Multipoint Neighbor Embedding

Adrian Lancucki(✉) and Jan Chorowski

Institute of Computer Science, University of Wrocław, Wrocław, Poland
{adrian.lancucki,jan.chorowski}@cs.uni.wroc.pl

Abstract. Dimensionality reduction methods for visualization attempt to preserve in the embedding as much of the original information as possible. However, projection to 2-D or 3-D heavily distorts the data. Instead, we propose a multipoint extension to neighbor embedding methods, which allows to express datapoints from a high-dimensional space as sets of datapoints in a low-dimensional space. Cardinality of those sets is not assumed a priori. Using gradient of the cost function, we derive an expression, which for every datapoint indicates its remote area of attraction. We use it as a heuristic that guides selection and placement of additional datapoints. We demonstrate the approach with multipoint t-SNE, and adapt the $\mathcal{O}(N \log N)$ approximation for computing the gradient of t-SNE to our setting. Experiments show that the approach brings qualitative and quantitative gains, i.e., it expresses more pairwise similarities and multi-group memberships of individual datapoints, better preserving the local structure of the data.

Keywords: Manifold learning · Data visualization · t-SNE · Barnes-Hut algorithm

1 Introduction

The objective of dimensionality reduction is to construct a mapping from a high-dimensional dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ to a low-dimensional dataset $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ where typically $\mathbf{x}_i \in \mathbb{R}^S$, $\mathbf{y}_i \in \mathbb{R}^s$, and $S \gg s$. A special case of dimensionality reduction is visualization where $s = 2$ or $s = 3$.

Non-linear methods perform better than linear at visualization, because it is unlikely for high-dimensional data to lay in a linear subspace of such small dimensionality. The class of Neighbor Embedding (NE) algorithms [15] is especially useful in this task. We propose a multipoint extension to NE in which every datapoint $\mathbf{x}_i$ is embedded as a set of low-dimensional datapoints $Y_i = \{\mathbf{y}_i, \mathbf{y}'_i, \mathbf{y}''_i, \ldots\}$ of equal importance. Cardinality of Y_is is not assumed a priori, as the copies are added as necessary. Replication of each $\mathbf{y} \in Y_i$ is decided heuristically with a scoring function. We demonstrate the extension on the example of multipoint t-SNE, which is used during validation.

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 456–464, 2017.
DOI: 10.1007/978-3-319-64206-2_51

1.1 Stochastic Neighbor Embedding

In Stochastic Neighbor Embedding (SNE) [7], the data is modeled with a distribution P, where $p_{i|j}$ is a conditional probability of $\mathbf{x}_i$ picking $\mathbf{x}_j$ as its neighbor, calculated by centering Gaussians at each $\mathbf{x}_i$. The embedding $\mathbf{Y}$ is modeled with a probability distribution Q. Similarly, it is constructed by centering Gaussians at each $\mathbf{y}_i$ [7]

$$p_{i|j} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2\right)}, \quad q_{i|j} \quad = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)}. \tag{1}$$

The embedding is constructed by minimizing the cost defined as the sum of Kullback-Leibler divergences

$$C(Y) = KL(P||Q) = \sum_i \sum_j p_{i|j} \log \frac{p_{i|j}}{q_{i|j}}. \tag{2}$$

In t-SNE [10] Gaussian distribution in Q is replaced with Student's t distribution.

2 Visualization with Multipoint Embeddings

We demonstrate and analyze our approach with multipoint t-SNE. Because we embed individual datapoints as sets of datapoints, we need to modify Q. We replace the Euclidean metric in Q with a distance function, which relates sets Y_i with Y_j through minimum distance between their elements $d_{ij} = \min\{\|\mathbf{y}_k - \mathbf{y}_l\| : \mathbf{y}_k \in Y_i \wedge \mathbf{y}_l \in Y_j\}$. Please note that it breaks the triangle inequality and is no longer a metric. Elements of Y_i are not weighted, and therefore are of equal importance to the embedding. The gradient can be derived with the chain rule [10], differing only in $\partial d_{ij}^2/\partial \mathbf{y}_i$ calculated using a subderivative of min.

