16th European Signal Processing Conference (Eusipco'08), Lausanne, August~25-29 2008

UNSUPERVISED CLUSTERING ON MULTI-COMPONENT DATASETS: APPLICATIONS ON IMAGES AND ASTROPHYSICS DATA

L. Galluccio [1,2]*, O. Michel* [2]*, and 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] Hippolyte FIZEAU Laboratory, CNRS - University of Nice Sophia Antipolis
Faculté des Sciences - Parc Valrose, 06108 Nice Cedex 2 - France
Phone: +33 (0) 4.92.94.27.92 - Fax: +33 (0) 4.92.94.28.96 - e-mail: {gallucci, pcomon}@i3s.unice.fr, olivier.michel@unice.fr

ABSTRACT

This paper proposes an original approach to cluster multi-component data sets with an estimation of the number of clusters. From the construction of a minimal spanning tree with Prim's algorithm and the assumption that the vertices are approximately distributed according to a Poisson distribution, the number of clusters is estimated by thresholding the Prim's trajectory. The corresponding cluster centroids are then computed in order to initialize the Generalized Lloyd's algorithm, also known as K-means, which allows to circumvent initialization problems. Metrics used for measuring similarity between multi-dimensional data points are based on symmetrical divergences. The use of these informational divergences together with the proposed method lead to better results than some other clustering methods in the framework of astrophysical data processing. An application of this method in the multi-spectral imagery domain with a satellite view of Paris is also presented.

1. INTRODUCTION

Consider a set V of N data points in $\mathbb{R}^L$, which we wish to partition into K classes, without prior information (*i.e.* this is an unsupervised classification).

Each data point (actually a vector in $\mathbb{R}^L$) can be considered as a vertex in a graph. Our study restricts to acyclic totally connected graphs of minimal length : the minimal spanning tree (MST). An easy way to segment the graph is to realize some cuts in the connection graph. The segmentation of the graph in K clusters will lead to remove the $K-1$ largest connections. This method, based on single linkage clustering, is known to be unstable, mainly when the data contain outliers, or possibly when they are corrupted by noise.

In this communication, we propose to improve the popular generalized Lloyd's algorithm (referred to as K-means) by an automatic initialization of both the number of clusters and corresponding centroids. Various methods have been proposed to estimate the number of clusters present in a dataset, *e.g.* using statistical criteria like AIC, BIC, MDL, