Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space

Szymon Łukasik[1,2] and Piotr Kulczycki[1,2]

[1] Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland
[2] Department of Automatic Control and Information Technology, Cracow University of Technology
ul. Warszawska 24, 31-155 Cracow, Poland
slukasik@ibspan.waw.pl

Abstract. This paper investigates a possibility of supplementing standard dimensionality reduction procedures, used in the process of knowledge extraction from multidimensional datasets, with topology preservation measures. This approach is based on an observation that not all elements of an initial dataset are equally preserved in its low-dimensional embedding space representation. The contribution first overviews existing topology preservation measures, then their inclusion in the classical methods of exploratory data analysis is being discussed. Finally, some illustrative examples of presented approach in the tasks of cluster analysis and classification are being given.

Key words: multidimensional datasets, dimensionality reduction, topology preservation, cluster analysis, classification

1 Introduction

Recently, the subject of intelligent data analysis are predominantly high dimensional datasets with huge sample lengths. It is a result of growing amount of information stored in distributed data warehouses and fast development of data-intensive software applications frameworks [3]. The knowledge extraction and visualization of such datasets are difficult, mainly due to methodological obstacles of high dimensional data analysis. They are caused mainly by inherent properties of such datasets referred in bibliography as "curse of dimensionality" [16].

To overcome those issues numerous dimensionality reduction procedures have been proposed. Let X to denote $n \times m$ data matrix:

$$X = \left[x_1 \; x_2 \; ... \; x_m \right] \tag{1}$$

columns of which represent n dimensional sample elements for given real-valued probabilistic variable. Each dimension of such variable will be referred later in

this paper as a feature. The general aim of dimensionality reduction is a data transformation to its new $N \times m$ sized form, where N is significantly smaller than n. This can be achieved either by selecting most N significant features (feature selection) or by construction of a new set of N features (feature extraction) based on the initial ones. The second case is more general and will be under consideration here. Among feature extraction procedures one can distinct: linear methods where synthesis of resulting dataset Y is performed by linear transformation:

$$Y = A\,X \tag{2}$$

with A being a transformation matrix of size $N \times n$ and nonlinear techniques where data transformation can be described by a nonlinear function $g : R^n \to R^N$ (or if such functional relationship does not exist). Detailed study on performance of routines belonging to both of above mentioned classes can be found in [10].

The general goal of dimensionality reduction is removing dataset's redundant content, however at the same time its application can cause a loss of important information carried within its entries. The latter can be quantitatively evaluated using different preservation quality indices, measuring datasets structural deformation. Some of those indicators can be considered on per-element basis which directly allows to assess how well each element of the dataset was relatively preserved by the dimensionality reduction transformation. Such approach is investigated in this paper, along with a novel concept of using this index (named below as a elements weight) to improve the performance of intelligent data analysis procedures in the reduced feature space. The presented idea was suggested first in our previous contribution devoted to the novel, metaheuristic-based dimensionality reduction technique [8], as well as in [9], where it was experimentally assessed for raw stress measure.

The paper is organized as follows. Various topology preservation indices already presented in the bibliography of the subject are given in the following Section. The use of some of them, on per-element basis, for selected data analysis procedures in the reduced feature space is discussed in Section 3, along with experimental results given in Section 4. Finally, the last part of the contribution contains some concluding remarks on the introduced approach and planned further research.

2 Topology Preservation Measures

Let us, in equivalence to (1), consequently denote the representation of given dataset in the reduced feature space by $N \times m$ data matrix:

$$Y = \left[\, y_1 \; y_2 \; ... \; y_m \,\right] \tag{3}$$

For the purpose of subsequent analysis we also define Euclidean distances between two datasets elements i and j ($i, j \in \{1, 2, ..., m\}$) in the initial and reduced feature space (d_{ij} and δ_{ij} accordingly) as follows:

$$d_{ij} = \|x_i - x_j\|_{R^n} \tag{4}$$

$$\delta_{ij} = \|y_i - y_j\|_{R^N} \tag{5}$$

Dimensionality reduction procedures are often classified into two, not always clearly distinguished, groups, namely global and local techniques [14]. The former are characterized by an attempt to preserve global geometrical properties of the original data in its low-dimensional representation Y, while the latter are based on trying to keep the local neighbourhood relations found initially in X.

