Contextual Distance for Data Perception

Deli Zhao Zhouchen Lin Xiaoou Tang
Visual Computing Group, Microsoft Research Asia, Beijing, China
{i-dezhao,zhoulin,xitang}@microsoft.com

Abstract

Structural perception of data plays a fundamental role in pattern analysis and machine learning. In this paper, we develop a new structural perception of data based on local contexts. We first identify the contextual set of a point by finding its nearest neighbors. Then the contextual distance between the point and one of its neighbors is defined by the difference between their contribution to the integrity of the geometric structure of the contextual set, which is depicted by a structural descriptor. The centroid and the coding length are introduced as the examples of descriptors of the contextual set. Furthermore, a directed graph (digraph) is built to model the asymmetry of perception. The edges of the digraph are weighted based on the contextual distances. Thus direction is brought to the undirected data. And the structural perception of data can be performed by mining the properties of the digraph. We also present the method for deriving the global digraph Laplacian from the alignment of the local digraph Laplacians. Experimental results on clustering and ranking of toy problems and real data show the superiority of asymmetric perception.

1. Introduction

Given a set of data points, how are the structures of the data perceived? From human perception, we can easily identify two surfaces surrounded by noise points in Figure 1 (a). The correctly perceived structures in this figure consist of two separated surfaces and a set of noise points (Figure 1 (b)). In this paper, we aim at developing algorithms that can robustly detect structures of data.

1.1. Previous Work

Classical methods to structural analysis of data include principal component analysis (PCA) and multidimensional scaling (MDS) which perform dimensionality reduction by preserving global structures of data, and non-negative matrix factorization (NMF) [13] which learns local representations of data. K-means is also frequently employed to identify underlying clusters in data. Recently, Ding *et al*. [11, 12] showed the connection between PCA and K-means, and NMF and spectral clustering. The underlying assumption behind the above methods is that spaces where data points (or samples) lie in are Euclidean.

Non-Euclidean perception of data was established by Tenenbaum *et al*. [20] and Roweis *et al*. [17]. In their work, nonlinear structures of data were modelled by preserving global (geodesic distances for Isomap) or local (locally linear fittings for LLE) geometry of data manifolds. These two methods directed the structural perception of data in manifold ways [18].

In recent years, spectral graph partitioning has become a powerful tool for structural perception of data. The representative methods are the normalized cuts [19] for image segmentation and the algorithm proposed by Ng *et al*. [16] (NJW clustering) for data clustering. Meilă and Shi [15] showed the connection between spectral clustering and random walks. For traditional spectral clustering, the structure of data is modelled by undirected weighted graphs, and underlying clusters are found by graph embeddings. The theoretical feasibility of spectral clustering was analyzed in [23, 4]. The method was detailed in [4] on how to find the number of clusters from spectral properties of normalized weighted adjacency matrices.

For semi-supervised structural perception, tasks are to detect partial manifold structures of data, given one or more labeled points on data manifolds. Zhou *et al*. [27, 28] and Agarwal [1, 2] developed simple but effective methods of performing transductive inference (or ranking) on data manifolds or graph data. Belkin *et al*. [6] developed a comprehensive framework of manifold regularization for learning from samples.

1.2. Limitations of Existing Methods

However, there are two issues untouched in the existing spectral methods for the structural perception of data. The first is the noise tolerance of algorithms, and the second is the measure of distances. These two problems are tightly related. It was reported [5] that spectral methods in manifold learning are not robust enough to achieve good results when the structures of data are contaminated by noise points. We have also found that almost all toy experiments on spectral

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

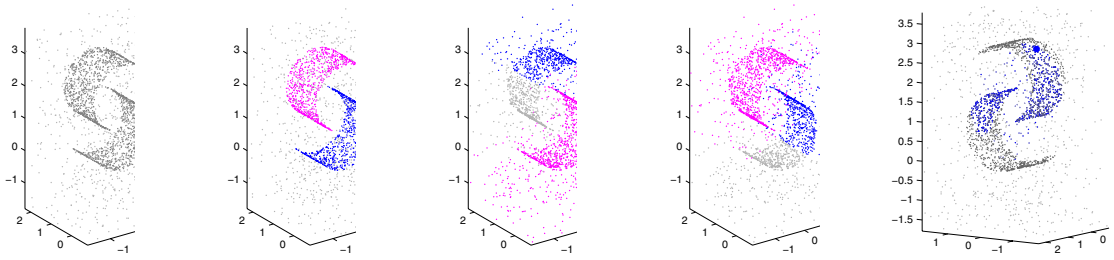


Figure 1. Two half-cylinders data and its clustering and ranking results by existing representative methods. (a) Randomly sampled points on two half cylinders. There are 800 points on each half cylinder. Then 800 noise points are mixed with these sample points. (b) Expected structures consist of two half cylinders and a set of dispersed noise points. (c) Clustering by NJW clustering. (d) Clustering by normalized cuts. (e) Ranking by Zhou's method. The free parameter α is set to be 0.1 (We find that a large α for Zhou's ranking yields bad ranking results. The same case occurs in Section 5.2.). The large dot is the randomly labelled point on one half-cylinder.

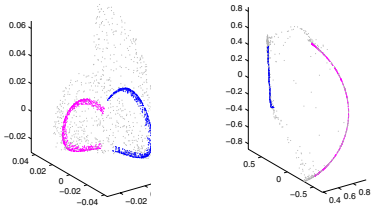


Figure 2. Embeddings of the two half-cylinders data. (a) NJW clustering. (b) Normalized cuts.

clustering and ranking in the existing papers are performed on clean data. In addition, traditional Euclidean based distances between two points may not cope with structural perception well.

