Local Intrinsic Dimensionality Based Features for Clustering

Paola Campadelli[1], Elena Casiraghi[1],
Claudio Ceruti[2], Gabriele Lombardi[1], and Alessandro Rozza[3]

[1] Dipartimento di Informatica, Università degli Studi di Milano,
Via Comelico 39-41, Milano, Italy
{campadelli,casiraghi,lombardi}@di.unimi.it
[2] Dipartimento di Matematica, Università degli Studi di Milano,
Via Saldini 50, Milano, Italy
claudio.ceruti@unimi.it
[3] Dipartimento di Scienze Applicate, Università degli Studi di Napoli Parthenope
Centro Direzionale di Napoli - Isola C4, Napoli, Italy
alessandro.rozza@uniparthenope.it

Abstract. One of the fundamental tasks of unsupervised learning is dataset clustering, to partition the input dataset into clusters composed by somehow "similar" objects that "differ" from the objects belonging to other classes. To this end, in this paper we assume that the different clusters are drawn from different, possibly intersecting, geometrical structures represented by manifolds embedded into a possibly higher dimensional space. Under these assumptions, and considering that each manifold is typified by a geometrical structure characterized by its intrinsic dimensionality, which (possibly) differs from the intrinsic dimensionalities of other manifolds, we code the input data by means of local intrinsic dimensionality estimates and features related to them, and we subsequently apply simple and basic clustering algorithms, since our interest is specifically aimed at assessing the discriminative power of the proposed features. Indeed, their encouraging discriminative quality is shown by a feature relevance test, by the clustering results achieved on both synthetic and real datasets, and by their comparison to those obtained by related and classical state-of-the-art clustering approaches.

Keywords: Local features, Intrinsic dimensionality, Dataset clustering, Multi-manifold structures.

1 Introduction

At the present, continuous technological advances allow to work on multiple sources to collect increasing amounts of informations, which are coded as vectors of numerical values usually called features. Therefore, a real dataset $\boldsymbol{X}_N$ usually comprises an high number N of D-dimensional feature vectors, that is $\boldsymbol{X}_N \equiv \{\boldsymbol{x}_i\}_{i=1}^N \equiv \{[x_1, \ldots, x_D]_i\}_{i=1}^N \subset \Re^D$. In the pattern recognition field

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 41–50, 2013.

it is often useful to partition $\boldsymbol{X}_N$ in C disjoint classes (generally called clusters) having somehow different peculiarities, thus obtaining: $\boldsymbol{X}_N \equiv \bigcup_{c=1}^{C} \boldsymbol{X}_c$ and $\forall i \neq j, i, j \in \{1, \cdots, C\}, \boldsymbol{X}_i \cap \boldsymbol{X}_j = \oslash$.

For this reason, unsupervised clustering techniques usually aim to create compact neighborhoods (clusters), by considering the clusters' internal homogeneity (similarity) and their external separation (dissimilarity). However, choosing the proper function to measure the similarity and the dissimilarity is often difficult. For this reason, and since clustering approaches are employed in a wide variety of fields, many techniques have been presented that differ for the employed similarity criteria and for the automatic algorithm used to identify the best partition. At the state-of-the-art, among the several recent surveys and comparative researches describing unsupervised biclustering or clustering methods, those reported in [18,7,25] are notable since they deeply consider the problem of clustering high dimensional data belonging to sets eventually having high cardinalities.

To cope with this kind of data, relevant literature works generally compute their lower dimensional projections through classical techniques, e.g. Principal Component Analysis (PCA) and its variants [2], and then apply either classical clustering techniques, such as K-means and its (fuzzy) variants [21], the Expectation Maximization (EM, [8]) algorithm, or algorithms specifically designed for the clustering problem to be handled. Furthermore, the similarity between points is often computed by employing the common Euclidean distance; yet, as explained in [1], this is a quite limited methodology that might not properly capture and express the typifying geometrical structure underlying each cluster, specially in case of high dimensional data. Consequently, though promising results have been obtained, the problem of high dimensional dataset clustering is still open.

For this reason, and based on the aforementioned considerations, recent clustering approaches change their perspective and view the c^{th} cluster as a set of points $\boldsymbol{X}_c = \{\phi_c(\boldsymbol{z}_{i,c})\}_{i=1}^{n_c} \subset \Re^D$ drawn from a low d_c-dimensional space (manifold) $\boldsymbol{\mathcal{M}}_c \subseteq \Re^{d_c}$ and embedded into an higher $D-$dimensional space $\Re^D$ ($d_c \leq D$) through a map $\phi_c(\cdot)$. Under this framework the dimensionality d_c of $\boldsymbol{\mathcal{M}}_c$, generally called intrinsic dimensionality (id), becomes a distinctive feature.

