Distance Transformation for Effective Dimension Reduction of High-Dimensional Data

Eniko Szekely and Eric Bruno and Stephane Marchand-Maillet

University of Geneva, Department of Computer Science
7 Route de Drize, Battelle Bat. A
1227 Geneva, Switzerland

Abstract. In this paper we address the problem of high-dimensionality for data that lies on complex manifolds. In high-dimensional spaces, distances between the nearest and farthest neighbour tend to become equal. This behaviour hardens data analysis, such as clustering. We show that distance transformation can be used in an effective way to obtain an embedding space of lower-dimensionality than the original space and that increases the quality of data analysis. The new method, called High-Dimensional Multimodal Embedding (HDME) is compared with known state-of-the-art methods operating in high-dimensional spaces and shown to be effective both in terms of retrieval and clustering on real world data.

1 Introduction

The difficulty of analysing high-dimensional data, mainly at global level, is a consequence of the relative equidistancy among distances. The effect of high dimensions on pairwise distances was investigated in [1], [2], [5], [8]. Results in [5] report that the distance from a query point Q to the nearest neighbour ($DMIN$) and farthest neighbour ($DMAX$) tend to become equal in ratio given that the condition is fullfilled:

Theorem 1. *If*

$$\lim_{D\to\infty} var\left(\frac{dist_D(\mathbf{X}_D, Q)}{E[dist_D(\mathbf{X}_D, Q)]}\right) = 0 \tag{1}$$

Then for every $\varepsilon > 0$ there exists $D > D_0$ such that:

$$\lim_{D\to\infty} P[DMAX \leq (1+\varepsilon)DMIN] = 1 \tag{2}$$

where $dist_D(\mathbf{X}_D, Q)$ is the distance between data points $\mathbf{X}_D$ and the query point Q and $E[dist_D(\mathbf{X}_D, Q)]$ is the expected value of the distances. The proof can be found in [5]. The problem of high-dimensional spaces was introduced by Bellman in 1961 and was called the “curse of dimensionality” [4].

Further research showed in [8] that even when distance tend to be equal in ratio, their absolute difference $DMAX - DMIN$ does not necessarily go to zero for all distance metrics. The above observations show that it is important to consider the behaviour of distances not only in terms of ratio, but of absolute difference too.

Another important aspect is that real world data often lies on complex manifolds with the intrinsic dimensionality lower than the original space. This is the typical case where dimension reduction methods can be used with success to find the low-dimensional embedding space. Many methods have been proposed to reduce the dimensionality and they can be roughly categorized into two main classes: global and local methods. Local methods aim to recover the local structure of the data - by mapping nearby points to nearby points - and hope to recover the global structure from local fits, while global methods try to recover the global topology at all scales - by mapping nearby points to nearby points and far away points to far away points. Local methods have generally proven to be more effective in many real cases. However they do not totally recover the global structure of the data. Moreover, in many datasets, data naturally follows a multimodal distribution, being organised into clusters. We therefore propose in this paper a dimension reduction method (High-Dimensional Multimodal Embedding) that combines local and global information to obtain an effective embedding. For this we perform a transformation of distances in order to avoid the equidistancy in ratio. The transformation will prove to be necessary for proper preservation of certain global properties, like clusters.

2 Related work

Dimension reduction is mainly employed in order to improve the analysis of the data and to simplify its processing. It is mostly based on the assumption that the data lies in subspaces of lower dimensionality than the original space. It is hoped that there exists a meaningful intrinsic dimensionality (d) of the data that is smaller than the original dimensionality (D), $d \ll D$.

Principal Component Analysis (PCA) is the most employed dimension reduction method in practice. It is a linear transform whose objective is to capture as much as possible from the variance in the data, but it is not designed to cope with non-Gaussian distributions and even less with clustered data. Multidimensional Scaling (MDS) [6] is a global approach to pairwise distance preservation, optimising a stress function over all distances. A widely used variant of MDS is Sammon Mapping [13] that increases the importance given to smaller distance values by using a self-normalisation procedure. As in high dimensions distances tend to be equal, global methods that use distances directly (e.g. MDS, Sammon Mapping) fail in preserving any agglomeration of data (e.g. clusters). The resulting embeddings are generally spherical.