To see the limitations of existing methods, we illustrate the results of clustering and ranking on the toy example shown in Figure 1 (a). We can see that NJW clustering[1] (Figure 1 (c)) and normalized cuts[2] (Figure 1 (d)) fail to detect the underlying clusters of the data and Zhou's ranking [27] (Figure 1 (e)) also yields the wrong transductive inference of the partial manifold structure. We further visualize the new representations of data in the 3-dimensional Euclidean space. These representations are produced by three eigenvectors used in NJW clustering and normalized cuts, respectively. As shown in Figures 2 (a) and (b), these two methods cannot separate the hidden structures and noise points.

To better understand the problem, we need the intuition of visual perception. Figure 3 (a) shows a simple set of nine points. The structure of the data is clear, which consists of two clusters identified by the '$\bullet$' markers (Cluster I) and the '$+$' markers (Cluster II). One can retrieve the information at first sight. However, the Euclidean distances between the point a and the other ones do not comply with

[1]For better visualization, we directly show the results without mapping feature representations onto the unit sphere. Figures 4 (a) and 9 (a) are treated in the same way.

[2]We run the Matlab codes of normalized cuts, available at http://www.seas.upenn.edu/~timothee.

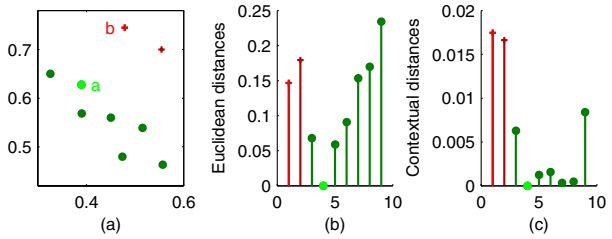


Figure 3. Two clusters of points and two kinds of distances. (a) The data set. Clusters I and II consist of the '$\bullet$' markers and the '$+$' markers, respectively. (b) Euclidean distances from the point a to the other points. (c) Contextual distances, computed using the coding length.

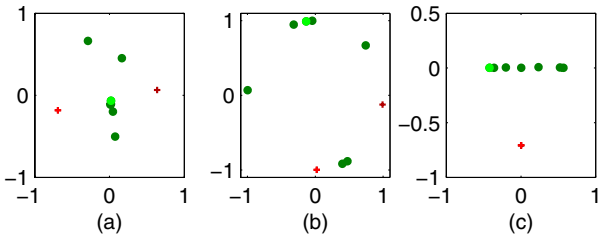


Figure 4. 2D representations of two-cluster data in Figure 3 (a). The presentations are derived by rows of two column eigenvectors of weighted adjacency matrices corresponding to the second and third largest eigenvalues. (a) By NJW clustering. (b) By normalized cuts. (c) By perceptual clustering.

our perception (Figure 3 (b)): the distances between point a and points $\#8, \#9, \#10$ in Cluster I are larger than those between point a and points $\#1, \#2$ in Cluster II. Figures 4 (a) and (b) illustrate that NJW clustering and normalized cuts mix the two clusters.

From the above analysis and illustrations, we see that the Euclidean-based distances between two points cannot capture the 'correct' structure of clusters. This should be the main reason why traditional spectral methods cannot perform well on noisy data.

1.3. Our Contribution

To overcome the limitations of the existing methods, we contend that the structural perception of data should be per-

(a) (b) (c)

Figure 5. Illustration of a simple cognitive psychological experiment on testing the influence of expectation on perception.

formed using local contexts. More specifically, our work is different from previous ones in two aspects.

(1) The distance is no longer defined for every two sample points and the distance is no longer Euclidean based either. Rather, the distance is defined within contextual sets only and the contextual distance between two points is defined by the difference of their contribution to the integrity of the geometric structure of the contextual set, where the contextual set consists of a point and its nearest neighbors to provide the context for the point.

(2) Furthermore, a digraph is built on the undirected data to model the asymmetry of perception, which is induced by the asymmetry of contextual distances. As a result, structural perception can be performed by mining the properties of the digraph. Thus, the applications of digraph theory can be extended from networks and the web to general multi-dimensional data.

With contextual distances and digraph embeddings, structures of data can be robustly retrieved even when there is heavy noise.

2. From Pairwise Points to Contextual Sets

2.1. A Contextual View on Data Perception

We start with the two clusters in Figure 3 (a). Consider the perceptual relationship between point b and Cluster I. It makes sense to say that point b is an outlier with respect to Cluster I. This is based on the observation that the set of dot points has a consistent global structure. We consider point b as an outlier by a comparison between the underlying structures of point b and Cluster I. Equivalently, we retrieve the structural information of point b *unconsciously* by taking Cluster I as reference. Therefore, we conclude that *the structural perception is relative and context-based*. An isolated point itself is not an outlier, but it may be an outlier when its neighboring points are taken as reference. Thus, the set of contextual points should be taken into account in order to compute distances compatible with the mechanism of human perception.

2.2. Cognitive Psychological Evidence

Our viewpoint that structural perception is relative and context-based is also supported by cognitive psychology. Bruner and Minturn [8] carried out a famous experiment on testing the influence of expectation on perception. For example, what is the central pattern in Figure 5 (a)? We perceive 13 in the context of numbers (Figure 5 (b)), whereas we perceive B in the context of letters (Figure 5 (c)). This implies that the same physical stimulus can be perceived differently in different contexts. This proves that the perceptual relationship between two sample points heavily relies on the contextual sets in which they belong to.

