Measuring embeddings distortion using KNN graphs

Mykhailo Sakevych
Johannes Kepler University

ABSTRACT

Numerous dimensionality reduction methods have been developed to facilitate visual data exploration. These methods enable data analysts to represent multidimensional data in a 2D or 3D space while retaining as much relevant information as possible. However, it is not possible for these tools to preserve all structures simultaneously, and they inevitably introduce some distortions. As a result, various criteria have been introduced to evaluate the overall quality of a map, primarily based on the preservation of neighborhoods. These global indicators are currently used to compare different maps and identify the most suitable mapping method and its hyperparameters. However, relying solely on aggregated indicators can obscure the local distribution of distortions. In this paper we present an alternative way to quantify and visualize a local space distortion introduced by a map.

1 INTRODUCTION

Dimensionality reduction is a technique used to reduce the number of variables or features in a dataset. It aims to preserve the most important information in the original dataset while reducing the noise and redundancy in the data [8]. There are several popular dimensionality reduction algorithms, such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP).

Distortion of Dimensionality Reduction Algorithm (DRA) can be measured in many different ways. Using a k-nearest neighbor graph to represent high-dimensional data and the reduced-dimensional data after applying DRA can be useful for efficient computation, capturing the local structure of the data, providing a visual representation of the data, and evaluating the performance of DRA in preserving the local structure of the data [3].

There are many graph-based metrics you can use to evaluate the distortion of DRA, such as:

- Graph Reconstruction Error: This measures the difference between the original high-dimensional graph and the graph constructed from the reduced-dimensional data.
- Neighborhood Preservation: This measures how well the neighbors of a vertex in the high-dimensional graph are preserved in the reduced-dimensional graph.
- Dimensional Ranking: This measures the degree to which distances between points are preserved in the reduced-dimensional space.
- Visualization Quality: This measures the degree to which the reduced-dimensional space can be visualized in a way that captures the structure of the high-dimensional data.

The method of computing the shortest paths between every pair of vertices in both the high-dimensional and reduced-dimensional graphs, using the Floyd-Warshall algorithm, is a useful way of evaluating the distortion of a DRA. This approach provides a quantitative measure of the difference between the two spaces and helps identify regions where the DRA may be struggling to preserve the local structure of the data. By computing the distortion quotient for each vertex and averaging it across all edges in the graph, this method can provide an overall measure of the distortion in the reduced-dimensional space, which can help in comparing different DRAs or choosing the most suitable DRA for analysis.

1.1 Notations:

Δ_{ij} is the Minkowski distance in the data space between two points X_i and X_j, and P_{ij} is the rank of X_j in the neighbourhood of X_i, meaning that X_j is the P_{ij}^{th} nearest neighbour of X_i, with $P_{ii} = 0$ by convention. Conversely to distances, ranks are not necessarily symmetric (P_{ij} may be different from P_{ij}). Distances δ_{ij} and ranks ρ_{ij} may be defined equivalently in the embedding space.

In the scope of this article, the neighbourhood $n_k(i)$ of the i-th point in a given space is defined as the set of its k nearest neighbours. To be more precise, graph $G = (V, E)$ is a k-neighbors graph on the set of points X iff:

- Every point in the space X has corresponding vertex in the graph G. I.e. $V \leftrightarrow X$.
- v_i and v_j are connected iff $\rho_{ij} \leq k$.

2 RELATED WORK

Currently, various indicators are employed to evaluate the quality of a map. Typically, these indicators involve comparing distances Δ_{ij} and D_{ij}, which are utilized in stress functions of numerous dimensionality reduction techniques , as well as ranks ρ_{ij} and r_{ij}, which are utilized in several rank-based quality criteria [4]. Ranks are often preferred due to their resilience to norm concentration in high-dimensional spaces.

3 METHOD

Assume we have dataset $\mathbb{X}$ of size $(n \times m)$, where n is number of data points and m is number of features.

Let's denote G_h as a k-neighbors graph built on the original dataset $\mathbb{X}$ and G_l as a KNN-graph built on the reduced dimension dataset $DRA(\mathbb{X})$.

Additionally we pick k such that graphs G_h and G_l are connected.

Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

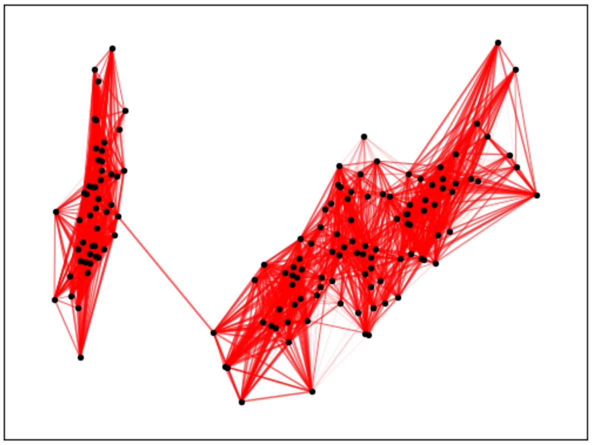

Figure 1: KNN graph G_l built after applying DRA

Finally we compute the shortest path between every pair of vertices in both G_h and G_l using the Floyd-Warshall algorithm [2].