The *manifold assumption* has been the key for some of the main algorithms developed recently (e.g. Isomap [15], Locally Linear Embedding (LLE) [12, 14]). The high-dimensional data is assumed to lie on a *manifold* of lower dimensionality than the original space (e.g. Swiss Roll, a 2D-manifold embedded in a 3D-space). A manifold is a space in which local neighbourhoods ressemble a Euclidean space, but the global structure is generally more complex. The local

Euclideanity of manifolds only justifies the use of the Euclidean distance for local neighbourhoods. Larger distance values may be estimated using “geodesic” distances computed *along* the manifold and not *through* the manifold. Unfortunately, whereas the geodesic distance captures well the shape of the manifold, it is not designed to discriminate between clusters.

A global method (e.g. MDS, Isomap) tries to recover the global structure of the data. On the contrary, a local method (e.g. LLE, Laplacian Eigenmaps) aims to recover the local structure of the data and hopes to recover the global structure from local fits. LLE preserves the linear reconstruction of a point from its neighbours. Laplacian Eigenmaps (LE) [3] embed points in the low-dimensional space with respect to the eigenvectors of the Laplacian matrix. SSammon [10] modifies Sammon Mapping so that only the first neighbours of a point contribute to the stress function. DD-HDS [9] uses a sigmoid function, similar to Curvilinear Component Analysis [7], to also allow only the shorter distances to contribute to the error. Among all the above-cited dimension reduction methods, Laplacian Eigenmaps is the only method tailored for clustered data, since it uses Laplacian graphs, as in spectral clustering, for the embedding.

3 High-Dimensional Multimodal Embedding

Local neighbourhoods can carry important information especially when data lies on manifolds. This information can be captured by absolute differences between distances that was shown to be meaningful even in high dimensions for certain distance metrics, like the Euclidean distance [8]. However local fits can not always capture the global structure of the collection, that becomes particularly useful in clustering tasks. Moreover, at global level, distances tend to be equal in ratio, which hardens cluster discrimination. To avoid the negative effects of the curse of dimensionality [4], we propose to apply a transformation of distances, prior to the embedding in the low-dimensional space. Distances are transformed by a step scaling function, such that distances between similar points are scaled down by a scaling factor. Two points will be considered similar if they are neighbours in the original space. After the distance transformation, a distance-based embedding is used to find the low-dimensional space. In this new space, distance measurements become more meaningful and thus allow for improved further data analysis.

3.1 Model

Let $\mathcal{X}$ be a set of N data points $\mathbf{x}_i$ in the original high-dimensional space $\mathbb{R}^D$. We search for a low-dimensional space $\mathbb{R}^d$ $(d \ll D)$ where points are to be embedded.

In high dimensions, data often lies on manifolds. On manifolds two points are consider similar if they are neighbours. In the following, we approximate the sought manifold by a k-nearest neighbour graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_k)$ built using the Euclidean distance.

Given graph relationships, two points become similar if they are neighbours in $\mathcal{G}$, i.e. they are connected in $\mathcal{G}$.

Let $\mathcal{P} = \mathcal{X} \times \mathcal{X}$ be the set of all pairs of points $(\mathbf{x}_i, \mathbf{x}_j)$ for $\mathbf{x}_i \neq \mathbf{x}_j \in \mathcal{X}$ and $\mathcal{P}_1$ be the subset of similar points:

$$\mathcal{P}_1 = \{(\mathbf{x}_i, \mathbf{x}_j) | \ \exists \ e_{ij} \in \mathcal{E}_k\} \tag{3}$$

and

$$\mathcal{P}_2 = \mathcal{P} \setminus \mathcal{P}_1 \tag{4}$$

The matrix of distances between the N points is given by $\mathit{\Delta} = (\delta_{ij}) \in \mathbb{R}^{N \times N}$, where δ_{ij} is the Euclidean distance in the original space between point $\mathbf{x}_i$ and point $\mathbf{x}_j$.

Given the similarity between points, we want to scale down distances between points that are similar. Different transformation functions (linear or nonlinear) can be applied but here we will report results using the following step function f_s :

$$\delta^*_{ij} = f_s(\delta_{ij}) = \begin{cases} \dfrac{\delta_{ij}}{\lambda}, & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_1; \\ \delta_{ij}, & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_2. \end{cases} \tag{5}$$

where δ^*_{ij} are the new transformed distances and λ is a scaling factor with $\lambda \geq 1$. Here, distances between similar points are scaled down with the constant scaling factor λ, while distances between points not considered as similar (not neighbours on the manifold/graph), remain unchanged.