2.3. The Contextual Distance

In this section, we present the general definition of contextual distances[3]. It is only defined within contextual sets of points.

Let $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be the set of m sample points in $\mathcal{R}^n$. The contextual set $\mathcal{S}_i$ of the point $\mathbf{x}_i$ consists of $\mathbf{x}_i$ and its nearest neighbors in the Euclidean distance sense, i.e., $\mathcal{S}_i = \{\mathbf{x}_{i_0}, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}\}$, where $\mathbf{x}_{i_j}$ is the j-th nearest neighbor of $\mathbf{x}_i$ and K is the number of nearest neighbors. Here and in the sequel, we set $i_0 = i$.

As we are interested in the geometric structure of $\mathcal{S}_i$, we may have a structural descriptor $f(\mathcal{S}_i)$ of $\mathcal{S}_i$ to depict some global structural characteristics of $\mathcal{S}_i$. We notice that if a point $\mathbf{x}_{i_j}$ complies with the structure of $\mathcal{S}_i$, then removing $\mathbf{x}_{i_j}$ from $\mathcal{S}_i$ will not affect the structure much. In contrast, if the point $\mathbf{x}_{i_j}$ is an outlier or a sample in a different cluster, then removing $\mathbf{x}_{i_j}$ from $\mathcal{S}_i$ will change the structure significantly. This motivates us to define (1) as the contribution of $\mathbf{x}_{i_j}$ to the integrity of the structure of $\mathcal{S}_i$, i.e., the variation of the descriptor with and without $\mathbf{x}_{i_j}$:

$$\delta f_{i_j} = |f(\mathcal{S}_i) - f(\mathcal{S}_i \setminus \{x_{i_j}\})|, \qquad j = 0, 1, \ldots, K, \quad (1)$$

where $|\bullet|$ denotes the absolute value for a scalar or a kind of norm for a vector. The descriptor $f(\mathcal{S}_i)$ is not unique. However, $f(\mathcal{S}_i)$ needs to satisfy the structural consistency among the points in $\mathcal{S}_i$, in the sense that δf_{i_j} is relatively small if x_{i_j} is compatible with the global structure formed by sample points in $\mathcal{S}_i$ and relatively large if not. Then the contextual distance from $\mathbf{x}_i$ to $\mathbf{x}_{i_j}$ is defined as

$$p(\mathbf{x}_i \to \mathbf{x}_{i_j}) = |\delta f_i - \delta f_{i_j}|, \qquad j = 0, 1, \ldots, K, \quad (2)$$

where the notation $\to$ emphasizes that the distance is from $\mathbf{x}_i$ to $\mathbf{x}_{i_j}$. Obviously, $p(\mathbf{x}_i \to \mathbf{x}_{i_j}) \geq 0$ and the equality holds if $j = 0$.

The contextual distance $p(\mathbf{x}_i \to \mathbf{x}_{i_j})$ defined above is consistent with our contextual view on structural perception. The set $\mathcal{S}_i$, consisting of the point $\mathbf{x}_i$ and its nearest neighbors $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}\}$, is taken as the context for computing the distances from $\mathbf{x}_i$ to its neighbors. The relative perception is modelled by investigating how much the structure of $\mathcal{S}_i$ changes by removing a point from $\mathcal{S}_i$. It is worth

[3] Precisely speaking, the contextual distance defined here is a kind of dissimilarity instead of a formal distance in the mathematical sense. In order to compare with the traditional Euclidean distance, however, we still name it by the distance.

noting that the asymmetry is the special nature of the contextual distance defined in (2), because $p(\mathbf{x}_i \to \mathbf{x}_{i_j})$ is not necessarily equal to $p(\mathbf{x}_{i_j} \to \mathbf{x}_i)$ as in the extreme case $\mathbf{x}_i$ may even not be in the contextual set of $\mathbf{x}_{i_j}$. The contextual distance heavily relies on the structural characteristic of the contextual set.

2.4. Examples of Contextual Set Descriptors

In this section, we present some examples of contextual set descriptors which are applied for computing the contextual distances.

2.4.1 Trivial descriptor

In fact, the Euclidean distance is a special case of our contextual distance. Let $K = 1$, and $f(\mathcal{S}_i) = \gamma\mathbf{x}_i + (1-\gamma)\mathbf{x}_{i_1}$, where $\gamma < 0$ or $\gamma > 1$, and the norm in (1) be the Euclidean norm $\|\bullet\|$. Then we have $p(\mathbf{x}_i \to \mathbf{x}_{i_1}) = \|\mathbf{x}_i - \mathbf{x}_{i_1}\|$. Therefore, the contextual distance coincides with the Euclidean distance in this special case.

2.4.2 Geometric Descriptor: Centroid

Here we present a simple yet effective descriptor of $\mathcal{S}_i$ by its centroid. Let $K > 1$ and $\bar{\mathbf{x}}_i(\mathcal{S}_i)$ denote the centroid of $\mathcal{S}_i$, i.e., $\bar{\mathbf{x}}_i(\mathcal{S}_i) = \frac{1}{K+1}\sum_{j=0}^{K}\mathbf{x}_{i_j}$. $\bar{\mathbf{x}}_i(\mathcal{S}_i)$ is a type of simple globally geometric characterization of $\mathcal{S}_i$. Removing $\mathbf{x}_{i_j}$ will cause relatively larger shifting of the centroid than the other elements in $\mathcal{S}_i$ if it is not compatible with the underlying global structure of $\mathcal{S}_i$. So an alternative descriptor of the set is $f(\mathcal{S}_i) = \bar{\mathbf{x}}_i(\mathcal{S}_i)$, which is a vector-valued descriptor.

