ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Mapping without visualizing local default is nonsense

Sylvain Lespinats[1] and Michael Aupetit[1]

1- CEA, LIST, Multisensor Intelligence and Machine Learning Laboratory. F-91191 Gif-sur-Yvette, France.

Abstract. High-dimensional data sets are often embedded in two-dimensional spaces so as to visualize neighborhood relationships. When the map is effective (i.e. when short distances are preserved) it is a powerful way to help an analyst to understand the data set. But, mappings most often show defaults and the user is then led astray. According to this notion, a mapping should not be considered when its overall quality is not good enough. Many imperfect mappings can however be exploited by informing the user of the nature and level of defaults. In this work, we propose to visualize local indices trustworthiness and continuity for that purpose.

1 Mappings

When dealing with high-dimensional data the capacity to visualize the "spatial organization" is a powerful way to help an analyst to understand the data set. Its use provides a critical benefit for extracting information or can (for example) lead the user to the most suitable analysis method. This point is the main purpose of dimensionality reduction methods (also called mapping methods). Especially, Multi Dimensional Scaling (MDS) is the set of methods (including [2, 3, 7, 8, 10]) designed to embed data in a low-dimensional Euclidean space (the so called "output space") while preserving "maximally" the distances observed between data in the original space with a special attention to short distances. A main difference between methods can be found in the mean used to quantify the level of interest of each distance (that is to say, how much a given distance can be considered as "short").
Using dimensionality reduction on a given dataset involves an underlying hypothesis: it assumes that data lies on (or close to) a low-dimensional submanifold of the original space. Following this line, the dimensionality of such a manifold is named "intrinsic dimension" of the dataset and should be chosen as the dimension of the output space. To limit the distortions Multidimensional scaling method unfolds the submanifold and delivers it on the output space.

2 Defaults

If the dimensionality reduction is successful, the resulting map allows one to observe the local organization of the dataset. But, naturally, a good result cannot always be

reached. For example, if the intrinsic dimension of the dataset is clearly higher than the dimension of the output space, the map cannot be correct. Other features can lead to unavoidable defaults. Especially, when the topology of the manifold and of the output space are different (for example, if data is on the surface of a sphere and the output space is a plane, just as in Fig. 1). To avoid these problems, some authors [5, 6] propose to try mapping data onto non-planar output spaces (such as spheres or tori). It should however be noticed that visualizing such output spaces is less easy, and in practice we often ignore which topology fits the best with the analyzed data.
In general, when dealing with real data, mappings most often come with defaults. In such a case, the user is lead astray. According to this line of thinking, when the several existing methods for evaluating the overall quality of the mapping do not indicate a fair enough result, the map should not be considered. We however think that many imperfect mappings can still be exploited. Indeed, in real life, we are used to working with rough maps. The best known example is the planishere (Fig. 1).

Fig. 1: Earth map. The surface of the 3-dimensional sphere is displayed onto a 2-dimensional space. Dashed lines emphasize tears.

The Earth map is indeed an imperfect mapping. As anyone knows, several tears must necessary be introduced. In the present example, the Pacific Ocean and the poles are torn. This map shows then important defaults and, following the previous reasoning, ones could advice rejecting its use. However, everyone is used to (correctly) inferring geographic information from the planisphere. The difference between the planisphere and mapped data in a data-mining process is that we have knowledge about default of the Earth map while we have nothing similar when analyzing unknown data. If, similarly, we had information about nature and position of defaults, a reasonably useful inference from the map to original data would be possible.

3 Local evaluation of default

3.1 Our contribution

Even if several global methods have been proposed to evaluate the overall quality of a map, very few methods exist for evaluating it locally. In a previous paper [4], we set up for that purpose a new method based on cost functions of Curvilinear Component

Analysis [2] and Sammon's mapping [7]. In another paper, one of us proposed to visualize the default from a chosen point of view by coloring the map according to the distance from a data point selected by the user [1]. In the present paper, we propose to localize a classical measure of mapping default: trustworthiness and continuity [9, 10], which allows us to visualize the defaults straight in the relevant location on the map. So that, the map provides at the same time an estimation of the relative position of the data and the local quality of this estimation.

