On Order-Constrained Transitive Distance Clustering

Zhiding Yu*, Weiyang Liu†, Wenbo Liu*‡, Yingzhen Yang§, Ming Li‡, B. V. K. Vijaya Kumar*
*Dept. of Electrical and Computer Engineering, Carnegie Mellon University
†School of Electronic and Computer Engineering, Peking University, P.R. China
‡SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University
§Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

Abstract

We consider the problem of approximating order-constrained transitive distance (OCTD) and its clustering applications. Given any pairwise data, transitive distance (TD) is defined as the smallest possible "gap" on the set of paths connecting them. While such metric definition renders significant capability of addressing elongated clusters, it is sometimes also an over-simplified representation which loses necessary regularization on cluster structure and overfits to short links easily. As a result, conventional TD often suffers from degraded performance given clusters with "thick" structures. Our key intuition is that the maximum (path) order, which is the maximum number of nodes on a path, controls the level of flexibility. Reducing this order benefits the clustering performance by finding a trade-off between flexibility and regularization on cluster structure. Unlike TD, finding OCTD becomes an intractable problem even though the number of connecting paths is reduced. We therefore propose a fast approximation framework, using random samplings to generate multiple diversified TD matrices and a pooling to output the final approximated OCTD matrix. Comprehensive experiments on toy, image and speech datasets show the excellent performance of OCTD, surpassing TD with significant gains and giving state-of-the-art performance on several datasets.

Introduction

Clustering has been and continues to remain one of the most fundamental machine learning problems. Today, with the fast growth of digital media and storage, the growing speed of annotation capability can hardly match the explosive increase of data. In many problems and applications where supervised information is difficult to obtain or even not available, clustering presents an important unsupervised learning approach to analyze the useful latent patterns. In speech processing for example, NIST recently organized the i-vector Machine Learning Challenge (Greenberg et al. 2014) where one is required to design speaker verification systems trained on an unlabeled i-vector development dataset. Among the participating works, clustering techniques are widely used as indispensable learning components. In computer vision, a number of important applications such as unsupervised image segmentation (Shi and Malik 2000) and image categorization (grouping) (Feng et al. 2014; Hu et al. 2014) can be naturally formulated as clustering problems.

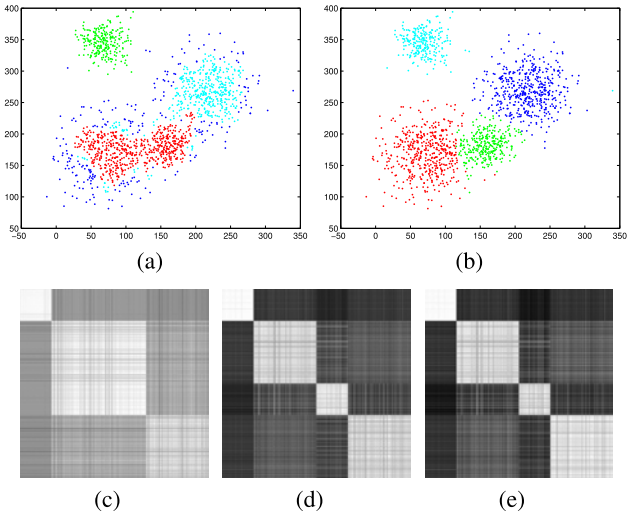


Figure 1: A toy example with densely aligned Gaussian distributed clusters. SVD (a) Clustering result with TD. (b) Clustering result with OCTD (Min). Similar result is given by OCTD (Mean) and is omitted. (c) (d) and (e) respectively correspond to the (negative) distance matrices of pairwise rows in TD, OCTD (Min) and OCTD (Mean) after SVD.

A wide variety of clustering algorithms were proposed in the past decades. Some famous early clustering methods include the k-means algorithm (Lloyd 1982), fuzzy clustering (Bezdek 2013), hierarchical clustering (Sibson 1973; Defays 1977), and mode seeking (Cheng 1995; Comaniciu and Meer 2002). More recently, with the increase of computer memory and computation power, more sophisticated methods including the families of spectral clustering (Ng et al. 2002; Shi and Malik 2000; Zelnik-Manor and Perona 2004) and subspace clustering (Elhamifar and Vidal 2009; Lu et al. 2012; Liu et al. 2013; 2013; Peng, Zhang, and Yi 2013; Peng, Yi, and Tang 2015; Hu et al. 2014; Li et al. 2015) were proposed. Spectral clustering often includes an eigen-decomposition on the affinity matrix and has wide applications in image segmentation. On the other hand, sub-

space clustering assumes a low dimensional subspace for each cluster. it is therefore very successful in applications such as facial image clustering and motion segmentation where the subspace assumption is satisfied.

