Neurocomputing 205 (2016) 247–263

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

ELMVIS+: Fast nonlinear visualization technique based on cosine distance and extreme learning machines

Anton Akusok [a,b,*], Stephen Baek [a], Yoan Miche [c,d], Kaj-Mikael Björk [e], Rui Nian [f], Paula Lauren [g], Amaury Lendasse [a,b,*]

[a] Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, USA
[b] The Iowa Informatics Initiative, The University of Iowa, Iowa City, USA
[c] Bell Labs, Nokia, Espoo, Finland
[d] Department of Information and Computer Science, Aalto University School of Science, FI-00076, Finland
[e] Risklab at Arcada University of Applied Sciences, Helsinki, Finland
[f] School of Information Science and Engineering, Ocean University of China, Qingdao, China
[g] School of Engineering and Computer Science, Oakland University, Rochester, USA

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form
14 March 2016
Accepted 29 April 2016
Communicated by G.-B. Huang
Available online 11 May 2016

Keywords:
Visualization
Nonlinear Dimensionality Reduction
Cosine Distance
Extreme Learning Machines
Big Data
Projection

ABSTRACT

This paper presents a fast algorithm and an accelerated toolbox[1] for data visualization. The visualization is stated as an assignment problem between data samples and the same number of given visualization points. The mapping function is approximated by an Extreme Learning Machine, which provides an error for a current assignment. This work presents a new mathematical formulation of the error function based on cosine similarity. It provides a closed form equation for a change of error for exchanging assignments between two random samples (called a swap), and an extreme speed-up over the original method even for a very large corpus like the MNIST Handwritten Digits dataset. The method starts from random assignment, and continues in a greedy optimization algorithm by randomly swapping pairs of samples, keeping the swaps that reduce the error. The toolbox speed reaches a million of swaps per second, and thousands of model updates per second for successful swaps in GPU implementation, even for very large dataset like MNIST Handwritten Digits.

1. Introduction

High-dimensional data is ubiquitous in the modern world, but it stays virtually impenetrable for human analysis, except for images or audio. Thus data visualization [1] stays a demanded area of research. For the exploratory data analysis of an arbitrary high dimensional data, a suitable visualization should be created. It is commonly restricted to two or three dimensions, which are easier to show, but for the visualization to be useful it must be representative of the original data.[2]

The naive dimensionality reduction method is variable (feature) selection, but a few selected variables could present only a part of the data structure, if any. Other dimensionality reduction methods optimize a selected criterion, with different criteria resulting in two different algorithms.

Linear dimensionality reduction methods such as Principal Components Analysis (PCA) [2] and linear Multidimensional Scaling (MDS) [3] yield the same results, as proven in [1]. Their criterion is variance maximization which works for datasets with linear dependencies, but the general performance may be poor.

If the variables are relevant but correlated (which is often the case), the dimensionality of data is higher than necessary. Then the same data could be explained by a smaller set of transformed variables, and is said to lie on a manifold [1]. As an example, one can imagine a camera rotating around an object at a fixed distance, then the pictures of that camera would lie on a 2-dimensional manifold (sphere), while their actual dimensions would be much higher. Many nonlinear dimensionality reduction methods, including those listed in the next section, aim to find and unfold such a manifold using various cost functions and training algorithms. Even PCA would find a manifold in the data, if the data is linear. Manifolds are commonly found by preserving the neighborhood in original and reduced spaces. Topology-preserving

* Corresponding authors at: Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA, USA.
E-mail addresses: anton-akusok@uiowa.edu (A. Akusok), amaury-lendasse@uiowa.edu (A. Lendasse).

[1] https://github.com/akusok/elmvis

[2] This paper is an extension of a publication accepted at the ELM'15 conference.

http://dx.doi.org/10.1016/j.neucom.2016.04.039
0925-2312/

methods that use graph distances, like Curvilinear Distance Analysis (CDA) [4,5], normally provide excellent results for un-foldable manifolds.

In a very high dimensional space, neighborhood rank is a weak metric [6]. This is caused by an empty space phenomenon [7] and the curse of dimensionality, studied thoroughly in [6]. The problem comes from the change of the distribution of distances between points in space as the dimensionality goes up. Distances between points in a dataset are typically normally distributed. With the increase of a space dimensionality, the mean of that normal distribution increases whereas the variance stays the same. It causes the distribution to concentrate around some value, and reduces the distance differences between various ranked neighbors, making the nearest neighbors unstable already at 10–20 dimensions [6]. These cases require a nonlinear dimensionality reduction method with general cost function without other assumptions. The Extreme Learning Machine (ELM) [8,9] is a popular [10] fast [11] version of Artificial Neural Networks that provides a required non-linear basis for deriving such methods [12]. It is used in ELM-based visualization methods ELMVIS [13] and its improvement ELMVIS+ is presented in this paper. They use Mean Squared Error (MSE) or cosine distance of ELM-reconstructed data accordingly, while the non-linearity of ELM provides the desired nonlinear projection.