To measure the quality of the global-based mapping one can use simple raw stress used in many variants of Multidimensional Scaling [1]:

$$S_R = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (d_{ij} - \delta_{ij})^2 \tag{6}$$

as well as its normalized form provided by Sammon [12]

$$S_S = \frac{1}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d_{ij}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{7}$$

which puts less emphasis on large distances.

Stress-based indices are the most commonly used, however in many practical problems it is sufficient to evaluate the preservation of distances order rather than their exact values. Spearman's rho [13] can be employed in that case as it estimates the correlation of rank order data. In the context of dimensionality reduction, this coefficient can indicate how well the corresponding low-dimensional embedding preserves the order of pairwise distances between the original data points converted to ranks. Spearman's rho is calculated by using the following equation:

$$\rho_{SP} = 1 - \frac{6 \sum_{p=1}^{M} (r_{p_d} - r_{p_\delta})^2}{M^3 - M} \tag{8}$$

where $M = m(m-1)/2$ is a total number of distances subjected to the comparison and r_{p_d} , r_{p_δ} are the ranks (with $p = 1, 2, ..., M$) of pairwise distances sorted in ascending order for both, initial and reduced feature space. Spearman's rho value equal to 1 is equivalent to perfect preservation of distances' order (in general $\rho_{SP} \in [-1, 1]$).

Local mappings are usually evaluated using neighbourhood graph preservation. While, there exists numerous methods to analyze it e.g. Konig measure [6], simple one-parameter Mean Relative Rank Error (MRRE) index [7] will be presented here.

Let $\mathcal{N}_k(x_i)$ to represent a group of k-nearest neighbors of x_i, and $R^i_{j_d}$, $R^i_{j_\delta}$ be the ordered rank of distances d_{ij} and δ_{ij} respectively, defined for a set of all distances between element i and a rest of the dataset. MRRE is then defined as follows:

$$MRRE = \frac{1}{C} \sum_{i=1}^{m} \sum_{x_j \in \mathcal{N}_k(x_i)} \frac{|R^i_{j_d} - R^i_{j_\delta}|}{R^i_{j_d}} \tag{9}$$

with the following normalizing factor C:

$$C = m \sum_{p=1}^{k} \frac{|2p - m - 1|}{p} \tag{10}$$

which ensures that $MRRE \in [0, 1]$. This measure can be found similar to continuity, and it vanishes to zero if nearest neighbours of each data point appear in the same order in both spaces [7]. In this initial study we consider MRRE with $k = 11$.

For more detailed description and comparison of above mentioned measures one could refer to [4]. The next Section of this contribution will be devoted to the use of some of them in data analysis procedures performed in the reduced feature space.

3 Proposed approach

Dimensionality reduction in general can significantly modify some data elements' relative position. Consequently the performance of data mining procedures defined in the reduced feature space can be seriously affected. Thus, it would be useful to synthesize individual measure which could serve as a quantitative index of how well each point of the dataset was relatively preserved by the dimensionality reduction transformation. This index element's weight w_i could be then used for exploratory data analysis procedures in the reduced feature space. Investigating if weights in such form are beneficial for data mining procedures performed in the space with reduced dimensionality constitutes one of the goals of this study.

To define the weight for each dataset element it is crucial first to directly evaluate a contribution w_i^* this elements embedding brings to the selected topology preservation index (note that these auxiliary coefficients do not have to sum up to the value of general index). Given the form of already introduced topology preservation measures these per-element factors could be defined accordingly, for raw stress:

$$w_i^* = S_{R_i} = \sum_{j=1}^{m} (d_{ij} - \delta_{ij})^2 \tag{11}$$

Sammon stress:

$$w_i^* = S_{S_i} = \frac{1}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d_{ij}} \sum_{j=1}^{m} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{12}$$

Spearmans rho (this time with $r_{p_d}^i$ and $r_{p_\delta^i}$ representing ranks of distances from p to the element i):

$$w_i^* = 1 - \rho_{SP_i} = \frac{6 \sum_{p=1}^{m} (r_{p_d}^i - r_{p_\delta}^i)^2}{M^3 - M} \tag{13}$$

and Mean Relative Rank Error:

$$w_i^* = MRRE_i = \frac{1}{C} \sum_{x_j \in \mathcal{N}_k(x_i)} \frac{|R^i_{j_d} - R^i_{j_\delta}|}{R^i_{j_d}} \tag{14}$$

Consequently weights are to be calculated using w_i^* values obtained from (11-14) and performing additional normalization:

$$w_i = \frac{m(w_i^*)^{-1}}{\sum_{i=1}^{m}(w_i^*)^{-1}} \tag{15}$$

for $i = 1, ..., m$ to ensure that

$$\sum_{i=1}^{m} w_i = m \ . \tag{16}$$

Please note that if $w_i^* = 0$ one should replace it with $\min_{j=1,...,m} w_j \neq 0$.