Good clustering methods in general should maximally reveal both the similarity between intra-cluster data and the dissimilarity between inter-cluster data. Finding such good methods is nontrivial as the distribution of a cluster often shows much more dynamic shapes than convex ones. The difficulty is further aggravated when cluster ambiguity is present (noisy data and overlapping clusters). Often, it is not that hard to design methods that can target specific scenarios. But it becomes considerably harder for a single method to adaptively handle all situations, including both varying cluster shapes and ambiguities. Spectral clustering received much attention for its ability to address clusters with arbitrary shapes. Besides spectral clustering, an elegant alternative with strong flexibility on cluster structure is the transitive distance clustering (also known as path based clustering) (Fischer and Buhmann 2003b; Yu et al. 2014). TD is an ultrametric which reveals the strength of connectivity between pairwise points. The TD between any pairwise samples is defined as the smallest possible "gap" on the set of paths[1] that connect the samples, where a "gap" of a path is the largest edge along it. In other words, two samples are considered to be strongly correlated if they are connected by at least one path with a very small gap. Such definition renders clustering methods based on TD very strong flexibility on cluster shapes.

A known problem of TD is the degraded performance on noisy data caused by short links. The TD metric also easily overfits and loses necessary regularization on cluster structure, reducing discriminative cluster information on "thick" clusters. Unfortunately, path-like cluster structures are not commonly seen in non-synthetic datasets. Even for facial image datasets with manifold structures, most of them do have relatively "thick" clusters. To address this problem, we propose the order-constrained transitive distance - a novel dissimilarity measure more robust than TD by reducing the maximum path order. Considering the intractability of finding the OCTD, we approximate it by a min/mean-pooling over a set of diversified TD matrices generated from random samplings (without replacement). Following (Yu et al. 2015), a top-down grouping with SVD is conducted on the OCTD matrix to give the clustering result. Fig. 1 gives a toy example with densely aligned clusters. Clustering with TD fails in this example due to the cluster ambiguity, while clustering with the proposed OCTD (with min-pooling) correctly identifies the clusters.

Our major contributions are summarized as follows: 1. We extend the current transitive distance framework with constrained path orders and propose a novel order-constrained transitive distance. 2. We propose an approximation framework to efficiently approximate OCTD. 3. Comprehensive experiments indicate that the proposed method significantly outperforms TD. The rest of the paper will describe the proposed algorithm and its properties in details.

[1]Sequences of non-repeated intermediate samples and edges.

Related Works

A number of previous literatures investigated the problem of clustering with TD. Several major works include the connectivity kernel (Fischer, Roth, and Buhmann 2004), the transitive distance closure (Ding et al. 2006) and the transitive affinity (Chang and Yeung 2005; 2008).

(Fischer and Buhmann 2003a) proposed bagging TD clustering with resamplings and label-level maximum likelihood fusion of multiple clustering results. It was shown that bagging can effectively reduce clustering errors caused by noise as resampling tend to filter out noisy samples. (Yu et al. 2015) proposed a generalized transitive distance (GTD) framework with minimum spanning forest and max-pooling to incorporate more robustness. GTD was shown to be more robust in finding weak contours in image segmentation, obtaining state-of-the-art image segmentation performance on the BSDS-300 dataset.

While Our work is to some extend related to both methods, it also differs considerably in many aspects. Unlike the grouping-level encoding of clustering robustness (Fischer and Buhmann 2003a), we seek to directly output a robust distance measure before grouping. In addition, a top-down grouping approach is used instead of the bottom-up agglomerative one in (Fischer and Buhmann 2003a). Our method also significantly differs from GTD as min and mean-poolings are used instead of max-pooling, allowing much more intense perturbations. This leads to more significant boost of performance in data clustering tasks.

Transitive Distance with Constrained Order

Transitive distance is an ultrametric defined to shorten the intra-cluster distances on long cluster structures. It is defined as follows with respect to path connectivity:

Definition 1. *Given certain pairwise samples (x_p, x_q) and the edge weights $d(e)$, the transitive distance is defined as:*

$$D_{td}(x_p, x_q) = \min_{\mathcal{P}\in\mathbb{P}} \max_{e\in\mathcal{P}} \{d(e)\}, \tag{1}$$

where $\mathbb{P}$ is the set of paths connecting x_p and x_q with at most n nodes (including x_p and x_q). In addition:

$$\max_{e\in\mathcal{P}} \{d(e)\} = \max_{(x_u,x_v)\in\mathcal{P}} \{d(x_u, x_v)\}. \tag{2}$$

An ultrametric is guaranteed to have a feasible Euclidean embedding in another space. The projected cluster structures in the TD embedded space can become much more compact (Yu et al. 2014). Essentially, TD implicitly builds the following non-linear mapping similar to spectral clustering:

$$\phi : V \subset \mathbf{R}^l \mapsto V' \subset \mathbf{R}^s. \tag{3}$$

As a result, clustering in the TD embedded space can handle highly non-convex cluster structures.

