Clustering on Manifolds with Dual-Rooted Minimal Spanning Trees

Laurent Galluccio, Olivier J.J. Michel, Pierre Comon

CLUSTERING ON MANIFOLDS WITH DUAL-ROOTED MINIMAL SPANNING TREES

L. Galluccio [1], *O. Michel* [2], *P. Comon* [1]

[1] I3S Laboratory, CNRS - University of Nice Sophia Antipolis
2000, route des Lucioles Les Algorithmes - BP.121- 06903 Sophia Antipolis Cedex - France
[2] Gipsa - Lab - UMR 5216 - Grenoble INP
961 rue de la Houille Blanche - BP 46 - 38402 Saint Martin d'Hères Cedex - France
Phone: +33 (0) 4.92.94.27.92 - e-mail: {gallucci, pcomon}@i3s.unice.fr, olivier.michel@gipsa-lab.inpg.fr

ABSTRACT

In this paper, we introduce a new distance computed from the construction of dual-rooted minimal spanning trees (MSTs). This distance extends Grikschat's approach [7], exhibits attractive properties and allows to account for both local and global neighborhood information. Furthermore, a function measuring the probability that a point belongs to a detected class is proposed. Some connections with diffusion maps [8] are outlined. The dual-rooted tree-based distance (DRPT) allows us to construct a new affinity matrix for use in a spectral clustering algorithm, or leads to a new data analysis method. Results are presented on benchmark datasets.

1. INTRODUCTION

Data clustering is the task of partitioning a set of data into non-overlapping subsets, without using any prior knowledge, such that patterns belonging to a same cluster share more similarity with each other than with patterns belonging to different clusters [15]. Such problems are commonly encountered in statistics, data mining, pattern recognition, image segmentation and bio-informatics [17]. Although many strides were achieved in this area, there remains many open issues. Hierarchical clustering, graph partitioning algorithms and k-means [10] for instance are among the most populars ones (see e.g. [17, 15] for a more exhaustive state of art). More recently, a new class of clustering methods based on some graph theory notions has emerged: the spectral clustering algorithms [11]. As in other methods, little success is found if clusters do not form convex subsets or are not well separated or even overlapping. Furthermore the presence of noise or outliers leads to dramatically decreased performances in general. Our methods exhibit improved performances in this context.

A crucial issue in clustering problems concerns the choice of an affinity measure between data points. We will restrict the scope of this paper to the case where data points are made of numerical features. Many situations cannot be efficiently addressed by methods using Euclidean distances to measure similarities between data points. Consider for instance an Euclidean space and two imbricated non convex clusters. Two points from cluster 1 may be more separated from each other than e.g. 2 points from the neighboring borders of cluster 1 and cluster 2 respectively. In such a case, no linear form will correctly classify the data from the set of pairwise distances. This makes the motivation for introducing more geometrically descriptive similarity measures. In their seminal work, Grikschat et al. [7] proposed a method inspired by some recent research on diffusion graphs [8], establishing connections between diffusion process on manifolds and random walks on finite data sets. Grikschat's method is based on symmetrically growing MSTs rooted at each pairs of points, by Prim's algorithm [12]; the hitting time of the two MSTs measures the affinity between points[1]. Dual rooted trees hitting time allows to describe global as well as local geometrical properties of the data set. In this paper, we introduce a slight modification of Grikschat's method, that confers new appealing properties. The new proposed distance is applied for both clustering and data analysis tasks. Additionally, a probability estimate that a point belongs to the different clusters is inferred from the proposed distance.

In Section 2.1, MST definitions and Prim's construction algorithm are briefly sketched. Dual-rooted MST (drMST) principles and drMST based distance and its properties are introduced in Section 2.2. Applications in the framework of clustering is presented in section 2.3; relation to existing method is proposed in Section 2.4, and some data exploratory application is presented in section 2.5. Section 3 presents results on both synthetic and real datasets.

2. METHODOLOGY

2.1 MST and Prim's algorithm

Let $V = \{v_1, v_2, \ldots, v_n\}$ denote a sample of data points in $\mathbb{R}^l$ having unknown Lebesgue multivariate density λ. The goal is to partition V into K clusters. Let $P = \{C_1, \ldots, C_K\}$ stand for a set of clusters.

