Evaluating Dissimilarity Measures for Topology Preservation Indices Used in Multidimensional Data Analysis

Szymon Łukasik[1,2]

[1]Faculty of Physics and Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: {slukasik,kulpi}@agh.edu.pl

Piotr Kulczycki[1,2]

[2]Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {slukasik,kulpi}@ibspan.waw.pl

***Abstract*—Dimensionality reduction constitutes a process of selecting/extracting most important features from the initial dataset to obtain its compact representation. It is commonly required for a majority of large datasets tackled by contemporary tools of data analysis. While a process of dimensionality reduction usually creates more usable representation of given data set, it may also result in perturbing relations between individual sample elements. To overcome this problem indices of so called topology preservation can be used. They may be employed in subsequent data analysis tasks to reduce the impact of dataset's structural deformation. The paper studies selected dissimilarity measures which can be used to construct these indices. We evaluate their usability and performance for selected benchmark datasets and data mining problems.**

I. Introduction

CONTEMPORARY data analysis is dealing with complex datasets characterized by huge volume, heterogeneous structure and other methodological obstacles, like missing values or the requirement of tedious data preprocessing [1]. Alleviating the problem of data dimensionality remains of critical importance. It is due to inherent properties of highly dimensional datasets referred in bibliography as "curse of dimensionality" [2] and its grave impact on the result of data analysis.

Conventionally a procedure of data dimensionality reduction is introduced to reduce the influence of aforementioned phenomena. Let X to denote $n \times m$ data matrix:

$$X = \left[\begin{array}{cccc} x_1 & x_2 & ... & x_m \end{array} \right]. \tag{1}$$

Each of n columns of this matrix will be referred later in this paper as a feature and single row as a sample element or a case. The aim of dimensionality reduction is a data transformation to its new $N \times m$ sized form Y, where N is significantly smaller than n. It is usually achieved through feature extraction, i.e. construction of a new set of N features based on the initial ones.

It can be deducted that the general goal of dimensionality reduction is removing dataset's redundant content. Still concurrently a loss of important information carried within its entries can be observed. In our previous contributions we proposed to quantitatively evaluate dataset's structural deformation on per point basis using different preservation quality indices [3], [4]. It allows to assess how well each element of the dataset was relatively preserved by the dimensionality reduction transformation. The goal of this paper is to study the impact of underlying dissimilarity measure on the performance of data analysis procedures employing topology preservation indices in the reduced feature space.

The paper is organized as follows. Various topology preservation indices already presented in our previous research are covered in the following Section. It is ensued by a description of dissimilarity metrics which were used to construct those indices. The use of some of them, on per-element basis, for selected data analysis procedures in the reduced feature space is discussed in Section 4, along with some preliminary experimental results given in Section 5. Finally, the last part of the contribution contains some concluding remarks on the introduced approach and planned further research.

II. Topology Preservation Indices

Let $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, +\infty)$ represent a dissimilarity measure which quantifies the degree to which objects differ from each other. It satisfies the conditions of nonnegativity, reflexivity and symmetry. If additionally triangle inequality is fulfilled the measure can be referred to as a distance or a metric [5].

Let us denote d_{ij} as a dissimilarity between x_i and x_j and δ_{ij} as a corresponding dissimilarity in the reduced feature space, i.e. between y_i and y_j. The quality of relative preservation of the element i can then be evaluated with simple raw stress used in many variants of Multidimensional Scaling [6]:

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

or with its normalized form provided by Sammon [7]:

$$S_{S_i} = \frac{1}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d_{ij}} \sum_{j=1}^{m} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{3}$$

In many situations it is adequate to measure the preservation of distances order rather than their exact values. Spearman's rho [8] can be used for that purpose as it estimates the correlation of rank order data. This coefficient, in the context of dimensionality reduction, can therefore indicate how well the corresponding low-dimensional embedding preserves the order of pairwise distances between the original data points converted to ranks [3]. Spearman's rho value equal to 1 is equivalent to perfect preservation of distances' order (in general $\rho_{SP} \in [-1, 1]$). For evaluating structural deformation it's modified non-negative $\rho^*_{SP_i}$ form will be used here. It is calculated by using the following equation:

$$\rho^*_{SP_i} = 1 - \rho_{SP_i} = \frac{6\sum_{p=1}^{m}(r^i_{p_d} - r^i_{p_\delta})^2}{M^3 - M} \tag{4}$$

where $M = m(m-1)/2$ is a total number of distances subjected to the comparison and $r^i_{p_d}$ and $r^i_{p_\delta}$ represent ranks of distances from p to the element i for both, initial and reduced feature space.

Fourth measure, namely Mean Relative Rank Error (MRRE) index, which is used in our research evaluates neighbourhood graph preservation. 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. Then MRRE on per-point basis is defined as follows:

$$MRRE_i = \frac{1}{C} \sum_{x_j \in \mathcal{N}_k(x_i)} \frac{|R^i_{j_d} - R^i_{j_\delta}|}{R^i_{j_d}} \tag{5}$$

with the corresponding normalizing factor C:

$$C = m\sum_{p=1}^{k} \frac{|2p - m - 1|}{p} \tag{6}$$

which ensures that $MRRE_i$ falls in $[0, 1]$ range. We used MRRE with $k = 11$ in our experiments.

