Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Programme

Visualization of high-dimentional data using t-SNE algorithm

Linear Algebra final project report

Authors:
Serhii Brodiuk
Andrii Yurkiv

24 January 2019

Abstract

Hello this is abstract

1 Introduction

Data visualization is one of the most important parts of applied data analysis nowadays. Without proper visualization it's hard, or sometimes even impossible to inteprete the data in the appropriate way or make reasonable inferences regarding its structure. Dealing with datasets that contain more than 3 attributes start causing problems, because standard visualization techniques doesn't ap ply to the data in 4-dimentional space. For humans, it's even hard to perceive graphical information in 3-dimentional space, that's why the majority of classical techniqes of data visualization are restricted to two dimentions.

It should be noted that one of the main purposes of visualization of high-dimentional datasets is to determine clusters inside the data which may be an important basement for further data analysis.

2 Motivation

Modern data often contains hundreds or even thousands of features. And classical visualization approaches cannot be used to determine patterns or clusters in it. This is the reason why so many different techniques for visualization high-dimentional data have been proposed. Majority of these techniques merely provide tools to display multidimentional data in two or three-dimentional form and do not give any inferences regarding the inner structure of the data. This severely limits the applicability of those techniques to real-world data, which sometimes contain too much features to reduce them nicely.

Another approach to this problem lies in the field of dimentionality reduc tion. We can apply classical plotting methods (for example, scatterplots) for the data obtained by decreasing the dimention of the original dataset. Dimentionality reduction techniques differ substantially from visualization techniques because they are aimed not to make data more visually appealing or understandable, but to preserve as much of the significant structure of the high-dimentional data as possible in its low-dimentional representation [1].

3 Problem setting

The most important goal of visualization of multidimentional data is to correctly identify clusters of similar data points. This is crucial when we are working on unsupervised learning problem and need to know for sure what number of distinct clusters can our data set be divided in without losing valuable information about inner structure of our data.

As we will see in this report, many classical visualization algorithms correctly identify the relationships between data points and position similar ones together. But at the same time from visualizations those algorithms provide it's hard to tell where one class of points ends and other begins (boundaries between classes are not well-defined). And while this is not a problem for small data sets, for large ones it can cause serious obstacles for correct

definition of of the number and location of data clusters.

t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm is aimed to overcome those limitations by identifying the relevant patterns using the approach in which similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability [2].

4 Related work

5 Approach to Solution

6 Solution

7 Evaluation

Since t-SNE is primaraly visualization technique, in order to evaluate its performance and visualizational capacity we need to compare visualization produced by t-SNE with the ones obtained using other non-parametric visualization techniques for multidimentional data, such as Multidimentional scaling using SMACOF (Scaling by MAjorizing a COmplicated Function) algorithm, Isomap, Locally Linear Embedding, Principal Components Analysis and Laplacian Eigenmaps.

7.1 Experiments

In our work we experimented with 2 different datasets: Olivetti faces dataset and handrwitten digits dataset.

Olivetti faces dataset contains 400 64×64 grayscale images of faces (10 images for each of 40 subjects). Those images were taken at different time with different datails varying, such as face expression, lighting and facial details (absence or presence of glasses). They are labeled according to the identity of the person depicted.

Handwritten digits dataset consists of 1797 8×8 grayscale images of handwritten digits of 10 classes. There are nearly 180 instances per digit class in this dataset.

In the original t-SNE paper [1], the authors use PCA to reduce dimentionality before feeding the data into algorithms that transform it into two-dimentional representation. They are doing it for the purpose of computation efficiency and it absolutely makes sence when we apply those algorithms to large datasets. But since the datasets which we use in our experiements are relatively small, preliminar dimentionality reduction doesn't make much difference in terms of computational costs in our case, so we decided to pass this step up.

For each of these datasets, there is information about the class of each data point. Class information is only used for visualization purposes and is not considered during calculation of the spatial coordinates of the points in two-dimentional space. We decided

Technique	Parameters
t-SNE	perplexity = 5, 1000 iterations
MDS (SMACOF)	-
Isomap	neighbors = 5
LLE	neighbors = 12
PCA	-
Laplacian Eigenmaps	-

Table 1: Parameters of the algorithms

to visualize the whole images instead of points because this kind of graph is much more understandable and provides better means of evaluating how well the mapping preserves the similarities within each class.

8 Conclusions