The ELMVIS+ represents data visualization as an assignment problem [14] of data samples to the same number of given visualization points, which are fixed. An ELM model learns the de-projection of visualization points back into the original data space, where a cost function is calculated. The optimization task is to find the best assignment between the two sets of samples, i.e. the best order of data samples for a fixed order of visualization points.

An original assignment problem is a challenging NP-hard [14] optimization task, similar to an open loop travelling salesman problem [15]. The ELMVIS+ methodology uses two improvements: a new cost function that can be updated very fast for the position exchange of two data samples, and a greedy optimization approach by changing only two assignments at a time which reduces complexity to $\mathcal{O}(N^2)$. In total, they provide a fast and useful method of data visualization onto arbitrary fixed set of points in the visualization space. The method has only one hyper-parameter, that is the number of neurons in the ELM model, and the local optimum problem may be solved by multiple re-runs of the method. The new cost function works for very high-dimensional data.

The rest of the paper is organized as follows. Section 2 gives an overview on the state-of-the-art. It also introduces reference methods to the reader. Section 3 describes the ELM algorithm and its adaptation for computation and fast update of a cost function for visualization. Section 4 presents experimental comparison with other methods on various datasets, while Section 5 analyses performance and large datasets results. Section 6 concludes on the work done, discusses about improvements compared to the original ELMVIS and directions of future research.

2. State-of-the-art

Various methods can be utilized for a data visualization task. A common assumption in dimensionality reduction, and especially in data visualization, is that the original data points lie on a low-dimensional manifold. If the assumption holds, then the points of a manifold may be mapped onto a low-dimensional visualization space with small information loss.

The visualization methods may be divided into two major groups, separated by whether they try to keep distances of topology structure. Distance-preserving methods include Multidimensional Scaling (MDS) [3], which gives the same solution as PCA; Sammon's mapping [16]; Curvilinear Component Analysis (CCA) [17]; Isomap [18,19]; Curvilinear Distance Analysis (CDA) [5], and Kernel PCA [20]. Topology preserving methods are Self-Organizing Maps (SOM) [21]; Generative Topographic mapping (GTM) [22]; Locally Linear Embedding (LLE) [23]; Laplacian Eigenmaps [24,25], Isotop [26] and Neighbor Retrieval Visualizer (NeRV) [27]. Out of these, the three benchmark methods selected are PCA, SOM and NeRV.

2.1. Visualization quality measures

There are different ways to measure and compare the quality of a visualization. The Mean Squared Error (MSE) of reconstruction came from the dimensionality reduction, and is a universal measure of quality. However, it requires a reversed projection from visualization to the original data space, which not all the methods can provide. So other quality measures are often used.

One of the common measures is precision and recall of a projection. It comes from the classification task, where the definitions are

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

The visualization task has no classes, but they are created manually by setting all points within a certain neighborhood as $+1$ class, and the others as -1 class [28], as shown in Fig. 1. As in visualization both precision and recall depend on the size of a neighborhood used for their calculation, which is not the case in classification, other similar measures are used: continuity is similar to precision, and trustworthiness to recall [29].

Another method, called Mean Relative Rank Error (MRRE), is a neighborhood preservation ratio. Based on the ideas from, among others, [30–32], and refined by [1], this measure displays the average normalized error in ranking within k nearest neighbors. The normalization puts the measure in range between 0 and 1, where 0 corresponds to the perfect match of the first k neighbors, and 1 to the replacement of the first k neighbors by the most distant k points. Depending on which space the closest k neighbors are chosen for calculation, two *MRRE's* exist: $MRRE_{\mathbf{X}\to\mathbf{V}}(k)$ can be compared to continuity, and $MRRE_{\mathbf{V}\to\mathbf{X}}(k)$ to trustworthiness.

And the last but not the least, a plot of visualized points may be used as a measure of goodness [1]. This is especially true if data points can be observed directly such as with images, then the user can estimate the quality of clustering by simply browsing the visualized data.

2.2. Principal components analysis

Principal Components Analysis (PCA) is a linear method, which has an exact and relatively fast solution. Given the dataset $\mathbf{X}$ with N samples as rows of $\mathbf{X}$ and d features as columns of $\mathbf{X}$, PCA decomposes the covariance matrix $\mathbf{C}_{xx}$ into eigenvectors $\mathbf{U}$ and eigenvalues $\mathbf{\Lambda}$. Eigenvalues of $\mathbf{\Lambda}$ are ranked from largest to smallest, and the corresponding eigenvectors in $\mathbf{U}$ are placed accordingly.

$$\mathbf{C}_{xx} = \mathbf{X}^T\mathbf{X} = \mathbf{U}^T\mathbf{\Lambda}\mathbf{U} \tag{3}$$

$$\mathbf{V} = \mathbf{X}\mathbf{U}_{:,1:k} \tag{4}$$

where $\mathbf{V}$ are points in the visualization space, and $\mathbf{U}$ has only the first k columns.

PCA projects data points to the dimensions of the largest variance. Its advantages are simplicity, robustness and lack of parameters. The main drawback of PCA is its linearity which captures a linear manifold, but nonlinear manifolds would be squashed using PCA.