3.2 Local trustworthiness and local continuity

Venna and Kaski define trustworthiness and continuity [9, 10] as:

$$Trustworthiness_k = 1 - \frac{2}{N \times k \times (2N - 3k - 1)} \sum_i \sum_{j \in U_k(i)} (r(i,j) - k)$$

$$Continuity_k = 1 - \frac{2}{N \times k \times (2N - 3k - 1)} \sum_i \sum_{j \in V_k(i)} (r^*(i,j) - k)$$

where:
* k is the number of considered neighbors (and has to be chosen by the user).
* N is the number of data samples.
* $r(i,j)$ is the neighborhood rank (1 for the first neighbor, 2 for the second neighbor, etc…) of data point j from points i point of view in original space.
* $r^*(i,j)$ is the neighborhood rank of data point j from data point i point of view in output space.
* $U_k(i)$ is the set of data that are one of the k-nearest neighbors of data point i in the output space but not in the original space.
* $V_k(i)$ is the set of data that are one of the k-nearest neighbors of i in the original space but not in the output space.

We define local trustworthiness and local continuity as:

$$Trustworthiness_k(i) = 1 - \frac{2}{N \times k \times (2N - 3k - 1)} \sum_{j \in U_k(i)} (r(i,j) - k)$$

$$Continuity_k(i) = 1 - \frac{2}{N \times k \times (2N - 3k - 1)} \sum_{j \in V_k(i)} (r^*(i,j) - k)$$

where $Trustworthiness_k(i)$ (respectively $Continuity_k(i)$) corresponds to the local trustworthiness (respectively continuity) on data point i. Trustworthiness (respectively continuity) is the average of local trustworthiness (respectively local continuity) on every data point.

3.3 Coloring Voronoï cells

In order to make the set of N local measures construable, we propose to visualize them straight onto the map, by coloring the Voronoï cells of the points. The use of Voronoï cells has been proposed in [1] when it provides views easier than coloring

each data points. Each mapping is duplicated twice in order to display both of the indexes (local continuity and local trustworthiness).

4 Example

4.1 Several maps

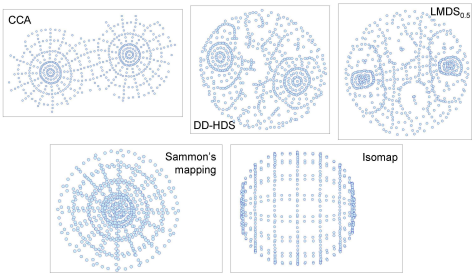

Fig. 2: 2-dimensional mappings obtained from 3-dimensional data using several dimensionality reduction methods: curvilinear component analysis (CCA) [2], Sammon's mapping [7], Isomap [8], data-driven high dimensional scaling (DD-HDS) [3], and local MDS with a λ-parameter = 0.5 ($LMDS_{0.5}$) [10].

Let us suppose that we face unknown data. Some dimensionality reduction methods are used to visualize the dataset (Fig. 2). But, how can we trust what we see? Even from such low dimensional original data (the original dataset is only 3-dimensional), inferring their structure from the different mappings is obviously cumbersome (or even useless); at least as long as we have not identified the mapping defaults first.

4.2 Local evaluation

The positions of gray points (Fig. 3) allow a comparison between global qualities of each mapping according to trustworthiness and continuity. In the present example, no map can be set aside: there is no map having lower trustworthiness and continuity than another one. Moreover, we cannot clearly decide which map to use from trustworthiness and continuity.

However, maps with Voronoï cells colored according to the previously presented method allow visualizing defaults locally. We can observe that Isomap and Sammon's mapping have provided maps with almost no fairly displayed areas: trustworthiness is very low most everywhere, and continuity is somewhat low in most of maps too. DD-HDS and Local MDS (with $\lambda = 0.5$) reached maps with large no-default zones but the repartition of defaults makes maps difficult to understand.

The result of CCA is more intuitive: the only defaults are tears which lie all around the map. We can trust the "continuious" area and their local topology. Data points lying on border of the map should then be connected. The proximity measure can then be used to explore the exact neighborhoods of points [1] and catch the local connected address.