The Proposed Definition of OCTD

While TD reduces the intra-cluster distances, inter-cluster samples can also be dragged much closer. Such disadvantage becomes particularly obvious when clustering ambiguities and noises are present since short links of path are easily formed upon them. Our key observation in this paper is

that constraining the maximum path order can regularize the path set $\mathbb{P}$ and significantly reduce such short links. Thus we consider the following order-constrained TD:

Definition 2. *Given certain pairwise samples (x_p, x_q) and the edge weights $d(e)$, the order-constrained transitive distance is defined as:*

$$D_{octd}(x_p, x_q) = \min_{\substack{\mathcal{P}\in\mathbb{P}, \\ \mathcal{O}(\mathcal{P})<L}} \max_{e\in\mathcal{P}}\{d(e)\}, \tag{4}$$

where $\mathcal{O}(\mathcal{P})$ denotes the order of path $\mathcal{P}$.

Diversified Spanning Graphs with Samplings

Given the TD defined on a graph G, an elegant property is that the transitive edges (gaps) always lie on the minimum spanning tree (MST) of G (Fischer and Buhmann 2003b). Therefore, finding TD has a practical solution despite its seemingly difficulty. Such property, however, no longer holds when the maximum path order is constrained. As a result, an alternative framework is needed to approximate the order-constrained TD that we want.

Inspired by the work of Nyström method (Drineas and Mahoney 2005; Williams and Seeger 2001), we propose a random sampling based approximation for TD. Suppose $\mathbf{X}_S^{(t)} = \{\mathbf{x}_i | i \in S^{(t)}\}$ denotes the sampled data each time and $\mathbf{X}_R^{(t)} = \{\mathbf{x}_i | i \in S^{(t)C}\}$ the rest of the data. We use $\mathbf{X}_S^{(t)}$ to construct a spanning graph $G_S^{(t)} = (V, E_S^{(t)})$ from the original complete graph $G = (V, E)$. The first step contains a kernel density estimation with a Gaussian kernel:

$$\hat{p}(\mathbf{x}_i) = C\sum_{j=1}^{N} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}), \tag{5}$$

where σ is the bandwidth and C is a normalization constant such that $\sum_{i=1}^{N} \hat{p}(\mathbf{x}_i) = 1$. The bandwidth parameter is automatically estimated as:

$$\hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathrm{knn}(\mathbf{x}_i, k)\|_2. \tag{6}$$

For most experiments in this paper, k is fixed to 10. With the estimated density at each data location, we randomly sample multiple subsets $\mathbf{X}_S^{(t)}$ without replacement, each contains $M < N$ data. The data are sampled with probabilities proportional to the estimated density.

To construct the corresponding spanning graph $G_S^{(t)}$ from G given every sampled subset $\mathbf{X}_S^{(t)}$, we first construct a clique (fully connected graph) on $\mathbf{X}_S^{(t)}$:

$$G_C^{(t)} = (\tilde{V}^{(t)} = S^{(t)}, E_C^t = \{e(i,j) | i, j \in S^{(t)}\}). \tag{7}$$

The spanning graph $G_S^{(t)}$ is then defined by connecting the rest of non-sampled data to the closest ones in $\mathbf{X}_S^{(t)}$:

$$G_S^{(t)} = (S, E_S^{(t)} = \{E_C^{(t)}, E_N^{(t)}\}), \tag{8}$$

where $E_N^{(t)} = \{e(i, nn(i, S\backslash S^{(t)})) | i \in S^{(t)}\}$.

There are several reasons for random samplings. First, multiple samplings will provide diversified spanning graphs and order-constrained path sets to approximate OCTD. Second, certain level of regularization is incorporated by sampling more important samples to form the major frameworks of spanning graphs. Third, the final ensemble of multiple samplings is expected to bring additional robustness.

Approximating OCTD with Spanning Graphs

Here we show that the TD matrix obtained on $G_S^{(t)}$ is order-constrained and is therefore an approximation of OCTD.

Theorem 1. *The maximum possible path order on the spanning graph $G_S^{(t)}$ is upper bounded by $|S^{(t)}| + 2$.*

Proof: There are three possible cases at both ends of the path: 1. Two sampled nodes; 2. A sampled node and a non-sampled node; 3. Two non-sampled nodes.
Case 1: We use contradiction to prove that $\mathcal{O}(\mathcal{P}) \leq |S^{(t)}|$. Suppose $\mathcal{O}(\mathcal{P}) > |S^{(t)}|$, by definition at least one node from non-sampled set is part of the path. This indicates that the node from non-sampled set is connected to at least two other nodes, which contradicts to our original setup.
Case 2: It is easy to prove that $\mathcal{O}(\mathcal{P}) \leq |S^{(t)}| + 1$ using contradiction similar to **Case 1**.
Case 3: One can prove that $\mathcal{O}(\mathcal{P}) = 2$ if two nodes share the same nearest node in the sampled set, and $\mathcal{O}(\mathcal{P}) \leq |S^{(t)}|+2$ otherwise, again with contradiction.

Theorem 1 states that the approximated pairwise transitive distance obtained on every spanning graph satisfies a constraint on the path order. This forms one of the core theoretical bases of our proposed framework.

Theorem 2. *For any pair of nodes, the number of connecting paths on $G_S^{(t)}$ is upper bounded by $(|S^{(t)}| - 2)!$*

Proof: There are three cases: 1. Non-sampled nodes sharing the same nearest sampled node; 2. A non-sampled node with a nearest sampled node. 3. Other situations.
Case 1 & 2: It is easy to prove there is only one path.
Case 3: When both are sampled nodes, any non-sampled nodes can not be part of the path. Since $G_S^{(t)}$ is a clique, a path can be formed by non-repeatedly selecting one out of $|S^{(t)}| - 2$ nodes, which leads to $(|S^{(t)}| - 2)!$ possibilities. When having one or two non-sampled nodes, the non-sampled node is connected to the clique with only one edge, which does not contribute any additional path candidates. Again we have $(|S^{(t)}| - 2)!$ possibilities.

Theorem 3. *The transitive distance obtained on $G_S^{(t)}$ is lower-bounded by the order-constrained transitive distance obtained on the original fully connected graph G:*

$$D_{td}^{(t)}(x_i, x_j | G_S^{(t)}) \geq D_{octd}(x_i, x_j | G). \tag{9}$$

Proof: This is a conclusion from the fact that $G_S^{(t)}$ is a subgraph of G, therefore the sets of connecting paths in $G_S^{(t)}$ is only a subset of that in G. Based on the definition of TD, we can prove the above theorem.

Pooling with Multiple Subgraphs

Given the set of T diversified TD matrices computed from spanning graphs, we seek to ensemble them and output a final distance matrix. A natural way of ensemble to consider is the min-pooling, which is computed as:

$$D_{octd1}(x_i, x_j) \triangleq \min_t D_{td}^{(t)}(x_i, x_j | G_S^{(t)}). \quad (10)$$

For OCTD obtained by min-pooling, one has the following approximation optimality theorem:

Theorem 4. *Given the set of $D_{td}^{(t)}(x_i, x_j|G_S^{(t)})$, min-pooling gives the optimal approximation of $D_{octd}(x_i, x_j|G)$.*

Proof: There are two alternative ways we can look into this: 1. According to **Theorem 3**, the OCTD is the lower bound of every diversified TD matrices. Therefore, min-pooling gives the optimal approximation. 2. Computing pairwise transitive distance in $D_{td}^{(t)}(x_i, x_j|G_S^{(t)})$ is looking for the smallest possible gaps among a set of order-constrained paths. Performing min-pooling basically equals to extending the set of connecting paths to the union of those in $G_S^{(t)}$, and therefore optimally complies with the definition of TD.

We will denote the computed distance with min-pooling as **OCTD (Min)**. Note that **OCTD (Min)** is no longer an ultrametric or even a metric since the metric triangle inequality may not hold any more. Such pairwise distances violate metricity and cannot be naturally embedded in a vector space (Roth et al. 2003). We therefore also consider an alternative strategy with mean-pooling:

$$D_{octd2}(x_i, x_j) \triangleq \frac{1}{T} \sum_{t=1}^{T} D_{td}^{(t)}(x_i, x_j | G_S^{(t)}). \quad (11)$$

We will denote the computed distance with mean-pooling as **OCTD (Mean)**. The mean pooling no longer strictly follows the definition of TD, yet the obtained distance is still a reasonable approximation of OCTD and shares many similar clustering properties.

Lemma 1. *$D_{td}^{(t)}(x_i, x_j|G_S^{(t)})$ is an ultrametric.*

This is a proved conclusion from (Fischer, Roth, and Buhmann 2004). The lemma indicates that $D_{td}^{(t)}(x_i, x_j|G_S^{(t)}) \leq \max(D_{td}^{(t)}(x_i, x_k|G_S^{(t)}), D_{td}^{(t)}(x_j, x_k|G_S^{(t)})), \forall\{i, j, k\}$. We will use this lemma to obtain the following theorem.

