DENSITY GEODESICS FOR SIMILARITY CLUSTERING

Umut Ozertem, Deniz Erdogmus[*]

Oregon Health & Science University
Computer Science and Electrical Eng. Dept.
Portland, OR 97239, USA

Miguel Á. Carreira-Perpiñán

University of California, Merced
Electrical Eng. and Computer Science Dept.
Merced, CA 95343, USA

ABSTRACT

We address the problem of similarity metric selection in pairwise affinity clustering. Traditional techniques employ standard algebraic context-independent sample-distance measures, such as the Euclidean distance. More recent context-dependent metric modifications employ the bottleneck principle to develop path-bottleneck or path-average distances and define similarities based on geodesics determined according to these metrics. This paper develops a principled context-adaptive similarity metric for pairs of feature vectors utilizing the probability density of all data. Specifically, based on the postulate that Euclidean distance is the canonical metric for data drawn from a unit-hypercube uniform density, a density-geodesic distance measure stemming from Riemannian geometry of curved surfaces is derived. Comparisons with alternative metrics demonstrate the superior properties such as robustness.

Keywords: Affinity based clustering, similarity clustering, context dependent distance measure

1. INTRODUCTION

We consider the problem of clustering a dataset using pairwise-affinity methods, which are based on an affinity or similarity function $w(\mathbf{x}, \mathbf{y})$ that defines how close two data points $\mathbf{x}, \mathbf{y}$ are, as opposed to feature-based methods such as k–means, which work directly with the feature vectors $\mathbf{x}$ and $\mathbf{y}$. Examples of pairwise-affinity methods are hierarchical (agglomerative and divisive) clustering [1] and spectral clustering [2], among others. They have the advantage of dealing more easily with clusters of complex shapes, since they do not impose a model (e.g. Gaussian-shaped clusters) on the data. While different methods use the affinities in different ways (e.g. sequentially merging data points in agglomerative clustering vs projecting the data on the eigenspace of an affinity matrix in spectral clustering), the definition of affinity is of paramount importance in all them, and it is the focus of this paper.

Most work uses an affinity function $w(\mathbf{x}, \mathbf{y})$ that depends only on the feature vectors for the points $\mathbf{x}$ and $\mathbf{y}$, and more specifically on the distance between $\mathbf{x}$ and $\mathbf{y}$. For example, the popular Gaussian affinity $w_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(- \|\mathbf{x} - \mathbf{y}\| / \sigma^2\right)$, which also introduces a scale parameter σ that offers more flexibility but has to be tuned to each particular dataset, often by trial and error, for the method to succeed. We call this a *context-independent* affinity function, in that it does not depend on the rest of the data set (other than $\mathbf{x}$ and $\mathbf{y}$). However, this is a very limited representation of the metric structure of the feature space, which in our view is determined by the dataset. Refer to fig. 1. On the left, we have 3 points which are equidistant from each other, and so $w_\sigma(A, B) = w_\sigma(A, C) = w_\sigma(B, C)$. However, in the context of the dataset shown on the right (represented as a density rather than a finite dataset) and from a clustering point of view, it makes intuitively little sense for the points to be equidistant. In order to go from A to B we have to traverse a low-density region, while from A to C there are paths (not necessarily straight) traversing high-density regions. Our objective is to define *context-dependent* affinity functions where the context is given by the entire dataset (a global context). While this might be achieved by working directly with the dataset points (e.g. by defining a particular type of graph with appropriate edge weights), we aim at a more general framework informed by differential geometry concepts where we define a density-dependent metric on the feature space. By basing our metric on a *probability density function* $p(\mathbf{x})$, applicable to any point $\mathbf{x}$ in the space, we are also able to define affinities between any pair of points in the space, not just between the points in the dataset. In practice, the density is estimated from the data either nonparametrically (e.g. with a kernel density estimate) or parametrically (e.g. with a Gaussian mixture trained by EM). A second component of our affinity definition is that of *paths*, more specifically *geodesics*—paths that extremise a cost functional. This captures the idea of fig. 1(right) where, of the many paths joining A and C, we care about the one that is as short as possible in Euclidean distance while traversing high-density regions; this will be made specific in section 2. In summary, we propose a context-dependent affinity function $w_p(\mathbf{x}, \mathbf{y})$ defined as the largest path affinity of all paths joining $\mathbf{x}$ and $\mathbf{y}$, where the path affinity depends on the path and the density p, and is essentially the line integral along the path of a function of p. Thus, we call this idea *affinity by density geodesics*.