Let $G = (V, E)$ be an undirected graph where $E = (e_{ij} : e(v_i, v_j), (i, j) \in (1, \ldots, N))$ denotes a set of undirected edges between vertices of V. The weight w_{ij} of an edge measures the dissimilarity between two vertices v_i and v_j.

A spanning tree $\mathscr{T}$ through the set of vertices V is a connected acyclic graph which passes through all the N vertices v_i, $i \in \{1, \ldots, N\}$ in the set. The minimal spanning tree (MST) is the tree which has the minimal weight

$$L_{N,\gamma}(V) = \min_{\mathscr{T}} \sum_{e \in \mathscr{T}} w_i j$$

A common choice for w_{ij} is $w_{ij} = |e|^\gamma$, $\gamma \in (0, l)$, where e is the Euclidean distance between vertices. The tree of minimal

[1] Hitting time is defined there as the number of iterations until the two subtrees collide.

power weighted length enjoys many interesting properties (see e.g. [5]). However, in this paper the only assumptions made for the weight $w_i j$ are $w_{ii} = 0$ and $w_{ij} = w_{ji}$. We apply the Prim's algorithm [12], whose complexity is $O(N \log(N))$. Prim's algorithm is a greedy procedure for growing trees by recursively connecting a new vertex to the existing subtree. At each iteration, the new vertex among the unconnected vertices is chosen, such that the edge which connects the new vertex to the subtree has a minimal weight. The procedure is iterated until no unconnected vertex remains. The resulted tree is unique[2], i.e., independent of the initial vertex of the graph, acyclic (no loop) and of minimal weight.

2.2 Dual Rooted Prim Tree

In [7], Grikschat et al. propose a graph-based distance measure between two vertices v_i and v_j to be the hitting-time of the two Prim subtrees simultaneously grown, rooted at v_i and v_j. A slight modification is proposed here consisting in competitive growing : at each step of the tree growing procedure, only one of the two Prim subtrees is grown, namely the one for which the new edge has minimal weight. As in [7], this process continues until the two subtrees collide. However, the number of vertices connected within each subtree are no longer identical. Let N_{iter} denote the hitting time of the subtrees.

The tree obtained by the union of the two Prim subtrees is referred to as **D**ual **Roo**ted **P**rim **T**ree (DRPT) (Fig. 1). The DRPT rooted in $v_i \in V$ and in $v_j \in V$ will be noted $DR(v_i, v_j)$.

Different distances measures $d(v_i, v_j)$ can be computed based on $DR(v_i, v_j)$:
• the hitting time of the two sub-MSTs

$$d_{iter}(v_i, v_j) \quad = \quad N_{iter}, \tag{1}$$

• the length of the final tree constructed

$$d_{leng}(v_i, v_j) \quad = \quad \sum_{iter=1}^{N_{iter}} w_{iter}, \tag{2}$$

• and the weight of the final edge connected

$$d_{max}(v_i, v_j) \quad = \quad \max_{iter \in [1, N_{iter}]} w_{iter}. \tag{3}$$

All these distances measures (1, 2, 3) enjoy the properties of being metrics in the mathematical sense [3].

This DRPT (Fig. 1) enjoys many interesting properties, some of which are used in the rest of the paper.

Property 2.1 *For a given couple of vertices $\{v_1, v_2\}$ serving as roots of two subtrees $\mathcal{T}_1$ and $\mathcal{T}_2$, the last constructed edge, which connects the two subtrees together, of weight noted w_{last} is always the largest (with maximum weight) among the set of all edges from both subtrees.*

Property 2.2 *Let $d(v_1, v_2) = w_{last}$ the weight of the largest edge among all the edges involved on the subtrees rooted at vertices v_1 and v_2, d is a distance.*

[2]The symmetry property $w_{ij} = w_{ji}$ insures unicity of the resulting graph, assuming furthermore that there is no ties in the similarity matrix.

[3]They are symmetric, positive, and satisfy the triangular inequality; proofs are developed with many details in [6].