Theorem 5. *OCTD (Mean) is a metric.*

Proof: Using **Lemma 1**, we have the triangle inequality:

$$\begin{aligned} D_{octd2}(i, j) \leq & \frac{1}{T} \sum_t \max(D_{td}^{(t)}(i, k), D_{td}^{(t)}(j, k)) \\ \leq & \frac{1}{T} \sum_t (D_{td}^{(t)}(i, k) + D_{td}^{(t)}(j, k)) \\ = & D_{octd2}(i, k) + D_{octd2}(k, j) \end{aligned} \quad . \quad (12)$$

Other properties such as non-negativity, symmetry and coincidence axiom are easy to prove and omitted.

Top-Down Clustering with SVD

With the obtained approximated OCTD distances, we follow (Yu et al. 2015) to perform top-down clustering where SVD is used for the low rank approximation and noise reduction of the distance matrix. For clustering with K clusters, the eigenvectors of SVD with the K largest eigenvalues are selected to form an $N \times K$ matrix U, followed by k-means over the rows of U to generate the final clustering labels.

Experimental Results

In this section, we describe the details of a comprehensive set of experiments, ranging from toy datasets to the widely used datasets of both image and speech.

Results on Toy Datasets

A set of challenging toy examples are used to test the algorithm performance. Our proposed methods are compared with two popular spectral clustering methods, which are spectral clustering (SC) (Ng et al. 2002) and normalized cuts (Ncut) (Shi and Malik 2000). Euclidean distance k-means (Kms (Euc)) and TD+SVD (Yu et al. 2015) are also used as baselines in addition to SC and Ncut.

To reduce the influence of fluctuated performance from k-means due to different initializations, the k-means grouping stages in SC, k-means, TD and the proposed methods are repeated 10 times. The result with the minimum distortion is selected. Euclidean distance input is used for all methods. The affinity matrices for SC and Ncut are then computed on this Euclidean input with a Gaussian kernel. For TD and the proposed methods, the edge weights of constructed graphs are also based on the Euclidean distance.

The clustering results of the comparing methods are visualized in Fig. 2 and Fig. 3. In addition, quantitative results of different methods are listed in Table 1. The parameters for every method on every example is tuned to optimize it clustering result. The proposed methods are not very sensitive to parameters as a set of fixed parameters can be easily found to work well on most examples. Only a few requires more detailed tuning of parameters.[2]

One could see that OCTD (Min) and OCTD (Mean) obtain the best results on most toy examples. While both methods maintain characteristics similar to TD by showing similar correct results on "Compound", they significantly improved the algorithm robustness over TD on a number of other examples where clustering ambiguities exist.

Results on Image Datasets

We also report our results on several widely used image datasets and describe the experimental setup, including the preprocessing of images and the parameters of our methods.

[2]For OCTD (Min) and OCTD (Mean), having a sample rate of 0.3 and 500 diversified TD matrices works well on most examples. Increasing or decreasing these parameters does not change the results too much. Special tunings are only required on "Pathbased" and 'Spiral' where elongated structures exist. The sample rates on the two examples are increased to 0.8. In addition, the KNN number for bandwidth estimation on "Pathbased" is reduced to 2.

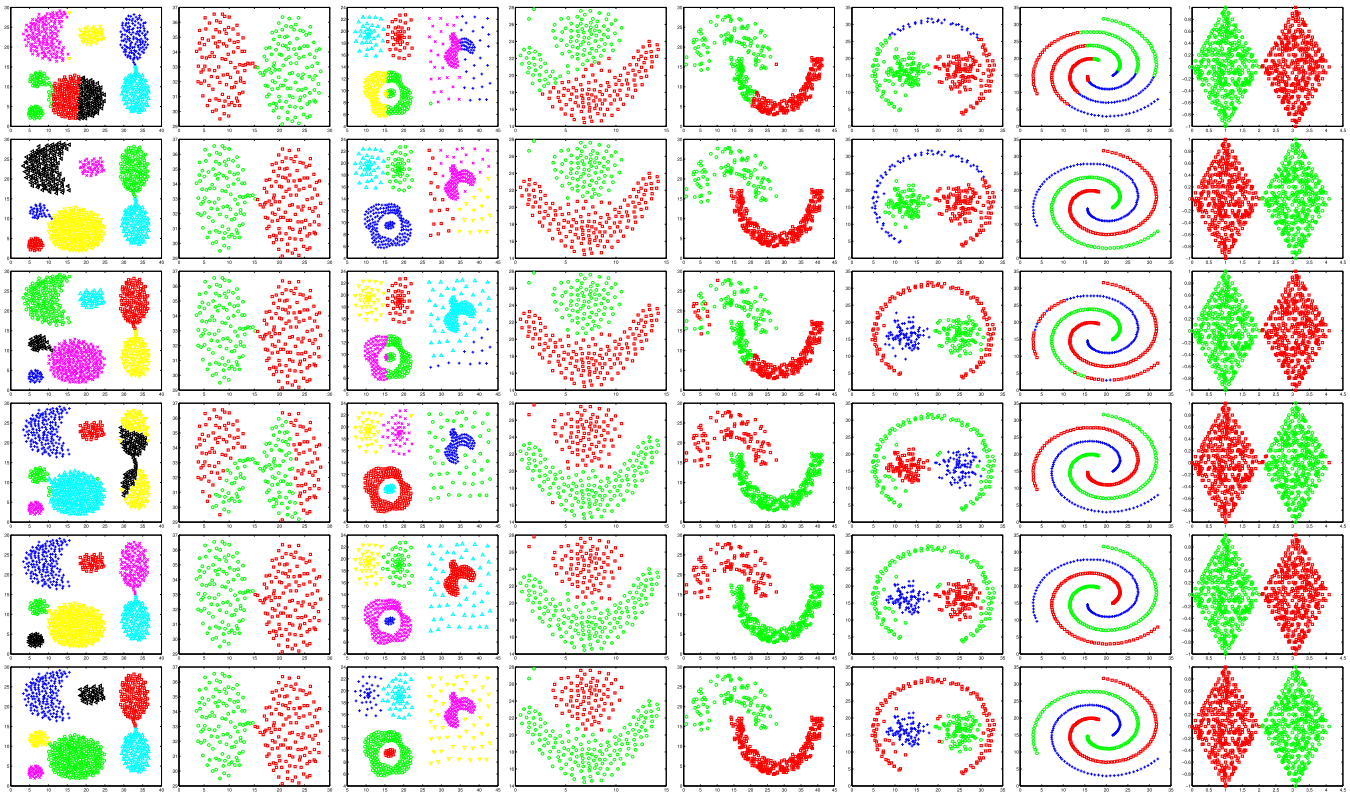
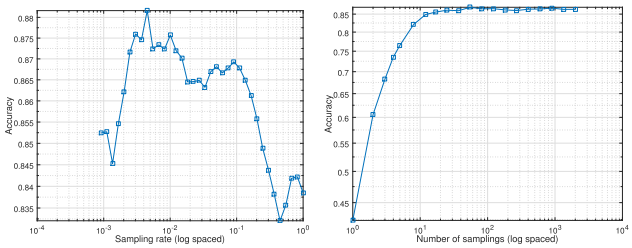


Figure 2: Results of comparing methods on toy examples with varying cluster shapes (Best viewed in color). Row 1-6 respectively correspond to Kms (Euc), SC, Ncut, TD+SVD, OCTD (Min) and OCTD (Mean). Names of examples are respectively "Aggregation", "Bridge", "Compound", "Flame", "Jain", "Pathbased", "Spiral" and "Two Diamonds".



(a) (b)

Figure 4: Clustering performance with different parameters. (a) Varying the sampling rate and fixing $T = 500$. (b) Varying T and fixing the sampling rate to be 0.06.

The Extended Yale B dataset (ExYB) contains 2414 frontal-faces (192×168) of 38 subjects. We use the complete dataset and resize the images to 55×48. For preprocessing, PCA whitening with 99% of energy is used.

For the AR face dataset (Martınez and Benavente 1998), we follow the exact setting of (Peng, Yi, and Tang 2015) where a subset of 50 male subjects and 50 female subjects is chosen. The subset contains 1400 cropped faces (55×40) which are not occluded. Again PCA whitening with 98% of energy is used to preprocess the images.

Table 2: Quantitative results of comparing methods on image datasets. Accuracies are measured with %.

Method	Kms	SC	Ncut	TD	OCTD (Min)
ExYB	44.74	87.28	83.76	82.81	**90.64**
AR	64.29	80.64	87.29	83.85	**88.28**
USPS	64.38	82.94	82.38	54.31	**85.13**

The USPS dataset contains 9298 16×16 handwritten digit images. We use the whole dataset and similarly perform PCA whitening with 98.5% of energy.