This conceptual framework guarantees that each cluster is strongly typified by the geometrical structure characterizing the cluster as a whole. This intrinsic structure is the one inherited by the feature space $\boldsymbol{\mathcal{M}}_c$ in $\Re^{d_c}$ from which the points of the cluster are supposed to be drawn. Therefore, the clustering of $\boldsymbol{X}_N$ can be achieved by multi-manifold clustering techniques, aimed at identifying the C intersecting manifolds underlying $\boldsymbol{X}_N$, and being (possibly) uniquely identified by their distinctive id (where $d_1 \neq \ldots \neq d_C$). To this aim, most of the few works proposed in literature [11,3,23,24,12] code each point as a vector of local id estimates[1], or local features related to them, with the aim of capturing the geometrical structure underlying the neighborhood of the coded point. Indeed, the discriminative power of these features allows to obtain promising results by employing classical clustering algorithms.

[1] The local id relative to one point is the id estimate computed on its neighborhood.

Considering the works in literature, one of the most related to ours is that described in [3], where the proposed clustering method (hereinafter referred to as NS) is based on local id estimates obtained by exploiting a modified version of the id estimator described in [6]. Precisely, in [6] the authors work on the k-nearest neighbors ($k << N$) of each point to estimate its local id, and then cluster the dataset by employing all the estimated local ids. Note that, as highlighted in Section 4 of [20], when dealing with high id datasets a strong underestimation problem affects most of the neighborhood based id estimators, causing unreliable id estimations and consequent inaccurate clusterings. To reduce this problem in [3] an effective "Neighborhood Smoothing" procedure is employed.

In this paper we describe some features that can be conceptually viewed as local id estimates and local characteristics of the underlying manifold portion (see Section 2); note that these features have been also exploited by effective global id estimators [17,5,4]. In Section 3 we show how they can be effectively exploited by classical clustering algorithms; indeed, a feature relevance test, promising clustering results on both synthetic and real datasets, and their comparison with those achieved by state-of-the-art clustering techniques (see Section 4), show the discriminative quality of the proposed features also when applied to high dimensional points characterized by both high and low ids.

2 Local id-Based Features

In this section we describe the three local features we exploited for dataset clustering; they are derived from two relevant global id estimators [17,5,4], that compute the global id estimate that characterizes the manifold from which the dataset $\boldsymbol{X}_N$ is assumed to be drawn.

The first local feature is successfully employed when estimating the (global) id by the well-known **Maximum Likelihood Estimator** for id (MLE [17]). The rationale of our choice is that this feature, referred as $\hat{d}(\boldsymbol{x}_i, k)$ in the following, can be theoretically viewed as a local id estimate computed in the k-neighborhood of each point $\boldsymbol{x}_i \in \boldsymbol{X}_N$. More precisely, it is computed by treating the neighbors of each point $\boldsymbol{x}_i \in \boldsymbol{X}_N$ as events in a Poisson process and considering the Euclidean distance $r^{(j)}(\boldsymbol{p}_i)$ between the query point $\boldsymbol{x}_i$ and its j^{th} nearest neighbor as the event's arrival time. Since this process depends on the id that characterizes the underlying manifold's portion, MLE estimates it by maximizing the log-likelihood of the observed process. In practice $\hat{d}(\boldsymbol{x}_i, k)$ is computed as:

$$\hat{d}(\boldsymbol{x}_i, k) = \left(\frac{1}{k} \sum_{j=1}^{k} \log \frac{r^{(k+1)}(\boldsymbol{x}_i)}{r^{(j)}(\boldsymbol{x}_i)} \right)^{-1}$$

The other two local features used by our clustering approaches, referred as $\hat{\nu}(\boldsymbol{x}_i, k)$ and $\hat{\tau}(\boldsymbol{x}_i, k)$ in the following, collect informations related to the distribution of the pairwise angles in the k-neighborhood of each $\boldsymbol{x}_i \in \boldsymbol{X}_N$. They have been introduced and used by the global id estimators proposed in [5,4] since they express further, and different, informations about the local geometry of the unknown manifold's portions underlying each data neighborhood. Indeed, their exploitation has shown to improve the reliability of the computed id estimates since they allow to reduce the underestimation problem [20] that affects most of the id estimators when applied to high id data (for details see [5,4]).