The next Section of this contribution will discuss possible dissimilarity measures which can be used to construct aforementioned topology preservation indices.

III. Selected Distance Measures

Traditional data mining algorithms employ the distance measure defined by the Euclidean metric:

$$d(z, v) = \sqrt{\sum_{i=1}^{n}(z_i - v_i)^2}, \tag{7}$$

where $z = (z_1, \ldots, z_n)$ and $z = (z_1, \ldots, z_n)$.

As an alternative Manhattan distance defined by:

$$d(z, v) = \sum_{i=1}^{n} |z_i - v_i|, \tag{8}$$

is also frequently used.

Fractional p-norm constitutes a generalization of (7-8) formulated as follows:

$$d(z, v) = \left(\sum_{i=1}^{n} \|z_i - v_i\|^p\right)^{1/p}, \tag{9}$$

with $p \in (0, 1)$. Such distance measures were found to perform better for some instances of data mining problems in case of multidimensional data [9]. In practical applications experimental determination of the exact value for p is recommended [10].

Finally cosine dissimilarity which is defined by:

$$d(z, v) = 1 - \frac{\sum_{i=1}^{n} z_i v_i}{\sqrt{\sum_{i=1}^{n} z_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}, \tag{10}$$

can be also used for evaluating the degree in which z and v differ from each other. Cosine dissimilarity for $z, v \in \mathbb{R}^+$ is bounded in $[0, 1]$ that is why it is most commonly used in high-dimensional positive spaces [11].

IV. Topology Preservation Measures for Data Analysis in the Reduced Feature Space

Indices presented in Section 2 (namely S_{R_i}, S_{S_i}, $\rho^*_{SP_i}$ and $MRRE_i$) can be treated as weights w^*_i indicating the adequacy of dataset's element i after dimensionality reduction. Cases with higher weight might be perceived as more adequate. Values of final weights w_i ensuring useful property:

$$\sum_{i=1}^{m} w_i = m\ . \tag{11}$$

are to be calculated using w^*_i values and formula

$$w_i = \frac{m(w^*_i)^{-1}}{\sum_{i=1}^{m}(w^*_i)^{-1}} \tag{12}$$

for $i = 1, ..., m$. If $w^*_i = 0$ it should be replaced with $\min_{j=1,\ldots,m} w^*_j \neq 0$.

Weights in such form can be used directly in data mining procedures executed for datasets of reduced dimensionality. However more invasive routine can be also applied. It involves neglecting (by setting $w_i = 0$) those cases for which w_i falls below elimination threshold W.

In our previous contribution [3] we have demonstrated 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. In the first case the influence of topology preservation ratio in the reduced feature space can be included in the second stage of clustering algorithm, i.e. establishing cluster centers. Similarly modified nearest neighbor algorithm assigns calculated weighted distance from the investigated element $\tilde{x}$ and assigns it to a class which nearest neighbour of $\tilde{x}$ from the training set belongs to.

V. Experimental Results

To evaluate the impact of used dissimilarity measure on the performance of data analysis procedures employing topology preservation indices we performed experiments involving clustering and classification of five reduced multidimensional datasets taken from the UCI Machine Learning Repository [12].

TABLE I
EXPERIMENTAL DATASETS DESCRIPTION

Dataset	*m*	*n*	*N*	Classes
glass	214	9	4	6
wine	178	13	5	3
WBC	683	9	4	2
vehicle	846	18	10	4
seeds	210	7	2	3

Dimensionality reduction was performed using Principal Components Analysis. We used values of embedding dimension N established in previous experiments. All tests where repeated 30 times.

The initial experiments were conducted to evaluate the distribution of weight values calculated from (2-5) using all dissimilarity measures described in Section 3. 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 exemplary results of those studies for *seeds* dataset and raw stress are shown on Figure 1. It can be seen for this specific dataset that cosine dissimilarity measure yields higher concentration of weights' values. Not surprisingly for all Minkowski metrics raw stress values distribution seems similar, with Euclidean distance causing the occurrence of the lowest values of weights. This would lead in this case to the highest intensity of sample reduction when the option with neglecting points having $w_i < W$ would have been chosen. The specific shapes of these distributions differ, however the main outcome remains the same for all datasets – chosen dissimilarity measure severely affects the weights distribution.

The second phase of experiments concerned clustering and classification. We used K-means clustering with its accuracy measured using Rand index value I_C calculated versus class labels. Nearest-neighbour classification algorithm was used in the second phase of experiments with average classifier accuracy I_K during 5-fold cross validation under close scrutiny. Mean and standard deviation of both performance indicators are being reported here (in "mean $\pm$ standard deviation" notation). Preliminary results for MRRE-based weights and neglecting cases with $w_i < 0.4$ are enclosed in Table 2.

