A DIMENSIONALITY REDUCTION APPROACH FOR THE VISUALIZATION OF THE CLUSTER SPACE: A TRUSTWORTHINESS EVALUATION

Andreea Griparis [(1)], *Daniela Faur*[(1)] *and Mihai Datcu* [(1)(2)] *

[(1)]University *Politehnica* of Bucharest, UPB,
Applied Electronics and Information Engineering, Bucharest, Romania
[(2)]German Aerospace Center, DLR, Oberpfaffenhofen, Germany

ABSTRACT

The data mining systems solve the problem of handling Earth Observation archives counting on a feature vectors based description of the data. Increasing the dimensionality of the feature vectors would offer an effective perspective of the dataset's content. The modern systems provide visual exploration of data projecting their high-dimensional feature space in a 3-D space. The dimensionality reduction methods represent the main way to achieve such representation. Several dimensionality reduction methods have been proposed to identify the mapping, bot not all of them retain the same dataset properties. In order to compare their performance, the development of formal measures like "Trustworthiness" or the measures based on Co-ranking matrix was mandatory. These measures objectively evaluate the similarity between the structure detected in the original and the reduced space. In this paper we evaluate six dimensionality reduction methods using "Trustworthiness" and "Continuity" measures. In this regard three datasets have been used: an artificial one and two remote sensing datasets. Each of them have been described by a high-dimensional feature space.

***Index Terms*—** dimesionality, reduction, visualization, evaluation, trustworthiness, continuity

1. INTRODUCTION

The remote sensing sensors provide Terabytes of Earth Observation (EO) images daily. To handle these archives, a feature extraction process is compulsory. Descriptors used for data representation result in a high-dimensional feature space. Usually, the interpretation of datasets based on their feature space requires methods for multidimensional data visualization such as parallel coordinates technique [1] or dimensionality reduction (DR) methods.

State of the art literature considers the DR methods as a solution for various issues such classification, visualization and compression of high-dimensional data. A great number of DR methods have been considered in going from the multidimensional space to the low-dimensional one. Naturally, most of the DR techniques are linear but, due to the need to handle more complex, nonlinear data, a consistent number of nonlinear algorithms have been suggested. Due the fact that low-dimensional space cannot faithfully represents all relations between datasets items in the high-dimensional space, it is difficult to decide which relations should be preserved. Therefore, there are several different visualizations for a given dataset according to the objectives of DR methods. For the researchers, it is a challenging task to decide which method is suitable for the current application. Accordingly to this, graphical or quantitative methods was developed to evaluate the DR techniques. Graphical methods consist in visualization of low-dimensional data identifying the proportion of the data topology conserved by the algorithm. Mostly, the quantitative measures are related to specific objectives of further data analysis e.g. classification.

Pearson' correlation R and Spearman's Rho are the measures mentioned in [2] aiming to asses the performance of seven DR methods applied on two image datasets captured by a robot. Motivated by the fact that relationships between dataset items can be given by their distances from the rest of the data, different measures are computed based on the dataset co-ranking matrix [3] [4]. Other two quantitative measures, described by "Trustworthiness", the proportion of points nearest neighbor and "Continuity" were used in [5] to evaluate the low-dimensional representation of five artificial and five real datasets. The artificial datasets are represented by the Swiss roll dataset, the helix dataset, the twin peaks dataset, the broken Swiss roll dataset and the high-dimensional dataset while a subset with 5000 handwritten digits from MNISTS dataset, the drug discovery dataset (HIVA), and three image datasets (COIL20, NiSIS, ORL) account for the five real ones. The study concluded that manifold learners perform well on data that forms a low-dimensional representation, such as the frequently used Swiss roll dataset but it cannot be generalized to more complex datasets.

In accordance with the above considerations, we have evaluate six DR methods measuring the "Trustworthiness"

*Research work done in the frame of VATEO project "Visual Analytics Tool for Earth Observation Images", financed by UEFISCDI, Project number 300/2014 - PN-II-PT-PCCA-2013-4-0536.

