ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Recent methods for dimensionality reduction: A brief comparative analysis

Diego H. Peluffo [1], John A. Lee [1,2], and Michel Verleysen [1] *

1- Machine Learning Group - ICTEAM,
Université catholique de Louvain

2- Molecular Imaging Radiotherapy and Oncology - IREC,
Université catholique de Louvain

Abstract. Dimensionality reduction is a key stage for both the design of a pattern recognition system or data visualization. Recently, there has been a increasing interest in those methods aimed at preserving the data topology. Among them, Laplacian eigenmaps (LE) and stochastic neighbour embedding (SNE) are the most representative. In this work, we present a brief comparative among very recent methods being alternatives to LE and SNE. Comparisons are done mainly on two aspects: algorithm implementation, and complexity. Also, relations between methods are depicted. The goal of this work is providing researches on this field with some discussion as well as criteria decision to choose a method according to the user's needs and/or keeping a good trade-off between performance and required processing time.

1 Introduction

Dimensionality reduction (DR) allows the extraction of lower dimensional, relevant information from big collections of data aimed at improving the performance of a pattern recognition system or allowing for intelligible data visualization. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$. Classical DR approaches were conceived following an intuitive criterion, such as variance preservation (principal component analysis - PCA) or distance preservation (classical multidimensional scaling - CMDS) [1]. Nowadays, more developed, recent methods are aimed at preserving the data topology. Such a topology is very often given by a data-related graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian eigenmaps [2]. On the other hand, once normalized, it can also represent a probability distribution. The latter is the case of the methods based on divergences such as stochastic neighbour embedding [3].

This work presents a brief comparative overview of recent, dimensionality reduction methods emerging as alternatives to Laplacian eigenmaps and stochastic neighbour embedding. Among them, locally linear landmarks for manifold learning [4], elastic

*J.A. Lee is a Research Associate with the FRS-FNRS (Belgian National Scientific Research Fund). This work is funded by FRS-FNRS (Belgian National Scientific Research Fund) project 7.0175.13 DRedVis.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

embedding [5] and methods based on mixtures of divergences [6, 7]. Comparative analysis is done mainly on two aspects: algorithm implementation, and complexity. As well, relations between methods are depicted. This work is gathers some criteria and discussion on how to choose among methods according to the user's needs, and/or the trade-off between performance and required processing time.

The rest of this paper is organized as follows: Sections 2 and 3 outline the studied methods. The comparative analysis is presented in section 4. Finally, section 5 draws the discussion and final remarks.

2 Spectral methods

A popular spectral approach for DR is Laplacian Eigenmaps (LE) introduced in [2], which is aimed at minimizing local distances. The LE's cost function can be written as $\sum_{n,m=1}^{N} w_{nm}\|\boldsymbol{x}_n - \boldsymbol{x}_m\|$, where $\boldsymbol{W} = [w_{nm}]_{1\leq n\leq N}$ is the similarity matrix and $\|\cdot\|$ stands for Euclidean distance. Alternatively, we can express LE's formulation as

$$E_{\mathrm{LE}}(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\top}) \text{ s. t. } \quad \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\top} = \boldsymbol{I}_d, \quad \boldsymbol{X}\boldsymbol{D}\mathbf{1}_N = \mathbf{0}_d, \tag{1}$$

where $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{W}\mathbf{1}_N)$ is the degree matrix and $\boldsymbol{L}$ is the graph Laplacian matrix given by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. LE's constraints facilitates the solution leading to a generalized eigenvalue problem. Along this line, the embedded data is then the d smallest vector eigenvectors of normalized Laplacian $\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$. Very recently, a fast algorithm to perform LE was introduced in [4]. Instead of using the whole input data $\boldsymbol{Y}$, this approach approximates the solution by using only a subset of L data points (landmarks) $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{L\times N}$. Furthermore, landmark projections are constrained to be locally linear, such that $\boldsymbol{Y} \approx \widetilde{\boldsymbol{Y}}\boldsymbol{Z}$, being $\boldsymbol{Z}$ the projection matrix. The embedded data is obtained by enforcing it to fulfill the same local linearity property so that $\boldsymbol{X} \approx \widetilde{\boldsymbol{X}}$. By replacing this approximation in Eq. 1, we can easily demonstrate that embedded data is now the eigenvectors of $\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^{\top}$ multiplied by $\widetilde{\boldsymbol{Z}}$. In addition, to determine $\boldsymbol{Z}$, the authors propose to solve the simple problem $\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}\boldsymbol{Z}\|^2$ subject to linear conditions.