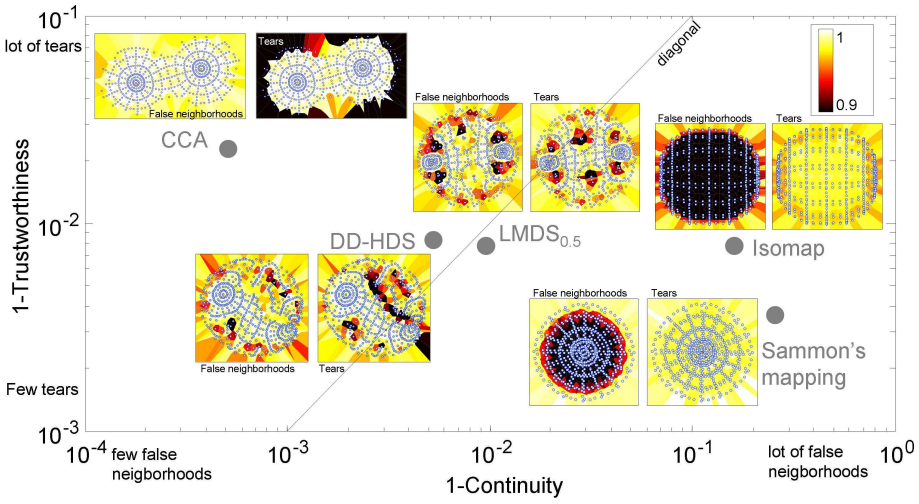

Fig. 3: five mappings of a 3-dimensional data (fig 4) projected on a 2-dimensional space. Each map corresponds to a grey circle in the (1-continuity) - (1-trustworthiness) graph (the lower the circle, the less false neighborhood – the more the circle is to the left, the less tear). At the side of each circle, the corresponding map is shown twice with local trustworthiness and local continuity displayed as color of Voronoï cells. Left inserts: local trustworthiness (dark for low trustworthiness, light for high trustworthiness), dark areas show false neighborhood zones. Right inserts: local continuity, dark areas show torn zones. Color-scales are the same for every insert and are shown at the top right corner.

4.3 **True data**

Original data points lie on the surface of a 3-dimensional sphere (Fig. 4).

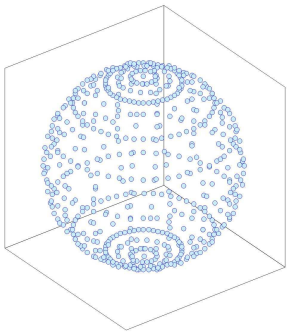

Fig. 4: Original 3-dimensional data.

5 Conclusion

In the present paper, we firstly highlight how critical the visualization of defaults is. Secondly, we adapted the "continuity" and "trustworthiness" measures proposed by Venna and Kaski [9, 10] to make them quantifying defaults locally. Lastly, we show level of default by coloring Voronoï cells rather than items to allow a better visualization of the results. We claim that visualizing defaults is necessary for a practical use of multidimensional scaling. At that time, we work on a tool to display the local trustworthiness and local continuity using a single map.

[1] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques" Neurocomputing, vol. 10, no. 7-9 pp. 1304–1330, 2007.

[2] P. Demartines and J. Hérault, Curvilinear component analysis: A selforganizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Netw., 8(1): 148–154, 1997.

[3] S. Lespinats, M. Verleysen, A. Giron and B. Fertil, DD-HDS: a tool for visualization and exploration of highdimensional data, IEEE Trans. Neural Netw., vol. 18, no. 5, pp. 1265-1279, 2007.

[4] S. Lespinats, and M. Aupetit, False neighbourhoods and tears are the main mapping defaults. How to avoid it? How to exhibit remaining ones? Quality Issues, Measures of Interestingness and Evaluation of data mining models, QIMIE09, April 2009, pp. 55-65. Bangkok (Thailand), 2009.

[5] J. X. Li, "Visualization of high-dimensional data with relational perspective map," Inf. Visualization, vol. 3, pp. 49–59, 2004.

[6] V. Onclinx, V. Wertz, M. Verleysen, Nonlinear data projection on non-Euclidean manifolds with controlled trade-off between trustworthiness and continuity Neurocomputing, Elsevier, Vol. 72, Issues 7-9, pp. 1444-1454, 2009.

[7] J. W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput., vol. C-18, no. 5, pp. 401–409, 1969.

[8] Tenenbaum J.B., de Silva V., and Langford J.C., A global geometric framework for nonlinear dimensionality reduction, Science, vol. 290, pp. 2319–2323, 2000.

[9] J. Venna and S. Kaski, Neighborhood preservation in nonlinear projection methods: An experimental study, In G. Dorffner, H. Bischof, and K. Hornik, editors, Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, pp. 485-491, Berlin, 2001. Springer.

[10] J. Venna and S. Kaski, Local multidimensional scaling, Neural Networks, vol. 19, no. 6-7, pp. 889-899, 2006.