Annals of Data Science
https://doi.org/10.1007/s40745-024-00548-x

A New Kernel Density Estimation-Based Entropic Isometric Feature Mapping for Unsupervised Metric Learning

Alaor Cervati Neto[1] · **Alexandre Luís Magalhães Levada**[1] · **Michel Ferreira Cardia Haddad**[2]

Received: 13 June 2023 / Revised: 25 April 2024 / Accepted: 14 May 2024

Abstract

Metric learning consists of designing adaptive distance functions that are well-suited to a specific dataset. Such tailored distance functions aim to deliver superior results compared to standard distance measures while performing machine learning tasks. In particular, the widely adopted Euclidean distance may be severely influenced due to noisy data and outliers, leading to suboptimal performance. In the present work, it is introduced a nonparametric isometric feature mapping (ISOMAP) method. The new algorithm is based on the kernel density estimation, exploring the relative entropy between probability density functions calculated in patches of the neighbourhood graph. The entropic neighbourhood network is built, where edges are weighted by a function of the relative entropies of the neighbouring patches instead of the Euclidean distance. A variety of datasets is considered in the analysis. The results indicate a superior performance compared to cutting edge manifold learning algorithms, such as the ISOMAP, unified manifold approximation and projection, and t-distributed stochastic neighbour embedding (t-SNE).

Keywords Dimensionality reduction · Metric learning · ISOMAP · Kernel density estimation · Kullback–Leibler divergence

✉ Alexandre Luís Magalhães Levada
alexandre.levada@ufscar.br

Alaor Cervati Neto
alaor_c_neto@yahoo.com.br

Michel Ferreira Cardia Haddad
m.haddad@qmul.ac.uk

[1] Department of Computing, Federal University of São Carlos, São Carlos, Brazil

[2] Department of Business Analytics and Applied Economics, Queen Mary University of London, London, UK

Published online: 06 July 2024

1 Introduction

In the big data era, high-dimensional data are increasingly more common in many fields. Real-world examples include genomics, healthcare, audio signal, digital photograph, and financial datasets, among others. Dimensionality reduction refers to the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Such a reduced representation aims to preserve the intrinsic dimensionality of the data. Dimensionality reduction methods include linear and nonlinear approaches. Linear-based methods are mostly incapable to appropriately capture intricate nonlinear data patterns. Recently, many related nonlinear methods have been introduced, including methodological frameworks inspired in metric learning [1–6].

Dimensionality reduction-based metric learning replaces widely adopted distance measures (e.g. Euclidean distance) through an adaptive manner. The resulting intrinsic distance function aims to be a more suitable, condensed, and insightful representation of the dataset. Among other machine learning tasks, this is particularly convenient to address classification problems [7–9]. The isometric feature mapping (ISOMAP) is one of the pioneering algorithms for dimensionality reduction metric learning [10]. It consists of a global method that adopts the multidimensional scaling (MDS) [11] technique to find an embedding of a pairwise distance matrix onto an Euclidean space.

In the last decade, many extensions to the ISOMAP have been proposed to overcome limitations of the original algorithm. The robust kernel ISOMAP method intends to address two relevant caveats, namely the generalization property and topological stability [12]. Another methodological extension refers to the landmark-ISOMAP (L-ISOMAP), which improves the scalability of the ISOMAP. This is achieved by performing most of its computations on a subset of points, called landmarks [13]. It searches to a minimum set cover of the neighbourhoods along the k-nearest neighbours (KNN) graph, removing observations that belong to neighbour sets of other points. The self-organizing incremental neural network L-ISOMAP extension addresses the problem of automatically selecting an appropriate number and position of landmarks, reducing short-circuit errors [14]. In addition, in the path-based ISOMAP, a low-dimensional embedding is computed through a path-mapping algorithm instead of preserving pairwise geodesic distances, which decreases time and memory complexity [15].

A recent work [16] introduces a method for nonlinear dimensionality reduction, in which a manifold is built considering smooth geodesics. It aims to address problems where manifold measurements are either sparse or noise corrupted. This methodological innovation indicates an improved performance with embedding of face images and handwritten digits compared to the ISOMAP algorithm. In addition, the parallel transport unfolding (PTU) refers to a geometrical approach to evaluate geodesic distances of discrete paths, which may be adopted to devise faster variants of the ISOMAP—e.g. the L-ISOMAP [17]. As the MDS algorithm is a crucial step within the ISOMAP, more efficient variations have been proposed to decrease its computational complexity, from quadratic to quasi-linear [18]. Further recent studies report that the ISOMAP may be applied to embed subsets of data within one dimension. It details the conditions under which high density clusters in the original space are guaranteed to be separable onto the 1-D embedding [19].

The approximate representation of the manifold by an KNN graph is one of the limitations of the ISOMAP. The edges of the KNN network are commonly weighted, adopting the pointwise Euclidean distance. This is particularly sensitive to noisy data and outliers. While manifold learning methods provide reasonable results for a variety of datasets, their computational complexity increase concurrently with their adaptability to nonlinearity or noise in the data [20]. Thus, replacing the Euclidean distance with a more suitable measure should provide a more robust data classification, without significantly increasing computational costs.