2.4.3 Informative Descriptor: Coding Length

The coding length [14] $L(\mathcal{S}_i)$ of a vector-valued set $\mathcal{S}_i$ is the intrinsic structural characterization of the set. This motivates us to exploit $L(\mathcal{S}_i)$ as a kind of scalar-valued descriptor of $\mathcal{S}_i$, i.e., $f(\mathcal{S}_i) = L(\mathcal{S}_i)$. The definition of $L(\mathcal{S}_i)$ is presented in Appendix. The allowable distortion ε in $L(\mathcal{S}_i)$ is a free parameter and $L(\mathcal{S}_i)$ is not very sensitive to the choice of ε. Here we empirically choose $\varepsilon = \sqrt{\frac{10n}{K}}$.

Figure 3 (c) illustrates the contextual distances from point a to the others. We see that the distances to Cluster II are much larger than those to Cluster I. Hence the contextual distances are much closer to what a human perceives.

3. Digraph Modelling: Bringing Direction to Data

The asymmetry of contextual distances among points naturally induces a digraph model to the data. This brings direction to the undirected data. It is worthwhile to note that the method of digraph modelling presented here is also applicable for general asymmetric or directed metrics.

3.1. Digraph on data

We may build a digraph for $\mathcal{S}$. Each point in $\mathcal{S}$ is a vertex of the digraph. A directed edge is put from $\mathbf{x}_i$ to $\mathbf{x}_j$ if $\mathbf{x}_j$ is one of the K nearest neighbors of $\mathbf{x}_i$. The weight $w_{i\to j}$ of the directed edge is defined as

$$w_{i\to j} = \begin{cases} e^{-\frac{[p(\mathbf{x}_i \to \mathbf{x}_j)]^2}{\sigma^2}}, & \text{if } \mathbf{x}_j \text{ is a nearest neighbor of } \mathbf{x}_i, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\to$ denotes that the vertex i points to the vertex j, and σ is a free parameter. The direction of the edge from $\mathbf{x}_i$ to $\mathbf{x}_j$ arises because the distance between them is asymmetric. Locally, the point $\mathbf{x}_i$ is connected to its nearest neighbors $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}\}$ by a K-edge directed star (K-distar). Hence the induced digraph on the data is composed of m K-distars. Let $\mathbf{W} \in \mathcal{R}^{m\times m}$ denote the weighted adjacency matrix of the weighted digraph, i.e., $\mathbf{W}(i,j) = w_{i\to j}$. $\mathbf{W}$ is asymmetric. Thus the structural information of the data is embodied by the weighted digraph, and data mining reduces to mining the properties of the digraph. We summarize the algorithm of digraph modelling below.

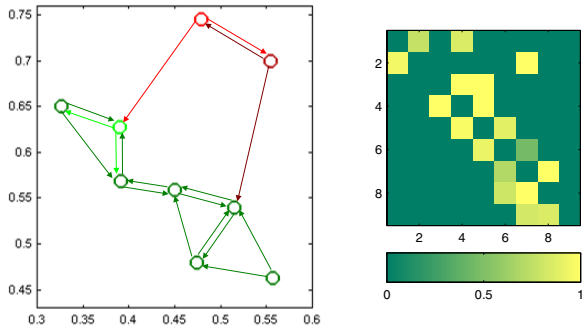
Algorithm of digraph modelling on data

Given a set of data $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, the digraph can be constructed as follows:

1. Search K nearest neighbors $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}\}$ for each sample point $\mathbf{x}_i$, where K is a parameter.
2. Compute contextual distances $p(\mathbf{x}_i \to \mathbf{x}_{i_j})$ according to formula (2).
3. Form the weighted adjacency matrix $\mathbf{W}$ according to formula (3).

Here, we present the approach of estimating σ in (3). Suppose that $\{p_1, \ldots, p_s\}$ are the s contextual distances that randomly selecting from r local contexts (r points along with their nearest neighbors). Obviously, we have $s = r(K+1)$. Let $\bar{p} = \frac{1}{s}\sum_{i=1}^{s} p_i$ and $\sigma_p = (\frac{1}{s}\sum_{i=1}^{s}(p_i - \bar{p})^2)^{\frac{1}{2}}$. The estimator of σ is given by $\sigma = \bar{p} + 3\sigma_p$.

A simple induced digraph on the two-cluster data is illustrated in Figure 6 (a). The asymmetry of the associated weighted adjacency matrix is shown in Figure 6 (b).



(a) (b)

Figure 6. (a) Induced digraph of the two-cluster data in Figure 3 (a). Two nearest neighbors are searched for each point, i.e., $K = 2$. (b) Visualization of the associated $\mathbf{W}$.

3.2. Global Digraph Laplacian and Alignment of Local Digraph Laplacians

When the data are modelled by a digraph, data processing reduces to mining the properties of it, which are in general revealed by the digraph Laplacian. Therefore, we need to derive the Laplacian of the digraph. It can be obtained by the alignment of local digraph Laplacians defined on local data patches. The procedure is as follows.

