MATH5473/CSIC5011 Final project: Dimension Reduction Techniques for MNIST Clustering

Wai Ming Chau wmchau@connect.ust.hk

Department of Mathematics, HKUST

1. Introduction

Recent research has highlighted the challenges associated with the curse of high dimensionality, especially in the context of the clustering problem. In addition to the significant computational burden, data becomes sparse in high dimensional space, and traditional techniques for indexing and algorithms are often inefficient and ineffective as the concepts of proximity, distance, or nearest neighbour may not even be qualitatively meaningful, exacerbating the challenges associated with the curse of high dimensionality.

Therefore, dimensionality reduction techniques are important in converting high-dimensional data into lower-dimensional data by extracting meaningful features. This simplification helps clustering algorithms identify patterns and groupings by measuring the similarity between data points based on their features.

In this project, we will convert high-dimensional data into two-dimensional data and apply the K-means clustering algorithm to obtain clusters from the reduced-dimension data. Next, we conduct an analysis to determine the effectiveness of various dimension reduction techniques and their impact on the performance of the K-means clustering algorithm.

2. Data

In this project, we are working with the MNIST dataset, consisting of 60,000 examples of handwritten digits ranging from 0 to 9. Each digit is represented by a 28 × 28 matrix with a unique handwriting style. Our intuition suggests that there should be 10 clusters present in the dataset, which we will aim to identify. We will randomly select 5000 examples from the dataset of 60000 and apply the algorithm to each subset. We will repeat this process 10 times.

3. Methodology

Dimension Reduction Techniques

We will apply the following dimension-reduction techniques to convert high-dimensional data into two-dimensional data:

1. **Principal component analysis (PCA)** identifies the underlying linear structure of the data with the highest variance.
2. **Linear Discriminant Analysis(LDA)** is a supervised technique that projects data onto a lower-dimensional space to maximize class separation based on linear discriminant functions.
3. **Isometric Feature Mapping (ISOMAP)** preserves the geodesic distances in high-dimensional space to capture intrinsic geometric structures in a low-dimensional representation.
4. **Locally Linear Embedding (LLE)** preserves local linear relationships between nearby points in high-dimensional space for a low-dimensional representation.
5. **Laplacian Eigenmaps (LAP)** preserves global relationships between all points in high-dimensional space for a low-dimensional representation using the graph Laplacian.
6. **t-Distributed Stochastic Neighbor Embedding (t-SNE)** models similarity between nearby points in high and low-dimensional spaces to capture local structure in a low-dimensional representation.

K-means clustering algorithm

K-means algorithm measures how similar or different things are using Euclidean distance. It groups things together based on how close they are to each other. It keeps adjusting the groups until they are as accurate as possible and don't change much.

3. Methodology (cont.)

Clustering Analysis

We will calculate the average of three metrics to evaluate the performance of clustering using various dimensionality reduction techniques in 10 numerical experiments.

1. **Purity** measures the homogeneity of the clusters with respect to the ground truth labels by counting the number of data points from the most common class in each cluster, then summing these counts over all clusters and dividing them by the total number of data points.

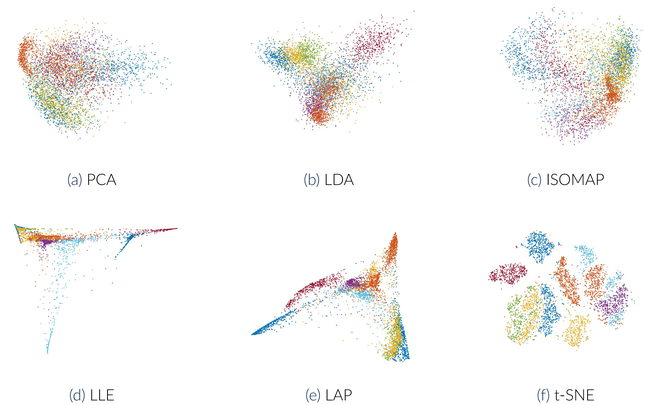
$$\frac{1}{N}\sum_{k}\max_{j}|\text{cluster}_k \cap \text{class}_j|$$

