CSE 519 PROJECT PROGRESS REPORT

Evaluating Dimensionality Reduction Methods

Introduction

As the data is increasing, it becomes more difficult to process the real time data as it is incomplete, inconsistent and unorganized. The main challenge is to find the transformation to intrinsic dimension with minimal loss while keeping significance intact. A Dimensionality Reduction technique is the process of reducing the number of variables of a high-dimensional dataset, which on reduction preserve the majority of information from the original dataset. These dimensionality reduction techniques are applied in latent knowledge discovery, feature analysis, machine learning algorithms and visualization. Dimensionality reduction techniques can sufficiently bring down the higher dimensional data to lower dimensional representations without significant loss of latent information. Different comparative studies comparing different dimensionality reduction methods are currently being addressed in the research field.

In this paper, we discuss how we can compare different dimensionality reduction algorithms based on the features like loss of quality and quantitative measures that will help us understand the application and working of each of those algorithms. We have considered the following dimensionality reduction techniques: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbor Embedding(t-SNE), kernel-PCA, IVIS, Uniform Manifold Approximation and Projection(UMAP), and Trimap. Some of the features that we have used to study the above-mentioned techniques include Mapping higher dimensional data into a 2d Mapping grid for visualization, Mantel test, Qualitative measures and quantitative analysis.

New Reduction Techniques Added

UMAP:

Umap is a novel manifold learning technique for dimensionality reduction. Umap is based on Riemannian geometry and algebraic topology. UMAP can be thought of a two-step process. In the first step, a particular weighted k-neighbor graph is constructed. In the second phase, a low-dimensional layout of the graph is computed. Umap performs better at preserving global aspects of the high dimensional data as compared to t-SNE. It uses binary-cross entropy as a cost function instead of the KL divergence used in t-SNE. This leads to a huge change in the global data preservation. UMAP does not use normalization for high dimensional as well as low dimensional probabilities.

IVIS:

Ivis is a machine learning algorithm for reducing dimensionality of very large datasets. It preserves global data structures in a low-dimensional space, adds new data points to existing embeddings using a parametric mapping function, and scales linearly to millions of observations.

TRIMAP:

TriMap is a dimensionality reduction technique based on triplet constraints that preserves the global accuracy of the data. The main idea behind TriMap is to capture higher orders of structure with triplet information (instead of pairwise information used by t-SNE and LargeVis), and to minimize a robust loss function for satisfying the chosen triplets. Tri-Map is fast and provides comparable runtime on large datasets

Dataset

1. **MNIST**
   The MNIST dataset of database of handwritten digits, published by *LeCun*, which consists of a training set of labelled 60,000 examples with 10,000 test data. Each of the image in the dataset is of shape of 28x28 and grayscale color space with 10 labels representing digits from 0-9.
   Example:

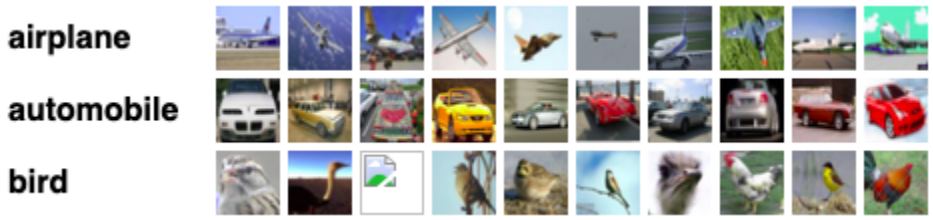
2. **Fashion-MNIST**
   The Fashion-MNIST is a dataset consisting of 60,000 training & 10,000 test examples published by *Zalando Research*. Each of the image in the dataset is a 28x28 grayscale image with associated label from 10 classes.
   Example:

3. **COIL-20**
   Columbia **O**bject **I**mage **L**ibrary is a database of 1440 images, which contains grayscale images of 20 objects, rotated in 360 degrees presenting different poses.
   Example:

4. **CIFAR-10**

   A dataset consisting of 60000 32*32 color images of 10 different classes.
   There are 10 categories present of everyday objects like airplanes, dogs etc.