NE methods have plausible interpretation as physical systems of springs [6]. The gradient of t-SNE can be decomposed into a difference of terms interpreted as attractive and repulsive forces [9,15]

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4\sum_j \frac{\partial d_{ij}^2}{\partial \mathbf{y}_i}\frac{p_{ij}}{(1+d_{ij}^2)} - 4\sum_j \frac{\partial d_{ij}^2}{\partial \mathbf{y}_i}\frac{q_{ij}}{(1+d_{ij}^2)} = F_{attr} - F_{rep}. \tag{3}$$

Because of the heavy tail of Student's t distribution used to construct Q, the repulsive force F_{rep} acting on $\mathbf{y}_i$ comes from all datapoints. Conversely, because of the light tail of the Gaussian used to construct P, the attractive force F_{attr} comes only from those datapoints, which are close neighbors of $\mathbf{x}_i$ in $\mathbf{X}$.

Optimization of t-SNE cost function in Eq. (2) is hard, as the function is highly non-convex [7,10]. Minimization of the cost with gradient methods moves the datapoints slightly at every gradient update. In practice, a misplaced datapoint $\mathbf{y}_i$ might be repulsed by dissimilar datapoints with a force which balances

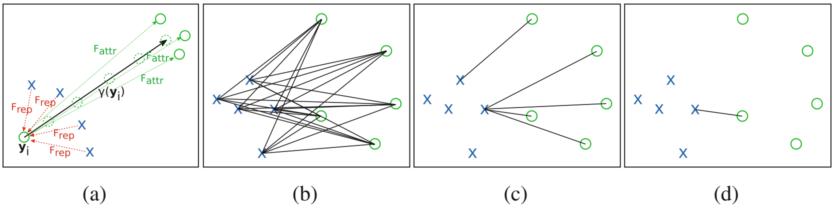


Fig. 1. Relationships between two sets Y_i (blue Xs) and Y_j (green Os) which embed $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. Similar datapoints are in the same color. (a) A datapoint prevented from reaching the area of attraction (pointed by $\gamma(\mathbf{y_i})$) by dissimilar datapoints, (b) Redundant pairwise relations, (c) Relations limited to, for every $y_i \in Y_i$, only the closest $y_j \in Y_j$, (d) The correct relation as the shortest distance between Y_i and Y_j. Best viewed in color.

the attractive force from a distant, similar datapoint (Fig. 1a). Thus the optimization might get stuck in a local optimum, where forces reach a spurious equilibrium.

Several strategies were proposed to remedy this problem, helping a datapoint to take a leap over dissimilar datapoints with a small increase in the cost in Eq. (2). For instance, a temporary dimension might be added to the embedding, weight of the datapoint might be lowered, or the whole embedding might be kept close early in the optimization [7,10].

2.1 Replication of Datapoints

Instead of encouraging the datapoints to move to remote areas, we propose a two-stage heuristic process, which happens during optimization. Firstly, we heuristically recognize obstructed datapoints, and place their copies in their remote areas of attraction. Secondly, we recognize and discard unused copies. This way a datapoint might be copied, but also moved by first being copied and then discarded.