3.1 Distortion quotient

Let the shortest path between vertices v_i and v_j in the higher/lower dimension be D_{ij} and d_{ij} correspondingly. Then we can compute the distortion quotient of vertex i as

$$\gamma i = \frac{1}{n}\sum_{j\neq i}\frac{D_{ij}}{d_{ij}}$$

3.2 Overview of resulting graphs

In Figure 1, we can see G_l. Nodes are located on 2D plane using $DRA(X)$ as coordinates. The width of an edge represents $L2$ distance between two points.

4 EXPERIMENTS

4.1 Considered datasets

To illustrate our method, we consider two datasets. The first consists of 50 samples of 3 different species of iris with their measurements of sepal length, sepal width, petal length, petal width [1, 5]. The second is the dataset, which consists of 318 daily reports on 15 different types of losses(headcount, weapons, etc) of Russian army during the war in Ukraine [6].

4.2 Preprocessing

Upon analyzing the Russian loss dataset, two distinct outliers emerge - namely the second day and the 67th day of the war.

The second day appears to be an outlier due to the observed number of losses of Armored Personnel Carriers (APCs), which is either the cumulative result of the unrecorded losses on the first and second days or can be attributed to the most productive day for the Ukrainian military.

Similarly, on the 67th day, the total count of losses for both vehicles and fuel tanks is seemingly the sum of all previously recorded as well as unrecorded losses.

As result of analysis, we will remove those points on postanalysis.

There was no preprocessing for Iris dataset.

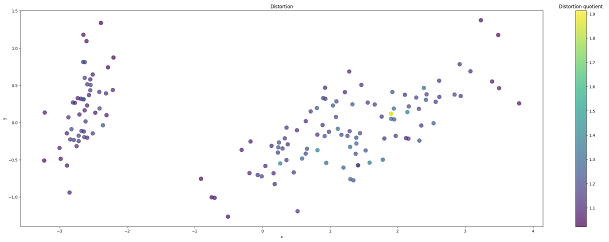

Figure 2: Scatterplot of applying PCA on the Iris dataset

4.3 Visualization approach

In this section we take a closer look on visualisation approach. All plots are generated using our implementation of the algorithm. code (higher resolution images and interactive plots are available here).

In Figure 2, we can see a set of points, these are result of PCA projection of Iris dataset. The x and y coordinates of the dots represent the reduced dimensions of the original dataset. The dots are colored based on a distortion quotient.

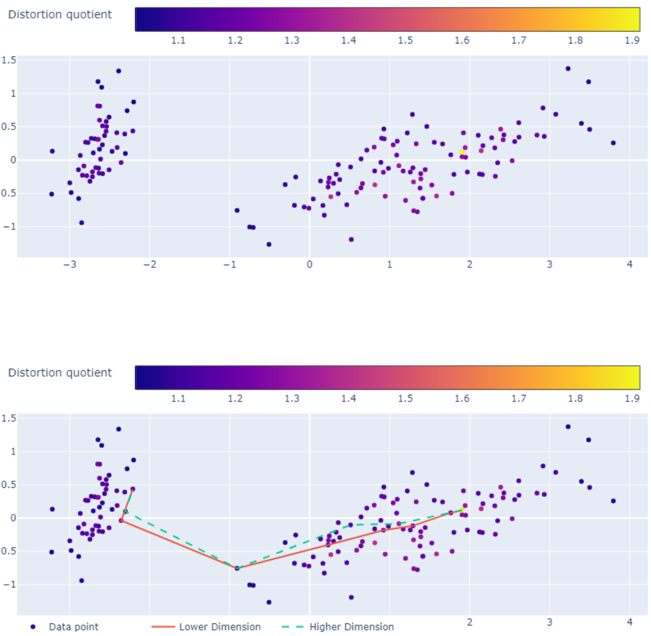

Figure 3: Example of interaction with path building interface

In Figure 3 is the interface of interactive plot. It allows you to click on two points consequently to build shortest path in graph between them in higher and lower dimensional spaces.

Figure 4 displays a cluster of data points, which correspond to the outcome of Singular Value Decomposition (SVD) projection applied to the Russian loss dataset.

4.4 Interpretation guidelines

Highest distortion quotient in Figure 2 we can spot in 148 entry. Analysing distances distortion we can see that shortest path from

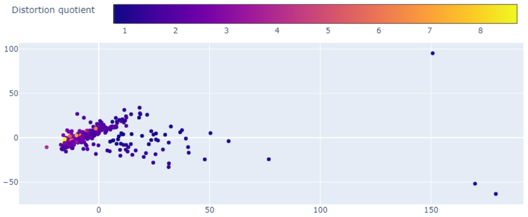

Figure 4: Russian Losses DRA distortion (SVD method)

point 148 to point 115 in higher dimension equals to 0.3 and in lower to 0.004 and in all other points there is no such significant difference.

Analysing Figure 4 we can state that days with the highest distortion quotients are 274th, 286th, and 310th. They seem to have the number of cruise missiles as the sum of previously unrecorded cruise missiles and cruise.