The theoretical basis of these features is expressed by the following theorem proved in [5,4]:

Theorem 1. *Given two independent random unit vectors* $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ *in* $\Re^d$*, drawn from a uniform distribution on* S^{d-1}*, for increasing values of* d *the concentration parameter* τ *of the von Mises (VM) distribution describing the angle* θ *between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *converges asymptotically to the dimensionality* d*.*

Taking into account this theorem, it is possible to consider local neighborhoods in $\boldsymbol{X}_N$ to capture the information provided by the concentration of pairwise angles. Practically, for each point $\boldsymbol{x}_i \in \boldsymbol{X}_N$ its k nearest neighbors $\bar{\boldsymbol{X}}_k^i$ are identified, they are centered to obtain $\hat{\boldsymbol{X}}_k^i = \left\{\boldsymbol{x}_j - \boldsymbol{x}_i : \forall\, \boldsymbol{x}_j \in \bar{\boldsymbol{X}}_k^i\right\}$, and then used to compute:

$$\theta(\boldsymbol{x}_z, \boldsymbol{x}_j) = \arccos \frac{\boldsymbol{x}_z \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_z\|\|\boldsymbol{x}_j\|}. \tag{1}$$

Employing Equation (1) the $\binom{k}{2}$ angles of all the possible pairs of vectors in $\hat{\boldsymbol{X}}_k^i$ are computed to compose the vector $\hat{\boldsymbol{\theta}}_i = \{\theta(\boldsymbol{x}_z, \boldsymbol{x}_j) : \forall\, \boldsymbol{x}_z,\ \boldsymbol{x}_j \in \hat{\boldsymbol{X}}_k^i\}_{1 \le z < j \le k}$. Considering Theorem 1, each component of $\hat{\boldsymbol{\theta}}_i = \left[\theta_1, \cdots, \theta_{\binom{k}{2}}\right]$ follows a VM pdf of parameters $\nu(\boldsymbol{x}_i, k)$ and $\tau(\boldsymbol{x}_i, k)$; therefore, the Maximum Likelihood (ML) of the population direction $\nu(\boldsymbol{x}_i, k)$ equals the sample mean direction:

$$\hat{\nu}(\boldsymbol{x}_i, k) = \mathrm{atan}_2\left(\sum_{j=1}^{\binom{k}{2}} \sin\theta_j, \sum_{j=1}^{\binom{k}{2}} \cos\theta_j\right)$$

where $\mathrm{atan}_2(x, y)$ is the arc tangent of y/x.

Likewise, the ML of the concentration parameter $\tau(\boldsymbol{x}_i, k)$ equals the estimate $\hat{\tau}(\boldsymbol{x}_i, k)$ calculated as a solution of $\eta(\boldsymbol{x}_i, k) = \frac{I_1(\tau(\boldsymbol{x}_i,k))}{I_0(\tau(\boldsymbol{x}_i,k))} \equiv A(\tau(\boldsymbol{x}_i, k))$, where I_v is the modified Bessel function of the first kind with order v, and $\eta(\boldsymbol{x}_i, k)$ is the norm of the sample mean vector defined in [22] as:

$$\eta(\boldsymbol{x}_i, k) = \sqrt{\left(\frac{1}{\binom{k}{2}} \sum_{j=1}^{\binom{k}{2}} \cos\theta_j\right)^2 + \left(\frac{1}{\binom{k}{2}} \sum_{j=1}^{\binom{k}{2}} \sin\theta_j\right)^2}$$

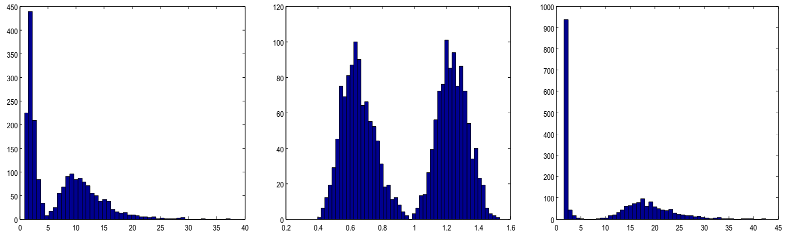
Being A a non invertible function, in [5,4] $A^{-1}(\eta(\boldsymbol{x}_i, k))$ is approximated by:

$$\hat{\tau}(\boldsymbol{x}_i, k) = \widetilde{A}^{-1}(\eta(\boldsymbol{x}_i, k)) = \begin{cases} 2\eta(\boldsymbol{x}_i, k) + \eta(\boldsymbol{x}_i, k)^3 + \frac{5\eta(\boldsymbol{x}_i,k)^5}{6} & \eta(\boldsymbol{x}_i) < 0.53 \\ -0.4 + 1.39\eta(\boldsymbol{x}_i, k) + \frac{0.43}{1-\eta(\boldsymbol{x}_i,k)} & 0.53 \le \eta(\boldsymbol{x}_i, k) < 0.85 \\ \frac{1}{\eta(\boldsymbol{x}_i,k)^3 - 4\eta(\boldsymbol{x}_i,k)^2 + 3\eta(\boldsymbol{x}_i,k)} & \eta(\boldsymbol{x}_i, k) \ge 0.85 \end{cases}$$

3 The Clustering Approaches

In this section we describe how we employ either the classical Expectation Maximization algorithm (EM, [8]) or a simple variant of the Label Propagation Algorithm (LPA, [19]) to cluster input datasets coded by the geometrical features described in Section 2; we underline that our choice of using classical and simple clustering techniques is motivated by the fact that our aim in this research work is to assess the discriminative ability of the features we are proposing.

Precisely, we consider input sets composed of C clusters, that is $\boldsymbol{X}_N = \{\boldsymbol{x}_i\}_{i=1}^N = \{\{\phi_c(\boldsymbol{z}_{i,c})\}_{i=1}^{n_c}\}_{c=1}^C \subset \Re^D$ (n_c is the cardinality of the c^{th} cluster, and $N = \sum_{c=1}^C n_c$), and we assume that the C clusters are composed of independent identically distributed points $\boldsymbol{z}_{i,c}$ drawn from C different low-dimensional manifolds (possibly characterized by different ids) embedded in the higher dimensional space $\Re^D$ by (possibly different) proper maps ϕ_c ($\phi_1 \neq \phi_2 \neq \cdots \neq \phi_C$). Our aim is to exploit $\hat{d}(\boldsymbol{x}_i, k)$, $\hat{\nu}(\boldsymbol{x}_i, k)$, and $\hat{\tau}(\boldsymbol{x}_i, k)$ to cluster $\boldsymbol{X}_N$.



(a) $\hat{d}(\boldsymbol{x}_i, k)$ (b) $\hat{\nu}(\boldsymbol{x}_i, k)$ (c) $\hat{\tau}(\boldsymbol{x}_i, k)$

Fig. 1. The three local parameters computed on points drawn by two manifolds having id $= 2$ and id $= 14$ embedded in $\Re^{30}$, and using $k = 30$

The first three clustering approaches we experimented employ only one of the aforementioned features. Precisely, the algorithm transforms each point $\boldsymbol{x}_i \in \boldsymbol{X}_N \subset \Re^D$ in a unique real value $y_i \in \Re$ by computing the local feature being used; this allows to obtain a 1-dimensional set $\{y_i\}_{i=1}^N = \boldsymbol{Y}_N \subset \Re$. If we consider points belonging to different manifolds, each feature tends to be distributed as a mixture of gaussian distributions (see Figure 1). For this reason we chose to employ EM on $\boldsymbol{Y}_N$. Precisely, being $h(s)$ the pdf of the event y_i and assuming that h is the sum of C normal distributions $h(s) = \sum_{c=1}^C \omega_c \mathcal{N}(y_i|\mu_c, \sigma_c)$, $\omega_c = {}^{n_c}/_N$, EM estimates from $\boldsymbol{Y}_N$ the means $\{\mu_1, ..., \mu_C\}$ and the standard deviations $\{\sigma_1, ..., \sigma_C\}$ of the C normal distributions. Employing the estimated values, each point $\boldsymbol{x}_i$ (coded as y_i) is associated to the cluster $c \in \{1, ..., C\}$ maximizing the probability $g_c(y_i, \mu_c, \sigma_c)$, i.e. the probability of a given y_i to belong to the c^{th} cluster. $g_c(y_i, \mu_c, \sigma_c)$ is defined as follows:

$$g_c(y_i, \mu_c, \sigma_c) = \frac{1}{\sigma_c\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \mu_c}{\sigma_c}\right)^2}$$

The three algorithms obtained by employing $\hat{d}(\boldsymbol{x}_i, k)$, $\hat{\nu}(\boldsymbol{x}_i, k)$, and $\hat{\tau}(\boldsymbol{x}_i, k)$ are called $\mathtt{EM}_{\hat{d}}$, $\mathtt{EM}_{\hat{\nu}}$, and $\mathtt{EM}_{\hat{\tau}}$, respectively. These techniques depend on the number k of the neighbors considered when computing the features, and on the parameter C of the `EM` algorithm (the number of clusters).