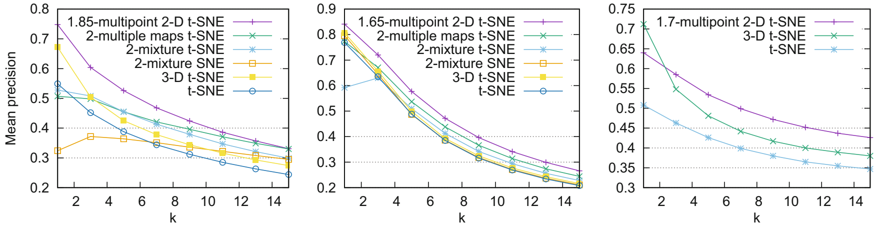
Formally, for each $\mathbf{y}_i$, we compute its **potential** $\gamma(\mathbf{y}_i)$

$$\gamma(\mathbf{y}_i) = 4 \sum_{p_{ij} > q_{ij}} \frac{\partial d_{ij}^2}{\partial \mathbf{y}_i} (p_{ij} - q_{ij}). \tag{4}$$

We derive γ by multiplying the cost function of t-SNE in Eq. (3) by the $(1 + d_{ij}^2)$ term to cancel it out, because it promotes closer pairs of datapoints. Interestingly, γ becomes the gradient of the cost function of symmetric SNE (excluding the $p_{ij} > q_{ij}$ restriction). It also has a physical interpretation, under which $(p_{ij} - q_{ij})$ may be treated as a spring constant, and $\partial d_{ij}^2 / \partial \mathbf{y}_i$ as spring length [5]. We interpret vector $\gamma(\mathbf{y}_i)$ as the direction of a promising region, and $\|\gamma(\mathbf{y}_i)\|$ as the magnitude of its attraction.

We score each datapoint $\mathbf{y}$ with $\|\hat{\gamma}(\mathbf{y})\|$, and replicate a fraction of top scoring ones. The copies are initialized one by one, through line search along the direction of $\gamma(\mathbf{y}_i)$, so that they would minimize the cost in Eq. (2) (Fig. 1a). Placement may be approximate, as the optimization will be continued. For instance, 10% of datapoints might be replicated every 100 iterations up to total of 1000 iterations.

During **cleanup**, we discard misplaced datapoints, which experience little attractive force. Probability mass $\sum_j p_{ij}$ induces attractive force F_{attr} which acts on Y_i in Eq. (3). Because each p_{ij} is modeled by exactly one $\mathbf{y}_k \in Y_i$, we can distribute it among all elements in Y_i and normalize to a probability distribution r_{ki}. It can be interpreted as probability of $\mathbf{y}_k \in Y_i$ being the closest datapoint from Y_i to a random neighbor of Y_i. All datapoints with low r_{ki} should be discarded, as they experience small attractive force, and thus will likely be repelled from the embedding into infinity. Distribution r may also be interpreted as a weighting function for datapoints in Y_i.



(a) Word associations (b) NIPS co-authorships (c) Word2vec

Fig. 2. Precision (in information retrieval sense) of reconstruction of k-NNs (higher is better). Decimal numbers denote how many maps/mixture components remained at the end (1.85 denotes ending up with 185% of initial datapoints etc.). For fairness of comparison, we select indexes of k highest entries in Q for a given $\mathbf{y}_i$, and k highest in P for a $\mathbf{x}_i$. The method is equivalent to selecting k closest datapoints w.r.t. the Euclidean distance for SNE and t-SNE, as their P and Q neighborhood functions monotonically decrease with the distance. Best viewed in color.

2.2 Fast Gradient Approximation

Barnes-Hut (B-H) algorithm for t-SNE [9,15] approximates a low-dimensional n-body simulation [1], allowing to compute the gradient of the cost in $\mathcal{O}(N \log N)$ steps instead of $\mathcal{O}(N^2)$. The space is partitioned with a quad- or oct-tree, traversed in a depth-first manner. Nodes (cells) which meet the B-H condition [1] are treated as approximate summaries of its contents. Direct application of the B-H algorithm to multipoint t-SNE would overestimate F_{rep} by counting all pairwise relations (Fig. 1b) instead of only those between closest of datapoint replicas (Fig. 1d). We prevent it in two ways.