We take the AR face dataset and investigate how the performance changes by varying the sampling rate and the number of samplings T. Each is repeated 10 times to return the averaged performance curve. The results are shown in Fig. 4. As the sampling rate goes up, the clustering accuracy first increases and then starts to decrease. Regularization by constraining the path order clearly improves performance and a trade-off between regularization and flexibility is preferred. Also, as T increases, so does clustering accuracy until it saturates. This shows that a positive correlation exists between approximating OCTD and clustering performance.

We compare the performance of different methods with cosine distance used as the distance measure input. The

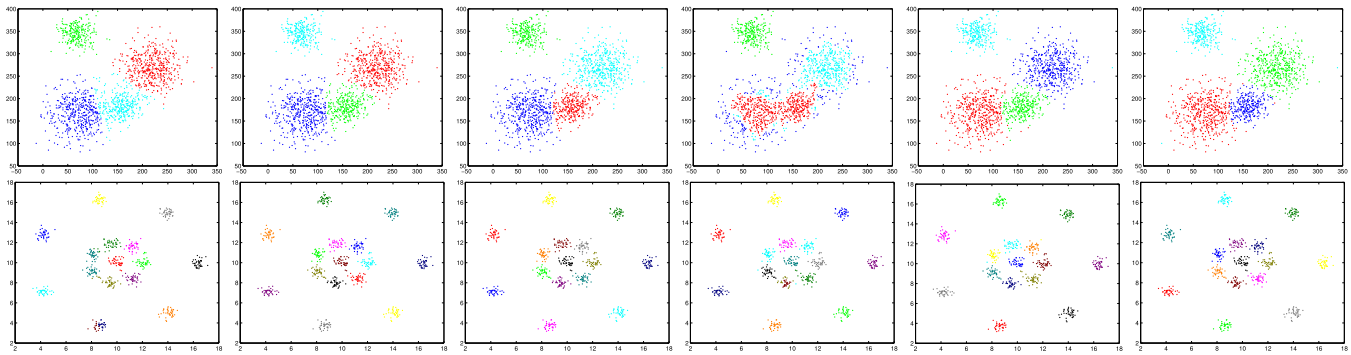


Figure 3: Results of comparing methods on toy examples with densely aligned Gaussian distributions (Best viewed in color). Column 1-6 respectively correspond to K-Means, SC, Ncut, TD+SVD, OCTD (Min) and OCTD (Mean). Names of examples are respectively "Gaussian" and "R15".

Table 1: Quantitative results of comparing methods on toy datasets. Accuracies are measured with %.

Method	Aggregation	Bridge	Compound	Flame	Jain	Path.	Spiral	TwoDiam.	Gaussian	R15
Kms (Euc)	93.91	99.14	83.21	83.75	78.28	74.58	33.97	**100**	93.13	92.5
SC	99.37	99.14	91.73	97.92	**100**	87.63	100	**100**	95.2	**99.67**
Ncut	99.37	99.14	86.72	98.75	77.48	**98.66**	87.18	**100**	**95.8**	**99.67**
TD+SVD	87.94	60.78	99.5	98.75	**100**	96.99	**100**	99.25	78.6	92.33
OCTD (Min)	**99.87**	**99.57**	**99.75**	**100**	**100**	96.66	**100**	**100**	95.33	99
OCTD (Mean)	99.75	**99.57**	**99.75**	98.33	**100**	96.32	**100**	**100**	**95.8**	**99.67**

Table 3: Quantitative results of comparing methods on speech datasets. Accuracies are measured with %.

Method	Kms (Euclid)	Kms (Cos)	SC	Ncut	TD+SVD	OCTD (Min)	OCTD (Mean)
NIST 04	66.32	81.49	83.32	80.49	77.17	**84.9**	84.51
NIST 05	72.99	77.08	74.3	76.1	72.86	**77.87**	73.04
NIST 06	79.84	86.43	80.72	84.4	87.07	**88.29**	83.47
NIST 08	74.52	78.58	**81.51**	62.65	74.13	77.91	78.81
NIST Combined	70.85	78.97	76.21	71.66	72.07	**80.89**	77.24
Switch Board	86.03	90.80	87.79	80.83	78.73	87.53	**90.88**

bandwidth parameters of both SC and Ncut are tuned to output the best results. For OCTD (Min), we also vary the number of random samplings and the sampling rate to optimize the results. Table 2 shows the clustering performance of the baselines and OCTD (Min). A significant gain over TD and other baselines is obtained by the proposed method.

Results on Speech Datasets

Finally, we first conduct large scale clustering experiment on several speech datasets. The NIST and Switch Board datasets are formed by extracting the i-vectors under the framework of (Li and Narayanan 2014)[3]. I-vectors from Switchboard form the "Switchboard" dataset containing 11587 500-dimensional i-vectors and 1052 identities. The rest from NIST SRE form the "NIST" dataset containing 21704 i-vectors and 1738 identities. Note that the NIST dataset is the combined set of NIST 04, 05, 06 and 08.