To improve the results obtained by the aforementioned clustering methods, we combine the three features $\hat{d}(\boldsymbol{x}_i, k)$, $\hat{\nu}(\boldsymbol{x}_i, k)$, and $\hat{\tau}(\boldsymbol{x}_i, k)$, to obtain, for each $\boldsymbol{x}_i \in \boldsymbol{X}_N \subset \Re^D$, a 3-dimensional vector $\boldsymbol{y}_i \in \Re^3$. We can then apply the same procedure described above to assign each point to the cluster that maximizes the probability $g_c(\boldsymbol{y}_i, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. This approach, called $\mathtt{EM}_{\hat{d},\hat{\nu},\hat{\tau}}$, still depends on the parameters k and C.

Finally, to relax the dependence with respect to the parameter C we substitute `EM` with the `LPA` variant proposed in [19], obtaining $\mathtt{LPA}_{\hat{d}}$, $\mathtt{LPA}_{\hat{\nu}}$, $\mathtt{LPA}_{\hat{\tau}}$, and $\mathtt{LPA}_{\hat{d},\hat{\nu},\hat{\tau}}$. Briefly, this version of `LPA` automatically determines the number of clusters by an iterative process that assigns each $\boldsymbol{x}_i \in \boldsymbol{X}_N$ to the cluster to which the maximum number of its k nearest neighbors belong. To this aim, every point is initialized with a unique label and the labels are left to propagate through the network of points; as the labels propagate, densely connected clusters are formed, and they continue to expand until it is possible to do so.

4 Experimental Results

To obtain a preliminary assessment of the discriminative quality of the proposed features, we initially employed them to augment the dimensionality of points in the real datasets described below, and we apply a classical feature selection technique provided by WEKA [13].

The considered real datasets are: the `Yeast` dataset [15], which is composed of 1484 points in $\Re^8$ representing yeast proteins organized into 10 classes according to their positions in cells; the `Segmentation` dataset [10], which is composed of 2310 points in $\Re^{19}$ describing pixels randomly drawn from a dataset of 7 (classes) outdoor images; the `MNIST` test dataset [16] containing 10000 grey-level images of size 28×28 representing hand-written digits from 0 to 9 (10 classes). To use the `MNIST` dataset, we downsampled its images to the size 12×12, we vectorized them, and we appended the 576 gradient and curvature features described in [9], thus obtaining a dataset composed of 10000 points in $\Re^{720}$.

Note that, according to [14] and to the results we obtained by employing the `id` estimators described in [20], the `MNIST` clusters are characterized by `id` values higher than those of `Segmentation` and `Yeast`; indeed, we obtained `id` values between 2 and 4 for the `Segmentation` clusters, between 2 and 8 for the `Yeast` clusters, whilst the `id` values of the `MNIST` clusters are in the range $[7, 14]$.

As mentioned before, after augmenting the dimensionality of each vector in each dataset by appending the proposed local features, each dataset is processed by using the `RankSearch` feature selector (with `GainRatio` as a feature evaluator[2]) proposed in WEKA. We evaluated the relevance of each feature by separately applying `RankSearch` on 10 disjoint sets (folds) randomly generated from each dataset, and counting the number of times each feature is selected. According to the obtained results, $\hat{\nu}(\boldsymbol{x}_i, k)$ seems to be not relevant since it is never selected, while both $\hat{d}(\boldsymbol{x}_i, k)$ and $\hat{\tau}(\boldsymbol{x}_i, k)$ seem to be relevant (see Table 1); specifically, $\hat{\tau}(\boldsymbol{x}_i, k)$ seems important when considering high `id` datasets, whilst $\hat{d}(\boldsymbol{x}_i, k)$ is always selected for the low `id` ones. Nevertheless, as can be noticed in the following experiments, when coupled with $\hat{\tau}(\boldsymbol{x}_i, k)$ and $\hat{d}(\boldsymbol{x}_i, k)$, $\hat{\nu}(\boldsymbol{x}_i, k)$ allows to improve the clustering results.

Table 1. Percentage of times each feature has been selected as relevant in the 10 fold experiments

Dataset	`id`	$\hat{d}(\boldsymbol{x}_i, k)$	$\hat{\nu}(\boldsymbol{x}_i, k)$	$\hat{\tau}(\boldsymbol{x}_i, k)$
`MNIST`	$[7, 14]$	0%	0%	100%
`Segmentation`	$[2, 4]$	100%	0%	70%
`Yeast`	$[2, 8]$	100%	0%	100%

At this stage, we proceeded with tests on datasets generated by composing two or three clusters, which belong to either the `MNIST` test dataset, or to the synthetic datasets generated by the tool proposed in [14] (see Table 2). Precisely, to use the samples in the `MNIST` test dataset, we simply vectorized the 28×28 digit images obtaining samples in $\Re^{784}$. Though these points belong to 10 clusters (one cluster per digit) each being typified by a (probably) specific `id`, all the samples are already embedded in $\Re^{784}$; therefore, no embedding procedure is needed. To reduce the number of possible combinations, we run our clustering tests by randomly choosing one cluster (digit 1 images), and combining it with the other clusters to form all the possible cluster couples and triplets (obtaining 9 datasets of cluster couples and 36 of triplets).