The embedding of the points into a lower-dimensional space is performed on the transformed distances using Sammon Mapping. We chose Sammon Mapping as HDME seeks to preserve both small and large distance values as faithfully as possible. Hence, the presence of a weighting factor in the embedding function to escape from privileging large distance values is desirable. This is typically the case in the Sammon Mapping stress function:

$$E_{SM} = \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}} \tag{6}$$

where $\mathbf{y}_i, \ i = 1..N$ are the projected points in the embedded space ($\mathbf{y}_i \in \mathbb{R}^d$) and d_{ij} are the distances in the low-dimensional space between projected points $\mathbf{y}_i$ and $\mathbf{y}_j$. Replacing the original distances δ_{ij} in (6) with the scaled distances δ^*_{ij}, we obtain the embedding function for HDME:

$$E_H = \frac{1}{\sum_{i<j} \delta^*_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta^*_{ij})^2}{\delta^*_{ij}} \tag{7}$$

The presence of the non-linear weighting by $\dfrac{1}{\delta^*_{ij}}$ inside the sum naturally increases the importance of smaller distance values in the embedding, thus reinforcing the scaling performed on distances.

4 Experimental results

4.1 Data preparation and Evaluation Measures

We choose for the experiments the 20 Newsgroups dataset as it is a fairly high-dimensional dataset (44,764 dimensions). It contains approximately 20,000 articles from 20 different newsgroups divided into training (60%) and testing (40%). For the experiments, we use the data from the training set of 11,269 documents. The dataset is processed as follows: 1) stopwords are eliminated and stemming [11] is performed; 2) stems that appear too many times ($\geq$ 2000) are deleted. The documents are represented as feature vectors where each feature represents one stem. Thus the dimensionality of the original space is 44,764 dimensions (corresponding to the number of stems). Feature values are given by the frequency of appearance of stems in documents. In this feature space we normalize documents to unit length (to avoid giving too much weight to long documents) and we eliminate duplicate documents and documents with Euclidean norm equal to zero. The final collection contains 11,222 documents from 20 categories in the 44,764-dimensional space.

High-dimensional spaces suffer from the "curse of dimensionality" with negative impact on the quality of data analysis. In these spaces, operations such as clustering or retrieval often give poor results. We consider the ground truth to be provided by data labels. HDME is proposed as a method meant to improve data analysis through dimension reduction. Therefore, in the evaluation part, we will test results obtained with HDME against results obtained in the original space and with other dimension reduction methods. Since the tasks that suffer the most from the curse of dimensionality are global tasks, like clustering, for evaluation we choose measures that evaluate the global quality of the data, that is Mean Average Precision (MAP) and k-means purity. MAP is chosen as retrieval tasks become more and more present in real data applications. MAP combines Precision and Recall and is therefore sensitive to the entire ranking of the data. Thus, it represents a good indicator of the global topology of the collection. Clustering is one main application of data analysis. k-means was chosen for its wide utilisation, however we mention here that cluster shapes may not always be well approximated by the spherical Gaussianity of k-means. To evaluate the quality of the clusters, each cluster is assigned the label of the majoritary class. We then estimate the purity of the clusters, that is, the percentage of points that were assigned the correct (real) label. For each parameter values, k-means is run three times on the whole collection with k=20 (the number of categories). And the average purity over the three runs is reported in the experiments. Reporting the mean over multiple runs of the algorithm avoids falling into extreme cases.

In Table 1 we show the MAP obtained when eliminating certain stems according to the frequency of appearance in the dataset. One important observation is that the value of MAP is very low when no normalisation is performed on documents (first column). The highest value obtained is when we eliminate all stems that appear more than 2000 times. A few example of stems that appear more than 2000 times are 'write', 'about', 'other'. Still we see that results do not deteriorate drastically with the decrease in dimensions. Thus, choosing only those stems that appear more than 100 times and less than 2000 gives a MAP of 0.1916. This may be more appropriate when a too high dimensionality is preferably to be avoided.