Examples:



5. **Breast-Cancer Dataset**

   A dataset which contains Wisconsin (Diagnostic) Data where features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. These features describe characteristics of the cell nuclei present in the image

Quantitative Analysis

There are many different quality assessment measures for evaluating the performance of the DR algorithms. Most of the measures are used to evaluate the local-neighborhood-preservation or the overall-structure-holding performance(global) of the DR methods. Some of the local and global approaches are listed in the table.

Distance Preservation

Historically Distance preservation has been the first and most important criterion used to achieve a DR in a non-linear way. In an ideal scenario, conservation of pairwise distances between different vectors of a dataset ensures that the low dimensional embedding inherits the main geometric properties of the data, such as the overall shape. However, in case of non-linear DR methods, distances are not perfectly preserved. In short, different DR algorithms preserve different types of distances when it comes to dimensionality reduction. These can be pairwise distances, local structural distances and global structures distances etc. E.g. Of these are Kernel PCA, Isomap etc.

Topology Preservation

In the case of topology, the overall shape of the manifold is preserved ie the structural properties of the dataset. These techniques, that can preserve topological characteristics of the input data are also known as local preservation approach. A lattice is defined as a discrete representation of the topology. Most of the DR methods fall into two categories, one in which uses a predefined lattice, e.g. Self-Organizing Maps(SOM's) and Generative Topographic mapping, and the other one uses a data-driven lattice. The data-driven lattice allows the DR method to change the shape of the lattice or entirely build the lattice while the models are running. E.g of these are Locally Linear Embedding and Laplacian eigenmaps etc.

YEAR	NAME OF THE MEASURE	CRITERION
1964	Kruskal Stress Measure(S)	Global
1988	Spearman's Rho(Sb)	Local
2003	Classification error rate	Classification error
2007	Mean Relative Rank Errors(MRRE)	Local
2009	Co-ranking Matrix(Q)	Local
2011	The Relative Error(Re)	Global

TABLE1: Methods for evaluating the quality of DR algorithms

The degree of correspondence between the distances among points implied by MDS and the matrix input by the user is measured by a **stress** function called "**Kruskal Stress**".

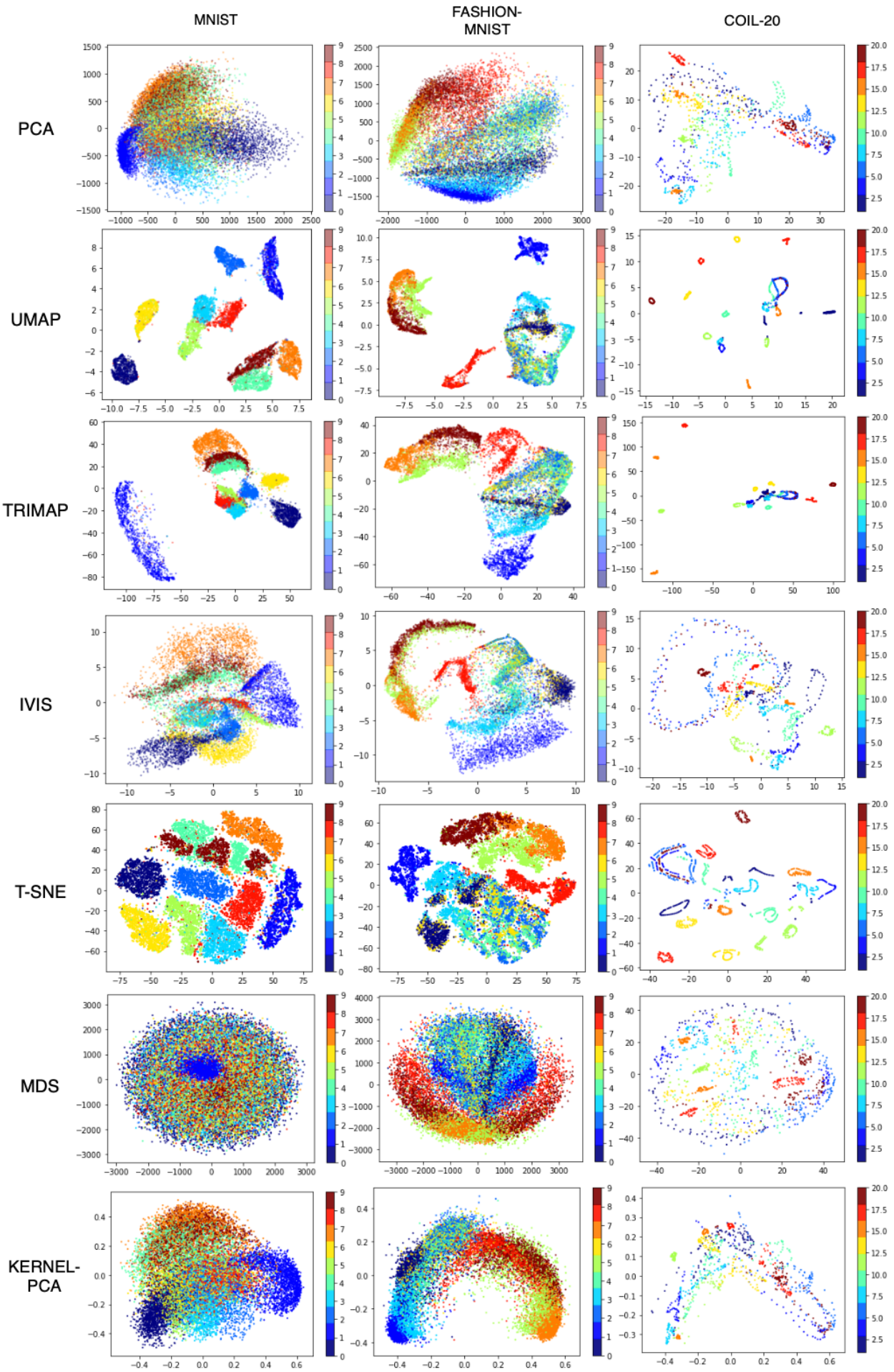
Spearman's Rho Siegel and Castellan presented one of the first measures to estimate the topology preservation (TP). This measure estimates the correlation of rank order data. That is, it tries to assess how well the corresponding projection preserves the order of pairwise distances between data-points in high-dimensional space.