Firstly, we compute $\partial C / \partial \mathbf{y}_i$ for all $\mathbf{y}_i \in Y_i$ in parallel. For each pair (Y_i, Y_j) only the closest $\mathbf{y}_k \in Y_j$ should interact with Y_i. We traverse the quadtree in a depth-first manner and assign each cell to the closest $\mathbf{y}_i \in Y_i$ if it meets the B-H condition. Instead of immediately updating the gradient with those cells, we store them in lists, keeping one list for each $\mathbf{y}_i \in Y_i$. When traversal is over, we sort the list for each $\mathbf{y}_i$ based on the distance of those cells from $\mathbf{y}_i$. However, F_{rep} is still overestimated (Fig. 1c).

Secondly, we correct the number k_i of datapoints in the ith traversed cell. For each $\mathbf{y}_j$ in a cell, we would like to check if any of its replicas already interacts with Y_i, and if so, subtract 1 from k_i. We call such situation a collision. Let α_i denote the estimated number of collisions within a cell and $k_c = \sum_{j<i} k_j$ total number of datapoints in previously processed cells. Then[1] $\alpha_i = k_i + N\left(N - \frac{1}{N}\right)^{k_c}\left[\left(1 - \frac{1}{N}\right)^{k_i} - 1\right]$. Knowing the exact number of collisions $\sum_j k_j - N$, we normalize α_i as as $\hat{\alpha}_i = \frac{\alpha_i}{\sum_j \alpha_j}\left(\sum_j k_j - N\right)$. The proposed approximation is applicable when repulsive weights are equal, i.e., every datapoint repels the others with the same force.

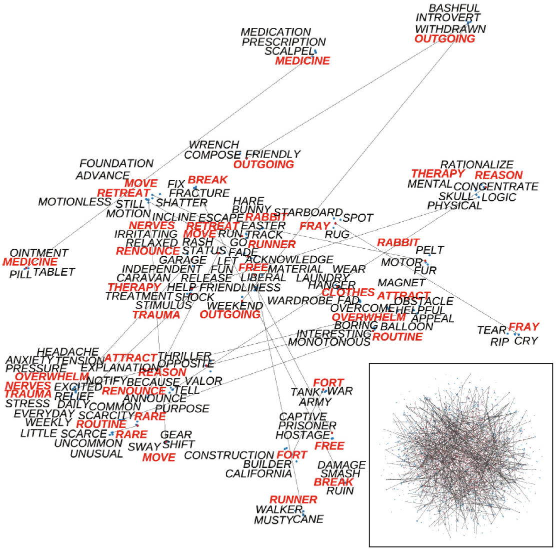


Fig. 3. Partial multipoint t-SNE embedding of word associations. Lines connect copies of the same datapoint (in red). 3-NNs were plotted for each copy. Overview of the complete embedding is shown in the lower right corner. Best viewed in color.

3 Related Work

To promote sparseness of the embedding, NE methods repulse all pairs of close datapoints. Weighted symmetric SNE (ws-SNE) [16] weights points in the repulsive term. Elastic Embedding (EE) [4] expresses the penalty with a simplified cost function.

[1] α_i amounts to expected number of collisions in a hash table of N locations when inserting k_i elements, after having inserted k_c elements.

Mixture SNE [7] embeds a single datapoint in multiple locations. The initial probability mass of a datapoint is distributed among its copies through weights. From the beginning of optimization, each datapoint has a fixed number of copies. Contrary, in our approach the copies are equally weighted, and their number varies.

Cook et al. [5] proposed visualization of pairwise similarities with mixtures of m 2-D maps, where each one is fitted with symmetric SNE. Multiview SNE [2] combines the data from t heterogeneous maps into a single embedding map. Each q_{ij} is a linear combination of input p_{ij}s, resulting in a single embedding of each datapoint.