Let $\{\mathbf{x}_{i_0}, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}\}$ be the neighborhood of $\mathbf{x}_i$ and the index set be $I_i = \{i_0, i_1, \ldots, i_K\}$, where $i_0 = i$. Suppose that $\tilde{\mathbf{Y}}_i = [\tilde{\mathbf{y}}_{i_0}, \tilde{\mathbf{y}}_{i_1}, \ldots, \tilde{\mathbf{y}}_{i_K}]$ is a kind of the representations yielded by the digraph embedding. The local weighted adjacency matrix $\mathbf{W}_i$ is a sub-matrix of $\mathbf{W}$: $\mathbf{W}_i = \mathbf{W}(I_i, I_i)$. The local transition probability matrix $\mathbf{P}_i$ of the random walk on the local digraph is given by $\mathbf{P}_i = \mathbf{D}_i^{-1}\mathbf{W}_i$, where $\mathbf{D}_i(u,u) = \sum_v \mathbf{W}_i(u,v)$ and zeros elsewhere. The corresponding stationary distribution vector π_i is the left eigenvector of $\mathbf{P}_i$ corresponding to 1, i.e., $\pi_i^T\mathbf{P}_i = \pi_i^T$ and $||\pi_i||_1 = 1$. Inspired by [9], we define an energy function on the global digraph as the following:

$$\mathcal{R}(\tilde{\mathbf{Y}}) = \frac{\sum_{i=1}^m \alpha_i}{\sum_{i=1}^m \beta_i}, \tag{4}$$

where

$$\begin{aligned} \alpha_i &= \tfrac{1}{2}\textstyle\sum_{u,v=0}^K \|\tilde{\mathbf{y}}_{i_u} - \tilde{\mathbf{y}}_{i_v}\|^2 \pi_i(u)\mathbf{P}_i(u,v), \text{ and} \\ \beta_i &= \textstyle\sum_{v=0}^K \|\tilde{\mathbf{y}}_{i_v}\|^2 \pi_i(v). \end{aligned} \tag{5}$$

With simple manipulations, we can write $\alpha_i = tr(\tilde{\mathbf{Y}}_i\mathbf{L}_i\tilde{\mathbf{Y}}_i^T)$ and $\beta_i = tr(\tilde{\mathbf{Y}}_i\mathbf{\Phi}_i\tilde{\mathbf{Y}}_i^T)$, where

$$\mathbf{L}_i = \mathbf{\Phi}_i - \frac{\mathbf{\Phi}_i\mathbf{P}_i + \mathbf{P}_i^T\mathbf{\Phi}_i}{2} \tag{6}$$

is the local digraph Laplacian defined on the i-th local patch and $\mathbf{\Phi}_i = \text{diag}(\pi_i)$.

The global Laplacian is obtained by aligning all the local Laplacians. To do so, let $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m]$, then for every i, $\tilde{\mathbf{Y}}_i$ should be a sub-matrix of $\tilde{\mathbf{Y}}$. So we can write $\tilde{\mathbf{Y}}_i = \tilde{\mathbf{Y}}\mathbf{S}_i$, where $\mathbf{S}_i$ is a binary selection matrix [4]. Thus we have

$$\alpha = \Sigma_{i=1}^m \alpha_i = \Sigma_{i=1}^m tr(\tilde{\mathbf{Y}}\mathbf{S}_i\mathbf{L}_i\mathbf{S}_i^T\tilde{\mathbf{Y}}^T) = tr(\tilde{\mathbf{Y}}\tilde{\mathbf{L}}\tilde{\mathbf{Y}}^T), \tag{7}$$

where

$$\tilde{\mathbf{L}} = \Sigma_{i=1}^m \mathbf{S}_i\mathbf{L}_i\mathbf{S}_i^T. \tag{8}$$

On the other hand, we have $\beta = \sum_i^m \beta_i = tr(\tilde{\mathbf{Y}}\tilde{\mathbf{\Phi}}\tilde{\mathbf{Y}}^T)$, where $\tilde{\mathbf{\Phi}} = \Sigma_{i=1}^m \mathbf{S}_i\mathbf{\Phi}_i\mathbf{S}_i^T$. Finally, $\mathcal{R}(\tilde{\mathbf{Y}})$ can be written as

$$\mathcal{R}(\tilde{\mathbf{Y}}) = \frac{tr(\tilde{\mathbf{Y}}\tilde{\mathbf{L}}\tilde{\mathbf{Y}}^T)}{tr(\tilde{\mathbf{Y}}\tilde{\mathbf{\Phi}}\tilde{\mathbf{Y}}^T)} = \frac{tr(\mathbf{Y}\mathcal{L}\mathbf{Y}^T)}{tr(\mathbf{Y}\mathbf{Y}^T)}, \tag{9}$$

where $\mathbf{Y} = \tilde{\mathbf{Y}}\tilde{\mathbf{\Phi}}^{\frac{1}{2}}$ is the embedding and $\mathcal{L} = \tilde{\mathbf{\Phi}}^{-\frac{1}{2}}\tilde{\mathbf{L}}\tilde{\mathbf{\Phi}}^{-\frac{1}{2}}$ is the global Laplacian.

Actually, the global Laplacian can be defined in a different yet simpler manner. Define the global transition probability matrix $\mathbf{P}$ as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}(u,u) = \sum_v \mathbf{W}(u,v)$ and zeros elsewhere. Let the stationary distribution of the random walk on the global digraph be π: $\pi^T\mathbf{P} = \pi^T$ and $||\pi||_1 = 1$, and $\mathbf{\Phi} = \text{diag}(\pi)$. In [9], the digraph Laplacian is defined as $\mathbf{L} = \mathbf{I} - \Theta$, where

$$\Theta = \frac{\mathbf{\Phi}^{\frac{1}{2}}\mathbf{P}\mathbf{\Phi}^{-\frac{1}{2}} + \mathbf{\Phi}^{-\frac{1}{2}}\mathbf{P}^T\mathbf{\Phi}^{\frac{1}{2}}}{2}. \tag{10}$$

It is derived by minimizing[5]

$$R(\tilde{\mathbf{Y}}) = \frac{1}{2}\frac{\sum_{u,v=1}^m \|\tilde{\mathbf{y}}_u - \tilde{\mathbf{y}}_v\|^2 \pi(u)\mathbf{P}(u,v)}{\sum_{v=1}^m \|\tilde{\mathbf{y}}_v\|^2 \pi(v)} \tag{11}$$

instead, which can be written as $R(\mathbf{Y}) = \frac{tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)}{tr(\mathbf{Y}\mathbf{Y}^T)}$. Note that the two energy functions defined in (4) and (11) are different. Therefore, the two global digraph Laplacians are different.