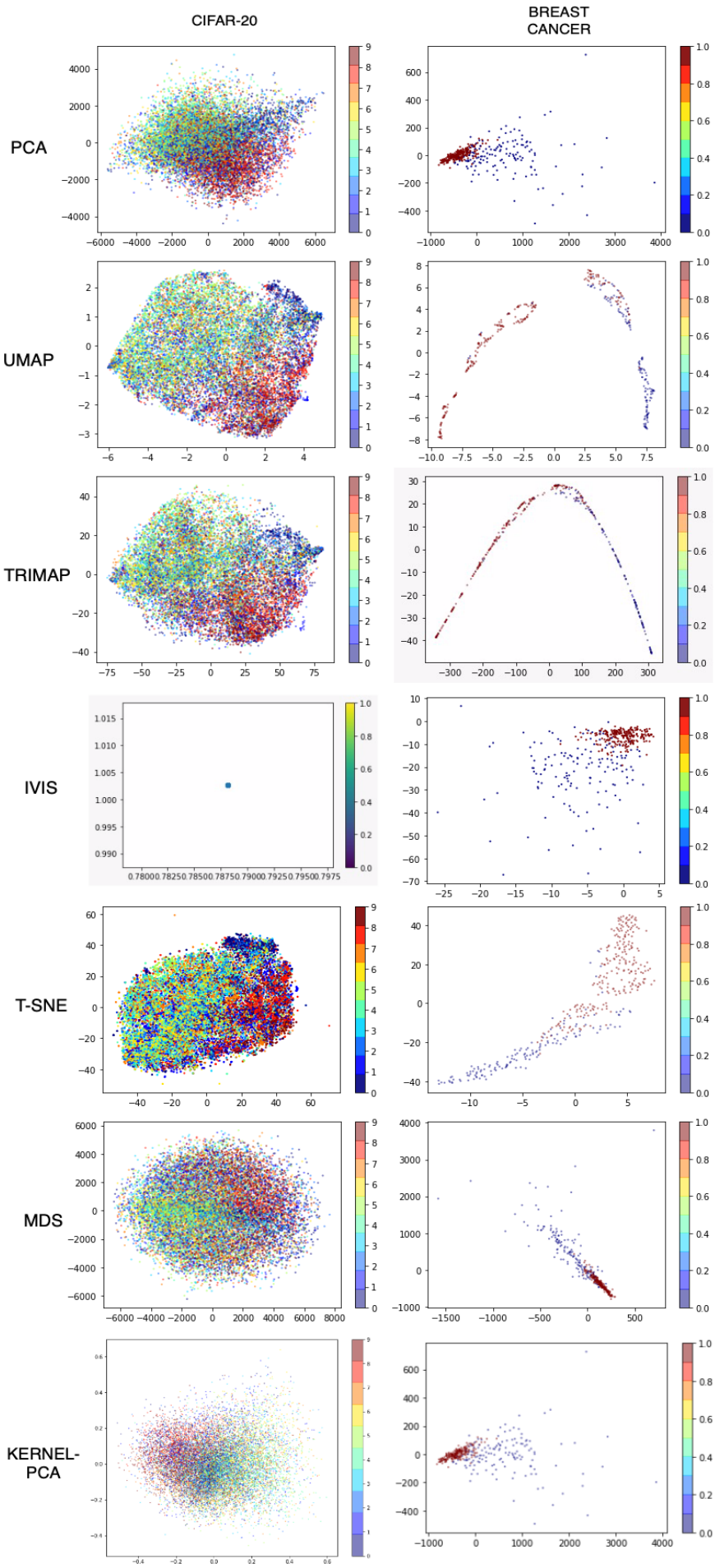
Mean Relative Rank Errors Lee and Verleysen developed a quality assessment measure, the mean relative rank errors (**MRRE**). It is based on ranks of pairwise Euclidean distances within local neighborhoods.