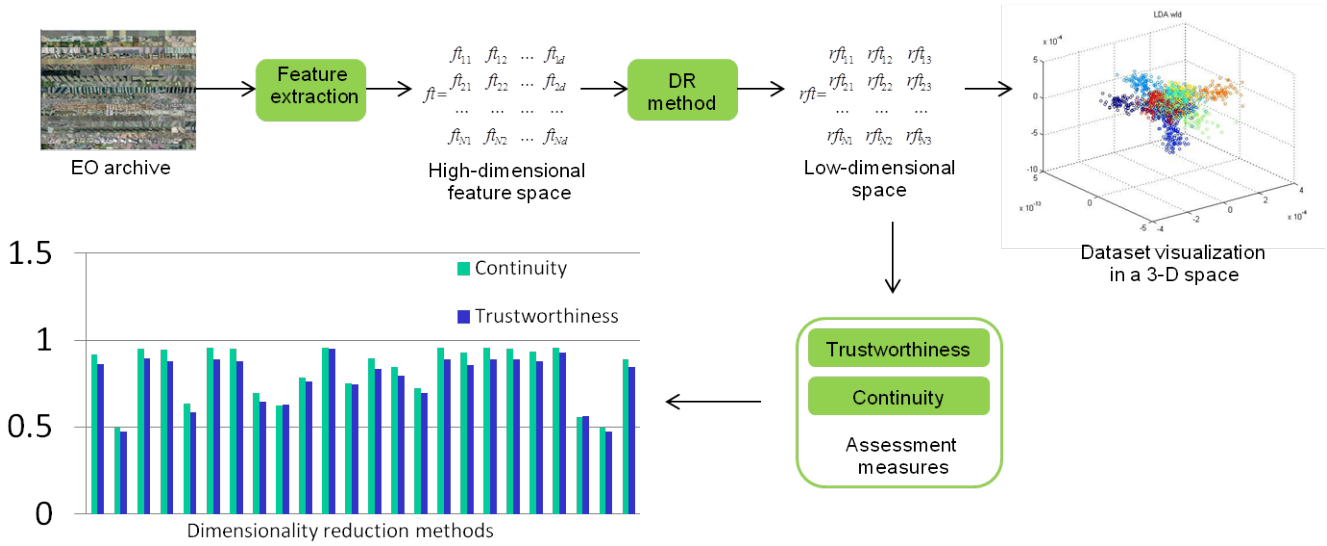

Fig. 1. The proposed methodology: it starts with the feature extraction process followed by the DR step while the evaluation of the projections performance ends the analysis.

and "Continuity" of their mapping in order to further use them to visual explore the EO archives in a 3-D space.

Further, the paper is structured as follows: Section 2 presents the proposed methodology, DR techniques (Subsection 2.1) and the quantitative measures used for their evaluation (Subsection 2.2). The results of the DR assessment are shown and discussed in Section 3 while Section 4 exposes the conclusions.

2. THE PROPOSED APPROACH

This study aims to objectively evaluate six DR methods, one linear and five non-linear, measuring the "Trustworthiness" and the "Continuity" for the low-dimensional representation of three datasets: one artificial, formed by six multidimensional Gaussian distributions and two real datasets, consisting in two remote sensing images. The envisaged methodology is displayed in Fig. 1. The dataset feature extraction process represents the first step in our analysis conducting in a high-dimensional feature space. In order to visualize the dataset in a 3-D projection we have reduced the feature space using a part of the DR library provided by [6]. Benefiting of the low-dimensional representation we measure the "Trustworthiness" and "Continuity" of the projection to evaluate the methods performance.

2.1. Dimensionality reduction

Considering a labeled EO archive expressed by patches tiled from a remote sensing image, a feature extraction process was performed. The dimensionality of the achieved feature space was reduced using Principal Component Analysis (PCA) [5], one of the traditional linear technique, and five non-linear DR methods: Kernel PCA (KPCA) [7], Diffusion Maps (DM) [8], Sammon mapping (S) [9] , Autoencoders (A) [10] and Locally Linear Coordination (LLC) [5].

Dimensionality reduction techniques project a dataset X with dimensionality D into a space Y with dimensionality $d(d < D)$ by preserving, as much as possible the geometry of the data.