Similarly, the synthetic datasets were created by merging two or three synthetic point sets (clusters), each comprising 1000 samples generated by the tool proposed in [14]. Note that each cluster is linearly embedded in a 40-dimensional space and the points belonging to the different clusters are concatenated, thus producing a point set containing either 2000 or 3000 points representing, respectively, 2 or 3 intersecting clusters. Furthermore, to reduce the possible combinations we randomly selected one of the synthetic sets (the $\boldsymbol{\mathcal{M}}_{10}$ with $D = 15$), and we intersected it with 1 or 2 synthetic sets to obtain all the possible couples and triplets (12 datasets of cluster couples and 66 of triplets). Note that, when the employed generator requires to set a dimensionality D (see Table 2), this parameter was set to 10 for the cluster selected as the second, and 20 for the third one.

To objectively evaluate the effectiveness of employing the proposed features for clustering, we compared the results achieved on both synthetic and real datasets to those obtained on raw data by well-known clustering techniques

[2] `GainRatio` evaluates the worth of a single feature by measuring the gain ratio with respect to the class, where the gain ratio is defined as: $\frac{H(Class)-H(Class|Feature)}{H(Feature)}$, being H the relative entropy function.

Table 2. Brief description of the 13 synthetic datasets employed in our experiments, where d is the `id` and D is the embedding space dimension. Note that for some datasets the parameter D should be selected by the user.

Name	d	D	**Description**
$\mathcal{M}_1$	$D-1$	D	Uniformly sampled sphere linearly embedded.
$\mathcal{M}_2$	3	5	Affine space.
$\mathcal{M}_3$	4	6	Concentrated figure, confusable with a 3d one.
$\mathcal{M}_4$	4	8	Nonlinear manifold.
$\mathcal{M}_5$	2	3	2-d Helix
$\mathcal{M}_6$	6	36	Nonlinear manifold.
$\mathcal{M}_7$	2	3	Swiss-Roll.
$\mathcal{M}_8$	12	72	Nonlinear manifold.
$\mathcal{M}_9$	20	20	Affine space.
$\mathcal{M}_{10}$	$D-1$	D	Uniformly sampled hypercube.
$\mathcal{M}_{11}$	2	3	Möebius band 10-times twisted.
$\mathcal{M}_{12}$	D	D	Isotropic multivariate Gaussian.
$\mathcal{M}_{13}$	1	13	Curve.

(`EM` and `LPA` variant) and by the multi-manifold clustering approach (`NS`) presented in [3]. Note that, for each method, the parameter settings were chosen to obtain the best mean results (see Table 3). To assess the compared clustering techniques we employed the following measure: $accuracy = \frac{\sum_{i=1}^{N}(\chi(l_i=\hat{l}_i))}{N}$, where χ is the indicator function, $\hat{l}_i$ is the label associated to the sample point x_i by the employed clustering approach, and l_i is the correct label for that point.

Table 3. The methods used in our experiments and the chosen parameters. k is the number of neighbors for each point, $k_{\texttt{LPA}}$ is the number of neighbors considered for Label Propagation, C is the number of clusters in the dataset, γ is the edge weighting factor, M is the number of Least Square (`LS`) runs, N is the number of re-sampling trials per `LS` iteration, Q is the number of different re-sampling values to be considered by `NS`.

Method	**Parameters**
Neighborhood Smoothing (`NS`)	$k = 20, \gamma = 1, M = 1, N = 10, Q = 10$
`EM`	$C = \{2, 3\}$
$\texttt{EM}_{\hat{d}}$	$k = 30,\ C = \{2, 3\}$
$\texttt{EM}_{\hat{\nu}}$	$k = 30,\ C = \{2, 3\}$
$\texttt{EM}_{\hat{\tau}}$	$k = 30,\ C = \{2, 3\}$
$\texttt{EM}_{\hat{d},\hat{\tau}}$	$k = 30,\ C = \{2, 3\}$
$\texttt{EM}_{\hat{d},\hat{\nu},\hat{\tau}}$	$k = 30,\ C = \{2, 3\}$
`LPA`	$k_{\texttt{LPA}} = 15$
$\texttt{LPA}_{\hat{d}}$	$k_{\texttt{LPA}} = 15$
$\texttt{LPA}_{\hat{\nu}}$	$k_{\texttt{LPA}} = 15$
$\texttt{LPA}_{\hat{\tau}}$	$k_{\texttt{LPA}} = 15$
$\texttt{LPA}_{\hat{d},\hat{\tau}}$	$k_{\texttt{LPA}} = 15$
$\texttt{LPA}_{\hat{d},\hat{\nu},\hat{\tau}}$	$k = 30,\ k_{\texttt{LPA}} = 15$