The reason why we expect density geodesics to work with clustering is that the corresponding affinities should make the cluster structure obvious. From fig. 1(right), the affinities between every pair of points in the left cluster (e.g. A, C) will be high (and similar in magnitude), while the affinities between points in different clusters will all be low. Thus, the affinity matrix will have a blocky aspect with near-zero across-cluster affinities and high within-cluster affinities (and so will its eigenvectors). A spectral algorithm will map the dataset to a structure consisting of widely separated, point-like clusters, which is trivial to cluster. Note, however, that this idea does not work for dimensionality reduction with pairwise-affinity methods such as Isomap [3] or Laplacian eigenmaps [4], because the internal metric structure of a cluster is almost completely erased.

Although our general formulation of clustering based on a metric induced by the data density is new to our knowledge, previous work has considered (separately) the ideas of paths between data points and of using the data density for clustering. In path-based clustering [5], one assumes a neighbourhood graph (e.g. k–nearest-

[*] This work is partially supported by NSF grants ECS-0524835, and ECS-0622239.

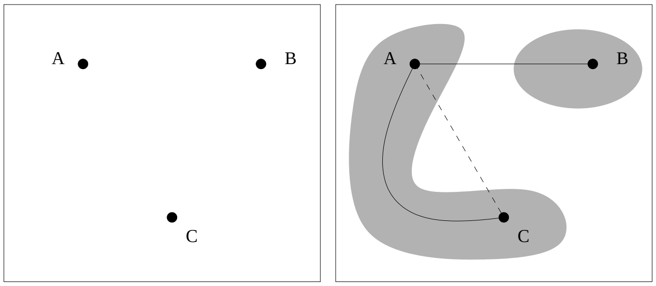

Fig. 1. Points A, B, C are equidistant in Euclidean distance (left), but the data distribution (right) suggests A and C are very close to each other and very far from B.

neighbours) on the dataset, and defines a distance (inverse affinity) as a minimax distance

$$\overline{d}_{nm} = \min_{P \in \mathcal{P}_{nm}} \max_{l=1,\dots,|P|} d(P_l, P_{l+1})$$

where $\mathcal{P}_{nm} =$ "all paths joining data points $\mathbf{x}_n$ and $\mathbf{x}_m$". Thus, the distance is the smallest bottleneck (= longest link) over $\mathcal{P}_{nm}$. Its motivation is, as in fig. 1(right), to obtain low across-cluster affinities—the bottleneck being the distance between the two clusters, which dominates other intermediate links. However, the minimax distance has the undesirable property that it does not change if we change the link distances (other than the bottleneck), and is sensitive to outliers. Also, for overlapping clusters, a bottleneck may not exist (particularly for large datasets), even if the density across clusters is much lower than within clusters. Besides, the minimax distance only applies to points in the dataset, unlike our affinity, which is defined on the entire space. Several modifications have tried to address the sensitivity to outliers, by using bagging [6] and by normalising link distances by the node degree [7]. In the method of [8], the minimax distance $\overline{d}_n m$ is modulated by a Gaussian, $\exp{(-\overline{d}_{nm}/\sigma^2)}$, and further the "max" is softened via another user parameter to interpolate smoothly between the two extreme cases of the maximum (= path bottleneck) and the mean (= path length). However, they applied this not to clustering but to transductive SVMs. Finally, the bottleneck geodesics [9] seek paths that are both short and dense. They define a local density estimate at each data point $\mathbf{x}_n$ as $\sum_{n \sim m} \|\mathbf{x}_n - \mathbf{x}_m\|^{-2}$ (where $n \sim m$ is a link in a neighbourhood graph). They then combine this with the path length in various ways, e.g. by normalising the link length by the density estimate, or by normalising the path length by the path bottleneck. Like the other methods, this yields affinities only for the dataset points.

An alternative approach to introducing context in the definition of pairwise affinities is by constructing better neighbourhood graphs. The usual graphs (k–nearest-neighbours, ϵ–ball, fixed grid for image segmentation) enforce global parameters for each data point no matter its context, so every point must have k neighbours or connect to all points at distance ϵ or less. In the perturbation approach of [10], multiple graphs are built on jittered duplicates of the dataset and then combined into an average graph that adapts locally to the data and is by construction more robust to noise. However, in this method point pairs such as (A, C) in fig. 1 are still assigned a low affinity, and the path in the graph counts distances but not density.