By either definition, $\mathbf{Y}^T$ corresponds to the c eigenvectors of the global Laplacian associated with the c smallest nonzero eigenvalues. For convenience, we adopt the *latter* definition for computation. In this case, the columns of $\mathbf{Y}^T$ are also the c nonconstant eigenvectors of $\mathbf{\Theta}$ associated with the c largest eigenvalues.

Note that for digraphs modelled by our method, there may exist nodes that have no inlinks. For instance, the bottom node of the digraph in Figure 5 (a) has no inlinks. Thus the elements in the corresponding column of the weighted adjacency matrix are all zeros (Figure 5 (b)). And such dangling nodes will not be visited by random walkers. To address this issue, we apply the approach [7, 1] by adding a perturbation matrix to the transition probability matrix

$$\mathbf{P} \leftarrow \beta\mathbf{P} + (1-\beta)\frac{1}{m}\mathbf{e}\mathbf{e}^T, \tag{12}$$

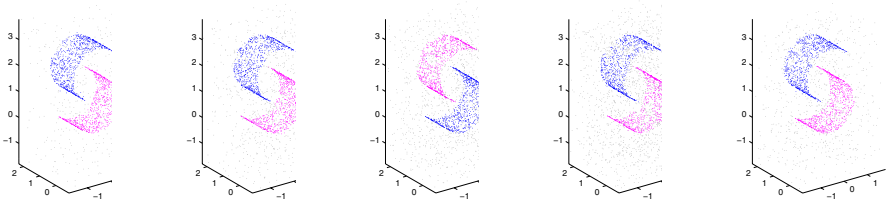
where $\mathbf{e}$ is an all-one vector and $\beta \in [0,1]$.

4. Applications: Clustering and Ranking

In this section, we present two applications of the proposed idea to unsupervised and semi-supervised learning associated with clustering and ranking, respectively. Given a graph and its weighted adjacency matrix, Ng *et al.* [16] proposed the clustering algorithm on an undirected weighted graph, and Zhou *et al.* [28] formulated the algorithms on how to perform clustering and ranking on a digraph. Recently, Agarwal [1] extended the principles of ranking on graph data. In effect, the clustering algorithms are performed on the nonlinear representations of the original samples that are derived by graph embeddings. Inspired by their work, we present the perceptual clustering and ranking algorithms.

[4] One can consult [26] for more details.

[5] One can refer to [9] to know the process of the similar deduction.



(a) (b) (c) (d) (e))

Figure 7. Perceptual clustering on two half-cylinders data. The number of clusters is set to three in advance. We first take mapped points nearest to the origin as the cluster of noise. Then GPCA [22] is employed to identify the remaining clusters. In this experiment, $K = 10$. (a) With 400 noise points. (b) With 800 noise points. (c) With 1200 noise points. (d) With 1600 noise points. The above results are based on coding length. (e) Results based on centroid.

Algorithm of perceptual clustering

1. Model the digraph of the data and form Θ in (10).
2. Compute the c eigenvectors $\{\mathbf{y}_2, \ldots, \mathbf{y}_{c+1}\}$ of Θ corresponding to the first c largest eigenvalues except the largest one. These eigenvectors form a matrix $\mathbf{Y} = [\mathbf{y}_2, \ldots, \mathbf{y}_{c+1}]$. The row vectors of $\mathbf{Y}$ are the mapped feature points of the data.
3. Perform clustering on the feature points.

Algorithm of perceptual ranking[a]

1. Model the digraph of the data and form Θ in (10).
2. Given a vector $\mathbf{v}$ whose i-th element is 1 if it corresponds to a labelled point and zeros elsewhere, compute the score vector $\mathbf{s} = (\mathbf{I} - \alpha\Theta)^{-1}\mathbf{v}$, where α is a free parameter in $[0, 1]$.
3. Sort the scores of $\mathbf{s}$ in descending order. The sample points with large scores are considered to be in the same class as the labelled point.

[a]Note that the ranking algorithm is inherited from Zhou's one [28].

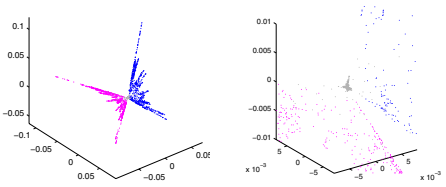
Figure 4 (c) shows the 2D representations of the two-cluster data in Figure 3 (a). We see that two clusters emerge in the perceptual feature space: Cluster I is mapped onto a line, and Cluster II is mapped nearly onto one point. This simple example illustrates the advantage of contextual distances in the structural perception.

5. Experiment

We compare the results of traditional algorithms (based on Euclidean distances) and our proposed algorithms (based on contextual distances) on clustering and ranking.