Table 4 shows the mean accuracies achieved on the synthetic and real datasets composed by two clusters, and those obtained on the datasets composed by three clusters[3]. It is possible to notice that $\texttt{EM}_{\hat{d},\hat{\nu},\hat{\tau}}$, which combines the information captured by the proposed local features, generally outperforms the other methods. Moreover, it is important to highlight that $\hat{\tau}(\boldsymbol{x}_i, k)$ is a very discriminative

[3] Note that the dataset points are coded by employing only the proposed features.

information, especially when facing datasets with high id, but it needs to be combined with $\hat{d}(\boldsymbol{x}_i, k)$ and $\hat{\nu}(\boldsymbol{x}_i, k)$ to effectively cope with datasets composed by clusters characterized by both high and low ids and eventually embedded in high dimensional spaces. This consideration is further shown by the lower accuracies achieved by coding the points combining $\hat{\tau}(\boldsymbol{x}_i, k)$ and $\hat{d}(\boldsymbol{x}_i, k)$ to obtain $\texttt{LPA}_{\hat{d},\hat{\tau}}$ and $\texttt{EM}_{\hat{d},\hat{\tau}}$. Note that we run these further tests since the feature selection approach has highlighted that these features are the most discriminative ones.

Table 4. Mean accuracies computed on synthetic cluster couples ($\boldsymbol{\mathcal{M}}_{10} + \boldsymbol{\mathcal{M}}_*$), real cluster couples ($\texttt{MNIST}_1 + \texttt{MNIST}_*$), synthetic cluster triplets ($\boldsymbol{\mathcal{M}}_{10} + 2\boldsymbol{\mathcal{M}}_*$), and real cluster triplets ($\texttt{MNIST}_1 + 2\texttt{MNIST}_*$). In boldface the best results have been highlighted.

Dataset	Measure	NS	EM	$\texttt{EM}_{\hat{d}}$	$\texttt{EM}_{\hat{\nu}}$	$\texttt{EM}_{\hat{\tau}}$	$\texttt{EM}_{\hat{d},\hat{\tau}}$	$\texttt{EM}_{\hat{d},\hat{\nu},\hat{\tau}}$	LPA	$\texttt{LPA}_{\hat{d}}$	$\texttt{LPA}_{\hat{\nu}}$	$\texttt{LPA}_{\hat{\tau}}$	$\texttt{LPA}_{\hat{d},\hat{\tau}}$	$\texttt{LPA}_{\hat{d},\hat{\nu},\hat{\tau}}$
$\boldsymbol{\mathcal{M}}_{10}+\boldsymbol{\mathcal{M}}_*$	*mean*	0.74	0.64	0.78	0.64	0.87	0.86	**0.88**	0.75	0.18	0.16	0.20	0.78	0.87
	std	0.12	0.17	0.19	0.07	0.16	0.17	0.15	0.20	0.05	0.03	0.07	0.25	0.18
$\texttt{MNIST}_1+\texttt{MNIST}_*$	*mean*	0.77	0.53	0.68	0.51	0.93	0.88	**0.95**	0.61	0.21	0.14	0.18	0.83	0.89
	std	0.08	0.01	0.06	0.01	0.02	0.12	0.03	0.17	0.04	0.04	0.05	0.17	0.14
$\boldsymbol{\mathcal{M}}_{10}+2\boldsymbol{\mathcal{M}}_*$	*mean*	0.34	0.46	0.63	0.52	0.71	0.71	**0.72**	0.63	0.06	0.05	0.06	0.64	0.65
	std	0.22	0.16	0.17	0.10	0.17	0.19	0.17	0.16	0.02	0.02	0.02	0.22	0.22
$\texttt{MNIST}_1+2\texttt{MNIST}_*$	*mean*	0.36	0.37	0.52	0.37	0.72	0.70	**0.74**	0.39	0.08	0.04	0.07	0.57	0.60
	std	0.22	0.01	0.03	0.03	0.08	0.10	0.08	0.08	0.04	0.02	0.03	0.12	0.08