Property 2.3 *The union of the subtrees $\mathcal{T}_1$ and $\mathcal{T}_2$ rooted at v_1 and v_2 respectively is the MST for the subset of vertices involved in one or the other subtree. This property is rather straightforward to prove, as a MST is unique and does not depend upon the root used for initializing Prim's algorithm.*

Property 2.4 *Property 2.1 above insures that any Prim's algorithm rooted at a vertex from $\mathcal{T}_1 \cup \mathcal{T}_2$ will connect all vertices of $\mathcal{T}_1 \cup \mathcal{T}_2$ before connecting a vertex outside $\mathcal{T}_1 \cup \mathcal{T}_2$. Then, by using property 2.2 above, it can therefore be concluded that*

$$\forall v_i \in \mathcal{T}_1, \forall v_j \in \mathcal{T}_2, d(v_i, v_j) = d(v_1, v_2)$$

and

$$\forall (v_i, v_j) \in [\mathcal{T}_1 \times \mathcal{T}_1] \cup [\mathcal{T}_2 \times \mathcal{T}_2], d(v_i, v_j) \leq d(v_1, v_2)$$

Property 2.5 *Let $\mathcal{R}_{v_2}^{v_1}$ stands for the relation, defined relatively to v_1 and v_2 by $v_i \mathcal{R}_{v_2}^{v_1} v_j$ if $d(v_i, v_j) \leq d(v_1, v_2)$.
$\mathcal{R}_{v_2}^{v_1}$ is trivially symmetric and reflexive. Transitivity of $\mathcal{R}_{v_2}^{v_1}$ is easily obtained as a consequence of properties 2.2 and 2.4. Therefore, $\mathcal{R}_{v_2}^{v_1}$ is an equivalence relation and the obtained clusters are equivalence classes wrt $\mathcal{R}_{v_2}^{v_1}$*

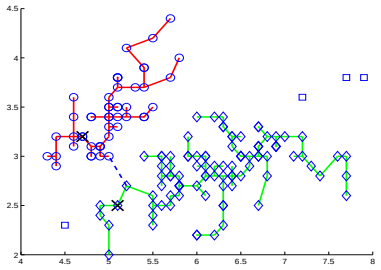


Figure 1: Dual rooted Prim tree built on a data set. Symbol X marks the rooted vertices. The dashed edge is the last connected edge.

It must be pointed out that two 'distances' are involved in the dual-rooted tree approaches : the first one is related to the weight w_{ij}, as introduced in section 2.1. The second is indexed on the MST grown on the vertex set from the knowledge of all w_{ij}.

When a new vertex is added in the process of growing trees, it is associated to an edge of minimal weight : this deals with local properties (neighborhood related) of the vertex set. Although Euclidean distances are commonly used for the w_{ij}, other dissimilarity measures may better fit the nature of the data at hand (e.g. information divergences if the data are spectra as presented later). Whatever the chosen function w_{ij}, its properties are encompassed in the construction of the tree, the DRPT distance properties 2.1 to 2.5 are preserved. More specifically, it is important to emphasize that the DRPT distance is a metric, whereas w_{ij} may be a semi-metric only. DRPT distances account for more 'global' features of the set V, as described e.g. by property 2.4.

2.3 Dual rooted trees-based distances for clustering

There exists a lot of clustering methods developed to partition a set of data, as mentioned in the introduction. Recently, spectral graph clustering algorithms [4] have received a lot of interests because of their properties and the quality of the results obtained [11, 2]. Basically, the algorithm starts with a neighborhood graph built on the dataset (either KNN-graph, ε-graph or even fully connected graph) and a distance matrix d ($d_{i,j} = d(v_i, v_j)$) is computed. This distance matrix is used to derive an affinity matrix commonly defined as:

$$A_{ij} = \exp\left(\frac{-d_{i,j}}{\sigma}\right).$$

The eigendecomposition of the normalized Laplacian (L) of the graph is realized:

$$D = diag\left(\sum A_{ij}\right), L = D^{-1/2} A D^{-1/2}.$$

A K-means algorithm is finally applied on the eigenvectors corresponding to the k largest eigenvalues, to exhibit the candidate clusters.

The usual distance measure $d_{i,j}$ used in the expression of A is the Euclidean distance. In [7], the authors proposed to use instead their graph-based distance. The obtained results overcome those obtained with the Euclidean distance, especially when the classes have non convex shapes. Following [7], we use DRPT distance together with spectral clustering algorithms to exhibit clusters.