In the present paper, it is introduced a new patch-based method that explores symmetrized Kullback–Leibler divergences between one-dimensional local densities, namely the kernel density estimation-based ISOMAP or KDE–ISOMAP method. It is computed in a nonparametric fashion for each local patch of the ϵ-neighbourhood graph, replacing the pointwise Euclidean distance. A pivotal contribution of this method refers to an increase in the robustness to noise within the ISOMAP. This is achieved as a consequence of replacing a pointwise similarity measure with a patch-based one. Computational experiments indicate that the KDE–ISOMAP yields superior clustering results in terms of the silhouette coefficient (SC) compared to state-of-the-art manifold learning algorithms, such as the ISOMAP, unified manifold approximation and projection (UMAP) [21], and t-distributed stochastic neighbour embedding (t-SNE) [22]. Moreover, the proposed method outperforms several competing established algorithms in terms of classification accuracy, suggesting its effective applicability on real-world data.

This paper is organized as follows. The ISOMAP method is detailed in Sect. 2. The kernel density estimation (KDE) approach is presented in Sect. 3. The new KDE–ISOMAP is described in Sect. 4. Tests and results are reported in Sect. 5. Lastly, Sect. 6 summarizes the main findings of the present work and future research possibilities.

2 Isometric Feature Mapping

The fundamental principle of the ISOMAP is to build a graph to approximate the underlying manifold by connecting the KNN within the input space. The ISOMAP may be broken down into three primary phases. Firstly, it is created an undirected proximity graph from the input data $\vec{x}_1, \ldots, \vec{x}_n \in R^m$ with the KNN or ϵ-neighbourhood rule, where the cost of the edge (v_i, v_j) is the Euclidean distance between the vectors $\vec{x}_i$ and $\vec{x}_j$. Subsequently, it is calculated the pairwise distance matrix D through n executions of either the Dijkstra algorithm or a single execution of the Floyd–Warshall algorithm. Lastly, the new coordinates of the points in an Euclidean subspace of R^d are determined by adopting the MDS technique, while preserving their distances.

Thus, the ISOMAP computes the shortest paths between each pair of vertices to find a mapping onto an Euclidean subspace of R^d, while preserving the geodesic distances between data points. The shortest paths in the KNN graph should consist of reliable estimates of the actual geodesic distances in the manifold.

2.1 Multidimensional Scaling

The coordinates of the n points $\vec{x}_r \in R^d$ for $r = 1, \ldots, n$ in an Euclidean subspace are recovered through the MDS from a square matrix $n \times n$ of pairwise distances, where the dimensionality output d consists of a user defined parameter [11, 23]. It is worth noticing that the elements of the pairwise distance matrix $D = \{d_{rs}^2\}$ are:

$$d_{rs}^2 = \|\vec{x}_r - \vec{x}_s\|^2 = (\vec{x}_r - \vec{x}_s)^T \, (\vec{x}_r - \vec{x}_s) \tag{1}$$

The Gram matrix of inner products is represented by the term B, where $B = \{b_{rs}\}$ and $b_{rs} = \vec{x}_r{}^T \vec{x}_s$. In order to determine the embedding, the MDS must obtain the matrix B from D. Considering that a translation of the samples retain their pairwise distances, it is crucial to assume that the data has zero mean. Otherwise, there would exist an infinite number of possible solutions. Manipulating the previous equation leads to the following:

$$d_{rs}^2 = \vec{x}_r^T \vec{x}_r + \vec{x}_s^T \vec{x}_s - 2\vec{x}_r^T \vec{x}_s \tag{2}$$

The average of an arbitrary column s of the matrix D is:

$$\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{r=1}^{n} \vec{x}_r^T \vec{x}_r + \vec{x}_s^T \vec{x}_s \tag{3}$$

While the average of an arbitrary row r is:

$$\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 = \vec{x}_r^T \vec{x}_r + \frac{1}{n}\sum_{s=1}^{n} \vec{x}_s^T \vec{x}_s \tag{4}$$

It is worth noticing that the average of all the elements in D is:

$$\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = \frac{2}{n}\sum_{r=1}^{n} \vec{x}_r^T \vec{x}_r \tag{5}$$

From (2), the term b_{rs} may be expressed as:

$$b_{rs} = \vec{x}_r^T \vec{x}_s = -\frac{1}{2}(d_{rs}^2 - \vec{x}_r^T \vec{x}_r - \vec{x}_s^T \vec{x}_s) \tag{6}$$

Combining Eqs. (3)–(5) yields the following:

$$b_{rs} = -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right) \tag{7}$$

Defining $a_{rs} = -\frac{1}{2}d_{rs}$, one may formulate the following equation:

$$a_{r.} = \frac{1}{n}\sum_{s=1}^{n} a_{rs} \quad a_{.s} = \frac{1}{n}\sum_{r=1}^{n} a_{rs} \quad a_{..} = \frac{1}{n}\sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs} \tag{8}$$

Which leads to:

$$b_{rs} = a_{rs} - a_{r.} - a_{.s} + a_{..} \tag{9}$$

In matrix notation, it is expressed as $B = HAH$, where:

$$H = I - \frac{1}{n}\vec{1}\vec{1}^T \tag{10}$$

The term H is the centring matrix. An eigendecomposition of the matrix B must be performed to find the coordinates of the points in R^d, which is:

$$B = V\Lambda V^T \tag{11}$$

where V is the matrix which columns are the eigenvectors of B, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix with the eigenvalues of B. Considering that $X^T X = V\Lambda V^T$, the intrinsic coordinates are expressed as follows:

$$B = \Lambda^{1/2} V^T \tag{12}$$

The computational complexity of ISOMAP is $O(n^3)$.

3 Kernel Density Estimation

The KDE is a nonparametric statistical technique to estimate the probability density function of a random variable [24, 25]. Let $\{x_1, \ldots, x_n\}$ be an i.i.d. sample from an 1-D random variable x, with unknown density function $f(x)$. The KDE of $f(x)$ is given by:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{13}$$

where $K(x)$ is the kernel function and h is the bandwidth—i.e., a parameter that controls the degree of smoothing of the density estimate. Several kernel functions have been proposed and successfully applied in many problems. The Gaussian, Epanechnikov, uniform, and triangular are among the most relevant cases. In the present work, Gaussian kernels are adopted. This is due to the fact that they are capable of providing a reasonable approximation for many distributions in the datasets included in the analysis. Moreover, Gaussian kernels should preserve properties that allow for some simplification of otherwise costly calculations [26].

3.1 Bandwidth Estimation Methods

The choice of the bandwidth h is pivotal for an appropriate estimation of the unknown density function. Large values of h result in over-smoothing, where $\hat{f}_h(x)$ becomes unimodal and with a large variance. Conversely, small values of h commonly lead to noisy data, yielding large variations in $\hat{f}_h(x)$ for points in the neighbourhood of x. The optimal bandwidth value is a trade-off between a smooth constraint and data fidelity.

In the case that both the kernel function and the unknown density are Gaussian, then the optimal bandwidth in terms of a minimum (IMSE) may be computed through the Silverman's rule of thumb [27]:

$$h_{SIL} = 0.9 \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-1/5} \tag{14}$$

where $\hat{\sigma}$ is the standard deviation of the samples, $IQR = Q_3 - Q_1$ is the interquartile range, and n is the sample size. In addition, still under the Gaussian assumption, the Scott's rule for bandwidth estimation rule is also optimal in terms of the IMSE [28], as follows:

$$h_{SC} = 3.49\hat{\sigma} n^{-1/3} \tag{15}$$

where $\hat{\sigma}$ is the standard deviation of the samples, and n is the sample size.

4 KDE-Based Entropic Isometric Feature Mapping

The parametric principal component analysis (PCA) metric learning algorithm is the main inspiration to devise the KDE–ISOMAP. The former computes the entropic covariance matrix of the data, adopting information-theoretic divergences between densities estimated in local patches along the neighbourhood graph [29]. The primary distinction between the existing ISOMAP and the proposed KDE–ISOMAP lies in the first phase of the algorithm.

The data matrix is denoted by $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$, where each column $\vec{x}_i \in R^m$ represents an observation. The $G = (V, E)$ refers to the ϵ-neighbourhood graph induced from X by creating an edge between each pair of samples $\vec{x}_i$ and $\vec{x}_j$ if $d_E(\vec{x}_i, \vec{x}_j) < \epsilon$, where $d_E(., .)$ is the regular Euclidean distance. Consequently, a patch P_i may be defined as the set formed by a sample $\vec{x}_i$ and its neighbourhood. For a sufficiently high sample density, it belongs to a single Euclidean subspace [30].

Let k_i be the number of $\vec{x}_i$ neighbours in G, then the patch P_i is the following $(k+1) \times m$ matrix:

$$P_i = \begin{bmatrix} \vec{x}_i \\ \vec{x}_{i1} \\ \vec{x}_{i2} \\ \vdots \\ \vec{x}_{ik_i} \end{bmatrix} = \begin{bmatrix} \vec{x}_i(1) & \vec{x}_i(2) & \cdots & \vec{x}_i(m) \\ \vec{x}_{i1}(1) & \vec{x}_{i1}(2) & \cdots & \vec{x}_{i1}(m) \\ \vec{x}_{i2}(1) & \vec{x}_{i2}(2) & \cdots & \vec{x}_{i2}(m) \\ \vdots & \vdots & \ddots & \vdots \\ \vec{x}_{ik_i}(1) & \vec{x}_{ik_i}(2) & \cdots & \vec{x}_{ik_i}(m) \end{bmatrix}$$

The k_i neighbours of $\vec{x}_i$ in the ϵ-neighbourhood network are denoted by $\{\vec{x}_{i1}, \ldots, \vec{x}_{ik_i}\}$. Consider each column of the matrix P_i being a sample of an 1-D random variable x_k, with a probability density function $f(x_k)$. These probability density functions $f(x_k)$ are estimated for $k = 1, \ldots, m$ for each patch of the ϵ-neighbourhood graph using the KDE—which is a nonparametric approach. Because each patch includes m 1-D densities and the graph contains n patches, the total number of nonparametric densities is nm, resulting in a computational cost of $O(n^3 m)$ to run the KDE–ISOMAP.

The relative entropies (KL-divergences) between the densities predicted for each pair of nearby patches P_i and P_j are used to replace the pointwise Euclidean distances in the edges $(\vec{x}_i, \vec{x}_j) \in E$. This is known as the entropic ϵ-neighbourhood graph. Having precisely m pairings of 1-D densities, there are m KL-divergences to compute for each pair of patches P_i and P_j. The KL-divergence of distributions $\vec{p} = [p_1, \dots, p_L]$ and $\vec{q} = [q_1, \dots, q_L]$, where L is the number of points (bins) employed in the KDE, may be calculated as follows:

$$D_{KL}(\vec{p}, \vec{q}) = \frac{1}{L} \sum_{i=1}^{L} p_i \log\left(\frac{p_i}{q_i}\right) \tag{16}$$

Moreover, the symmetrized KL-divergence may be found as detailed below:

$$\begin{aligned} D_s(\vec{p}, \vec{q}) &= \frac{1}{2}\left(D_{KL}(\vec{p}, \vec{q}) + D_{KL}(\vec{q}, \vec{p})\right) \\ &= \frac{1}{2}\left(H(\vec{p}, \vec{q}) - H(\vec{p}) + H(\vec{q}, \vec{p}) - H(\vec{q})\right) \\ &= \frac{1}{2}\left(H(\vec{p}, \vec{q}) + H(\vec{q}, \vec{p})\right) - \frac{1}{2}\left(H(\vec{p}) + H(\vec{q})\right) \end{aligned} \tag{17}$$

This reflects the average of the cross-entropies minus the average of the individual entropies. A vector of relative entropies $\vec{\Psi}_{ij}$ is built after computing the KL-divergences between the m pairings of 1-D densities in P_i and P_j:

$$\vec{\Psi}_{ij} = \left[D_s(\vec{p}_1, \vec{q}_1), \dots, D_s(\vec{p}_m, \vec{q}_m)\right] \tag{18}$$

Lastly, the weight of the edge $(\vec{x}_i, \vec{x}_j) \in E$ is replaced by:

$$w_{ij} = \vec{\Psi}_{ij}^T \vec{\Psi}_{ij} = \|\vec{\Psi}_{ij}\|^2 \tag{19}$$

This leads to the entropic neighbourhood graph. The second and third phases of the KDE–ISOMAP are the same ones adopted in the regular ISOMAP algorithm.

5 Results

Two sets of computational experiments are conducted to evaluate the performance of the proposed KDE–ISOMAP for dimensionality reduction-based metric learning. Firstly, a comparison of the clusters obtained after mapping the data onto a two-dimensional subspace, adopting the SC to measure the difference in terms of clustering fit. Secondly, after the same feature extraction process, a comparison is performed between the average classification accuracy of the following widely adopted supervised classifiers: KNN, support vector machines (SVM), and Bayesian classifier under the Gaussian hypothesis.

In the case that the underlying metrics are successfully learnt, this should be reflected through a substantial increase in the SC and clustering fit measures while examining many multivariate datasets. The proposed KDE–ISOMAP is directly evaluated with the following seven competing methods: PCA, kernel PCA (KPCA), ISOMAP, locally linear embedding (LLE), Laplacian eigenmaps (LAP), t-SNE, and UMAP. In addition, the KDE–ISOMAP is modelled considering three different variations: the KDE–ISOMAP with fixed bandwidth $\mathrm{h} = 0.1$ for all probability density functions (K-ISO-F), the KDE–ISOMAP with Silverman's rule for bandwidth estimation (K-ISO-SIL), and the KDE–ISOMAP with Scott's rule for bandwidth estimation (K-ISO-SC). In the experiments, the number of density points (bins) used in the KDE is set to $L = 256$.

All datasets included in the analysis are publicly available at openML.org/url, along with details on their respective number of instances, features, and classes. The results of the first set of experiments are reported in Table 1. It is worth noticing that in 28 out of 30 datasets, one of the KDE–ISOMAP versions yields the best SC value, corresponding to almost 93% of the cases. The averages and medians resulted from the proposed method outperform the existing methodological alternatives.

A nonparametric Friedman test is performed to check whether the superior results of the KDE–ISOMAP are statistically significant. The null hypothesis that all groups are identical is rejected ($p = 1.11 \times 10^{-16}$), considering a significance level of $\alpha = 0.05$. Moreover, a post-hoc Nemenyi test is applied to determine if groups are statistically different between themselves. At a significance level of $\alpha = 0.05$, this test indicates a considerably higher SC of the K-ISO-F, K-ISO-SIL, and K-ISO-SC compared to the PCA, KPCA, ISOMAP, LLE, LAP, t-SNE, and UMAP. The p values of these tests are reported in Table 2. There is no evidence that the K-ISO-F and K-ISO-SC vary in terms of the SC ($p = 0.965$). Similar results are reported for the K-ISO-F and K-ISO-SIL ($p = 0.982$) as well as the K-ISO-SIL and K-ISO-SC ($p = 0.949$).

Subsequently to the dimensionality reduction-based metric learning, in the second set of experiments 50% of the samples of each of the datasets are used to train three different classifiers. Those refer to the Bayesian classifier under the Gaussian hypothesis with different covariance matrices for each class (i.e., parametric and quadratic classifier), the SVM with no kernel (i.e., nonparametric and linear classifier), and the KNN with $K = 7$ (i.e., nonparametric and nonlinear classifier). The 50% remaining samples from the test set are classified using those three classifiers, being selected the classifier with the highest accuracy to assess how each metric learning method affects supervised classification. The findings of this analysis are reported in Table 3. Remarkably, one of the three variations of the proposed KDE–ISOMAP yields the largest classification accuracy in 26 out of 30 datasets, corresponding to 86% of the cases.

A nonparametric Friedman test is performed to determine whether the prevailing results in terms of classification accuracy achieved by the KDE–ISOMAP are statistically significant. There is strong evidence against the null hypothesis that all groups are identical, considering a significance level of $\alpha = 0.05$. A post-hoc Nemenyi test is also applied to determine whether the groups are equivalent. The test finds that the K-ISO-F, K-ISO-SIL, and K-ISO-SC provide a significantly higher classification accuracy compared to the PCA, KPCA, ISOMAP, LLE, LAP, t-SNE, and UMAP.

Table 1 Silhouette coefficients (SC) for clusters generated by the algorithms PCA, KPCA, ISOMAP, LLE, LAP, KDE–ISOMAP (K-ISO-F, K-ISO-SIL, K-ISO-SC), *t*-SNE, and UMAP for 30 openML.org datasets (2-D case)

Dataset/statistic	PCA	KPCA	ISO	LLE	LAP	KISO-F	KISO-SIL	KISO-SC	*t*-SNE	UMAP
iris	0.401	0.469	0.452	0.365	0.541	0.588	0.597	**0.619**	0.494	0.526
wine	0.526	0.610	0.547	0.242	0.750	0.742	**0.766**	0.765	0.556	0.605
prnn_crabs	0.040	0.030	0.037	0.022	0.028	**0.156**	0.117	0.130	0.038	0.048
happiness	−0.067	−0.062	−0.066	−0.068	−0.063	**0.000**	**0.000**	**0.000**	−0.064	−0.052
mux6	0.072	0.064	−0.015	**0.084**	−0.014	0.028	0.068	0.038	0.048	0.037
parity5	−0.062	−0.047	−0.048	−0.051	−0.036	**0.000**	**0.000**	**0.000**	−0.016	−0.053
Hayes-roth	−0.023	0.038	−0.010	−0.013	−0.013	0.090	0.160	**0.215**	−0.012	−0.013
aids	−0.022	−0.027	−0.027	−0.037	−0.033	**0.090**	0.054	0.054	−0.013	−0.018
pm10	0.000	0.002	0.000	0.000	0.000	**0.006**	0.004	0.002	0.000	0.000
strikes	0.007	0.008	0.004	0.002	0.007	**0.027**	0.025	0.025	0.008	0.019
disclosure_z	−0.002	0.006	−0.002	−0.001	−0.002	0.005	**0.008**	0.008	0.000	0.000
diggle_table_a2	0.406	0.409	0.450	0.328	0.304	**0.673**	0.645	0.639	0.431	0.199
Monks-problem	0.024	0.001	0.000	0.000	−0.002	**0.062**	0.04	0.036	0.029	−0.001
Breast-tissue	−0.029	−0.030	−0.017	−0.081	−0.018	−0.007	−0.012	−0.022	−0.024	−**0.001**
planning-relax	−0.002	−0.011	−0.004	0.003	−0.004	**0.076**	0.035	0.046	−0.006	−0.003
haberman	0.060	−0.024	0.061	−0.004	−0.032	**0.269**	0.175	0.188	−0.017	−0.027
KnuggetChase3	0.199	0.070	0.187	0.077	0.091	0.488	0.502	**0.510**	0.062	0.063

Table 1 continued

Dataset/statistic	PCA	KPCA	ISO	LLE	LAP	KISO-F	KISO-SIL	KISO-SC	t-SNE	UMAP
bolts	0.337	0.254	0.286	0.028	0.317	**0.573**	0.425	0.556	0.159	0.347
fl2000	0.180	0.043	0.119	0.073	0.025	0.128	**0.302**	0.301	0.045	0.014
Engine1	−0.133	−0.032	−0.149	−0.170	−0.168	−0.243	−0.049	**0.053**	−0.118	−0.091
fri_c2_100_10	0.099	0.059	0.083	0.021	0.093	**0.122**	0.113	0.117	0.096	0.092
Vineyard(2)	0.277	0.262	0.280	0.191	0.252	**0.415**	0.357	**0.415**	0.301	0.379
diabetes_numeric(2)	0.092	0.081	0.093	0.085	0.089	0.167	**0.171**	0.170	0.069	0.072
prnn_fglass	0.018	0.004	0.011	0.029	−0.009	**0.078**	0.017	0.039	0.052	0.026
parkinsons	0.130	0.155	0.114	0.002	0.242	0.191	**0.339**	0.249	0.193	0.203
Acute-inflammations(2)	0.278	0.315	0.266	0.113	0.081	**0.371**	0.367	0.367	0.247	0.355
blogger	0.036	−0.011	0.052	0.029	0.003	**0.398**	**0.398**	0.397	0.019	−0.033
prnn_viruses	0.371	0.118	0.112	0.496	0.232	**0.490**	0.358	0.363	0.213	0.079
analcatdata_creditscore	0.111	0.081	0.131	0.071	0.049	**0.389**	0.257	0.315	0.007	0.003
Confidence(2)	0.173	0.214	0.123	−0.142	−0.087	0.540	**0.546**	0.533	0.122	0.056
Average	0.117	0.102	0.102	0.056	0.087	0.230	0.226	**0.238**	0.097	0.094
Median	0.066	0.041	0.057	0.022	0.005	0.142	0.166	**0.179**	0.042	0.023
Minimum	−0.133	−0.062	−0.149	−0.170	−0.168	−0.243	−0.049	**−0.022**	−0.118	−0.091
Maximum	0.526	0.610	0.547	0.496	0.750	0.742	**0.766**	0.765	0.556	0.605
SD	0.165	0.167	0.164	**0.144**	0.192	0.246	0.228	0.230	0.164	0.176

The bold value indicates the best result in each row of the table (the highest score for all rows, except for the last row, which is the standard deviation and the minimum is the best)

Table 2 Results of the post-hoc Nemenyi tests for the silhouette coefficient (SC)

Method	K-ISO-F	K-ISO-SIL	K-ISO-SC
PCA	1.79×10^{-6}	1.99×10^{-6}	1.44×10^{-6}
KPCA	1.75×10^{-7}	1.97×10^{-7}	1.39×10^{-7}
ISOMAP	4.81×10^{-8}	5.43×10^{-8}	3.78×10^{-8}
LLE	1.20×10^{-11}	1.39×10^{-11}	8.95×10^{-12}
LAP	2.42×10^{-10}	2.77×10^{-10}	1.83×10^{-10}
t-SNE	5.43×10^{-7}	6.07×10^{-7}	4.35×10^{-7}
UMAP	1.56×10^{-7}	1.75×10^{-7}	1.24×10^{-8}

The p values for the tests performed are reported in Table 4. There is no evidence that the K-ISO-F and K-ISO-SC vary in terms of classification accuracy ($p = 0.550$). Similar results hold true regarding the K-ISO-F and K-ISO-SIL ($p = 0.508$) as well as the K-ISO-SIL and K-ISO-SC ($p = 0.949$).

A convenient attribute of the proposed KDE–ISOMAP refers to its strategy to address the out-of-sample problem in manifold learning. Most unsupervised metric learning algorithms are not capable of appropriately handling new samples that are not part of the training set. A natural choice is to include such new samples to the dataset and perform another full training round, which may be time consuming. It is worth noticing that the ISOMAP is directly related to the KPCA. In fact, the KPCA becomes the ISOMAP when the kernel matrix $K(\vec{x}_i, \vec{x}_j)$ is defined as minus one-half of the geodesic distance matrix [31]. Thus, it is possible to tackle out-of-sample instances thorugh the KDE–ISOMAP while adopting the same projection strategy of the KPCA.

A caveat of the KDE–ISOMAP method refers the specification of the parameter ϵ (radius), which determines the patch size—i.e., number of neighbours of a particular sample in the ϵ-neighbourhood graph. Tests report that the classification accuracy and SC are substantially affected by changes in this parameter. In the present work, the following strategy is employed: For each dataset, the complete graph is built by linking a sample to every other sample. Then, for each sample $\vec{x}_i$, the approximate distribution of the distances from $\vec{x}_i$ to any other sample $\vec{x}_j$ is computed. In addition, for each dataset the whole network is constructed by connecting each sample to every other sample. Subsequently, for each sample $\vec{x}_i$ the estimated distribution of the distances between any other sample $\vec{x}_j$ and itself is computed.

Previous works report that the percentiles p of distributions that fall within the range $P = [1, 20]$ provide the most suitable values of ϵ (radius). Thus, in the best model selection process, several percentiles of this distribution are tested as the radius that maximizes the classification accuracy among all values of $p \in P$. Differently from an KNN graph, this graph is not regular due to the fact that the degree of the vertices may be considerably different. It is worth mentioning that, besides using the class labels to perform a model selection, the feature extraction stage is fully unsupervised—i.e., the KDE–ISOMAP performs unsupervised metric learning.

To illustrate how the KDE-ISOMAP may improve clustering and classification accuracy by learning a suitable metric, Fig. 1 depicts scatterplots exploring the AIDS dataset—after reducing the number of features into two. In comparison with the

Table 3 Maximum accuracy between the KNN, SVM, and Bayesian classifiers after dimensionality reduction with the PCA, KPCA, ISOMAP, LLE, LAP, KDE-ISOMAP (K-ISO-F, K-ISO-SIL, K-ISO-SC), *t*-SNE and UMAP for 30 openML.org datasets (2-D case)

Dataset/statistic	PCA	KPCA	ISO	LLE	LAP	KISO-F	KISO-SIL	KISO-SC	*t*-SNE	UMAP
iris	0.960	0.866	0.920	0.973	0.840	0.973	**1.000**	**1.000**	0.986	**1.000**
wine	0.966	0.977	0.989	0.797	0.989	0.989	0.988	**1.000**	0.943	0.943
prnn_crabs	0.620	0.660	0.610	0.710	0.620	0.860	**0.900**	0.870	0.820	0.810
happiness	0.333	0.266	0.333	0.266	0.300	**0.567**	0.400	0.433	0.400	0.400
mux6	0.609	0.703	0.703	0.734	0.546	0.656	0.750	0.688	**0.812**	0.734
parity5	0.500	0.437	**0.625**	0.437	0.437	0.500	0.500	0.500	0.562	0.375
Hayes-roth	0.606	0.621	0.636	0.606	0.606	**0.818**	**0.818**	0.788	0.727	0.606
aids	0.480	0.480	0.440	0.480	0.480	0.560	**0.760**	0.600	0.520	0.520
pm10	0.532	0.532	0.536	0.512	0.512	0.564	0.576	0.560	**0.588**	0.536
strikes	0.638	0.648	0.648	0.661	0.661	0.757	0.767	0.770	**0.853**	0.750
disclosure_z	0.531	0.558	0.525	0.519	0.519	0.568	0.565	**0.586**	0.558	0.549
diggle_table_a2	0.877	0.974	0.916	0.883	0.929	0.968	0.968	**0.974**	0.948	0.948
Monks-problem	0.604	0.600	0.589	0.647	0.669	**0.766**	0.676	0.687	0.748	0.733
Breast-tissue	0.415	0.584	0.490	0.547	0.547	0.642	**0.660**	0.623	0.566	0.509
planning-relax	0.714	0.714	0.714	0.714	0.714	**0.725**	**0.725**	**0.725**	0.714	0.714
haberman	0.790	0.764	0.764	0.764	0.764	**0.797**	0.791	**0.797**	0.764	0.777
KnuggetChase3	0.804	0.793	0.793	0.793	0.793	**0.814**	**0.814**	0.804	0.804	0.793

