FEATURE SPACE DIMENSIONALITY REDUCTION FOR THE OPTIMIZATION OF VISUALIZATION METHODS

Andreea Griparis, Daniela Faur

Department of Applied Electronics and Information Engineering, Politehnica University of Bucharest, Romania

Mihai Datcu

German Aerospace Center, Earth Observation Center Oberpfaffenhofen, Germany

ABSTRACT

Visual data mining methods are of great importance in exploratory data analysis having a high potential for mining large databases. As the data feature space is generally n-dimensional, visual data mining relies on dimensionality reduction techniques. This is the case for image feature spaces which can be visualized by giving each data point a location in a three dimensional space. This paper aims to present a comparative study of several dimensionality reduction methods considering as input image feature spaces, in order to detemine an optimal visualization method to illustrate the separation of the classes. At the beginning, to check the performance of the envisaged method, an artificial dataset consisting of random vectors describing six, 20-dimensional Gaussian distributions with spaced means and low variances was generated. Further, two real images datasets are used to evaluate the contributions of dimensionality reduction algorithms related to data visualization. The analysis focuses on the PCA, LDA and t-SNE dimensionality reduction techniques. Our tests are performed on images for which the computed features include the color histogram and Weber descriptors.

Index Terms— visualization, classification, dimensionality reduction, features vectors

1. INTRODUCTION

The availability of continuously growing image archives has created a strong demand to develop and implement automatic systems for relevant information retrieval. Data mining is defined as the extraction of models or patterns from observed data, as part of more complex process of drawing high-level useful knowledge from low-level data, known as knowledge discovery in databases. As a consequence data visualization plays a key role in the knowledge discovery process due to its ability to illustrate reveal hidden relationships between data items and to offer relevant displays of inherent data traits.

There is a twofold approach to data visualization. The first one aims to identify methods for multidimensional data visualization such as parallel coordinates technique [1], iconographic or dense pixel displays [2] while the second aims to minimize the information loss which occurs during the dimensionality reduction process, e.g. PCA [3], LDA [4], t-SNE [5], KECA [6], NeRV [7] and IPCA [8].

The main idea behind these techniques is to transform a dataset X with dimensionality D into a new dataset Y with dimensionality d ($d < D$) by preserving, as much as possible the geometry of the data. Practically neither the geometry of the data manifold, nor the intrinsic dimensionality of the dataset X is known [1].

This paper aims to present a comparative study of several dimensionality reduction methods starting from image feature spaces so as to delineate an optimal image content visualization method. The dimensionality reduction assumes the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation corresponds to the native dimensionality of the data, e.g. the minimum number of parameters needed to account for the observed properties of the data [3].

2. THE PROPOSED APPROACH

Considering a multidimensional feature space describing an annotated (classified) dataset, three dimensionality reduction methods are used in order to attain a 3 D representation of the informational content of an images dataset. The dimensionality reduction methods used in this paper are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and t- Distributed Stochastic Neighbor Embedding (t-SNE).

Principal Component Analysis (PCA) is of the traditional methods for data reduction. It is a linear transformation based on the covariance matrix and its eigenvalues. The algorithm computes the covariance matrix and its eigenvalues, keeping only the d (number of dimension which reduces space) largest values. The eigenvectors associated with kept eigenvalues, form the transformation [3].

t- Distributed Stochastic Neighbor Embedding (t-SNE) is a method for converting high-dimensional data set into a matrix of pair-wise similarities. The goal is to preserve much of local structure of high dimensional data and to reveal global structure by converting Euclidian distance between

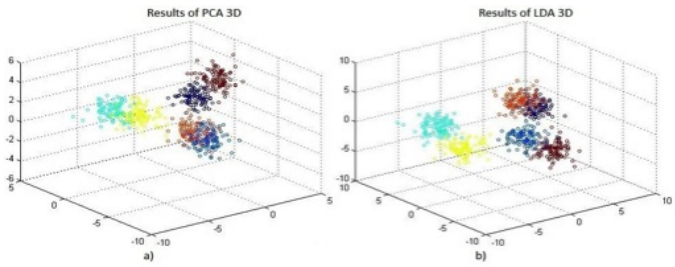

Fig. 1. Results of the artificial dataset dimensionality reduction. The colors represent classes. a) Results of PCA algorithm applied on the artificial dataset (20 D). b) Results of LDA algorithm applied on the artificial dataset (20 D).

data points into conditional probabilities (similarities) and to find the a low-dimensional representation that minimizes the mismatch between the D-dim and d-dim spaces. [5].