Parameter σ in the affinity matrix determines the horizon above which two vertices are considered to be extremely distant from each other and cannot belong to a common cluster. Although this parameter drastically influences the quality of the results, there is no broadly adopted strategy to determine its value [9] . In [7], the authors choose the maximum distance in d. In order to be more robust to the outliers, Schclar [13] proposed two heuristics for choosing σ: the median heuristic (median of d) and the max-min heuristic ($\max_i \min_j d_{ij}$). All these heuristics allow to define a global parameter. Based on this observation, Zelnik-Manor and Perona [18] have proposed to consider a local σ in the computation of the affinity matrix. The choice of σ depends on the neighborhood of each vertex: $\sigma_i = d(v_i, v_K)$, where v_k is the K th nearest neighbor of v_i. Therefore, the affinity matrix is changed into this new expression: $A_{ij} = \exp\left(\frac{-d(v_i,v_j)}{\sigma_i \sigma_j}\right)$. The main drawback of this approach is its sensitivity to the number K of neighbors, for which no heuristic exists.

In Fig.2, the Jaccard [17]. index is computed on the results obtained by applying the spectral clustering algorithm with the Euclidean distance on the Wine data set [1] for various values of σ. Note that σ (horizontal axis) is normalized by the median of the distance distribution, in order to insure independence of the results with respect to affine transform of the data. Then σ corresponds to the percentage of the median distance of d. This plot highlights the importance of the parameter σ in the clustering result.

2.4 Relation to Diffusion Maps, probability of membership

In many applications, data clusters may overlap each other and/or exhibit complicated non convex shapes. In such situation, "hard labeling" turns out to be not satisfactory enough. A crucial issue is then to introduce the probability that a given data point is a member of a detected cluster or of another. In this section, we introduce such a probability of membership, and a close relation to transition matrices introduced for diffusion maps [8] is presented.

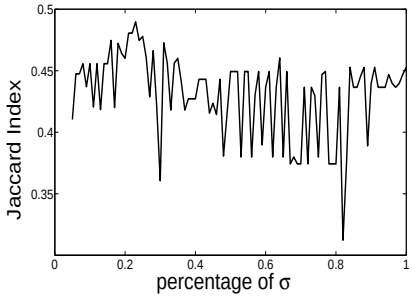


Figure 2: Jaccard index computed on the results obtained by applying the spectral clustering algorithm with the Euclidean distance on the Wine data set with σ varying.

For each vertex v, it is proposed to compute the probability of being a member of the cluster C_i as follows (4):

$$Proba(v \in C_i) = \frac{\sum_{v_i \in C_i} h\left(d(v, v_i)\right)}{\sum_{v \in \mathcal{V}} h\left(d(v, v_i)\right)}, \tag{4}$$

where h may be any integrable decreasing function of the distance measure $d(v, v_i)$.

A popular choice for h is the exponential function:

$$h(d(v, v_i)) = \exp\left(\frac{-d^2(v, v_i)}{\varepsilon}\right), \tag{5}$$

where ε stands for the characteristic decay length. Note that a discussion for choosing ε would use similar arguments as those developed for discussing σ in the previous section.

Euclidean distance is often chosen for d but the algorithm fails to correctly cluster V when the classes are either non-convex or lie on some non-linear manifold; it is proposed here to substitute DRPT distance to d. Actually DRPT properties allow to deal with non-convex clusters by following the shape of the clusters on the manifold (see [5]) and to account for both local and global feature of the data space, as explained previously (see Fig. 3). Let us emphasize that replacing d by the DRPT distance is made possible, as the latter is actually a metric. This could not make sense for Grikschat's distances for instance, as it is not a metric.

It is worth noticing that the expression of the probability measure (4) is similar to the expression of the probabilities entering the transition probability matrix of Lafon et al. [8] for constructing diffusion maps. The probability of diffusion from vertex i to vertex j is actually defined given by

$$M(v_i, v_j) = \frac{\exp\left(\frac{\|v_i - v_j\|^2}{\varepsilon}\right)}{\sum \exp\left(\frac{\|v_i - v_j\|^2}{\varepsilon}\right)}.$$

The diffusion map is given by the eigenelements of M, and clusters are issued by applying a simple (e.g. K-means) algorithm on the obtained map.

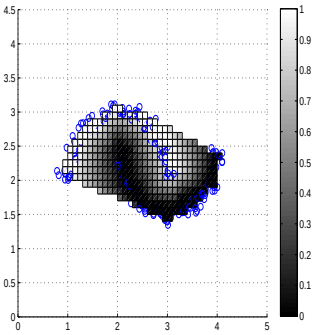


Figure 3: Probability membership map of the upper half-moon data set with the use of the DRPT distances.