Introduction of weights allows to take into account deformations in a dataset relative structure. Data elements with higher weight might then be treated as more adequate. Besides using weights values directly one can use them also to eliminate the influence of some badly deformed data elements. It can be performed by neglecting in weight-based data analysis procedures by setting w_i to 0 those elements for which associated weights fulfil the following condition: $w_i < W$, where $W \in \mathbb{R}^+$ can be referred to as elimination threshold. Then all other weights should be either normalized to keep (16) or alternatively set to 1. This second variant of proposed approach will be under investigation here. The following two subsections of the paper will discuss how the general weight-based scheme defined above can be utilized for two standard data mining algorithms: clustering with K-means procedure and nearest neighbour classification.

3.1 Use Case 1: K-means Clustering Algorithm

The task of cluster analysis is equivalent to such division of available data elements into subgroups (clusters) that elements belonging to each cluster are similar to each other and on the other hand there exist a significant dissimilarity between different clusters elements. The technique of data clustering considered here is based on a modification of the classic K-means algorithm. K-means is an iterative clustering algorithm which is aimed at minimizing sum-of-squares error i.e. sum of distances of dataset elements to their nearest cluster center $C_i = [c_1, c_2, ..., c_N]$, with $i \in 1, 2, ...K$. The procedure, in its standard form, includes a step of cluster assignment followed by clusters centers update [2]. The influence of topology preservation ratio in the reduced feature space can be included in the second stage of clustering algorithm. Each cluster center is then established using the following modified equation:

$$c_{ij} = \frac{1}{\sum_{y_l \in C_i} w_l} \sum_{y_l \in C_i} w_l y_{l_j} \ , \tag{17}$$

with $i = 1, ..., K$ and $j = 1, ..., N$. Consequently, such modified algorithm can be referred to as weighted K-means [5].

3.2 Use Case 2: Nearest Neighbour Classifier

Now let us consider the task of classification, that is designating element $\tilde{x} \in \mathbb{R}^n$ to one of the fixed class with known set of representative patterns (training set), similar to (1). Nearest neighbour classifier is a basic solution for this problem. The algorithm itself assigns element $\tilde{x}$ to a class which nearest neighbour of $\tilde{x}$ from the training set belongs to. Its modified variant, taking into account topology preservation, makes a similar decision on a basis of weighted distances i.e. divided additionally by weight w_i. This approach can be easily generalized for broader category of k-Nearest Neighbour Classifiers [11].

4 Experimental results

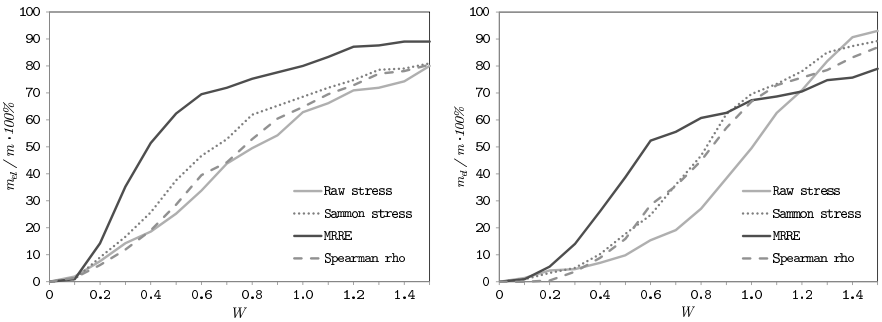
Proposed technique was preliminarily verified for data exploration procedures performed for five multidimensional datasets taken from the UCI Machine Learning Repository [15] listed in Table 1.

