Analysis and visualisation of high-dimensional data

Keywords : Data mining; Big data; Dimensionality reduction; Exploratory data analysis; Machine learning.

John A. Lee, Michel Verleysen

Abstract* – *Visualisation and processing of high-dimensional data can be made easier by dimensionality reduction, which attempts to represent data with few variables or features and minimal information loss. Our research projects investigate a particular family of state-of-the-art methods of dimensionality reduction, called stochastic neighbour embedding. Improved and more flexible loss functions, multi-scale neighbourhoods, and scalable implementation are our current objectives. Another line of research studies quantitative criteria to assess the quality of dimensionality reduction results.

Modern technology allows huge amounts of data to be collected, with many recorded observations, as well as many variables. High data dimensionality often means that interpretation by a human analyst is difficult. Similarly, automated processing is hindered by the so-called *curse of dimensionality*. This refers to a series of counter-intuitive phenomena that affects high-dimensional spaces. A typical example is the phenomenon of distance concentration. In high-dimensional spaces, distances are poorly discriminant and all points seem to lie far from each other.

In this context, methods of dimensionality reduction (DR) can ease both manual and automated analysis. For instance, DR can embed data in 3D or 2D spaces that can be easily be visualised and explored. Less drastic DR, from thousands or hundreds of dimensions to only a few dozens, can mitigate the curse of dimensionality in automated data processing tasks like regression, function approximation, etc. Linear DR has been known for quite a long time, with methods like principal component analysis, which projects data onto some lower-dimensional subspace. Recent methods are more powerful and can embed data in a nonlinear fashion [1]. To some extent, these methods can thus learn from data some underlying, nonlinear manifold, instead of a linear subspace. Several approaches can be followed to achieve this task: auto-associative neural networks, spectral graph embedding, kernel extensions of PCA, etc. Yet another well known principle consists in finding a low-dimensional representation of that best preserve pairwise relationships, like distances or similarities. Current research in our team investigates this last option. Pairwise similarities quantify the probability of two data points to be neighbours. Data embedding is then obtained by trying to reproduce in a low-dimensional space the neighbourhoods observed in the initial data space. Several ways to improve this method of *stochastic neighbour embedding* (SNE, [2, 5]) are investigated:

- Improvement of the loss functions [4]. These measure the discrepancy between the low-dimensional neighbourhoods and their high-dimensional counterparts [4]. Specifically, mixtures of divergence functions have improved the DR results.
- Multi-scale neighbourhoods [6]. Most SNE variants involve neighbourhoods with a single, fixed size, chosen arbitrarily by the user. Our approach explores combinations of neighbourhoods with multiple sizes, in order to capture data structure on all scales, from local to global. Here too experimental results have shown more faithful rendering of data.
- Scalable implementation of multi-scale SNE. This ongoing project aims at making DR applicable to large data sets within reasonable computation times.

Our team also studies quality assessment of DR [3, 6]. As an unsupervised learning task, DR is difficult to appraise and we have thus proposed several quality criteria, as well as a unifying framework. An ongoing project investigates scalable approximations of these criteria, in order to evaluate DR quality as fast as possible for very large data sets.

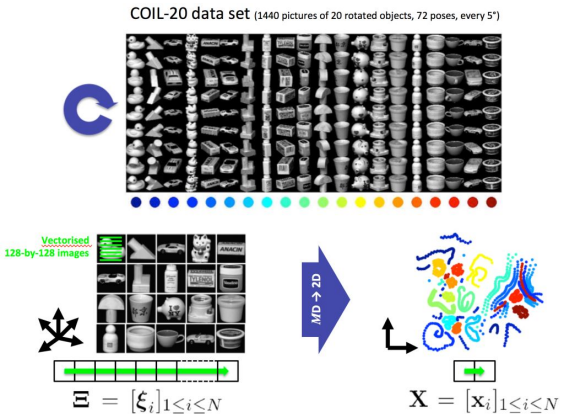

Figure 1: Visualisation of an image bank (COIL-20) with dimensionality reduction. Images are first vectorised and then embedded with an SNE-like method. Clusters appear for each object.

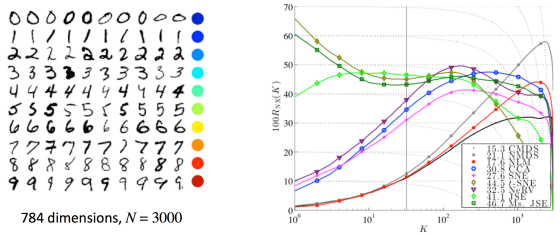

Figure 2: Quality assessment of several DR methods applied to images of handwritten digits. The higher the curve, the better the corresponding method preserves neighbourhoods observed in the initial data space.

References

[1] J. A. Lee, M. Verleysen. "Nonlinear dimensionality reduction." Springer, 2007.

[2] J. A. Lee, M. Verleysen. "Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants." Procedia Computer Science 4, 538-547.

[3] J. A. Lee, M. Verleysen. "Quality assessment of dimensionality reduction: Rank-based criteria." Neurocomputing, 2009, 72(7):1431-1443.

[4] J. A. Lee, E. Renard, G. Bernard, P. Dupont, M. Verleysen. 'Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation." Neurocomputing, 2013, 112:92-108.

[5] J. A. Lee, M. Verleysen. "Two key properties of dimensionality reduction methods." IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2014.

[6] J. A. Lee, D. H. Peluffo-Ordonez, M. Verleysen. "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure." Neurocomputing, 2015, 169:246-261.