2.5 Exploratory analysis of DRPT clustering

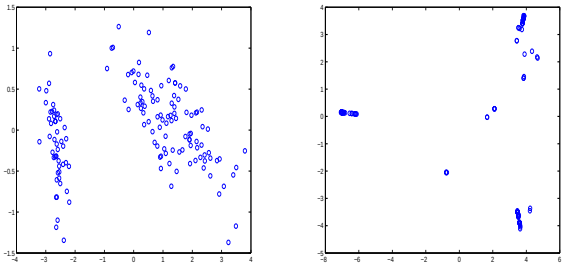
Several methods can be used to obtain an embedding of the data set into a low dimensional-space where the data can be easily explored. A popular method to achieve such representation consists in projecting the data onto a low dimensional Euclidean space, under the constraint that the Euclidean distances on the image space are as close as possible to the distances in the high dimensional original data space. This is the strategy adopted in the Multi Dimensional Scaling algorithm (MDS) [16] or Isometric mapping (Isomap) [14]. Note that Laplacian eigenmaps introduced by Belkin et al. [2] also provide a solution to this problem, exploited in spectral clustering algorithms. This section is focused on applying MDS to the DRPT distance matrix introduced previously.

MDS may be summarized by the following steps:

- First compute J: $J = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^t$. J is referred to as the double centering matrix.
- Normalize the row and column of d: introduce $L'_{ij} = -\frac{1}{2}Jd_{ij}J$.
- Compute the eigen-decomposition of L' and keep the k' largest eigenvalues λ_j and their corresponding eigenvectors μ_j.
- The new set of coordinates is given by computing $\sqrt{\lambda_j}\mu_j$.

The Iris data set consists in 150 points in 4-dimensions containing three clusters (one of which is well separated from the others and the two others exhibit interleave). Figure 4 shows this set embedded in a 2-dimensional space computed by MDS with the Euclidean distances (a) and with the DRPT distances (b). This clearly emphasizes the ability of DRPT distance to 'concentrate' the image vertices on the low dimensional space into three well separated clusters. No theoretical details will be given here, but this appears clearly as being a consequence of property 2.4 above. By applying a basic K-means algorithm in the low dimensional Euclidean space represented on Fig.4(b), a correct labeling score of 146/150 was obtained (136/150 for classical unsupervised clustering algorithms).

This simple experimentation allows us to attest the major importance of the distance measures computed on the data poin wts. The use of the dual-rooted trees-based distances better discriminates the data points into relevant clusters.



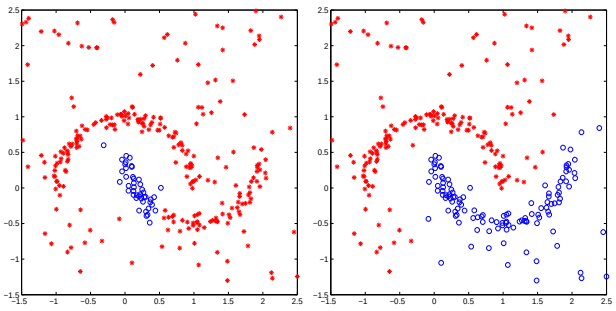
(a) (b)

Figure 4: Embedded Iris data set with MDS (a) Euclidean distance, (b) DRPT distances.

3. EXPERIMENTAL RESULTS

The performances of the proposed methods are illustrated on various data sets. The main features of the algorithm are tested on simulation data. The quality of the results are evaluated by computing the Jaccard index Let P^* be some known ground truth reference partition of the data and let P be the obtained partition. The Jaccard (J) index between P and P^* measures the similarity between the partitions. It is expressed as $J(P,P^*) = \frac{a}{a+b+c}$, where a is the number of pairs of points in V belonging to a same set in P and a same set in P^*, b is the number of pairs of points in V belonging to a same set in P and different sets in P^* and c is the number of pairs of points in V belonging to different sets in P and a same set in P^*. $J(P^*,P) = 1$ indicates a perfect match of the partitions.