4 Experiments

We demonstrate our approach with multipoint t-SNE. In experiments on small datasets, we compare multipoint t-SNE, multi-map t-SNE, mixture SNE and mixture t-SNE. The last one is created by switching Gaussian neighborhoods to Student's t neighborhoods in the distribution Q of mixture SNE. On larger datasets, we compare multipoint t-SNE with plain t-SNE, both using Barnes-Hut approximations.

In all experiments the datasets were reduced to $d = 50$ dimensions with Principal Component Analysis, and perplexity of Gaussians in the data space was set to 50. We optimize with Adam [8] using learning rate $\eta = 3.5$. In our experience it often works better than gradient descent with momentum. In subsequent stages of optimization, we reuse previous values of moments m, v. For mixture approaches, we found gradient descent with momentum and L-BFGS-B [3] to perform slightly better.

As multipoint t-SNE adds variables, it takes longer to converge, and we run the optimization for total of 2000 to 4000 gradient updates. In all experiments p_{ij} values were multiplied by 4 (or 12 in the case of large datasets [9]) for the first 250 gradient updates. In all experiments the datapoints were replicated in four stages, with replication of the top scoring 25% during each and an immediate cleanup of those $y_k \in Y_i$ with low probability mass $r_{ki} < 0.2$.

4.1 Pairwise Relationships Datasets

We adopt the word similarities dataset [13] (size 5000×5000) and the NIPS co-authorships dataset (size 1418×1418) from studies on multiple maps t-SNE and aspect maps [5,11]. Both datasets come in form of square matrices of pairwise relations. The former describes word similarities judged by human volunteers, the latter is a co-occurrence matrix of paper authors of two or more contributions accepted to NIPS in years 1988–2009. Quality of the embeddings can be scored with mean precision [11] or mean precision/recall [16] of k-nearest neighbors between $\mathbf{X}$ and $\mathbf{Y}$. For every datapoint in $\mathbf{X}$, k closest datapoints in $\mathbf{X}$ and $\mathbf{Y}$ are selected based on their Euclidean distances. For fairness of comparison, we select them based on their P and Q values. Both approaches are equivalent for SNE and t-SNE. However, the latter scores higher methods where each q is composed of many datapoints in Y, for instance through weights in mixture SNE.

Multipoint t-SNE achieves higher mean precision for $k < 15$ (Figs. 2a and b). This is due to better modeling of small neighborhoods. Moreover, 2-D visualizations with multipoint t-SNE are better in this regard than 3-D visualizations with t-SNE. Figure 3 shows a portion of the word association embedding, namely selected copies of datapoints and their 3-NNs. Words like *move*, *break*, *free* end up in different, natural contexts, or even among different parts of speech like *reason*.

4.2 Vector Datasets

Figure 4 shows embeddings of the COIL-20 dataset [14] (size 1440 × 16384). The dataset consists of B&W photos of a number of small objects, taken under different angles. Typically t-SNE embeds the images as closed, separated loops, sometimes torn or incomplete. In multipoint t-SNE, the lines between replicas connect certain areas of the embedding and pointing out imperfections, making for a more informative visualization.

Next, we visualize 10000 Google News dataset word embeddings[2] [12] of the most frequent words[3] and compare mean precision of multipoint t-SNE and t-SNE ($d = 2$ and $d = 3$), all using Barnes-Hut approximation with $\theta = 0.2$. Figure 2c shows the results. Multipoint t-SNE better preserves small neighborhoods in 2-D embeddings than t-SNE in 2-D and 3-D, outperforming them by a large margin.