5 Conclusions

In this paper we show that effective clustering results can be obtained by viewing dataset clustering as a multi-manifold clustering problem, where the dataset to be clustered is assumed to be drawn from a geometrical structure composed of several, eventually intersecting, clusters drawn from manifolds embedded into a higher dimensional space, and being characterized by (possibly) different ids. Under this assumption, we achieve promising clustering results by coding the input data by means of local id estimates and features related to them. The promising discriminative quality of the proposed features is shown by a feature relevance test, by the clustering results achieved on both synthetic and real datasets, and by their comparison to those obtained by related and classical state-of-the-art clustering approaches. Note that the proposed features have shown their discriminative power also when applied to a difficult problem such as the clustering of high dimensional datasets characterized by high and low ids.

To further assess the quality of the proposed features, our future works will be focused at their usage to deal with supervised classifion problems.

References

1. Bennett, R.S.: The Intrinsic Dimensionality of Signal Collections. IEEE Trans. on Information Theory IT-15(5), 517–525 (1969)
2. Bishop, C.M.: Bayesian PCA. In: Proc. of NIPS 11, pp. 382–388 (1998)
3. Carter, K.M., Raich, R., Hero, A.O.: On local intrinsic dimension estimation and its applications. IEEE Trans. on Signal Processing 58(2), 650–663 (2010)

4. Ceruti, C., Bassis, S., Rozza, A., Lombardi, G., Casiraghi, E., Campadelli, P.: DANCo: Dimensionality from Angle and Norm Concentration. ArXiv e-prints (June 2012)
5. Ceruti, C., Rozza, A., Bassis, S., Lombardi, G., Casiraghi, E., Campadelli, P.: DANCo: an intrinsic Dimensionalty estimator exploiting Angle and Norm Concentration. Submitted to Pattern Recognition Letters (2013)
6. Costa, J.A., Hero, A.O.: Learning intrinsic dimension and entropy of high-dimensional shape spaces. In: Proc. of EUSIPCO, pp. 231–252 (2004)
7. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Trans. Knowl. Data Eng. 16(11), 1370–1386 (2004)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the `EM` algorithm. J. of the Royal Statistical Soc.: Series B 39(1), 1–38 (1977)
9. Dong, J., Krzyzak, A., Suen, C.: Fast svm training algorithm with decomposition on very large data sets. IEEE Trans. on PAMI 27(4), 603–618 (2005)
10. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
11. Goldberg, A.B., Zhu, X., Singh, A., Xu, Z., Nowak, R.: Multi-manifold semi-supervised learning – learning when data lives on multiple, intersecting manifolds. In: Proc. of 12th International Conference on Artificial Intelligence and Statistics (2009)
12. Gong, D., Zhao, X., Medioni, G.: Robust multiple manifolds structure learning. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK (2012)
13. Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations 11 (2009)
14. Hein, M., Audibert, J.: Intrinsic dimensionality estimation of submanifolds in *Rd*. In: Proceedings of the ICML, pp. 289–296. ACM (2005)
15. Horton, P., Nakai, K.: A probablistic classification system for predicting the cellular localization sites of proteins. In: Intelligent Systems in Molecular Biology, pp. 109–115 (1996)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. of IEEE 86, 2278–2324 (1998)
17. Levina, E., Bickel, P.J.: Maximum likelihood estimation of intrinsic dimension. In: Proc. of NIPS, vol. 17(1), pp. 777–784 (2005)
18. Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comp. Biol. Bioinfor. 1(1), 24–45 (2004)
19. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76 (2007)
20. Rozza, A., Lombardi, G., Ceruti, C., Casiraghi, E., Campadelli, P.: Novel high intrinsic dimensionality estimators. Machine Learning Journal (May 2012)
21. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. Tech. Rep. 00034 (2000)
22. Upton, G.J.G.: Approximate confidence intervals for the mean direction of a von Mises distribution. Biometrika 73(2), 525–527 (1986)
23. Wang, Y., Jiang, Y., Wu, Y., Zhou, Z.: Local and structural consistency for multi-manifold clustering. In: Proceedings of IJCAI 2011. AAAI Press (2011)
24. Xiao, Y., Yu, J., Gong, S.: Intrinsic dimension induced similarity measure for clustering. In: Tang, J., King, I., Chen, L., Wang, J. (eds.) ADMA 2011, Part II. LNCS, vol. 7121, pp. 110–123. Springer, Heidelberg (2011)
25. Xu, R., Wunsch, D.: Survey of Clustering Algorithms. IEEE Trans. on Neural Networks 16(3), 645–678 (2005)