5 RESULTS

We presented a useful tools for the visualization and analysis of space distortion created by DRAs. Here are few example applications:

- Spotting severe distortions at a glance and analyzing their causes.
- A comparison of the DRA distortions on the dataset can result in a better understanding of the dataset and the suitability of DRAs.(Figure 5).
- Filtering dataset by distortion quotient.

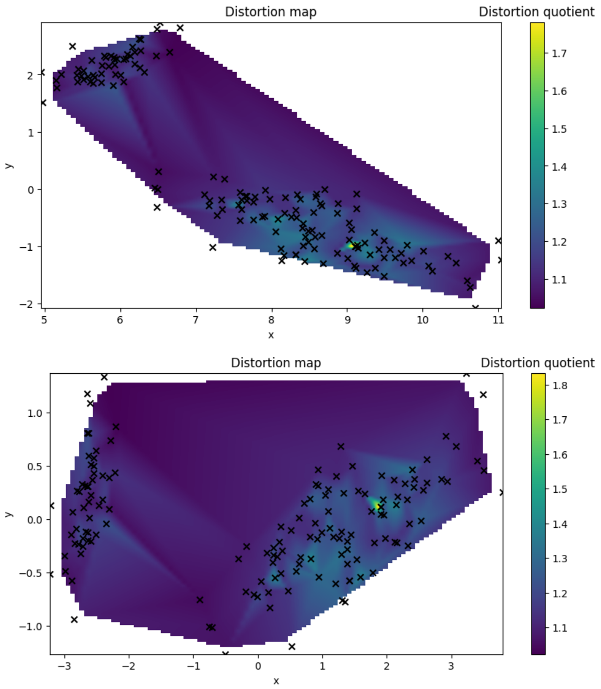

Figure 5: PCA(top) and SVD(bottom) applied to the Iris dataset

6 DISCUSSION AND FUTURE WORK

Distortion analysis is a valuable tool for assessing the strengths and weaknesses of DRAs and gaining insights into datasets. To develop such a tool, we can implement various features, such as a built-in labeling algorithm and dynamic box or lasso selection. However, there are certain problems that we need to overcome to make the tool more efficient.

One significant issue is that the algorithm's total complexity is mainly dominated by the Floyd-Warshall algorithm, which has a complexity of $O(N^3)$ [7]. Unfortunately, Python is relatively slow, and even with only 600 datapoints, it takes 5 minutes to process the dataset. To solve this problem, we can consider using some approximation instead of the exact shortest distances calculated by Floyd-Warshall, as this can significantly reduce the computational complexity and make the analysis more efficient.

Another problem is that the dynamic reconstruction of the graph after adding or removing some points can be extremely time-consuming. To address this issue, we can predict the distortion for new points and base the reconstruction process on the distortion coefficient after deleting a point. This can reduce the computational complexity and make the analysis more scalable and efficient.

In conclusion, by addressing the issues mentioned above and implementing the possible solutions, we can enhance the distortion analysis tool's performance and efficiency, making it more useful for analyzing and gaining insights into datasets.

7 CONCLUSION

Dimensionality reduction algorithms are powerful tools for data exploration and visualization, but they introduce distortions in the reduced-dimensional space. Various criteria have been developed to evaluate the quality of a map, primarily based on the preservation of neighborhoods, dimensional ranking, visualization quality, and graph reconstruction error. However, relying solely on global indicators can obscure the local distribution of distortions. Therefore, it is important to perform a local evaluation to ensure an accurate interpretation of the maps. The method of computing shortest paths using the Floyd-Warshall algorithm is a useful way of evaluating the distortion of a DRA, as it provides a quantitative measure of the difference between the two spaces and helps identify regions where the DRA may be struggling to preserve the structure of the data. Ultimately, the choice of the most suitable DRA depends on the specific characteristics of the data and the research questions at hand.

REFERENCES

[1] Edgar Anderson. 1936. The species problem in Iris. *Annals of the Missouri Botanical Garden* 23, 3 (1936), 457–509.

[2] Chandler Burfield. 2013. Floyd-warshall algorithm. *Massachusetts Institute of Technology* (2013).

[3] Benoît Colange, Laurent Vuillon, Sylvain Lespinats, and Denys Dutykh. 2019. Interpreting distortions in dimensionality reduction by superimposing neighbourhood graphs. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 211–215.

[4] Benoît Colange, Laurent Vuillon, Sylvain Lespinats, and Denys Dutykh. 2019. Interpreting Distortions in Dimensionality Reduction by Superimposing Neighbourhood Graphs. In *2019 IEEE Visualization Conference (VIS)*. 211–215. https://doi.org/10.1109/VISUAL.2019.8933568

[5] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

[6] Petro Ivaniuk. 2022. 2022 Ukraine Russia War. (2022). https://doi.org/10.34740/KAGGLE/DS/1967621

[7] Antonio E Mirino et al. 2017. Best routes selection using Dijkstra and Floyd-Warshall algorithm. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. IEEE, 155–158.

[8] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J. Appl. Sci. Technol. Trends* 1, 2 (2020), 56–70.