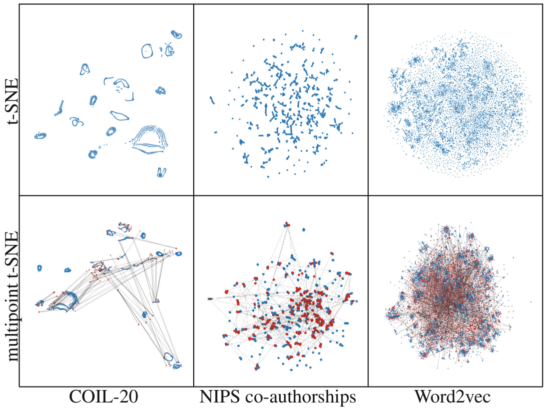


Fig. 4. Embeddings constructed with t-SNE and multipoint t-SNE. Copies of datapoints are shown in red. Lines connect copies of the same datapoints (only 10% of lines shown to avoid clutter). Copies of datapoints allow more clusters to form and reduce the number of satellite datapoints on the edges of embedding, by placing them in denser areas. Best viewed in color.

[2] Taken from https://code.google.com/archive/p/word2vec/.

[3] We exclude top 100 as mostly stop words or containing non-letter symbols.

Figure 4 compares t-SNE and multipoint t-SNE embeddings. Lines connecting datapoints with their copies form a dense network which connects remote areas of the embedding. Additional copies of datapoints allow small, isolated groups to form. They also seem to reduce the number of evenly spaced satellite datapoints on the verge of the embedding, which typically do not fit elsewhere.

5 Discussion

We have introduced the multipoint extension to NE methods and showed its effectiveness with multipoint t-SNE. It naturally extends NE methods through sensible addition of copies of certain datapoints during optimization. Moreover, the extension does not raise the complexity, and we have implemented the Barnes-Hut approximation.[4] The approach allows to preserve small neighborhoods better than multiple maps and mixture approaches, and even better than 3-D t-SNE. Embeddings constructed with multipoint t-SNE have fewer misplaced datapoints on the edge of the embedding, as well as crisper clusters datapoints inside. Datapoints are not weighted, but weights may be derived through analysis of their neighborhoods. Natural directions of future work include analysis of optimization stages and out-of-sample extension, i.e., embedding points not seen during training.

Acknowledgments. Adrian Lancucki was supported by local grant 0420/1710/16, and National Center for Research and Development (Poland) grant Audioscope (Applied Research Program, 3rd contest, submission no. 245755). Jan Chorowski was supported by National Science Center (Poland) grant Sonata 8 2014/15/D/ST6/04402. The authors also thank WCSS for computing power.

References

1. Barnes, J., Hut, P.: A hierarchical O(N log N) force-calculation algorithm. Nature **324**(6096), 446–449 (1986)
2. Xie, B., Yang, M., Tao, D., Huang, K.: m-SNE multiview stochastic neighbor embedding. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **41**(4), 1088–1096 (2011)
3. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
4. Carreira-Perpinán, M.A.: The elastic embedding algorithm for dimensionality reduction. In: ICML, vol. 10, pp. 167–174 (2010)
5. Cook, J., Sutskever, I., Mnih, A., Hinton, G.E.: Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics, pp. 67–74 (2007)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR. vol. 2, pp. 1735–1742. IEEE (2006)

[4] Source code to reproduce the experiments available publicly at https://github.com/alancucki/multipoint_tsne.

7. Hinton, G.E., Roweis, S.T.: Stochastic neighbor embedding. In: Advances in Neural Information Processing Systems, pp. 833–840 (2002)
8. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. van der Maaten, L.: Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)
10. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
11. van der Maaten, L., Hinton, G.: Visualizing non-metric similarities in multiple maps. Mach. Learn. **87**(1), 33–55 (2012)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Nelson, D.L., Mcevoy, C.L., Schreiber, T.A.: The University of South Florida Word Association, Rhyme, and Word Fragment Norms (1998)
14. Nene, S.A., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-20). Technical report CUCS-005-96 (1996)
15. Yang, Z., Peltonen, J., Kaski, S.: Scalable optimization of neighbor embedding for visualization. In: ICML, vol. 28, pp. 127–135 (2013)
16. Yang, Z., Peltonen, J., Kaski, S.: Optimization equivalence of divergences improves neighbor embedding. In: ICML, pp. 460–468 (2014)