ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Effects of sparseness and randomness of pairwise distance matrix on t-SNE results

Eli Parviainen

BECS, Aalto University, Helsinki, Finland

Abstract. We apply ideas from random graph theory to sparse pairwise distance matrices in dimension reduction. We use matrices with some short and some randomly chosen distances, and study effects of matrix sparseness and randomness on trustworthiness and continuity of t-SNE visualizations. The existing works have either concentrated on matrices with only short distances, or implemented heuristics with mixed distances without explaining the effects. We find that trustworthiness generally increases with randomness, but not without limit. Continuity is less affected, but drops if matrices become too random. Sparseness has little effect on continuity, but decreases trustworthiness. Decrease in quality appears sublinear, which suggests that sparse t-SNE could be made subquadratic in complexity without too much effect on quality.

1 Introduction

Many dimension reduction methods use a pairwise distance matrix, and this significantly restricts the number of data points which can be embedded. It would be tempting to reduce computational burden by using only part of the distances.

Since importance of local distances is widely recognized in dimension reduction research [1, 2, 3], a natural approach to sparsifying is to keep the local, short distances and discard the long distances. However, purely local approaches to multidimensional scaling are inadequate, since removing the large distances can decrease quality more than removing the short ones [4]. This finding may have resulted in generally pessimistic views regarding sparsification of distance matrices, judging from the small number of works on this topic. Some work on partially filled matrices or naturally sparse data has been done [5, 6, 7], but most sparse dimension reduction methods are based on landmarks [8, 9], a different approach, where only a subset of points is embedded, and locations for other points are inferred from the landmarks locations.

We show that part of distances can be discarded without too much impact on quality. The sparse pairwise matrix should contain both short-scale and long-scale distances. Results from such a matrix can give much better embedding results than a purely local approach.

We parameterize the mixture of local and global distances borrowing ideas from random graph theory [10]. We see the sparse distance matrix as a graph, with a link between points if the corresponding distance is present in the matrix. Some links between near neighbors are always kept, and some links point to randomly chosen, possibly far-away, neighbors. Randomness of the graph determines, how big proportion of the links is chosen randomly. We study connection

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

between randomness level and embedding quality, measured by trustworthiness and continuity.

We start by introducing our methods and data sets in Section 2. Experiments for studying randomness and sparseness are explained in detail in Section 3. Section 4 concludes the work.

2 Methods and data

Stochastic neighbor embedding with t-distributions (t-SNE) is a relatively new dimension reduction method developed by van der Maaten and Hinton [11]. T-SNE works by placing Gaussian (in the data space) or t-distributed (in the embedding space) kernels at data points, and using them to determine probabilities for points being neighbors. Embedding points are found so that neighborhood probabilities in the data space match those in the embedding space as closely as possible.

The original t-SNE determines kernel widths from a perplexity parameter, which is essentially a soft number of nearest neighbors. This means that widths can differ for different points. It is not obvious what is the correct way to choose kernel widths, or how to keep results comparable, when different neighbor links are used in different runs. To keep things simple, we use ϵ-neighborhoods, using same fixed kernel width everywhere. This lowers overall quality, but is suitable for our experimental setting, since we focus on randomness effects and not on absolute quality. Developing sparse t-SNE into a usable method, complete with a way of choosing kernel widths in the sparse setting, is left for future work.

Data sets We use three data sets in the experiments. We will only show results for the MNIST data set[1], but qualitatively similar results (with somewhat smaller sample sizes and smaller parameter ranges) were obtained using two more data sets, the USPS and Yale data[2]. MNIST and USPS are benchmark data sets of handwritten digits, and Yale data has face photographs.

Trustworthiness and continuity We measure visualization quality by two criteria, trustworthiness and continuity [13]. They count numbers of correct and incorrect points in a neighborhood of a given size, and also take into account how far from the neighborhood (in rank distance) a wrong point came from, or how far it should have been placed. Trustworthiness and continuity values at 5-neighborhoods are used.

3 Experiments

In the following, we identify data points with graph nodes, and pairwise distances with links between nodes. The number of data points is N, and L is the average

[1] MNIST and USPS, `http://www.cs.nyu.edu/~roweis/data.html`.