3 Divergence-based methods

Stochastic neighbor embedding (SNE) [3] minimizes the information divergence D between two distributions $\boldsymbol{P}_n = [p_{nm}]_{1\leq m\leq N}$ and $\boldsymbol{Q}_n = [q_{nm}]_{1\leq m\leq N}$ associated with the n-th point from observed and latent data, respectively. Then, using the Kullback-Leibler directed divergence D_{KL}, the SNE objective function is in the form:

$$E_{\mathrm{SNE}}(\boldsymbol{X}) = \sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}(\boldsymbol{P}_n\|\boldsymbol{Q}_n) = \sum_{n,m=1}^{N} p_{nm}\log\frac{p_{nm}}{q_{nm}}. \tag{2}$$

Defining $\delta_{nm} = \|\boldsymbol{y}_n - \boldsymbol{y}_m\|^2$ and $d_{nm} = \|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2$, distributions $\boldsymbol{P}_n$ and $\boldsymbol{Q}_n$ can be chosen as generalized, normalized nonsymmetric affinities in the form

$$p_{nm} = \frac{\exp\left(-\frac{1}{2}\delta_{nm}^2/\sigma_n^2\right)}{\sum\limits_{n\neq m'}\exp\left(-\frac{1}{2}\delta_{nm'}^2/\sigma_n^2\right)}, \quad \text{and} \quad q_{nm} = \frac{\exp\left(-\frac{1}{2}d_{nm}^2/\pi_n^2\right)}{\sum\limits_{n\neq m'}\exp\left(-\frac{1}{2}d_{nm'}^2/\pi_n^2\right)}, \tag{3}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

with $q_{nn} = 0$ and $p_{nn} = 0$. A symmetric version of SNE (SSNE) can be achieved by selecting full normalized affinities which can readily be obtained by slightly expressions in (3). In this case, rather than a restricted sum, all entries must be summed on the denominator in order to enforce that all normalized entries sum to 1. This can be done by guaranteeing that $\mathbf{1}_N^\top \boldsymbol{Q} \mathbf{1}_N = \mathbf{1}_N^\top \boldsymbol{P} \mathbf{1}_N = 1$. SNE-based methods suffer from reaching distorted and overlapped latent space, when d is smaller than the intrinsic dimension [5]. To cope with this issue, another variant raised, which is named t-SNE and consists of defining $\boldsymbol{Q}_n$ as a t-distribution [7]. Further recently, enhanced approaches have been proposed founded on the mixture of divergences. In [8], it is proposed a mixture by adding a regularization parameter β to balance *precision* and *recall* so: $(1-\beta)\,\mathrm{D}_{\mathrm{KL}}(\boldsymbol{P}_n \| \boldsymbol{Q}_n) + \beta\,\mathrm{D}_{\mathrm{KL}}(\boldsymbol{Q}_n \| \boldsymbol{P}_n)$. Similarly, in [6], a novel approach is introduced which mixes the divergences as $\mathrm{D}_{\mathrm{KL}}^{\beta} = (1-\beta)\,\mathrm{D}_{\mathrm{KL}}(\boldsymbol{P}_n \| \boldsymbol{S}_n) + \beta\,\mathrm{D}_{\mathrm{KL}}(\boldsymbol{Q}_n \| \boldsymbol{S}_n)$, where $\boldsymbol{S}_n$ is a distribution following the same mixture rule so that $\boldsymbol{S}_n = (1-\beta)\boldsymbol{P}_n + \beta \boldsymbol{Q}_n$. This divergence is used in the so-called Jensen-Shannon embedding (JSE), which aims then to minimize $E_{\mathrm{JSE}} = \sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}^{\beta}(\boldsymbol{Q}_n \| \boldsymbol{S}_n)$ [6].