Of investigated dissimilarity measures cosine coefficient seems to be the best choice for most of studied datasets. It offers stable, relatively high accuracy of both clustering and classification across tested datasets. It may be also noted however that using some dissimilarity measures for specific

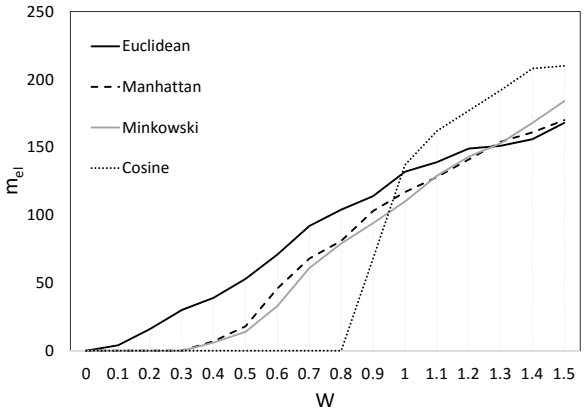

Fig. 1. Weights values distribution for *seeds* datasets and raw stress topology preservation indicator

TABLE II
CLUSTER ANALYSIS AND CLASSIFICATION ACCURACY FOR MRRE-BASED TOPOLOGY PRESERVATION INDEX AND $W = 0.4$

Dataset		Euclidean	Manhattan	Minkowski	Cosine
glass	I_C	68.6 $\pm$2.2	69.9 $\pm$2.2	70.8 $\pm$1.9	69.8 $\pm$2.3
	I_K	62.0 $\pm$2.7	62.2 $\pm$2.7	62.2 $\pm$2.7	61.5 $\pm$2.6
wine	I_C	70.8 $\pm$1.1	71.2 $\pm$1.1	71.3 $\pm$1.0	71.3 $\pm$1.1
	I_K	67.6 $\pm$2.9	68.8 $\pm$2.7	68.7 $\pm$2.7	72.3 $\pm$2.7
WBC	I_C	91.7 $\pm$0.7	78.6 $\pm$12.8	55.3 $\pm$3.5	92.9 0
	I_K	93.6 $\pm$1.0	65.1 $\pm$1.5	65.1 $\pm$1.5	95.6 $\pm$0.7
vehicle	I_C	63.8 $\pm$1.9	64.2 $\pm$1.8	64.0 $\pm$1.7	64.0 $\pm$2.1
	I_K	57.5 $\pm$1.5	58.0 $\pm$1.5	57.8 $\pm$1.5	57.5 $\pm$1.6
seeds	I_C	86.3 $\pm$1.3	87.4 $\pm$0.0	87.0 $\pm$0.3	87.6 $\pm$0.7
	I_K	89.7 $\pm$1.6	88.9 $\pm$1.8	89.3 $\pm$1.7	89.6 $\pm$1.6

datasets (e.g. WBC) causes concentration of weights below $W = 0.4$ and elimination of too many cases. It causes dramatic decrease in the clustering/classification accuracy.

VI. Conclusion

Building up on our previous studies the paper examined the impact of dissimilarity measures on the form and efficiency of topology preservation indices used in data analysis procedures for reduced feature space.

We found that the choice of dissimilarity measure is important for the efficiency of those procedures. It not only affects the distribution of synthesized weights but also their impact on the performance of clustering/classification for the reduced datasets.

It should be also noted that performed experiments generated huge amount of experimental data which has to be filtered and analyzed more carefully. That is why this study is planned to be enriched with more detailed analysis in the framework of the forthcoming follow-up contribution.

ACKNOWLEDGMENT

First author is thankful to Professor Rita A. Ribeiro and her team at Instituto de Desenvolvimento de Novas Tecnologias (UNINOVA), Portugal for providing helpful suggestions and inspiration supporting this contribution.

REFERENCES

[1] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. New York: John Wiley and Sons, 2011.

[2] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," *Lecture Notes in Computer Science*, vol. 3512, pp. 758–770, 2005.

[3] S. Łukasik and P. Kulczycki, "Using topology preservation measures for multidimensional intelligent data analysis in the reduced feature space," *Lecture Notes in Artificial Intelligence*, vol. 7895, pp. 184–193, 2013.

[4] P. Kulczycki and S. Łukasik, "An algorithm for reducing dimension and size of sample for data exploration procedures," *International Journal of Applied Mathematics and Computer Science*, vol. 24, pp. 133–149, 2014.

[5] I. Kononenko and M. Kukar, *Machine Learning and Data Mining*. Elsevier, 2007.

[6] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Heidelberg: Springer, 2010.

[7] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, pp. 401–409, 1969.

[8] C. Sammut and G. Webb, Eds., *Encyclopedia of Machine Learning*. New York: Springer, 2011.

[9] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," *Lecture Notes in Computer Science*, vol. 1973, pp. 420–434, 2001.

[10] D. François, V. Wertz, and M. Verleysen, "Choosing the metric: A simple model approach," in *Meta-Learning in Computational Intelligence*, ser. Studies in Computational Intelligence, N. Jankowski, W. Duch, and K. Grabczewski, Eds. Springer Berlin Heidelberg, 2011, vol. 358, pp. 97–115.

[11] A. Singhal, "Modern information retrieval: A brief overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, pp. 35–43, 2001.

[12] "UCI machine learning repository," http://archive.ics.uci.edu/ml/, access 30.11.2015.