Suyog Raj Joshi
CS-521
HW1
09/20/2024

Comparison of Haar Wavelet and PCA: Speed And Count Comparison

In this experiment, I compared the performance of three different transformations and their corresponding distance matrices:

I computed Euclidean distance, Wavelet Transform(Haar) and Principal Component Analysis(PCA) and considered Euclidean distance to be the ground truth. For the Haar Wavelet transform I used the first four coefficients and for the PCA I used the best four principal components.

The comparison is mainly focused on time taken to compute the distance matrices for each transformation and the percentage of times the relationship (greater than) between distances.

Methods

- I used google colab to perform this experiment.
- The given dataset was normalized using the z-normalize.
- The Euclidean distance was then calculated using the normalized data.
- A haar wavelet is applied and the first four wavelet coefficients were selected for each record. A distance matrix is computed using the Euclidean distance between the first four Haar coefficients of each record.
- PCA is applied and the best four principal components were selected. A distance matrix is computed using the Euclidean distance between the four principal components of each record.

Speed Comparison

The time taken to calculate each distance is measured and compared

Transformation	Time Taken (approx. in minutes)
Euclidean Distance	12
Haar Wavelet	11
PCA	16

Speed:

- **Euclidean Distance (12 minutes):** This is considered the ground truth and involves computing the pairwise Euclidean distances between records. Although straightforward, it involves calculating distances across 16,000 records in a 128-dimensional space, which is computationally expensive due to the number of records and dimensions.

- **Haar Wavelet (11 minutes):** The Haar wavelet transformation reduces the data by extracting only the first four coefficients for each record. This reduces the dimensionality, which in turn decreases the time needed to calculate the distance matrix compared to the full 128-dimensional space. This maybe the reason why Haar wavelet was slightly faster than Euclidean distance.

- **PCA (16 minutes):** PCA was slower than Haar wavelet and Euclidean distance. PCA involves more complex mathematical operations (e.g., eigenvalue decomposition) to select the four best principal components. Once the components are selected, the distance matrix calculation is faster due to the reduced dimensionality.

Summary of Speed:

- Haar Wavelet was the fastest in this experiment maybe because of its simpler transformation and dimensionality reduction.

- PCA was slower maybe due to the eigenvalue calculations.

Count Comparison

Transformation	Relationship Matches (%)
Haar Wavelet	60.45 %
PCA	55.30 %

Count Comparison:

- **Haar Wavelet (60.45%)**: The Haar wavelet matrix preserves about **60.45%** of the relationships compared to the Euclidean distance matrix. This means that for around 60.45% of the pairs, the distance relationships (whether one distance is greater than another) are consistent between the Haar and Euclidean matrices.

- **PCA (55.30%)**: PCA preserves **55.30%** of the relationships, which is worse than Haar. PCA focuses on maximizing the variance explained by the principal components, which helps in preserving more of the important relationships between data points compared to Haar.

Summary of Relationship Consistency:

- **HAAR Wavelet** performs better than PCA in terms of relationship consistency with the Euclidean matrix. This is because PCA selects the components that explain the most variance, making it more likely to maintain the structure of the data.

- **PCA** is slightly worse but still maintains a significant percentage of the relationships, making it a decent choice when we need computation at the cost of some loss in accuracy.

Conclusion:

- **Speed**: Haar wavelet is the fastest transformation method, followed by the Euclidean distance and PCA.

- **Relationship Consistency**: HAAR Wavelet retains more of the original relationships between distances compared to PCA, making it a better choice for accuracy, and is better for faster computation.