Table 3 continued

Dataset/statistic	PCA	KPCA	ISO	LLE	LAP	KISO-F	KISO-SIL	KISO-SC	*t*-SNE	UMAP
bolts	0.850	0.950	0.900	0.700	0.850	**0.950**	**0.950**	**0.950**	0.850	0.900
fl2000	0.647	0.676	0.647	0.647	0.617	0.676	0.676	**0.706**	0.647	0.617
Engine1	0.791	0.89	0.885	0.765	0.885	0.906	0.927	**0.938**	0.901	0.906
fri_c2_100_10	0.740	0.680	0.700	0.540	0.680	0.820	0.820	**0.860**	0.760	0.700
Vineyard(2)	0.846	0.807	0.807	0.846	0.846	**0.885**	**0.885**	**0.885**	0.807	0.807
diabetes_numeric(2)	0.681	0.590	0.681	0.681	0.636	0.681	0.681	**0.727**	0.681	0.590
prnn_fglass	0.757	0.719	0.757	0.635	0.710	0.738	**0.757**	0.710	0.728	0.710
parkinsons	0.897	0.897	0.816	0.836	0.806	0.898	**0.908**	**0.908**	0.897	0.846
Acute-inflammations(2)	**1.000**	**1.000**	**1.000**	**1.000**	0.967	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
blogger	0.660	0.680	0.760	0.740	0.680	**0.800**	**0.800**	0.780	**0.800**	0.650
prnn_viruses	0.839	0.806	0.741	0.774	0.774	0.839	**0.871**	**0.871**	0.774	0.774
analcatdata_creditscore	**0.840**	0.760	0.820	0.760	0.780	0.820	**0.840**	**0.840**	0.820	0.820
Confidence(2)	0.833	0.833	0.888	0.861	0.833	**0.917**	**0.917**	**0.917**	0.833	0.833
Average	0.712	0.716	0.721	0.694	0.700	0.782	**0.790**	0.783	0.760	0.728
Median	0.727	0.709	0.728	0.712	0.695	**0.807**	**0.807**	0.793	0.787	0.742
Minimum	0.333	0.266	0.333	0.266	0.300	**0.500**	0.400	0.433	0.400	0.375
Maximum	**1.000**	**1.000**	**1.000**	**1.000**	0.989	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
SD	0.169	0.172	0.162	0.160	0.164	**0.143**	0.151	0.154	0.145	0.167

The bold value indicates the best result in each row of the table (the highest score for all rows, except for the last row, which is the standard deviation and the minimum is the best)

Table 4 Results of the post-hoc Nemenyi tests for classification accuracy

Method	K-ISO-F	K-ISO-SIL	K-ISO-SC
PCA	6.84×10^{-6}	2.47×10^{-7}	3.47×10^{-7}
KPCA	5.56×10^{-6}	1.56×10^{-7}	2.21×10^{-7}
ISOMAP	4.57×10^{-6}	1.57×10^{-7}	2.20×10^{-7}
LLE	1.25×10^{-8}	2.10×10^{-10}	3.18×10^{-10}
LAP	8.24×10^{-10}	1.03×10^{-11}	1.61×10^{-11}
t-SNE	2.80×10^{-2}	4.00×10^{-3}	5.00×10^{-3}
UMAP	2.01×10^{-5}	8.43×10^{-7}	1.16×10^{-6}

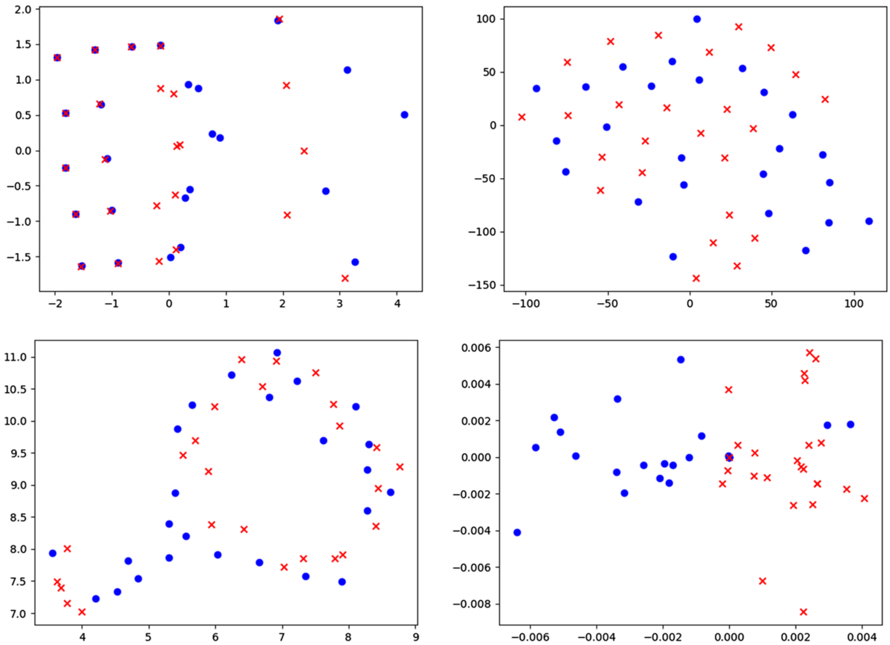

Fig. 1 Scatterplots for the AIDS dataset after dimensionality reduction. From left to right and top to bottom, these refer to the ISOMAP, t-SNE, UMAP, and KDE–ISOMAP. In cases where the number of samples is limited, the t-SNE and UMAP tend to perform below the expectations due to numerical optimization

original ISOMAP, t-SNE, and UMAP, the proposed KDE–ISOMAP provides less overlapping samples in terms of data discrimination. The two classes (male and female—represented as circles and crosses, respectively) are more clearly identified in the KDE-ISOMAP compared to popular competing methods.

6 Conclusion

It is highly desirable to overcome limitations of the Euclidean distance while extracting nonlinear characteristics from data. Manifold learning and unsupervised metric learning are commonly used interchangeably due to the fact that both aim to find intrinsic geometric structures of data. In the present work, the relative entropy between distributions estimated from patches along the ϵ-neighbourhood graph is introduced. This consists of an alternative to the Euclidean distance through an entropic ISOMAP based on the KDE method.