Unlike previous methods, the Linear Discriminat Analysis (LDA), also known as Fisher Discriminat Analysis, is a supervised classification aiming for an optimal class separation. This is achieved by searching for the projection that maximizes the separation between classes (covariance SB) and also minimizes the distance between the same classes items (variance SW). The transformation is formed by the correspondence eigenvectors of the n largest eigenvalues of the resulted matrix from covariance inverse matrix multiplied with variance matrix [5].

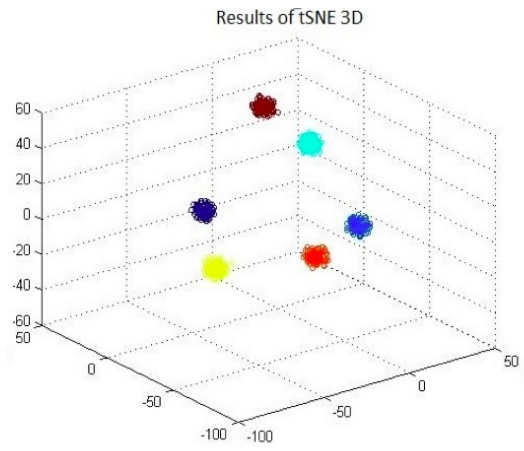

Fig. 2. Results of t-SNE algorithm applied on the artificial dataset (20 D)

A synthetic high dimensional dataset consisting of random vectors describing six, 20-dimensional Gaussian distributions with spaced means and low variances was generated to illustrate the algorithms performance. The PCA, LDA and t-SNE algorithms presented in [3] [4] [5] were applied on this database conducting to the visualizations depicted in the Fig. 1 and Fig 2.

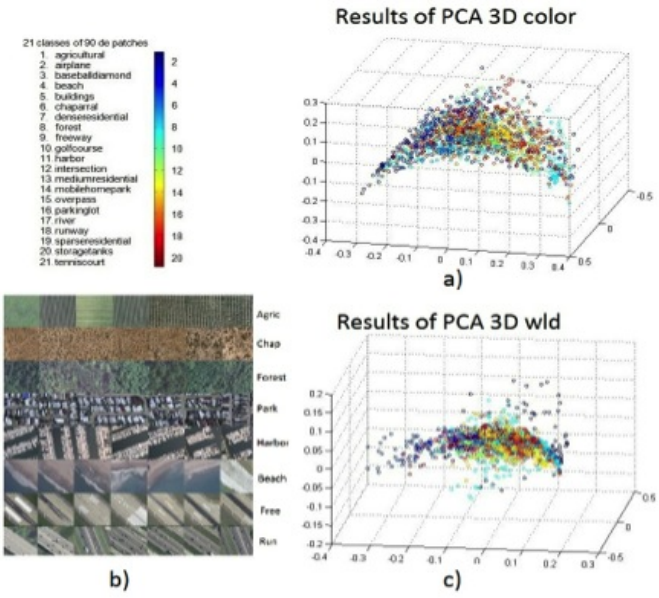

Fig. 3. a) Results of PCA algorithm applied on the feature space $192D$ defined by the color histograms of the 21 classes dataset. b) Patches samples of the 21 classes dataset.c) Results of PCA algorithm applied on the feature space $432D$ defined by the Weber Local descriptors of the 21 classes dataset.

Further, a database of real images (21 classes, 90 remote sensing images patches of 256×256 pixels) was considered. By the means of content based image annotation software described in [9], for each patch color histogram (color) and Weber Local descriptors (wld) [10] were computed. The resulted feature space is 192-dimensional for color histograms and 432-dimensional for Weber. Finally through PCA, LDA and t-SNE algorithms this space is reduced to three dimensions. Fig.3 and Fig.4 show the $3D$ space representation.

Inspecting the LDA projection of Weber Local descriptors image space in Fig.4 we can notice a transition from the "parking lot" (orange) to the "harbor" (green) and then further to the "beach" (blue) areas, the first two containing rectangular objects ordered in a certain direction and the last two sharing the water surface. The visualization results of t-SNE applied on Weber Local descriptors feature space reveals the same transition and, in addition, it can be seen the group of "runway" and "freewa" classes from the upper left corner differs by their distributions that follow orthogonal directions.

