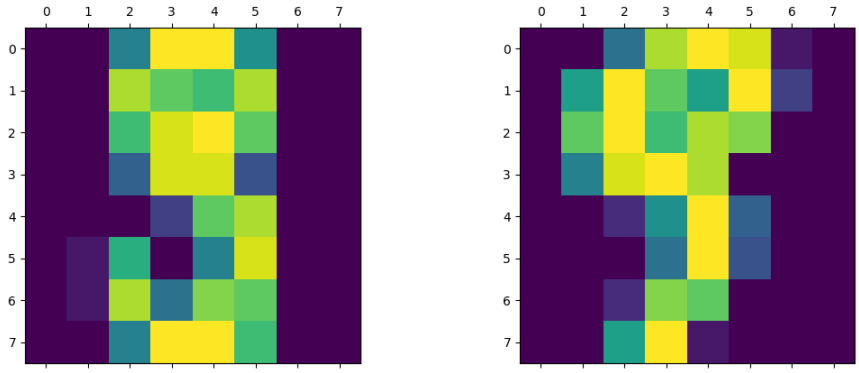


Figure 4: Pairwise Distance Matrix for Selected Samples



(a) digit with index 1658 (b) digit with index 1582

Figure 5: Images of the digits from the selected outliers

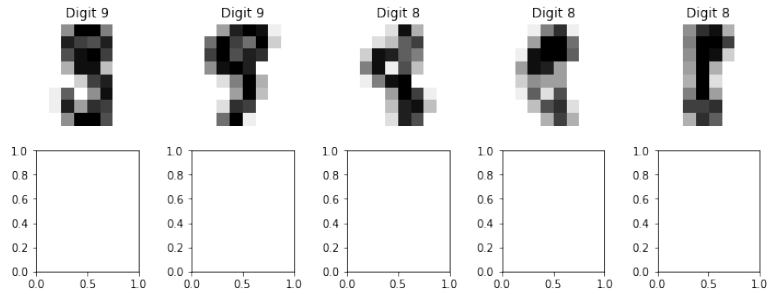


Figure 6: Images of all of the selected digits (outliers) from 2D t-SNE scatter plot of the digits dataset

References

1. Cleveland Becker. Brushing Scatterplots, September 1987. Publisher: Taylor Francis Group.
2. Rasmus Bro and Age K. Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014. Publisher: Royal Society of Chemistry.
3. Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, February 2021.