2. DENSITY GEODESICS

We consider the data probability density function (pdf) $p : \mathbb{R}^n \rightarrow \mathbb{R}^+$ as a metric-imposing mass distribution that distorts the Euclidean nature of $\mathbb{R}^n$. Therefore, affinity estimates of pairs of data based on Euclidean metrics are inaccurate. A geometrically consistent measure of distance in a pdf-warped space must properly take into account local and directional stretching effects imposed on the data when transforming its density to *p* from uniform . In a differential geometric framework, the infinitesimal length of a curve segment depends on its direction (tangent vector), the local stretching (Riemann metric of smooth pdf manifolds), and for clustering purposes, a measure that promotes large density while penalizing low density. The latter is especially useful for distinguishing clusters separated by valleys or gaps in data density. The former two are useful for making the metric locally invariant to stretching and is a natural consequence of change of variables for contour integration when transforming the data with a nonlinear map from uniform to *p* density.

Let $f : \mathbb{R}^n \rightarrow [0, 1]^n$ be an invertible map that transforms pdf *p* to a uniform one in the unit hypercube (note that such a mapping can always be constructed using conditional cumulative densities derived from *p* [11]). Due to change of variables, the Euclidean length of a curve in the unit-cube under a uniform density has to be modified to utilize a metric $\mathbf{M}(\mathbf{x}) = (\nabla^T \mathbf{f}(\mathbf{x}) \nabla \mathbf{f}(\mathbf{x}))^{-1}$. Assuming a density penalizing/promoting measure *h*, we define the length of a curve **c** in $\mathbb{R}^n$ under pdf *p* as

$$l_{\mathbf{c}} \doteq \int_0^1 h(p(\mathbf{c}(t)))(\dot{\mathbf{c}}^T(t)\mathbf{M}(\mathbf{c}(t))\dot{\mathbf{c}}(t))^{1/2}dt \qquad (1)$$

Under this definition of curve length, the distance between two points **x** and **y** is the length of the shortest curve (geodesic) connecting them: $d(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{c}} l_{\mathbf{c}}$ such that $\mathbf{c}(0) = \mathbf{x}$ and $\mathbf{c}(1) = \mathbf{y}$. The minimization of the functional in (1) requires calculus of variations and is computationally expensive. For a given iid data set $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with sufficiently large *N* drawn from *p*, a simple yet sufficient approximation to the geodesic could be determined by restricting the search to the paths over a fully- or partially-connected neighborhood graph. We utilize the ϵ-ball neighborhood graph where an edge is included in the graph if the length of the straight line connecting $\mathbf{x}_i$ to $\mathbf{x}_j$ as measured by (1), $l_{ij} < \epsilon$ and employ a shortest path search algorithm [12].

In pairwise affinity clustering, for sharp block-diagonal affinity matrices one needs to penalize geodesics that pass through relatively low density regions and promote geodesics that remain in high density regions. The bottleneck clustering principle [5] is also aimed at achieving this, however, the length of a curve is measured not as an integral over the curve but as an extremum of local length, which occurs at such valleys. By selecting monotonically decreasing $h : \mathbb{R} \rightarrow \mathbb{R}$ such that $\lim_{p \rightarrow 0} h(p) = \infty$ and $\lim_{p \rightarrow \infty} h(p) = 0$, this could be achieved. Alternatively, one could simply select the neutral measure of $h(p) = 1$.

2.1. Case Studies in 1-Dimension

We present three specific examples in 1D, constructed to illustrate certain aspects of the proposed metric, while avoiding the geodesic search step: (1) distances under a uniform density; (2) piecewise uniform clusters; and (3) distances under an arbitrary pdf *p*.

Uniform Density: Consider distance between two points a/ε and b/ε, where $\{a, b\} \in [0, 1]$, under a uniform density in the interval $[0, 1/\varepsilon]$. The function that maps this density to a uniform in $[0, 1]$ has slope ε, thus, the metric is $M(x) = 1/\varepsilon^2, x \in [0, 1]$. For the contour

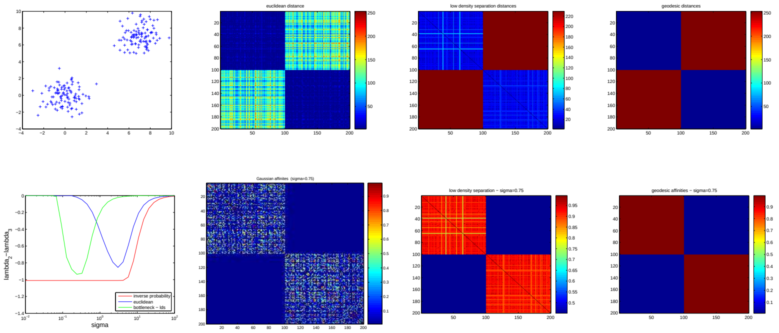

Fig. 2. Gaussians dataset

.