PCA is the most widely used DR method for many kind of application such face recognition, image compression, text analysis. It provides a linear transformation of the high-dimensional data using the largest d eigen vectors related to datasets covariance matrix. Being a linear method, it represents non-linear data in an inefficient manner.

To overcome this issue, PCA was extended to KPCA reconstructing the high-dimensional space using a kernel function. The KPCA principle is similar to PCA differing by matrix based on which are computed the eigen vectors. Exactly, the KPCA transformation consists in the largest d eigen vectors of the kernel matrix rather than covariance matrix aiming to preserve large pairwise distances [7].

DM is a non-linear mapping algorithm based on eigen vectors. Its goal is to identify a diffusion map consisting in the first d eigen vectors that retains the best possible pairwise diffusion distance. A diffusion distance represents the distance function between any two points achieved by the random walk on the graph [8].

The previous presented techniques focus on keeping the pairwise similarities but, for the data geometry, the small ones are more important. In this idea, several DR have been proposed, three of these: Sammon mapping, Autoencoders and LLC being evaluated in this study. The novelty of Sammon

Table 1. The "Trustworthiness" values of the DR performance for different datasets.

Dataset	PCA	KPCA	DM	S	A	LLC
Gauss (20D)	**0.96**	0.58	0.95	**0.96**	0.92	0.55
UCM WLD (462D)	**0.97**	**0.97**	0.95	0.67	0.93	0.60
UCM SPECTRAL (192D)	0.89	0.88	**0.90**	0.47	0.86	0.63
LANDSAT WLD (432D)	0.95	0.85	0.94	**0.96**	0.94	0.67
LANDSAT SPECTRAL (192D)	0.92	0.91	0.92	**0.96**	0.92	0.60

Table 2. The "Continuity" values of the DR performance for different datasets.

Dataset	PCA	KPCA	DM	S	A	LLC
Gauss (20D)	**0.96**	0.58	0.95	**0.96**	0.92	0.55
UCM WLD (462D)	**0.99**	**0.99**	0.96	0.68	0.97	0.58
UCM SPECTRAL (192D)	**0.96**	0.95	0.95	0.49	0.92	0.62
LANDSAT WLD (432D)	0.96	0.96	0.94	**0.97**	0.96	0.68
LANDSAT SPECTRAL (192D)	0.95	0.95	0.94	**0.97**	0.93	0.59

mapping lies on the cost function used to measure the structure of the original dataset reflected in the low-dimensional representation. This method consist in initialization of the low-dimensional dataset Y_0 by performing PCA on the original dataset X then, Y_i is updated using the gradient descent method of the cost function with respect to the Y_i until the convergence is achieved [9] [11].

An Autoencoder represents a feed-foreword neural network. The low-dimensional representation is provided by the node values of the middle hidden networks. The input and the output layers present a number of nodes equal to the original dataset dimensionality while the middle hidden layer had only d nodes. The algorithm computes the node values in order to minimize the difference between the input and the output of the network [10].

The last DR method evaluated in this paper is an automatic algorithm that performs a global alignment of local linear models mixture, named Locally Linear Coordination. For linear models arrangement, LLC uses an Expectation-Maximization algorithm [5].

2.2. Quantitative measures used for evaluation of the DR methods

The quality of the DR methods was computed measuring the "Trustworthiness" (M_t) and the "Continuity" (M_c) of the projection. A "trustworthy" projection is one that achieves same k-nearest neighbors of a data items both in original and low-dimensionality representation [12].

$$M_t(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k) \qquad (1)$$

$$A(k) = \begin{cases} \frac{2}{Nk(2N-3k-1)} & , \ k < \frac{N}{2} \\ \frac{2}{N(N-k)(N-k-1)} & , \ k \geq \frac{N}{2} \end{cases} \qquad (2)$$

where N represents the data samples, $r(i,j)$ - the rank of element j in the ordering according to the distance from i in the original space, $U_k(i)$ - the set of k-nearest neighbors of the data sample i in the low-dimensional representation but not in the original space and $A(k)$ scales the measure between zero and one.