The same processing stages were applied on a database consisting of 50×50 pixels patches tiled from LANDSAT 7 ETM+ formed by spectral indices (NDVI Normalized Difference Vegetation Index, NDBI Normalized Difference Build Up Index, MNDWI Modified Normalized Difference Water Index) of Bucharest, the number of classes being five.

The results of the PCA algorithm for dimensionality reduction for the considered feature space (color and wld) are

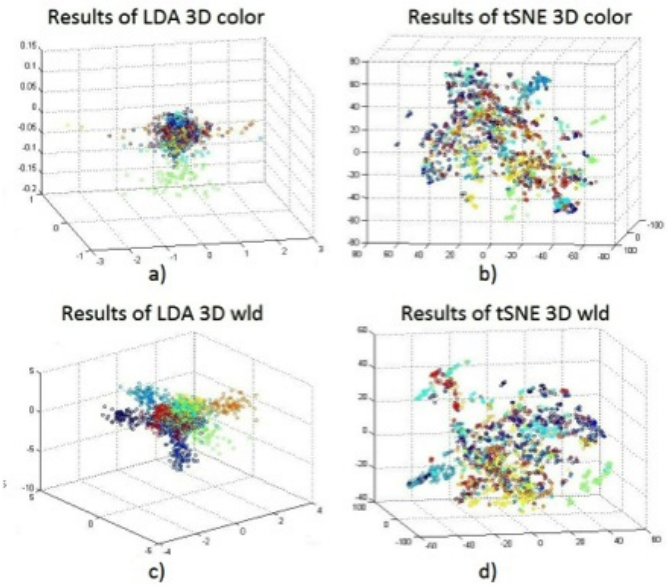

Fig. 4. a), b) Results of LDA/t-SNE algorithm applied on the feature space $192D$ defined by the color histograms of the 21 classes dataset. c), d) Results of LDA/t-SNE algorithm applied on the feature space $432D$ defined by the Weber Local descriptors of the 21 classes dataset.

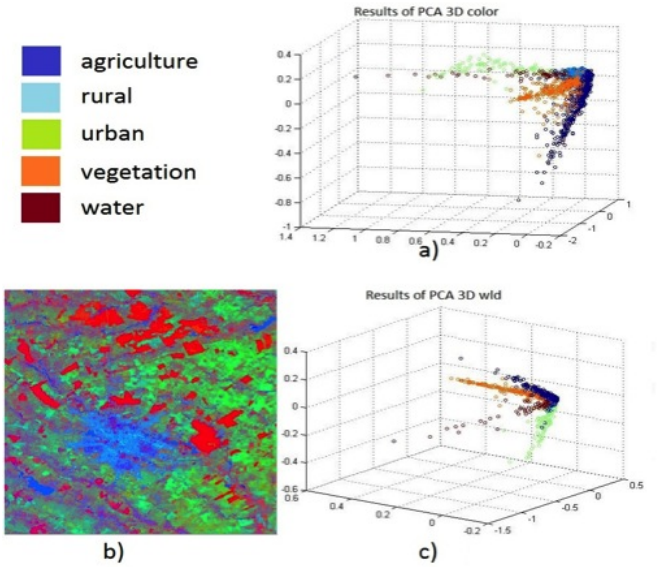

Fig. 5. a) Results of PCA algorithm applied on the $192D$ feature space defined by the color histograms of the LANDSAT 7 ETM+ dataset, b) Samples of the LANDSAT 7 ETM+ dataset patches, c) Results of PCA algorithm applied on the $432D$ feature space defined by the Weber Local descriptors of the LANDSAT 7 ETM+ dataset.