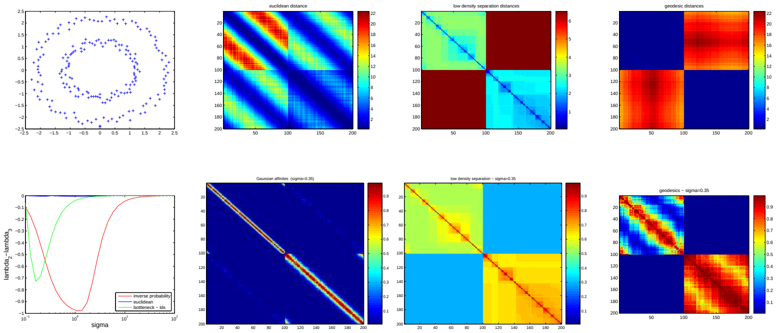

Fig. 3. Rings dataset

.

c connecting a to b, we have $\dot{c} = |b-a|/\varepsilon$. Assuming $h(p) = 1$ and substituting all expressions in (1), we obtain $l_{[a/\varepsilon,b/\varepsilon]} = |b-a|/\varepsilon^2$. The inverse square dependency of this metric on ε is intuitively interpreted as one factor coming from the scaling of the actual Euclidean distance and another from the local stretching of the density.

Piecewise Uniform Clusters: Consider a piecewise uniform density and its local metric

$$\begin{aligned} p(x) &= \begin{cases} 1/2-\varepsilon & \text{if } \ x \in [0,1] \cup [2,3] \\ 2\varepsilon & \text{if } \ x \in [1,2] \\ 0 & \text{otherwise} \end{cases} \\ M(x) &= \begin{cases} 1/(1-2\varepsilon)^2 & \text{if } \ x \in [0,1] \cup [2,3] \\ 1/(2\varepsilon)^2 & \text{if } \ x \in [1,2] \\ \infty & \text{otherwise} \end{cases} \end{aligned} \qquad (2)$$

For $\{a,b\} \in [0,1]$ and $c \in [2,3]$, we are interested in the distances $d(a,b)$ and $d(a,c)$ assuming $h(p) = 1$. For $d(a,b)$, we have $\dot{c} = |b-a|$ and the integral can be computed over the segment $[0,1]$ as $l_{[a,b]} = 2|b-a|/(1-2\varepsilon)$. For $d(a,c)$, the integral needs to be computed in three segments. Letting $c(t_1) = 1$ and $c(t_2) = 2$ for the curve connecting a to c, one can write $l_{[a,c]} = |b-a|\{2t_1/(1-2\varepsilon)+(t_2-t_1)/(2\varepsilon)+2(1-t_2)/(1-2\varepsilon)\}$. Note that $\lim_{\varepsilon\to 0} l_{[a,b]} = 2|b-a|$, while $\lim_{\varepsilon\to 0} l_{[a,c]} = \infty$.

Arbitrary PDF: For a univariate pdf $p(x)$, the invertible function that maps the random variable to a uniform in $[0,1]$ is its cumulative density function (cdf). Consequently, its derivative is the pdf, thus the metric becomes $M(x) = p^{-2}(x)$. For an arbitrary $h(p)$ measure, by substituting $M(x)$ and $\dot{c}(t) = (b-a)$ in (1), one can write $d(a,b) = |b-a| \int_0^1 h(p(a+(b-a)t))p^{-1}(a+(b-a)t)dt$. This shows that appropriate selection of $h(p)$ will modify the Euclidean distance in the desirable manner.

2.2. Omitting Directional Dependency Multidimensional Case

In implementation, numerical integration of (1) considering the local tangent length $(\dot{\mathbf{c}}(t)^T\mathbf{M}(\mathbf{c}(t))\dot{\mathbf{c}}(t))^{1/2}$ is computationally expensive. In certain situations, significant gain in speed are achieved by dropping this term from the integrand without much impact on performance. For x on the line segment that connects $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{l}_{ij}$, if we assume that the metric is $\mathbf{M}(\mathbf{x}) = \mathbf{I}\|\mathbf{x}_i - \mathbf{x}_j\|^2$, letting $\dot{\mathbf{c}}(t) = \|\mathbf{x}_i - \mathbf{x}_j\|^{-1}$, the integrand reduces to

$$l_{\mathbf{l}_{ij}} = \int_0^1 h(p(\mathbf{l}_{ij}(t)))dt \qquad (3)$$

This corresponds to assuming that the metric is simple scaling along the data connected by the graph, but employs different scaling factors depending on specific data pairs.