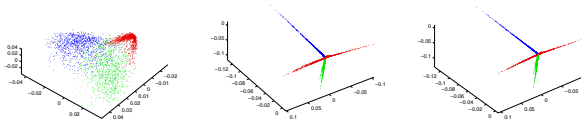
5.1. Clustering

On toy data. Figure 7 shows the results of perceptual clustering on the two half-cylinders data. We see that the perceptual clustering algorithm detects the real structures of the data. We observe an interesting phenomenon that dispersed points in the sample space will be mapped near the origin in the perceptual feature space. Therefore, the noise points can be identified as those points near the origin. Figures 8 (a) and (b) show the 3D representations of samples in the perceptual feature space. The two surfaces



(a) (b)

Figure 8. Embeddings of the two half-cylinders data in the perceptual feature space. (a) 3D representations of the data in Figure 7 (b). (b) Zoom-in view of (a).



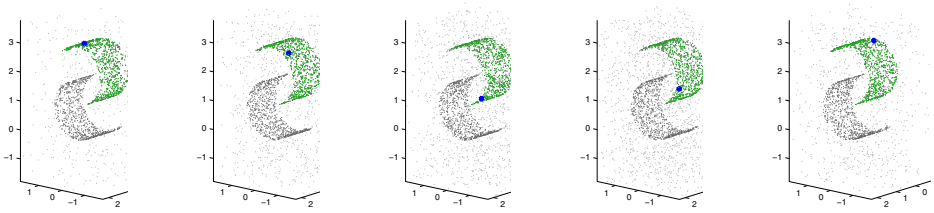
(a) (b) (c)

Figure 9. Visualization of handwritten digits clustering. Red dots represent the digit '1', green '2', and blue '3'. (a) NJW clustering. (b) Perceptual clustering I (based on coding length). (c) Perceptual clustering II (based on centroid). $K = 15$ for both perceptual clustering.

are mapped into two different linear subspaces and noise points are mapped around the origin.

On handwritten digits. We use all samples of digits 1, 2, and 3 in the test set of the MNIST handwritten digit database[6]. There are 1135, 1032, and 1010 samples, respectively. We directly visualize the representations of samples in the associated feature spaces instead of a quantified comparison as different clustering methods should be chosen for different distributions of mapped points. Besides, it is more intuitive for one to compare the distinctive characteristics of the involved algorithms by visual perception. As shown in Figure 9, the perceptual clustering algorithms yield more compact and clearer representations of clusters than the NJW clustering algorithm does. We observe that different clusters are mapped approximately into different linear subspaces by perceptual clustering. Such mixed linear structures can be easily identified by GPCA [22] and the method in [14]. For each underlying cluster, we find the farthest samples from the origin and the nearest from it in the

[6]http://www.cs.toronto.edu/~roweis/data.html



(a) (b) (c) (d) (e)

Figure 11. Perceptual ranking on two half-cylinders data. One point is randomly labelled on one of the half cylinders for each trial. Then we recolor the 800 points that correspond to the first 800 largest ranking scores to be green. In this experiment, $K = 10$ and $\alpha = 0.999$. (a) With 400 noise points. (b) With 800 noise points. (c) With 1200 noise points. (d) With 1600 noise points. The above are results based on coding length. (e) Result based on centroid.

(b)

Figure 10. Distribution of handwritten digits in the perceptual feature space. The first ten-column digits are the farthest from the origin and the second ten-column digits are the nearest to the origin. (a) Based on coding length. (b) Based on centroid.

perceptual feature space. The results are shown in Figure 10. As expected, noise samples are near to the origin and 'good' samples are far from it.

5.2. Ranking

On toy data. Figure 11 shows the results of perceptual ranking on the two half-cylinders data. The perceptual ranking algorithm accurately labels the points on the labelled surface. The results are robust against noise. In contrast, the result by Zhou's ranking is not satisfactory (Figure 1 (e)).

On family photos. The database used in this experiment is a collection of real photos of a family and its friends [10, 21]. The faces in photos are automatically detected, cropped, and aligned according to the positions of eyes. There are all together 980 faces of 26 persons. Figure 12 shows one cropped face of each person. We first apply the algorithm of local binary pattern (LBP) [3] to extract the expressive features, and then exploit dual-space LDA [24] to extract the discriminant features from the LBP features. Then Zhou's ranking and our perceptual ranking are performed, respectively. The ratio of the number of correctly ranked faces to the total number of faces in the first 50 ranked faces is considered as the accuracy measure. Specifically, let Z denote the ranked faces and z the correctly ranked ones. Then, the accuracy is defined as $\frac{z}{Z}$. Only the photos of five members in the family are ranked. For each person, the ranking experiment is performed for two hundred trials, and the mean accuracy is illustrated in Figure 13 (a), where perceptual ranking shows the superiority. Figures 13 (b) and (c) indicate that perceptual ranking is robust with the variations of α and K.

Figure 12. Family photos. The identities of first five photos in the first row are Mingming, mama, papa, grandma, and grandpa, respectively. The numbers of cropped faces of them are 153, 171, 152, 94, and 61, respectively.

6. Conclusion

We propose a new perspective on structural perception of data. We locally define contextual distances between nearby points based on the geometric descriptors of associated contextual sets. The asymmetry of contextual distances naturally induces a digraph on data to model the global structure of data, whose directed edges are weighted by the exponential function of contextual distances. As a result, the structural perception of data can be achieved by mining the properties of the digraph. We test the proposed asymmetric perception based algorithms on data clustering and ranking. Experiments show the superiority of our approaches.