Dims	all(no norm.)	all(with norm.)	<5000	<2000	5-2000	10-2000	50-2000	100-2000
MAP	0.0718	0.1706	0.1946	0.1982	0.1980	0.1979	0.1953	0.1916

Table 1. Mean Average Precision in the original space for the 11,269 documents.

HDME relies on the nearest neighbour graph built in the original space (44,764 dimensions). Table 2 gives the accuracy of the k-nearest neighbour for different values of k. The values show that despite pessimism concerning the meaningness of the nearest neighbour, real data often displays high quality of local information. This is generally due to the underlying patterns that govern real data applications, and that have low intrinsic dimensionality.

kNN	1	2	5	10	20	30	40	50
Acc.	83.59	79.25	79.34	78.33	77.09	76.23	75.37	74.86

Table 2. Accuracy of nearest neighbour in the original space.

4.2 Experiments

In the first experiment (Figure 1) data is embedded in a 2-dimensional space using various dimension reduction methods. Mean average precision and k-means accuracy for the 11,222 documents from Newsgroups were estimated. MAP is estimated for the original space with the euclidean and the geodesic distance, for HDME, LE, PCA and Sammon Mapping. k-means accuracy is estimated in the original space, for HDME, LE, PCA and Sammon Mapping. PCA is chosen as it is the most widely employed dimension reduction method in practice. Sammon Mapping is the baseline for HDME and LE gives the best results for clustered data, to our knowledge. The geodesic distance is chosen as it captures better than the euclidean distance the complex structure of high-dimensional data. However the geodesic distance often performs poorly in discriminating among data groups, especially in high dimensions. We vary $k = 1..150$ (the number of nearest neighbours used to build the graph) for multiples of 10. For HDME we also vary the scaling factor λ.

As a first observation, local-based methods (HDME, LE and geodesic) generally perform better. In terms of clustering accuracy, HDME performs fairly the same as the original space. Still, 2D spaces are very limiting, especially in our case of a complex dataset with 20 classes and more than 10,000 elements. The clear advantage of 2D spaces is that they allow for data visualisation.

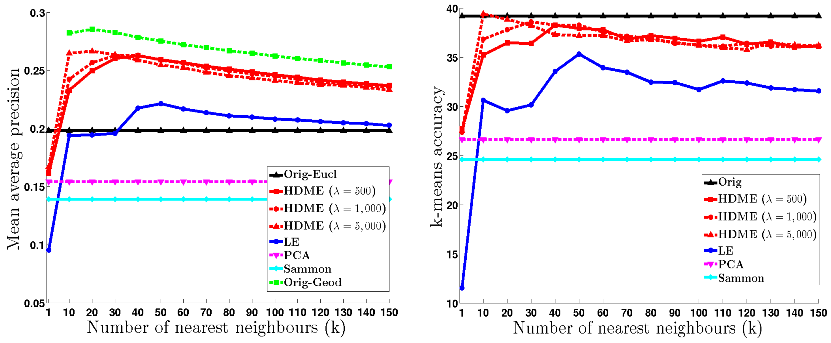


Fig. 1. MAP and k-means accuracy for the Newsgroups in a 2D space for different k (the number of nearest neighbours used to build the graph).

In the next experiment (Figure 2) we increase the dimensionality of the embedded space to 3, 4, 20, 50 and 100. For LE, HDME and the Geodesic Distance we show here results for $k = 20$ to build the nearest neighbour graph. Results show that dimension reduction can be very helpful for further data analysis as results are generally better than in the original space. Moreover HDME outperforms the other dimension reduction methods. We observe that an increase in the dimensionality of the embedded space (e.g. $d = 100$) results in lower quality of the analysis (with both MAP and k-means) for LE, PCA and Sammon, probably because a 100-dimensional embedded space starts to be influenced by the curse of dimensionality.