As an alternative to SNE methods, in [5], the Elastic Embedding (EE) is introduced. EE is aimed to optimize:

$$E_{\mathrm{EE}}(\boldsymbol{X}|\lambda) = \sum_{n,m=1}^{N} w_{nm}^{+} d_{nm}^{2} + \lambda \sum_{n,m=1}^{N} w_{nm}^{-} \exp(d_{nm}^{2}) = E_{\mathrm{EE}}^{+}(\boldsymbol{X}) + \lambda E_{\mathrm{EE}}^{-}(\boldsymbol{X}). \qquad (4)$$

Briefly put, this method attempts to involve the two objectives that SNE fulfills but in a simpler way. To this end, which is the key of this method, two graphs are used. Then, we have two kind of weighting coefficients w_{nm}^{+} and w_{nm}^{-} being the entries of attractive $\boldsymbol{W}^{+}$ and repulsive $\boldsymbol{W}^{-}$ affinity matrices, respectively. Both of them are positive semi-definite matrices. For simplicity, full graphs affinities are considered: $w_{nm}^{-} = \|\boldsymbol{y}_n - \boldsymbol{y}_m\|^2$ and $w_{nm}^{+} = \exp(-\frac{1}{2}\delta_n^2/\sigma^2)$. From Eq. (4), the gradient of E_{EE} can be written as: $\boldsymbol{G}(\boldsymbol{X}|\lambda) = 4\boldsymbol{X}(\boldsymbol{L}^{+} - \lambda \widetilde{\boldsymbol{L}}^{-}) = 4\boldsymbol{X}\boldsymbol{L}$, where $\widetilde{w}_{nm}^{-} = w_{nm}^{-} \exp(-d_{nm}^{2})$, $w_{nm} = w_{nm}^{+} - \lambda \widetilde{w}_{nm}^{-}$, and their corresponding Laplacians $\widetilde{\boldsymbol{L}} = \widetilde{\boldsymbol{D}} - \widetilde{\boldsymbol{W}}$ and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. Likewise, as calculated in LE, $\boldsymbol{L}^{+}$ is the non-normalized Laplacian and thus $\boldsymbol{L}^{+} = \boldsymbol{D}^{+} - \boldsymbol{W}^{+}$. In [5], to carry out the search for the suboptimal embedded solution $\boldsymbol{X}$, a gradient descent algorithm is used, which is powered via the spectral direction (SD) proposed in [9].

4 Comparative analysis

The brief comparative analysis presented here encompasses links between methods (Section 4.1), a discussion on algorithm implementation (Section 4.2) and some experimental results (Section 4.3).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

4.1 Links between methods

Relation between SNE and EE: Eliminating independent terms from $\boldsymbol{X}$, Equation (2) can be expanded as

$$E_{\mathrm{SNE}}(\boldsymbol{X}) = \sum_{n,m=1}^{N} p_{nm}\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2 + \sum_{n=1}^{N} \log \sum_{n\neq m} \exp(\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2). \tag{5}$$

Hence we can appreciate that by omitting the log operator and adding a homotopy parameter λ, E_{SNE} becomes the EE's cost function. Furthermore, EE is a variant of the elastic network applied to solve the traveling salesman problem as explained in [10].

Relation between SNE and LE: Recalling Equation (5), it is noticeable that, doing as in diffusion maps [11] which means using the normalized affinities so that $p_{nm} = w_{nm}$, the right hand side of the Equation is the same as the LE objective function.

Relation between EE and LE: The same as in SNE applies when comparing with EE. However, it is noteworthy that by setting $\lambda = 0$, EE does not reach the same embedding as LE since the optimization is different. EE's embedding is determined through a search and that of LE comes from a spectral decomposition under orthonormality assumptions.