[2] Extended Yale Face Database B [12],
`http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html`

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

number of links from a point to other points. Relationship of L and N determines *sparseness* of a matrix.

We create a continuum of sparse matrices. We start with a matrix with links from a point to its L nearest (in the sense of Euclidean distance) neighbors. A fraction of these links is rewired to point to randomly chosen nodes, making sure that the total number of links in the matrix remains the same. This is a similar construction as Watts and Strogatz used in their celebrated study [10] on small world graphs. The rewiring fraction, or *randomness*, is denoted with R. Individual nodes can have different numbers of neighbors after rewiring, but average number of neighbors is L, as the total number of links is fixed.

T-SNE cost is evaluated as usual, but only for nonzero entries of the sparse similarity matrix. This changes the time complexity from $O(N^2)$ to $O(NL)$.

In these experiments, we start with a full matrix, which we use to find the nearest neighbors. The sparse matrices are created by discarding entries from the full matrix. This is a systematic method, which should not introduce unexpected side effects, and which is therefore suitable for our experimental setting. In real use, a full matrix would not be used as an intermediate step. Instead, the nearest neighbors would be found with an approximate method, or the probability of linking two points would be based on their distance.

3.1 Effect of randomness

In the first experiment, we look at trustworthiness and continuity as function of randomness. We repeat the experiment for several sparseness levels, by fixing L and letting N change.

The regular connection pattern $R = 0$ % gives poor results. Trustworthiness levels around 0.5 are typical when a random set of points is compared to the original data. This is seen both in the 2D visualization in Fig. 1 an in the numerical results in Fig. 2.

When R increases, results get more trustworthy. Figure 2 shows an upward sloping line, and Fig. 1 has more clearly separated classes. When matrices are not very sparse (cases N=2000 and N=4000), the completely random matrix gives best results. When matrices are sparser, best results are obtained with $R = 80$ %. Continuity, on the other hand, stays at roughly the same level or only slightly increases, up to $R = 80$ %. When the matrix is made completely random, continuity drops. Matrix sparseness affects the steepness of the drop.

These observations are in line with the nature of the two criteria. Achieving high trustworthiness requires global information, i.e. long distance links, because different neighborhoods must be kept separate in the visualizations. A regular matrix has only local links. Each random link is a potential long distance link; therefore increasing proportion of random links increases trustworthiness. Continuity, more easily created by local constraints, is much less sensitive to matrix randomness. As long as points from a neighborhood stay together the result is continuous; it does not matter if the points mix with other neighborhoods.

High quality with $R = 80$ % and a drop with $R = 100$ % brings up an important point regarding the connection patterns: although global links are impor-

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

tant, also enough local information must be provided. Missing local information is suggested by different behavior of the continuity and trustworthiness criteria. Continuity, associated with local constraints, drops very clearly, whereas the more global trustworthiness is not affected as drastically.

Existence or steepness of the drop depends on matrix sparseness. When the matrix is not very sparse, some random links will necessarily be between near neighbors. This provides some local information, even in completely random matrices. The sparser the graph, the smaller the probability of hitting a pair of near neighbors by chance. Therefore, effects of having an insufficient number of local links are seen more clearly in sparser matrices.

3.2 Effect of sparseness

Our second experiment fixes randomness and lets the sparseness change. We use randomness $R = 80$ %, which gave the best results in the first experiment.

Our most important discovery is the form of dependence of trustworthiness on sparseness in Fig. 4. Trustworthiness increases rapidly in the beginning, but growth slows down when L approaches N. For $N = 3000$, using 2000 neighbors gives the same trustworthiness as the full matrix. Although numerical results for $L = 1000$ are lower, in the illustration in Fig. 3 it is difficult to visually tell the $L = 3000$ and $L = 1000$ results apart. Lines for larger N in Fig. 4 have not reached a constant level within the L range used, but also their growth decelerates with growing L. This suggests that the number of neighbors needed for certain quality can be a sublinear function of N.

This has promising practical implications. A sparse dimension reduction method is really useful only, if a reduction in computational complexity is achieved. A method scaling $O(NL)$ is still quadratic, if we must use an L which a linear function of N. Figures 3 and 4 suggest that for satisfactory results with sparse t-SNE, L may grow more slowly than N. This makes the overall complexity subquadratic.