Many different concepts and quality criteria for DR can be summarized using the **Co-ranking matrix** framework (Q), Several of the aforementioned methods (based on distance ranking in local neighborhoods like MRRE), are easily unified into an overall framework.

Visualization (2D plots)

We have enumerated the 2D representation of the reduction embedding space as tabular chart to give a better overview of performance of the reduction technique.


MNIST
FASHION-MNIST
COIL-20
PCA
UMAP
TRIMAP
IVIS
T-SNE
MDS
KERNEL-PCA

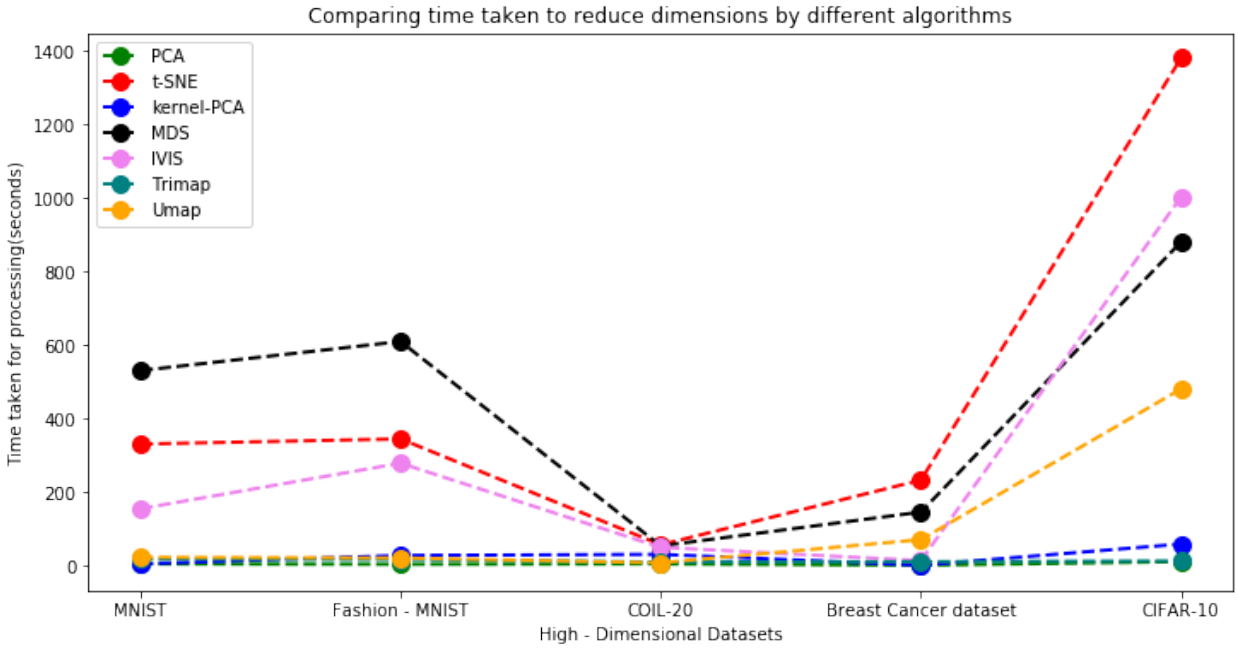

CIFAR-20
BREAST CANCER
PCA
UMAP
TRIMAP
IVIS
T-SNE
MDS
KERNEL-PCA

Quantitative Analysis

According to the Nature Paper, to analyze qualitatively, Author have formulated various methodologies to formalize the qualitative analysis. Following the same path, we also used similar methods. **First**, on the computational aspects of the running time of the reduction algorithms, we benchmarked all of our five datasets by time taken to complete the data reduction. **Second**, to assess the ability of each reduction algorithm to separate the datasets into defined 2D reduced embeddings, we trained a Random Forest classifier on the reduced dataset. **Third**, we analyzed each of seven dimensionality reduction algorithms on the basis of how each algorithm really have the capability to preserve the global structure by comparison of distances between all pairwise distances of high-dimension and reduced dimension with the help of Mantel test.

Let's dive into more detailed analysis below.