Simulated Data Sets: We consider the classical 'two moons' problem with ouliers. Spectral clustering method (with Euclidean distance) with a local scaling of σ succeeds in recovering the classes in the absence of outliers, but fails when outliers are present (V counts 150 data points and 100 outliers). Replacing the Euclidean distance by the DRPT distance leads much better results, as shown on figure 5 . For both cases, σ was chosen according to Zelnik-Malnor and Perona method. The performance of DRPT based approach comes from its ability to convey information from both local and global features of the analyzed set V.



(a) (b)

Figure 5: Two Moons perturbated by a random noise: Spectral clustering with (a) Euclidean distance, (b) DRPT distances.

Real data: The Iris and Wine data sets from the UCI machine learning repository [1] are used for benchmarking the proposed approach. Firstly, spectral clustering algorithm are applied to detect clusters, with Euclidean distances and DRPT based distances (with Euclidean weight used in the Prim's growing algorithm). The number of clusters is known a priori. As Wine dat set is made of a set of proportions of chemical elements, it behaves like a spectrum. Following [3], we propose to use a symmetrized Kullback information divergence (D_{kls}) for the weight function w_{ij}; this choice for w_{ij} leads to improved results as w_{ij} is better adapted to the nature of the data, although it is not a metric.

Secondly, we applied MDS algorithm to embed the data into a 2-dimensional Euclidean space where K-means can be used. The inter vertex distance matrices computed in their original (high dimensional) space are either using Euclidean metric or DRPT based distances.

Again, the proposed graph-based distances allow improved performances, especially in the case (Wine) where the weight w function is adapted to the data characteristics, and despite it is not a metric.

Table 1: Results obtained in terms of Jaccard Index for various datasets.

Methods	Iris	Wine
Spectral Clustering (Euclidean)	0.7445	0.4397
Spectral Clustering (d_{iter})	**0.8876**	**0.6627**
Spectral Clustering (d_{leng})	**0.8876**	0.4276
Spectral Clustering (d_{max})	0.5000	0.4276
Spectral Clustering (Grikschat [7])	**0.8876**	0.4499
MDS (Euclidean) + Kmeans	0.7016	0.4199
MDS (DRPT) + Kmeans	**0.8876**	0.5338

Remark : This choice to embed the data in a 2 dimensional space is not motivated by some theoretical properties but was set for sake of visualization. The determination of the optimal embedding dimension is not addressed in the present paper.

4. CONCLUSION

In this paper, we have presented some dual-rooted diffusion distances (DRPT) computed from the construction of dual-rooted MSTs. These distances exhibit appealing properties and allow to account for both local and global properties of the set to be clustered. As the new proposed distance is a metric, it allows us to introduce a function that measures the probability of a point to belong to the different classes, that brings some connections with diffusion maps. It allows furthermore to use non metric distance measures for growing trees on which the DRPT is based, which may leads to improved clustering performances in some cases ('spectrum-like' data). The usefulness of the new proposed distance is illustrated through some spectral clustering applications, and for some data exploratory analysis.

REFERENCES

[1] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances on 15th Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2001.

[3] C. I Chang. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Transactions on information theory*, 46(5):1927–1932, August 2000.

[4] F. R. K. Chung. *Spectral Graph Theory*. Number 92 in Conference Board on the Mathematical Sciences. American Mathematical Society, 1997.

[5] J. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, Aug 2004.

[6] L. Galluccio, O. Michel, P. Comon, M. Kliger, and A. O. Hero. Dual rooted trees based clustering. Technical report, Laboratoire I3S, CNRS-Université de Nice-Sophia Antipolis, 2010.

[7] S. Grikschat, J. A. Costa, A. O. Hero, and O. Michel. Dual rooted-diffusions for clustering and classification on manifolds. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[8] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1784–1797, 2006.

[9] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[10] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–287, Berkeley, 1967.

[11] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances on 15th Annual Conference on Neural Information Processing Systems*, volume 14, Vancouver, British Columbia, Canada, 2001.

[12] R. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36:1389–1401, 1957.

[13] A. Schclar. A diffusion framework for dimensionality reduction. In *Soft Computing for Knowledge Discovery and Data Mining*, pages 315–325. Springer, 2008.

[14] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[15] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.

[16] W. S. Torgerson. *Theory and methods of scaling.* Wiley, 1958.

[17] R. Xu and D. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, May 2005.

[18] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2005.