Similar to M_t, the "Continuity" of a projection quantifies its discontinuities.

$$M_t(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in V_k(i)} (\hat{r}(i,j) - k) \qquad (3)$$

where $V_k(i)$ represents the set of k-nearest neighbors of the data sample i in the original space but not in the low-dimensional representation and $\hat{r}(i,j)$ expresses the rank of element j in the ordering according to the distance from i in the low-dimensional representation.

In the ideal case, the measures (M_t, M_c) must be equal to one.

3. RESULTS OF THE EVALUATION

A synthetic high-dimensional dataset consisting in random vectors describing six, 20-dimensional Gaussian distributions with spaced means and low variances was generated to illustrate the algorithms performance.

To establish the quality of the DR applied on real datasets we used the UC Merced Land Use dataset (UCM) [13], consisting in 21 classes with 90 remote sensing images patches of 256×256 pixels. For each image, spectral signatures (SPECTRAL) and Weber Local Descriptors (WLD) [14], representing the n-dimensional feature vector, were computed. The resulted feature space is 192-D for SPECTRAL and 432-D for WLD. The same analysis was made on a database consisting of 50×50 pixels patches tiled from a LANDSAT 7 ETM + scene of Bucharest and is formed by spectral indices (NDVI Normalized Difference Vegetation Index, NDBI Normalized Difference Build Up Index, MNDWI Modified Normalized Difference Water Index), the number of classes being five.

Table 1 and Table 2 present the corresponding values of "Trustworthiness", respectively "Continuity" attained on the low-dimensional representations. The best performance achieved for each analyzed dataset is highlighted.

The M_t and M_c measures achieved on low-dimensional representations reveal that Sammon mapping is indicated for the synthetic and LANDSAT dataset while Principal Component Analysis is suitable for UCM dataset. The association between synthetic and LANDSAT datasets for Sammon projections can be explain by the fact that LANDSAT dataset is actually a semi-synthetic whose data consist of the overlap of the thematic bands, each band being computed as a combination of the spectral bands - a spectral index.

4. CONCLUSIONS

Considering the results of our analysis it can be noticed that the projections quality is highly related to the current task.

Our results are related to ones attained in [5]: Sammon and PCA has acquired the best performance. In fact for the majority DR methods,a score higher than 90% was achieved, LLC representing the exception.

Both results, our and those achieved in [5], show a strong relation between the "Trustworthiness" and "Continuity" values.

5. REFERENCES

[1] M. C. Ferreira de Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, July 2003.

[2] M. Naseer and S. Y. Qin, "Performance comparison of nonlinear dimensionality reduction methods for image data using different distance measures," in *Computational Intelligence and Security, 2008. CIS '08. International Conference on*, Dec 2008, vol. 1, pp. 41–46.

[3] John A. Lee and Michel Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomput.*, vol. 72, no. 7-9, pp. 1431–1443, Mar. 2009.

[4] Wouter Lueks, Bassam Mokbel, Michael Biehl, and Barbara Hammer, "How to evaluate dimensionality reduction? - improving the co-ranking matrix," *CoRR*, vol. abs/1110.3917, 2011.

[5] LJP Van der Maaten, EO Postma, and HJ Van den Herik, "Dimensionality reduction: A comparative review," *Technical Report TiCC TR 2009-005*.

[6] "http://lvdmaaten.github.io/drtoolbox/," .

[7] S. K. Joshi and S. Machchhar, "An evolution and evaluation of dimensionality reduction techniques - a comparative study," in *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*, Dec 2014, pp. 1–5.

[8] S. Lafon and A.B. Lee, "Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 9, pp. 1393–1403, Sept 2006.

[9] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401–409, May 1969.

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] Paul Henderson, "Sammon mapping," *Pattern Recognition Letters*, vol. 18, no. 11-13, pp. 1307–1316, 1997.

[12] Jarkko Venna and Samuel Kaski, "Visualizing gene interaction graphs with local multidimensional scaling," in *In this volume*, 2006, pp. 2–930307.

[13] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2010, GIS '10, pp. 270–279, ACM.

[14] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikinen, Senior Member, Xilin Chen, Senior Member, and Wen Gao, "Wld: A robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1705–1720, 2010.