Table 1: Used datasets description

Dataset	m	n	N	Classes	Class Description	Sample length
glass	214	9	4	6	*building_windows_float_processed*	70
					building_windows_non_float_processed	76
					vehicle_windows_float_processed	17
					containers	13
					tableware	9
					headlamps	29
wine	178	13	5	3	*producer_1*	59
					producer_2	72
					producer_3	47
WBC	683	9	4	2	*benign*	444
					malign	239
vehicle	846	18	10	4	*Opel*	212
					Saab	217
					bus	218
					van	199
seeds	210	7	2	3	*Kama*	70
					Rosa	70
					Canadian	70

Dimensionality reduction was performed using Principal Components Analysis. We used fixed values of embedding dimension N established in previous experiments. The accuracy of K-means clustering was measured using Rand index value I_C calculated versus class labels, whereas for nearest-neighbour classification average classifier accuracy I_K during 5-fold cross validation was under close scrutiny. All experiments involving aforementioned data mining procedures where repeated 30 times, with mean and standard deviation being reported here (in "mean ± standard deviation" notation).

The initial experiments were conducted to evaluate the distribution of weight values calculated from (11-14). It was computationally verified by setting $W = 0.1, 0.2, ..., 1.5$ and observing the percentage (relative to the sample size m) of dataset elements with weight values under W, labelled as m_{el}. The results of those studies are shown on Figure 1. It can be seen that the distribution of weight values is not linear. However, for all considered datasets less than 50% of sample elements are characterized by weights values below average i.e. $w_i < 1$. It was also observed that weighting scheme based on raw stress could tend to be conservative, whereas using MRRE would offer neglecting large part of the dataset even for a small value of W.



(a) (b)

Fig. 1: Weights values distribution for *seeds* (a) and *glass* (b) datasets

The next series of experiments was designed to investigate the performance improvement for clustering, when using weighted variant of K-means algorithm with varying values of threshold W. First, the standard K-means procedure was tested for the reduced dataset, with $I_c \cdot 100\%$ representing the Rand index value obtained at that stage. Then the algorithm supported by aforementioned weighting scheme was executed with multiple runs measuring the performance variation $\Delta I = (I_c - I_{c_W}) \cdot 100\%$ for different values of W. Figure 2 exhibits the results obtained at that stage for two selected datasets and different weighting schemes.

For all datasets introducing the weights based on topology preservation index improve clustering performance. For most cases it is also recommended to neglect some dataset elements here. Increasing W beyond 1 however, leads to spectacular decrease in clustering accuracy.

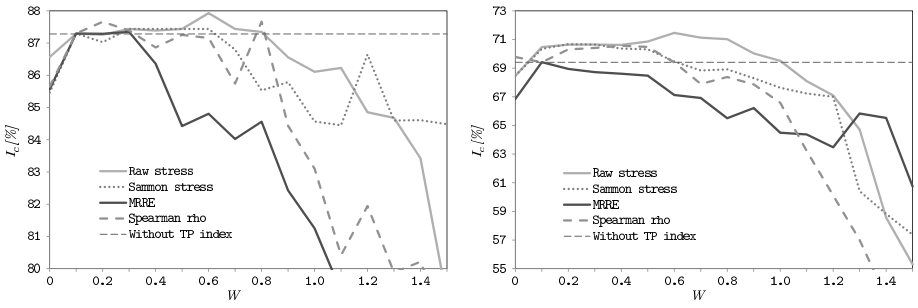


Fig. 2: Comparison of Rand indices obtained using K-means in reduced feature space with, and without topology preservation index weights for *seeds* (a) and *glass* (b) datasets

The summary of our results, concerning both cluster analysis and classification is given in Table 2. Along with standard data mining procedures their modified variants, with the best combination of topology preservation indices and W values, are being included in this comparison. Once again it is worth to note that using proposed approach is in general beneficial for the performance of data mining procedures performed in reduced feature space. It is particularly recommended to use elimination of deformed datasets elements for nearest-neighbor classifier. It is a result of weak robustness of this classifier to noisy training sample members, which thanks to approach being introduced here, can be neglected.

5 Conclusion

This paper introduces novel scheme designed for high-dimensional tasks of intelligent data analysis performed in the reduced feature space. Our proposal is based on an observation that dimensionality reduction affects the topological structure of the datasets. It is consequently suggested here to use measures of topology preservation to improve data analysis procedures performed thereafter.