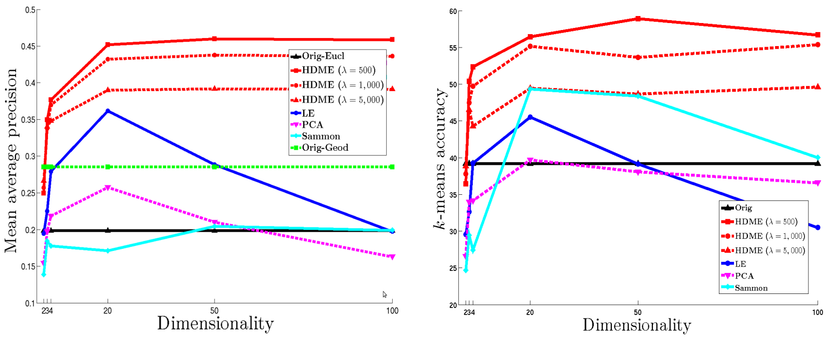
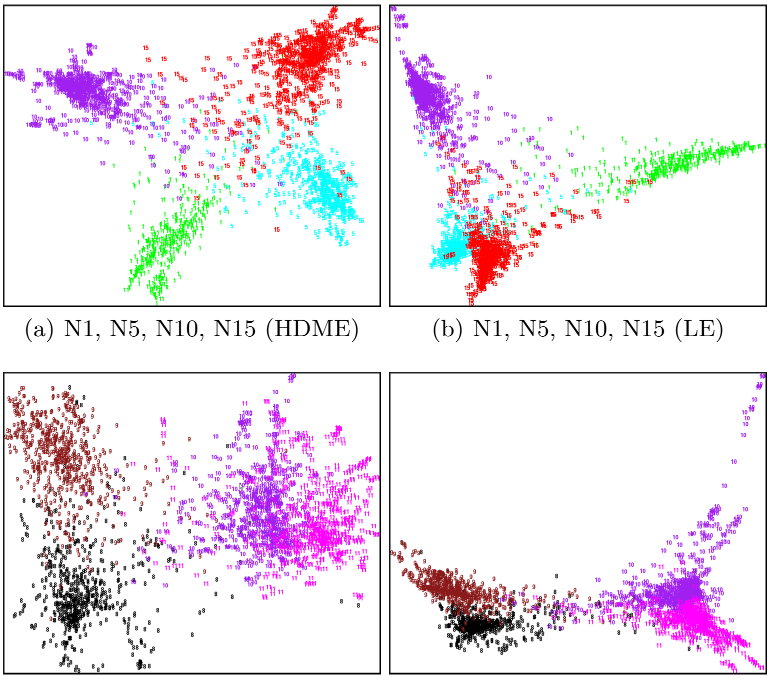


Fig. 2. MAP and k-means ($k = 20$) for different dimensionalities of the embedded space.

The last experiment gives two examples of mappings for the Newsgroups data: 1) the first 200 documents from each of the following categories: N8: rec.autos, N9: rec.motorcycles, N10: rec.sport.baseball, N11: rec.sport.hockey and 2) the first 200 documents from each of the following categories: N1: alt.atheism, N5: comp.sys.mac.hard-ware, N10: rec.sport.baseball, N15: sci.space. We apply the same processing as for the whole collection. Visualisation with HDME and LE of the embeddings are displayed in Figure 3 and MAP and k-means results in Figures 4 and 5. For the first dataset, HDME performs the best both in terms of retrieval and clustering, while for the second dataset, LE outperforms HDME in terms of retrieval, but not in terms of clustering. Various experiments performed on different datasets showed that LE performs well when the number of clusters is small as the eigenvectors computed on the Laplacian graph successfully represent the main directions corresponding to the real clusters, whereas HDME behaves particularly better than other methods when the number of real clusters increases.



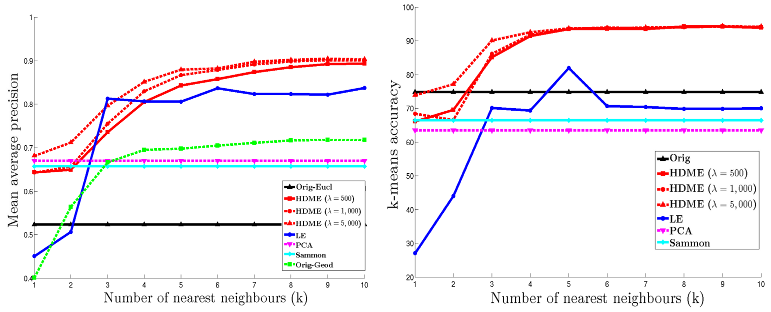
(a) N1, N5, N10, N15 (HDME)

(b) N1, N5, N10, N15 (LE)