4 Conclusions

The idea of reducing computation times of dimension reduction methods by discarding part of pairwise distances has been considered in the literature. The natural approach is to favor local accuracy by using only distances to nearest neighbors. It has been shown, however, that large distances have a large impact on quality, and therefore cannot be discarded. This has resulted in preconception that creating sparse dimension reduction methods by keeping all points but only part of distances would be next to futile.

In this work we brought new light on this overly simplified view. We pointed out that keeping the nearest neighbors is not the only way to sparsify a distance matrix. Indeed, it is much more advisable to replace some local links with random links. Proportion of random links used has a clear impact on results.

We discussed the effect of randomness, changing from a matrix with purely local links to a matrix with purely randomly chosen neighbors. Effects on trust-

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

worthiness and continuity of visualizations were measured. We found a general connection between increasing randomness and increasing trustworthiness. Completely random matrices do not work well, however. Trustworthiness is lower for completely random than highly random matrices. Continuity stays high until high randomness levels, but drops sharply for completely random matrices. This effect is reduced if the matrix is not very sparse. In such case, also random links will sometimes be between near neighbors, which helps to maintain high continuity.

Sparseness seems to have little effect on continuity of visualizations. Trustworthiness is affected, but we made a promising observation that the number of links needed for certain quality seems to be a sublinear function of N. This makes the overall operation subquadratic, an improvement over the quadratic full-matrix version of t-SNE. The lower complexity with not too much decrease in quality could stretch the feasible range of t-SNE use from thousands of points to at least tens of thousands.

References

[1] Sam Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

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

[3] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[4] Jed Graef and Ian Spence. Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, 86(1):60–66, 1979.

[5] Matthew Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proc. of IEEE Visualization '96*, pages 127–132, 1996.

[6] Manuel Martín-Merino and Alberto Muñoz. A new Sammon algorithm for sparse data visualization. In *Proc. of ICPR*, 2004.

[7] Lisha Chen and Andreas Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104:209–219, 2009.

[8] Vin de Silva and Joshua Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

[9] Kilian Q. Weinberger, Benjamin D. Packer, and Lawrence K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proc. of AISTATS*, 2005.

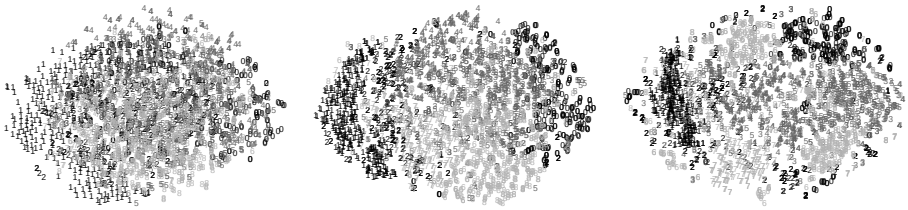
[10] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, 1998.

[11] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[12] A. S. Georghiades, P. N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[13] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proc. of ICANN*, volume 2130 of *LNCS*, pages 485–491, 2001.

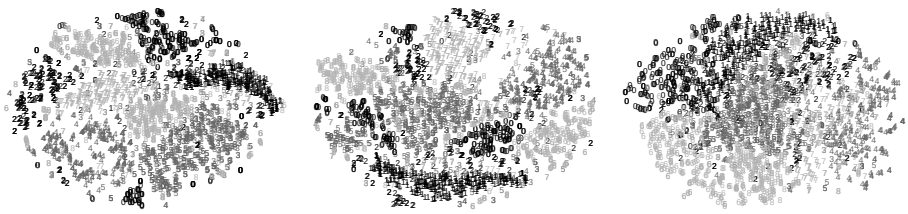
ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.



(a) R=0 % (b) R=80 % (c) R=100 %

Fig. 1: 3000 MNIST points, with $L = 450$ and varying R. Colors denote digit classes.

Fig. 2: Effect of randomness level, for different N and L=200.



(a) L=3000 (full matrix) (b) L=1000 (c) L=300

Fig. 3: 3000 points of MNIST data, with different average neighbor numbers. Randomness level is 80 %.

Fig. 4: Effect of average number of neighbors, for different N, with $R = 80$ %.