Comparing time taken by different algorithms

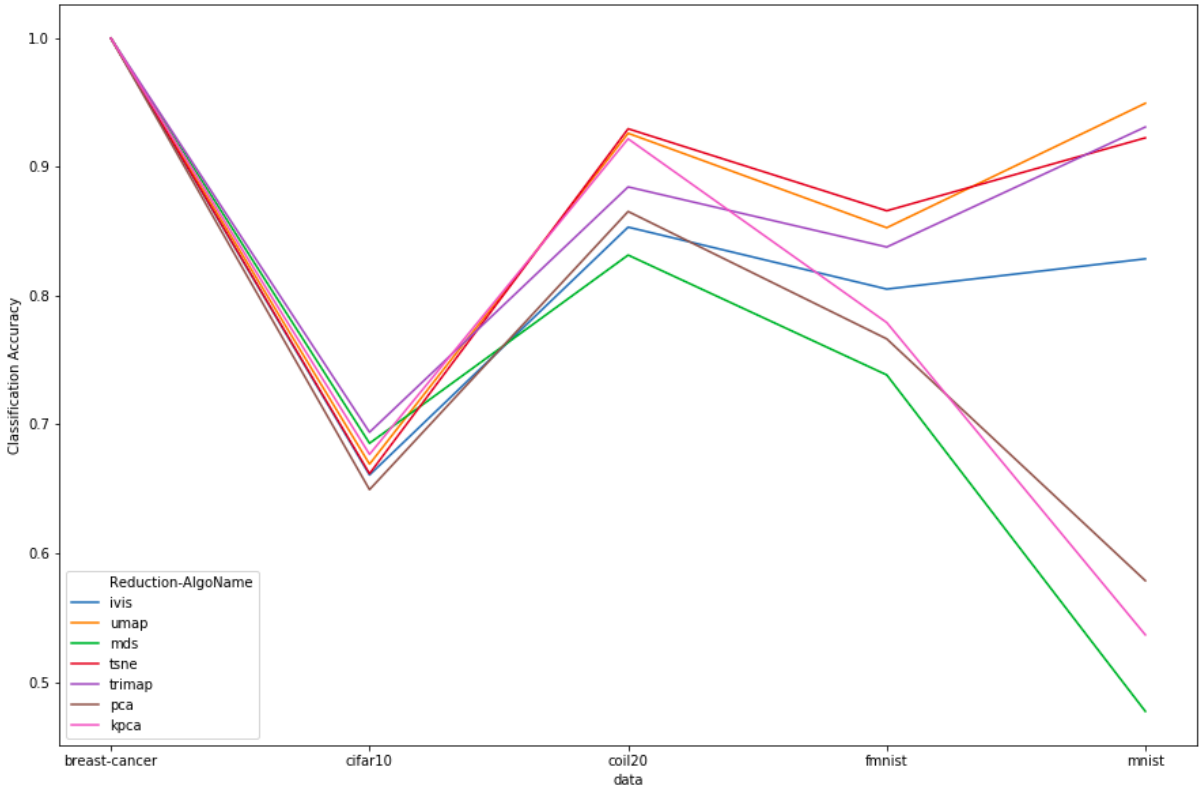


Observations:

The above plot gives a qualitative comparison on the time taken by different methods to fit a high-dimensional dataset. We applied 7 DR methods on the available datasets and compared the execution time for each method versus the others. As we can see clearly, MDS is usually slower than other algorithms. Even t-SNE is generally slower compared to other algorithms. This maybe because it does not have the scaling properties. Also, one of the fastest algorithms is the PCA, it usually takes the least amount of time to execute. Algorithms like Trimap, IVIS and kernel-PCA take average time in terms of execution. To gain a deeper insight and get more accurate comparison results, we will need to go for large datasets and then re-compare. Another noticeable algorithm is UMAP. Being recently in use, it is still somewhat slower than the PCA, but is comparatively much better to the t-distributed SNE algorithm.

Conclusion: By far, PCA proved to be the fastest algorithm when compared to the other 6 methods. While not competitive when compared to PCA, umap is the next best option to go for. It works very efficiently and given the quality of results umap gives, we think umap is a better option for dimensionality reduction.

Classification Accuracy



Methodology

Almost all the methods lead to almost 100% accuracy in the breast-cancer dataset. One of the reasons, we think this happened is because the dataset size was very small which is why all the methods achieved near-perfect accuracy. COIL-20 dataset: We can see the methods achieving accuracy score in the range 80-94%. MDS didn't return great results and could only achieve around 80% accuracy. t-SNE and UMAP were the best performers with both achieving around 93% accuracy. For the CIFAR-10 dataset, we see an unusual trend with all the methods failing to achieve a score greater than 75%. This might suggest some more in-depth analysis of the dataset, algorithms and data pre-processing. We also faced this anomaly for the MNIST dataset with PCA, kernel-PCA and MDS. MDS, probably, was the least expected to under-perform. It might have not been tuned correctly with the parameters, input pre-processed data due to which it achieved a score less than 50%. In the F-MNIST dataset, the different methods achieved an accuracy score in the range of 75% - 90%. t-SNE and UMAP were expected to perform well given the quality of results and achieved a score close to 90%. One thing to note was the performance of t-SNE and UMAP. We see that these two modern approaches out-perform most of the traditional methodologies. To get a better comparison between the two, we need to apply the two algorithms on larger datasets, use more quality features for dimensionality

reduction, data pre-processing and study different parameters of both the algorithms that will help us rank the two.