(c) N8, N9, N10, N11 (HDME)

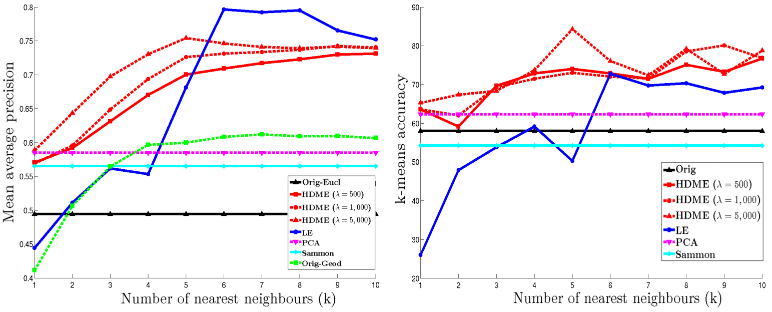
(d) N8, N9, N10, N11 (LE)

Fig. 3. The best mappings in terms of MAP with HDME and LE of the two subsets of the 20Newgroups data (a) HDME ($\lambda = 5000$, $k = 9$), (b) LE ($k = 10$), (c) HDME ($\lambda = 5000$, $k = 5$) and (d) LE ($k = 6$).



(a) MAP (b) k-means

Fig. 4. Results for the newgroups N1, N5, N10, N15.



(a) MAP (b) k-means

Fig. 5. Results for the newgroups N8, N9, N10, N11.

5 Conclusions

We showed in this paper that very high-dimensional data may be embedded in a different space that is not influenced by the curse of dimensionality and where distances start to be meaningful both at local and global level. In this transformed space data analysis can be performed with improved results. Existing dimension reduction methods are either local or global, whereas HDME wishes to preserve information at all levels, by using all distances. However, as distances are not meaningful in high dimensions at all scales, we perform a transformation/scaling based on the construction of a nearest neighbour graph. HDME, as a class of dimension reduction techniques tailored for structured datasets is valuable as a preprocessing for mining, retrieval, and visualisation.

HDME shows good performance in quality. The complexity in time is high, but once the embedding obtained (e.g. the 2D space) following computations become much faster than in the original space. Parameter estimation is important but not too sensible, as good results were obtained for a wide range of values. However, we plan to further investigate methods of automatically estimating the values of the parameter. The scaling factor depends on the dimensionality of the data, whereas the number of nearest neighbours depends on the number of documents in each class. Evaluation with global methods like MAP and k-means showed that HDME helps global analysis, like clustering and full ranked retrieval, while local evaluation measures like kNN, even is poorer than in the original space, still gave good results, showing that HDME still preserves well local information.

6 Acknowledgements

This work has been partly funded by SNF fund No. 200020-121842 in parallel with the Swiss NCCR(IM)2.

References

1. C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the International Conference on Database Theory*, pages 420–434, 2001.
2. C. C. Aggarwal and P. S. Yu. Redefining clustering for high-dimensional applications. *IEEE Transactions on Knowledge and Data Engineering*, 2002.
3. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.
4. R. Bellman. Adaptive control processes: A guided tour. Princeton University Press, 1961.
5. K. Beyer, J. Goldstein, R. Ramakrishnan, U., and Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, volume 1540, pages 217–235. Springer, 1999.
6. I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. 2005.
7. P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizig neural network for nonlinear mapping of data sets. In *IEEE Transactions on Neural Network*, volume 8. 1997.
8. A. Hinneburg, C. Aggarwal, and D. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th VLDB Conference*, 2000.
9. S. Lespinats, M. Verleysen, A. Giron, and B. Fertil. Dd-hds: a method for visualization and explorationof high-dimensional data. *IEEE Transactions on Neural Networks*, 18, 2007.
10. M. Martin-Merino and A. Blanco. A local semi-supervised sammon algorithm for textual data visualization. *Journal of Intelligent Systems*, 2008.
11. M. Porter. The porter stemming algorithm. `http://tartarus.org/~martin/PorterStemmer/`.
12. S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
13. J. W. Sammon. A nonlinear mapping for data structure analysis. In *IEEE Transactions on Computers*, volume C-18. 1969.
14. L. K. Saul, S. T. Roweis, and Y. Singer. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.
15. J. B. Tennenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.