4.2 Brief discussion on implementation and complexity

Implementation via SD: Methods such as EE, SNE and SSNE can be implemented in fast fashion via SD-based gradient descent search [5]. We denote the n-th embedded data point at iteration r as $\boldsymbol{x}_n[r] = \boldsymbol{x}_n[r-1] + \alpha[r]\boldsymbol{\varrho}_n[r]$. SD is aimed at determining the optimal direction $\boldsymbol{\varrho}_n[r]$ by incorporating a partial-Hessian strategy within the gradient descent heuristic [9]. Then, by design, Hessian is heavily exploited which is advantageous for subsequent developments since it can be be computed fast and has the suitable property to be positive semi-definite. As an intuitive condition, sought direction must hold that $\boldsymbol{B}[r]\boldsymbol{\varrho}_n[r] = -\boldsymbol{g}_n$, being $\boldsymbol{g}_n$ the column n of $\boldsymbol{G}(\boldsymbol{X}|\lambda)$ and $\boldsymbol{B}[r]$ any positive semi-definite matrix. SD consists of calculating the gradient of $E_{\mathrm{EE}}(\boldsymbol{X}|\lambda)$ following the direction of an underlying convex function which arises when $\lambda = 0$. Also, the calculation of SD is speeded up by using Cholesky decomposition. Namely, rather than calculating matrix directly with $\boldsymbol{\mathcal{P}} = -\boldsymbol{G}(\boldsymbol{X}|\lambda)(\boldsymbol{B})^{-1}$ (which is $\mathsf{O}(N^3D)$ when using conventional Gaussian-Jordan elimination), two solve triangular systems in the form $\boldsymbol{R}^\top\boldsymbol{R}\,\mathrm{vec}(\boldsymbol{P}) = -\mathrm{vec}(\boldsymbol{G})$ are solved, where $\boldsymbol{R}$ is the upper triangular matrix resulting from the Cholesky decomposition of $\boldsymbol{B}\otimes\boldsymbol{I}_d$. Latter calculation can be done in $\mathsf{O}(N^2d)$ with standard linear algebra routines. In addition, computation of $\boldsymbol{R}$ needs to be done only once at first iteration and its complexity is $\mathsf{O}(\frac{1}{3}N)$.

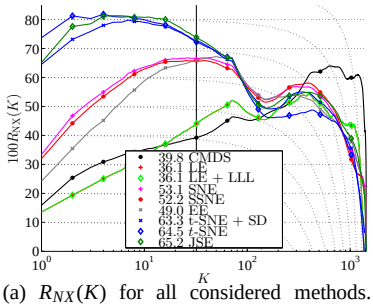
Implementation via a full gradient and Hessian: In [6], the search is done by using a full gradient calculated over the whole cost function (no approximations are done). In this case, the search is done via $\boldsymbol{x}_n[r] = \boldsymbol{x}_n[r-1] + \mu_n[r]\nabla E$, where $\mu_n[r]$ is an adaptive step size dependent on the Hessian. Given the nature of divergences, doing so can increase the complexity. Even more when using a mixture of divergences ($E = E_{\mathrm{JSE}}$), calculation of gradient and Hessian may be more expensive. Nonetheless, the advantage of this implementation is that scaling is considered in both high and low dimensional

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

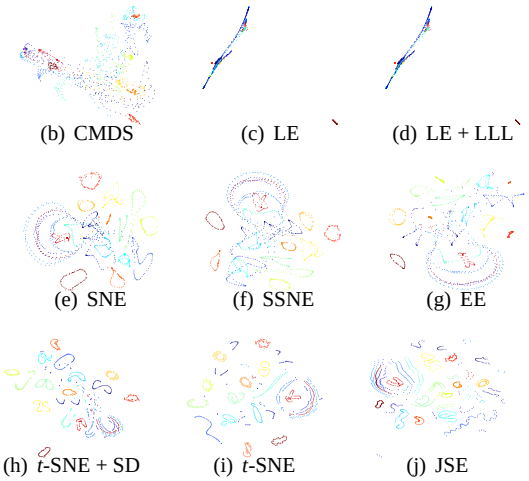
space. This provides a more modulated gradient and then a better tracking of the local structure of data during the optimization process.