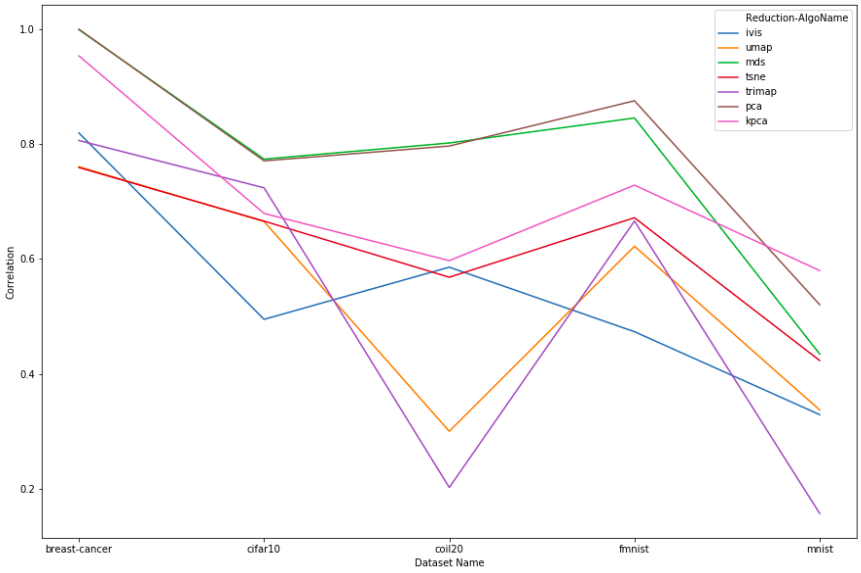
Mantel Test

To establish how well a reduction algorithm, preserve data structure in low-dimensional space, a Euclidean distance based pairwise distance matrix was created for all of the seven datasets from their original embedding space as well as reduced embeddings. The level of correlation between the original distance matrix and the distance matrices in the embedding spaces was then assessed using the **Mantel test**. The Mantel Test is the *Pearson's product-moment correlation coefficient* was used to quantitate concordance between original data and low-dimensional representations. We have used all the dataset that was used in the reduction technique which portrays the correct picture about good as well as bad examples of distance preservations.

The modified version of the Mantel test implementation takes two distance matrices (pairwise distances) and returns: correlation, the empirical p-value, and a standard score (z-score). We have chosen **10000** permutations and pearson correlation method to compute the correlations by default. For some of the datasets, due to lack of good computing resources, we reduced the number of permutations to 1000 (MNIST and Fashion-MNIST). Additionally, since many of the nearest neighbor algorithm suffer from poor indexing and slow retrieval, we have chosen the famous **Annoy** library from Spotify. **Annoy** creates large read-only file-based data structures that are mapped into memory so that many processes may share the same data.

Steps to complete Mantel Test:

1. Set the Annoy library to compute pairwise distances of input data
2. Compute the pairwise distance of original embedding space. For given n rows, it generated (n x n) pairwise distance matrix.
3. After reducing the embedding space to 2D, we calculate the pairwise distance of reduced dimension space, we again get the same (n x n) pairwise matrix back.
4. We pass these two obtained matrices to Mantel Test with 10000 permutations with upon performing two tail statistical analysis; which returns a correlation value, an acceptable z-score (because of large permutations test, it always comes less than 0.05).



Mantel Test Correlation Comparison

Key Takeaway Points from Mantel test:

1. It is interesting to see the low correlation in coil-20 and MNIST dataset, but at the same time it gives good accuracy with Fashion-MNIST dataset.
2. MDS is able to keep the high correlation but many of SOP (self-organizing maps) like UMAP and tri-map which gave good accuracy scores but they fail to keep the distance preserved in lower dimension.

Next Steps

1. In this paper, we covered the computation part of different methods to get to known(expected) results. We applied various DR techniques to the datasets and compared the complexities, loss of quality and other qualitative measures. Furthermore, we plan to get to the fundamentals of the DR methods and understand the reasoning for various behaviors exhibited by the methods.
2. Produce analysis for different types of datasets (categorical, count, numerical) in supervised and unsupervised setting.
3. A more detailed comparison between t-SNE and UMAP, since these two are the most talked dimensionality reduction techniques.
4. Explore more in the local, global structure preservation space to get better information
5. In quantitative analysis, we have seen that many of self-organizing maps are failing in Mantel test and we believe that they require more in-depth analysis.
6. We are planning to publish a pip package to showcase our analysis at the end.

References

1. Python Implementation of Mantel test: https://github.com/jwcarr/MantelTest
2. https://en.wikipedia.org/wiki/Mantel_test
3. https://core.ac.uk/download/pdf/148668147.pdf
4. https://bering-ivis.readthedocs.io/en/latest/embeddings_benchmarks.html
5. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144059
6. https://joncarr.net/s/a-guide-to-the-mantel-test-for-linguists.html
7. http://papers.nips.cc/paper/2141-global-versus-local-methods-in-nonlinear-dimensionality-reduction.pdf
8. https://ieeexplore.ieee.org/abstract/document/1316859
9. https://ieeexplore.ieee.org/abstract/document/6032745
10. https://github.com/spotify/annoy
11. https://arxiv.org/pdf/1910.00204.pdf
12. https://www.nature.com/articles/s41598-019-45301-0