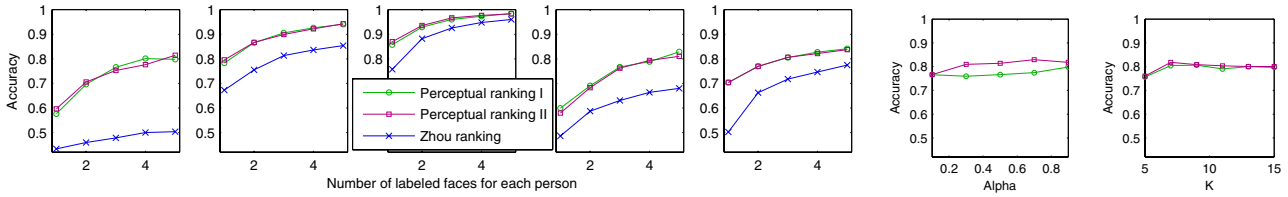
Acknowledgement

The authors would like to thank Yi Ma and John Wright for discussion and sharing their draft on classification via minimum incremental coding length (MICL) [25], and Steve Lin for his careful proofreading on the draft.

Appendix

Coding Length. Let $\mathbf{X}_i = [\mathbf{x}_i, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_K}]$ and $\bar{\mathbf{x}}_i = \frac{1}{K+1}\mathbf{X}_i\mathbf{e}$, where $\mathbf{e}$ is the $K+1$ dimensional all-one vector. Then the matrix of centered points is written as $\bar{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{x}}_i\mathbf{e}^T$, where T denotes the transpose of a matrix. The total number of bits needed to code $\mathcal{S}_i$ is

$$L(\mathcal{S}_i) = \frac{K+1+n}{2}\log\det(\mathbf{I} + \frac{n}{\varepsilon^2(K+1)}\bar{\mathbf{X}}_i\bar{\mathbf{X}}_i^T) + \frac{n}{2}\log(1 + \frac{\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i}{\varepsilon^2}), \quad (13)$$



(a) (b) (c)

Figure 13. Ranking results on family photos. (a) From left to right, the results correspond to the identities of Mingming, mama, papa, grandma, and grandpa. In this experiment, $K = 7$ and $\alpha = 0.9$ for both perceptual ranking, and $\alpha = 0.1$ for Zhou's ranking. (b) Variation of accuracy with α in the case of $K = 7$. (c) Variation of accuracy with K in the case of $\alpha = 0.9$. (b) and (c) are both the results of perceptual ranking on Mingming's photos.

where $\det(\bullet)$ is the determinant operator and ε is the allowable distortion. In fact, the computation can be considerably simplified by the commutativity of determinant

$$\det(\mathbf{I} + \frac{n}{\varepsilon^2(K+1)}\bar{\mathbf{X}}_i\bar{\mathbf{X}}_i^T) = \det(\mathbf{I} + \frac{n}{\varepsilon^2(K+1)}\bar{\mathbf{X}}_i^T\bar{\mathbf{X}}_i) \tag{14}$$

in the case of $K + 1 \ll n$. One can refer to [14] for more details.

References

[1] S. Agarwal. Ranking on graph data. In *ICML*, pages 25–32, 2006. 1, 5

[2] S. Agarwal, K. Branson, and S. Belongie. Higher order learning with graphs. In *ICML*, pages 17–24, 2006. 1

[3] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: application to face recognition. *PAMI*, 28:2037–2041, 2006. 7

[4] A. Azran and Z. Ghahramani. Spectral methods for automatic multiscale data clustering. In *CVPR*, pages 190–197, 2006. 1

[5] M. Balasubramanian and E. L. Schwartz. The Isomap algorithm and topological stability. *Science*, 295:7a, 2002. 1

[6] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. *JMLR*, 7:2399–2434, 2006. 1

[7] S. Brin, L. Page, R. Motwami, and T. Winograd. The Pagerank citation ranking: bringing order to the web. Technical Report 1999-0120, Computer Science Department, Standford University, Standford, CA, 1999. 5

[8] J. Bruner and A. Minturn. Perceptual identification and perceptual organization. *Journal of General Psychology*, 53:21–28, 1955. 3

[9] F. Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9:1–19, 2005. 5

[10] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *ACM Special Interest Group on Computer-Human Interaction (SIGCHI)*, 2007. 7

[11] C. Ding and X. F. He. K-means clustering via principal component analysis. In *ICML*, pages 225–232, 2004. 1

[12] C. Ding, X. F. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM Data Mining*, 2005. 1

[13] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. 1

[14] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Tran. on PAMI*, 2007. 4, 6, 8

[15] M. Meilă and J. B. Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001. 1

[16] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *NIPS*, pages 849–856, 2001. 1, 5

[17] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 1

[18] H. Seung and D. Lee. The manifold ways of perception. *Science*, 290:2268–2269, 2000. 1

[19] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Tran. on PAMI*, 22:888–905, 2000. 1

[20] J. Tenenbaum, V. Silva, and J. Langford. A gloabal geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000. 1

[21] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *International Conf. on Computer Vision and Pattern Recognition*, 2007. 7

[22] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis. *IEEE Tran. on PAMI*, 27:1–15, 2005. 6

[23] U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In *NIPS*, pages 13–18, 2004. 1

[24] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *International Conf. on Computer Vision and Pattern Recognition*, pages 564–569, 2004. 7

[25] J. Wright, Y. Tao, Z. Lin, Y. Ma, and H. Shum. Classification via minimum incremental coding length. *Submitted to IEEE Tran. on PAMI. The preprint is available at: http://perception.csl.uiuc.edu/˜jnwright*, 2007. 7

[26] D. Zhao, Z. Lin, and X. Tang. Laplacian PCA and its applications. In *International Conf. on Computer Vision*, 2007. 5

[27] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004. 1, 2

[28] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *ICML*, pages 1036–1043, 2005. 1, 5, 6