Computational analyses confirm two key points to consider the proposed KDE–ISOMAP as a competing alternative regarding widely adopted learning algorithms. Firstly, the nonlinear features of the KDE–ISOMAP may be more discriminative in supervised classification than features produced using state-of-the-art manifold learning algorithms. Secondly, the KDE–ISOMAP provides superior quality clustering compared to the same competing algorithms. In particular, the superiority of the proposed KDE–ISOMAP compared to the t-SNE and UMAP becomes clearer while exploring datasets with limited number of samples. This is due to the fact that these two existing algorithms require numerical optimization methods that are overly dependent upon the sample size.

The main contribution of the proposed KDE–ISOMAP framework refers to its robustness to the presence of noisy date and outliers. This is contrast to the pointwise Euclidean distance because of the fact that the KDE–ISOMAP adopts a patch-based distance function to measure similarity between samples. Furthermore, considering that a projection matrix may be created, tackling out-of-sample data becomes simpler due to the relationship between the KPCA and ISOMAP. In addition, it is worth mentioning the fact that the KDE–ISOMAP is suitable for small sample sizes, thus not requiring a large amount of data for convergence purposes. This is an advantage compared to autoencoders and other deep-learning-based algorithms.

Suggestions of future work include the adoption of further information-theoretic divergences and families of entropies, such as the Hellinger, Bhattacharyya, Cauchy–Schwarz, total variation divergences, Renyi, and Sharma–Mittal entropies. Moreover, a supervised version of the KDE–ISOMAP may be devised by combining the relative entropy and Euclidean distance. Such a framework would adopt a single similarity measure to weight the edge for neighbouring samples of a same class, whereas the sum of the distances may discourage the shortest paths from crossing that edge for neighbouring samples belonging to different classes.

Author Contributions All authors contributed equally to this work. The authors confirm responsibility for the following: study conception and design, data coding, analysis and interpretation of the results, and manuscript preparation.

Funding This study is partially funded by the the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil—Finance Code 001.

Data Availability All datasets used in the experiments are publicly available at www.openml.org.

Code Availability The source code of the Python implementation used to generate the results of this paper is available at: https://github.com/alexandrelevada/KDE_ISOMAP.

Declarations

Conflict of interest The authors declare there is no competing interests.

Ethical Statements The authors declare that this work is original and is not under review elsewhere.

References

1. Shi Y (2022) Advances in big data analytics. Springer, Singapore
2. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill/Irwin, New York
3. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer, London
4. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4:149–178
5. Van Der Maaten L, Postma EO, Herik HJ (2009) Dimensionality reduction: a comparative review. J Mach Learn Res 10(66–71):1–41
6. Fukunaga K (2013) Introduction to statistical pattern recognition. Elsevier, Amsterdam
7. Wang F, Sun J (2015) Survey on distance metric learning and dimensionality reduction in data mining. Data Min Knowl Discov 29(2):534–564
8. Li D, Tian Y (2018) Survey and experimental study on metric learning methods. Neural Netw 105:447–462
9. Wu W, Tao D, Li H, Yang Z, Cheng J (2021) Deep features for person re-identification on metric learning. Pattern Recognit 110:107424
10. Tenenbaum JB, Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
11. Cox T, Cox M (2000) Multidimensional Scaling (2nd ed.). Chapman and Hall/CRC
12. Choi H, Choi S (2007) Robust kernel isomap. Pattern Recognit 40(3):853–862
13. Shang F, Jiao LC, Shi J, Chai J (2011) Robust positive semidefinite L-isomap ensemble. Pattern Recognit Lett 32(4):640–649
14. Gan Q, Shen F, Zhao J (2014) An extended isomap for manifold topology learning with SOINN landmarks. In: 22nd international conference on pattern recognition (ICPR 2014), pp 1579–1584
15. Najafi A, Joudaki A, Fatemizadeh E (2016) Nonlinear dimensionality reduction via path-based isometric mapping. IEEE Trans Pattern Anal Mach Intell 38(7):1452–1464
16. Gajamannage K, Paffenroth R, Bollt EM (2019) A nonlinear dimensionality reduction framework using smooth geodesics. Pattern Recognit 87:226–236
17. Budninskiy M, Yin G, Feng L, Tong Y, Desbrun M (2019) Parallel transport unfolding: a connection-based manifold learning approach. SIAM J Appl Algebra Geom 3(2):266–291
18. Shamai G, Zibulevsky M, Kimmel R (2020) Efficient inter-geodesic distance computation and fast classical scaling. IEEE Trans Pattern Anal Mach Intell 42(1):74–85
19. Tasoulis S, Pavlidis NG, Roos T (2020) Nonlinear dimensionality reduction for clustering. Pattern Recognit 107:107508
20. Anowar F, Sadaoui S, Selim B (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Comput Sci Rev 40:100378. https://doi.org/10.1016/j.cosrev.2021.100378
21. McInnes L, Healy J, Melville J (2020) UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv: 1802.03426
22. Maaten LJP, Hinton GE (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579–2605
23. Borg I, Groenen P (2005) Modern multidimensional scaling: theory and applications, 2nd edn. Springer, New York
24. Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27(3):832–837
25. Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33(3):1065–1076

26. Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. Math Geosci 42:487–517. https://doi.org/10.1007/s11004-010-9276-7
27. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall/CRC, New York
28. Scott DW (1979) On optimal and data-based histograms. Biometrika 66(3):605–610
29. Levada ALM (2020) Parametric PCA for unsupervised metric learning. Pattern Recognit Lett 135:425–430
30. Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
31. Ham J, Lee DD, Mika S, Schölkopf B (2004) A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the twenty-first international conference on machine learning (ICML ’04). Association for Computing Machinery, New York, NY, USA, p 47

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.