3. EXPERIMENTAL RESULTS

In our experiments we concentrate on spectral clustering, but our definition is equally applicable to hierarchical clustering and other affinity methods. We compare the Euclidean distance (yielding the widely used usual Gaussian affinity), low density separation (LDS) algorithm [8], and density geodesics. All three examples have two

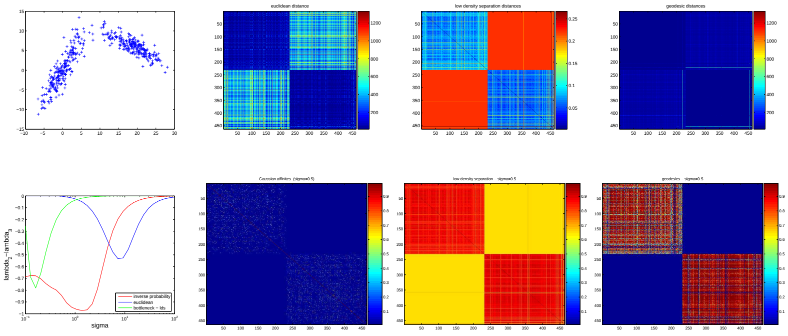

Fig. 4. Bridge dataset

.

natural clusters, and we use the eigengap between the second and third eigenvalues to evaluate the quality of the affinity matrix. As the difference between these two eigenvalues approaches to -1, the desired two cluster solution is more healthy.

In Figure 2, we present result for an easy clustering problem with two Gaussian clusters with 100 samples each. As all the clusters are well seperated, all three methods return blockwise distance matrices for this example. The peak performance of LDS is greater than Euclidean distances, yet still lower then density geodesics. The interval of sigma values that gives the *correct* clustering result is much wider for the density geodesics. As some inter-cluster Euclidean distances are less than some in-cluster Euclidean distances, ring dataset lays a harder clustering problem, and has been widely used in clustering papers. In this example, density geodesics also demonstrate superior performance as compared to LDS and Euclidean distance, with a bigger eigengap for a wider interval of σ. Results are given in Figure 3. One problem with bottleneck methods is that due to noise or heavytailed distrubutions a bottleneck may not exist. Here we show an example where two Gaussian clusters are next to eachother. The pdf drops significantly at the cluster boundary, yet there is no strong bottleneck due to the samples around the boundary. As shown in Figure 4, density geodesics provide a better eigengap for a wider range of σ in this example as well.

4. DISCUSSION

The problem of affinity measure selection is at the core of similarity-based clustering techniques. Various propositions include the usual algebraic distance measures as well as path-bottleneck geodesics along certain neighborhood graphs. Intuitively, the distance (inverse affinity) between two data points cannot be assessed without any regard to the context set forth by the distribution of other data samples. In principle, distances between pairs should be influenced strongly by the probability distribution of data. Stating from the postulate that Euclidean distances are canonical for data points drawn from unit-uniform densities, we develop a principled path-length measure rooted in Riemannian geometry. It has been demonstrated theoretically and through analytical calculations for specific one-dimensional case studies that the proposed distance metrics satisfy desirable invariance properties. For high dimensional cases, in order to reduce the computational complexity, the metric is simplified and the geodesic search problem is approximately solved utilizing shortest path search over a discretized graph. Experimental results demonstrate that the proposed distance metric yields blocky affinity matrices which will lead to clear cluster separations.

5. REFERENCES

[1] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[3] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 22 2000.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[5] B. Fischer and J. M. Buhmann. Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(4):513–518, April 2003.

[6] B. Fischer and J. M. Buhmann. Bagging for path-based clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(11):1411–1415, November 2003.

[7] H. Chang and D. Yeung. Robust path-based spectral clustering with application to image segmentation. In *Proc.Int. Conf. Computer Vision*, pages 278–285, Beijing, China, October 15–21 2005.

[8] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. *Proc. of Int. Workshop on Artificial Intelligence and Statistics* , pages 57–64, Barbados, January 6–8 2005.

[9] I. Omer and M. Werman. The bottleneck geodesic: Computing pixel affinity. *Proc. of the Computer Vision and Pattern Recognition*, pages 1901–1907, New York, NY, June 17–22 2006.

[10] M. Á. Carreira-Perpiñán and R. S. Zemel. Proximity graphs for clustering and manifold learning. *Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 225–232. MIT Press, 2005.

[11] A. Hyvarinen and P. Pajunen. Nonlinear independent component analysis: Existance and uniqueness results. *Neural Networks*, volume 12, no. 3, pages 429–440, 1999.

[12] E. Kreyszig. Advanced Engineering Mathematics. New York, Wiley, 1972.