Introductory studies on method's performance conducted for selected datasets show that it offers promising efficiency. Further research in this area will involve studying the influence of dimensionality reduction procedure being employed in the first step on the beneficial effect of the proposed weighting scheme. The effect of the technique under consideration on the performance of other intelligent

Table 2: Clustering and classification in reduced feature space - comparison of standard approach and our proposal ($I_c \cdot 100\%$ for clustering and I_k for classification were reported).

Procedure	***glass***	***wine***	***WBC***	***vehicle***	***seeds***
PCA, K-means	69.41 ±2.43	71.37 ±1.07	92.51 ±0.13	64.21 ±1.91	87.28 ±0.15
PCA, weighted K-means	71.46 ±1.88	71.40 ±0.98	93.21 ±0.00	64.21 ±1.80	87.93 ±0.00
Best W	W=0.6	W=0.2	W=0	W=0.1	W=0.6
TP weight type	Raw stress	Spearman rho	Sammon stress	Spearman rho	Raw stress
PCA, NN classifier	59.99 ±2.79	74.51 ±3.08	96.78 ±0.62	55.78 ±1.21	88.34 ±1.80
PCA, weighted NN	63.80 ±2,69	76.58 ±2.57	96.82 ±0.63	58,13 ±1.40	90.24 ±1.88
Best W	W=0.5	W=0	W=0.1	W=0	W=0.7
TP weight type	MRRE	Raw stress	MRRE	Sammon stress	Raw stress

data analysis procedures, e.g. neural-network classifiers will be also investigated. Furthermore the guidelines for choosing proper weighting function, as well as setting a value of W for given dataset will be worked out. Additional experiments, for larger datasets (representing for example documents' content or gene expression data), are likewise planned to be performed.

Acknowledgements

The first author would like to express his gratitude to National Laboratory of Pattern Recognition, Chinese Academy of Science and Professor Bao-Gang Hu in particular for their support throughout the research.

The study is co-funded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

This research was supported in part by PL-Grid Infrastructure.

References

1. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling: Theory and Applications. Springer, Heidelberg (2010).
2. Everitt, B. S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, New York (2011).

3. Furht, B., Escalante, A. (eds.): Handbook of Data Intensive Computing. Springer, Heidelberg (2011).
4. Karbauskaite, R., Dzemyda, G.: Topology Preservation Measures in the Visualization of Manifold-Type Multidimensional Data. Informatica, vol. 20, pp. 235–254 (2009).
5. Kerdprasop, K., Kerdprasop, N., Sattayatham, P.: Weighted K-Means for Density-Biased Clustering. Lecture Notes in Computer Science, vol. 3589, pp. 488–497 (2005).
6. Konig, A.: Interactive Visualization and Analysis of Hierarchical Neural Projections for Data Mining. IEEE Transactions on Neural Networks, vol. 11/3, pp. 615–624 (2000).
7. Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, New York (2007).
8. Łukasik, S., Kulczycki, P.: An Algorithm for Sample and Data Dimensionality Reduction Using Fast Simulated Annealing. Lecture Notes in Artificial Intelligence, vol. 7120, pp. 152–161 (2011).
9. Łukasik, S., Kulczycki, P.: Using Topology Preservation Measures for High-Dimensional Data Analysis in a Reduced Feature Space (in Polish). Technical Transactions, vol. 1-AC, pp. 5–16 (2012).
10. Maaten, L.J.P.v., Postma, E.O., Herik, H.J.: Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005 (2009).
11. Parvin, H., Alizadeh, H., Minati, B.: A Modification on K-Nearest Neighbor Classifier. Global Journal of Computer Science and Technology, vol. 10, pp. 37–41 (2010).
12. Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, vol. 18, pp. 401–409 (1969).
13. Sammut, C., Webb, G.I. (eds.): Encyclopedia of Machine Learning. Springer, New York (2011).
14. Silva, V.D., Tenenbaum, J.B.: Global versus Local Methods in Nonlinear Dimensionality Reduction. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems, vol. 15, pp. 705–712. MIT Press, Cambridge (2003).
15. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
16. Verleysen M., François D.: The Curse of Dimensionality in Data Mining and Time Series Prediction. In: Cabestany, J., Prieto, A., Sandoval, F. (eds.) Computational Intelligence and Bioinspired Systems. LNCS, vol. 3512, pp. 758–770. Springer, Heidelberg (2005).