[3]The i-vectors are trained on Switchboard II part1 to part3 and NIST SRE 04, 05, 06, 08 corpora on the telephone channel.

For the experiment, no data preprocessing is conducted. The cosine distance is used as the distance input for all methods except Euclidean k-means. Again the parameters of baselines, including the bandwidths of SC and Ncut are tuned to output the best performance. We fix the sample rate of OCTD (Min) to 0.06 and OCTD (Mean) to 0.2, while the random sampling numbers of both methods are set to 2000.

Table 3 shows the results of comparing methods on speech datasets. One could see that overall OCTD (Min) works best and both OCTD (Min) and OCTD (Mean) show significant gains over TD and other baselines.

Conclusion

In this paper, we propose the concept of using order-constrained transitive distance for data clustering and an efficient approximation framework with random samplings to extract it. The proposed method shows many nice theoretical properties, while demonstrating very promising practical performance in a comprehensive set of experiments.

References

Bezdek, J. C. 2013. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Chang, H., and Yeung, D.-Y. 2005. Robust path-based spectral clustering with application to image segmentation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, 278–285. IEEE.

Chang, H., and Yeung, D.-Y. 2008. Robust path-based spectral clustering. *Pattern Recognition* 41(1):191–203.

Cheng, Y. 1995. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 17(8):790–799.

Comaniciu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(5):603–619.

Defays, D. 1977. An efficient algorithm for a complete link method. *The Computer Journal* 20(4):364–366.

Ding, C.; He, X.; Xiong, H.; and Peng, H. 2006. Transitive closure and metric inequality of weighted graphs: detecting protein interaction modules using cliques. *International journal of data mining and bioinformatics* 1(2):162–177.

Drineas, P., and Mahoney, M. W. 2005. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research* 6:2153–2175.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2790–2797. IEEE.

Feng, J.; Lin, Z.; Xu, H.; and Yan, S. 2014. Robust subspace segmentation with block-diagonal prior. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3818–3825. IEEE.

Fischer, B., and Buhmann, J. M. 2003a. Bagging for path-based clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25(11):1411–1415.

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

Fischer, B.; Roth, V.; and Buhmann, J. M. 2004. Clustering with the connectivity kernel. *Advances in Neural Information Processing Systems* 16:89–96.

Greenberg, C. S.; Bansé, D.; Doddington, G. R.; Garcia-Romero, D.; Godfrey, J. J.; Kinnunen, T.; Martin, A. F.; McCree, A.; Przybocki, M.; and Reynolds, D. A. 2014. The nist 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*.

Hu, H.; Lin, Z.; Feng, J.; and Zhou, J. 2014. Smooth representation clustering. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3834–3841. IEEE.

Li, M., and Narayanan, S. 2014. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification. *Computer Speech & Language* 28(4):940–958.

Li, B.; Zhang, Y.; Lin, Z.; Lu, H.; and Center, C. M. I. 2015. Subspace clustering by mixture of gaussian regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2094–2102.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(1):171–184.

Lloyd, S. P. 1982. Least squares quantization in pcm. *Information Theory, IEEE Transactions on* 28(2):129–137.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012*. Springer. 347–360.

Martınez, A., and Benavente, R. 1998. The ar face database. *Rapport technique* 24.

Ng, A. Y.; Jordan, M. I.; Weiss, Y.; et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2:849–856.

Peng, X.; Yi, Z.; and Tang, H. 2015. Robust subspace clustering via thresholding ridge regression. In *AAAI Conference on Artificial Intelligence (AAAI)*, 3827–3833. AAAI.

Peng, X.; Zhang, L.; and Yi, Z. 2013. Scalable sparse subspace clustering. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 430–437. IEEE.

Roth, V.; Laub, J.; Kawanabe, M.; and Buhmann, J. M. 2003. Optimal cluster preserving embedding of nonmetric proximity data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25(12):1540–1551.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8):888–905.

Sibson, R. 1973. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal* 16(1):30–34.

Williams, C., and Seeger, M. 2001. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, 682–688.

Yu, Z.; Xu, C.; Meng, D.; Hui, Z.; Xiao, F.; Liu, W.; and Liu, J. 2014. Transitive distance clustering with k-means duality. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 987–994. IEEE.

Yu, Z.; Liu, W.; Liu, W.; Peng, X.; Hui, Z.; and Kumar, B. V. K. V. 2015. Generalized transitive distance with minimum spanning random forest. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2205–2211. AAAI Press.

Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *Advances in neural information processing systems*, 1601–1608.