4.3 Experimental results and discussion

For shorthand notation, *t*-SNE using SD is denoted as *t*-SNE + SD. Likewise, LE via Locally linear landmarks is denoted as LE + LLL. JSE and *t*-SNE are implemented via a full gradient scheme. Both SD and full gradient implementations involve a backtracking line search. To form the similarity matrices, given a perplexity parameter K, the relative bandwidth parameter σ_n is estimated regarding its distribution $\boldsymbol{P}_n$ so that the entropy over neighbors of such distribution is approximately $\log K$. This is done by a binary search as explained in [5]. For experiments, we set $K = 30$. Also, for LE + LLL, the number of landmarks is $L = 500$, and $\lambda = 100$ for EE. Regularization parameter β for JSE is set to be $1/2$. The methods are tested over the well-known database COIL20 image bank holding $N = 1440$ data points (20 objects in 72 poses/angles) with $D = 128^2$. To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [6] is used, which is ranged within the interval $[0, 1]$. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N - 1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). Overall results regarding AUC$R_{NX}(K)$ are shown in Fig. 1. As well, the resultant embedded spaces reached by each method are depicted.



(a) $R_{NX}(K)$ for all considered methods. The value of AUC is shown in the legend besides the method's name.



(b) CMDS (c) LE (d) LE + LLL

(e) SNE (f) SSNE (g) EE

(h) *t*-SNE + SD (i) *t*-SNE (j) JSE

Fig. 1: Results are shown regarding the quality measure $R_{NX}(K)$. The curves and their AUC (a) for all considered methods are depicted, as well as the embedding data (b)-(j).

5 Discussion and final remarks

By one hand, spectral methods, in general, attempt to preserve the global structure. Particularly, CMDS exhibiting a pronounced peak on large neighbors. LE + LLE re-

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

sembles the LE's behaviour. Then we can say that LLL is a good alternative to initialize LE. In addition, LLL can also mean a significantly degreasing of the processing time if $\mathsf{O}(N^2d) + \mathsf{O}(\frac{1}{3}N) + \mathsf{O}(L^3) < \mathsf{O}(N^3)$. This inequality depends heavily on the number of landmarks, then determining an optimal number of landmarks is a crucial stage aiming to get a good trade-off between processing time and performance (how much it resembles the LE's performance). By the other hand, SNE-like methods perform a better embedding preserving smaller neighbours (local structure). We can notice that SNE, SSNE and EE have a similar performance. In this case, SD makes that SNE and EE behave as a symmetrized version due to strong assumption on the gradient calculation. On the contrary, *t*-SNE + SD performs a better embedding since *t*-distributed probabilities may improve the separation of underline clusters despite of biasing the gradient. Indeed, *t*-SNE + SD accomplishes a similar $R_{NX}(K)$ shape and AUC in comparison with *t*-SNE. JSE outperforms the remaining considered methods due to both the divergence type, and the identical similarity definition in the high-dimensional and low-dimensional space.

This work gathers some key aspects to compare dimensionality reduction methods. Namely, relations between them, algorithm implementation, and complexity/processing time. Very recent methods were studied such as elastic embedding, locally linear landmarks for laplacian eigenmaps and Jensen-Shanon embedding. Discussion and hints provided here may facilitate users to chose a method according the trade-off between performance and complexity.

References

[1] Ingwer Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 833–840, 2002.

[4] Max Vladymyrov and Miguel Á Carreira-Perpinán. Locally linear landmarks for large-scale manifold learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 256–271. Springer, 2013.

[5] Miguel A Carreira-Perpinán. The elastic embedding algorithm for dimensionality reduction. In *ICML*, volume 10, pages 167–174, 2010.

[6] John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 2013.

[7] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[8] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research*, 11:451–490, 2010.

[9] Max Vladymyrov and Miguel Carreira-Perpinan. Partial-hessian strategies for fast learning of nonlinear embeddings. *arXiv preprint arXiv:1206.4646*, 2012.

[10] Richard Durbin, Richard Szeliski, and Alan Yuille. An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 1(3):348–358, 1989.

[11] Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012.