2. **Normalized mutual information** measures the similarity between the ground truth labels and the labels assigned by a clustering algorithm while considering their mutual information and normalizing it by the entropy of each label set.

$$\text{NMI}(\text{cluster}, \text{class}) = \frac{2I(\text{cluster}; \text{class})}{H(\text{cluster}) + H(\text{class})}$$

4. Results

Method	PCA	LDA	ISOMAP	LLE	LAP	t-SNE	NONE
Purity	0.4101	0.5247	0.4563	0.5873	0.5397	0.7458	0.5758
Normalized mutual information	0.3544	0.4652	0.3575	0.6030	0.5421	0.7038	0.4912



(a) PCA (b) LDA (c) ISOMAP

(d) LLE (e) LAP (f) t-SNE

Figure 1. Embedded coordinates with different dimension reduction techniques from one sampled data set

The table displays different clustering metrics computed using various dimension reduction techniques, as well as metrics computed using the full-dimensional data without any reduction. We also provide a visual representation of the clusters through the two-dimensional embedded data, facilitating their analysis.

The results suggest that t-SNE is the most effective clustering method, performing better than other reduction techniques and even full-dimensional clustering. LLE, LDA, and LAP show similar performance to full-dimensional clustering, while PCA and ISOMAP perform worse.

5. Analysis

Clustering algorithms rely on both **Between-Class Variability** (BCV), which measures the degree to which groups are spread apart from each other, and **Within-Class Variability** (WCV), which measures how tightly they are grouped. Ideally, we want the underlying data set to have high BCV and low WCV. However, based on the graph of embedded coordinates and clustering metrics, it appears that BCV has a more significant impact on the effectiveness of these techniques.

According to Figure 1, t-SNE is effective in separating different classes from each other, making it easy for the k-means algorithm to identify distinct classes. Conversely, PCA, LDA, and ISOMAP fail to separate the classes, leading to significant mixing of classes. LLE and LAP perform better than these methods in separating the classes to some extent, but their performance is still inferior to t-SNE. Although LLE and LAP have very low WCV, their relatively low BCV results in poorer clustering outcomes compared to t-SNE.

To support our visual observations of the embedded coordinates, we will compute the following quantity:

$$\text{BCV} = \sum_{i=1}^{k} n_i(\boldsymbol{m}_i - \boldsymbol{m})^T(\boldsymbol{m}_i - \boldsymbol{m}), \quad \text{WCV} = \sum_{i=1}^{k}\sum_{\boldsymbol{x}\in\text{Class}_i}(\boldsymbol{x} - \boldsymbol{m}_i)^T(\boldsymbol{x} - \boldsymbol{m}_i)$$

, where k is the number of classes, n_i is the number of instances in class i, $\boldsymbol{m}_i$ is the mean vector of instances in class i and $\boldsymbol{m}$ is the mean vector of all instances. To ensure a fair comparison of BCV and WCV across different dimensionality reduction methods, we normalize each reduced dimension dataset before calculating the BCV and WCV of different embedded coordinates.

Method	PCA	LDA	ISOMAP	LLE	LAP	t-SNE
BCV	134.2739	189.0304	215.1110	339.5823	328.4660	422.3671
WCV	86.3030	56.3350	110.0640	43.8792	53.3823	92.3314

According to the table, BCV shows a stronger positive correlation with the clustering metric than WCV. Typically, a higher BCV indicates better clustering performance. However, when different dimensional reduction techniques yield similar BCV results, WCV appears to play a role in determining clustering performance. For instance, LDA has a smaller BCV than ISOMAP but performs better in clustering. This could be due to the fact that ISOMAP has a very high WCV, whereas LDA has a lower WCV.

(a) LDA (b) LLE (c) LAP

Figure 2. The clustering result of the embedded coordinates

In addition to BCS and WCS, the geometric structure of the reduced dataset can also impact clustering performance. For example, the reduced dataset from LLE and LAP has a non-spherical geometry with elongated or irregular clusters, making it difficult for k-means clustering to accurately identify the clusters (see Figure 2). As a result, even though the BCS of LLE and LAP are larger than that of LDA, their clustering performance is similar.

References

[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001*. Springer Berlin Heidelberg, 2001.

[2] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr. Boston.*, 12(4):461–486, July 2008.