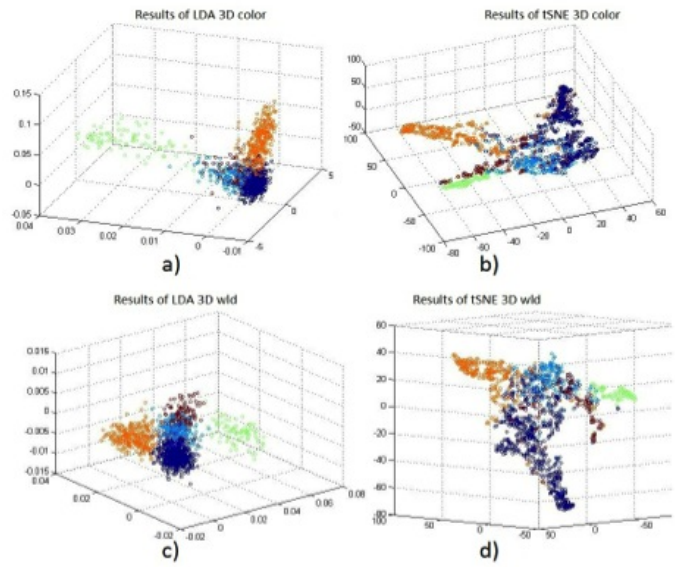

Fig. 6. a), b) Results of LDA/t-SNE algorithm algorithm applied on the $192D$ feature space defined by the color histograms of the LANDSAT 7 ETM+dataset. c), d) Results of LDA/t-SNE algorithm applied on the $432D$ feature space defined by the Weber Local descriptors of the LANDSAT 7 ETM+ dataset.

provided in Fig.5 a) and c). The $3D$ space resulted after LDA and t-SNE reduction is presented in Fig.6. Interpreting the figures it can be noticed that the LDA algorithm applied on color histogram performs a good separation of the "urban" and "vegetation" classes. All resulted projections reveal that rural areas exhibit the transition to all the other classes, due to the fact the "rural" patches contains small regions from all classes.

3. COMMENTS AND CONCLUSIONS

Dimensionality reduction algorithms transform a dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining, as much as possible, the geometry of the data. Usually, neither the geometry of the data manifold, nor the inherent dimensionality d of the original dataset is known. Hence, dimensionality reduction is an ill-posed problem that can only be solved by assuming certain properties of the data. Analyzing the results one can observe that the t-SNE algorithm leads to a more compact group of objects belonging to the same class, allowing a facile isolation of the classes and if necessary, a further efficient information retrieval. This result is independent of the database properties and images type. Fig.1 reveals that all three methods preserve the spatial relationship between the items of the artificial database. For the artificial dataset, t-SNE seems to be the most appropriate method for dimensionality reduction while LDA is the best

for the real datasets when is applied on the Weber Local descriptors feature space. PCA provides the best results in the case of spectral indices LANDSAT 7 ETM+ image described by the color histograms feature space.

In conclusion, the PCA, LDA and t-SNE algorithms can be used for multidimensional database visualization. Their performance is directly dependent on the right choice of the database suitable descriptors. This observation is also valid for Fig.6 where the dimensionality reduction of Weber Local descriptors feature space leads to a better separation of classes compared to color histograms feature space that does not contain the patches orientation. The number of classes also affects the quality of results. The comparison of the second and third experiment results shows that the number of classes influences the performance of the algorithm. The data set used in the second experiment was grouped in 21 classes, some of them being similar ("buildings" class is similar to "dense residential", "medium residential" and "sparse residential", the "storage tanks" can be confused with "baseball diamond" and "runway" comes close to "overpass" and "freeway") fact that leads to an increased uncertainty.

Finally, for real datasets the best results are provided by LDA also, an a priori labeling of the dataset would be effective.

4. ACKNOWLEDGEMENT

Research work done in the frame of VATEO project (Visual Analytics Tool for Earth Observation) funded by the UEFISCDI under the National Plan for Research, Development and Innovation PN II Partnerships Program.

5. REFERENCES

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

[2] D.A. Keim, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 1, pp. 1–8, Jan 2002.

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

[4] Max Welling, "Fisher linear discriminant analysis," *Max Welling's Classnote in Machine Learning*, 2009.

[5] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *The Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.

[6] R. Jenssen, "Entropy-relevant dimensions in the kernel feature space: Cluster-capturing dimensionality reduction," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 30–39, July 2013.

[7] Jarkko Venna and Samuel Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, 2007, pp. 572–579.

[8] Kevin M. Carter, R. Raich, W.G. Finn, and A.O. Hero, "Information-geometric dimensionality reduction," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 89–99, March 2011.

[9] C.O. Dumitru, Shiyong Cui, D. Faur, and M. Datcu, "Data analytics for rapid mapping: Case study of a flooding event in germany and the tsunami in japan using very high resolution sar images," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 8, no. 1, pp. 114–129, Jan 2015.

[10] 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.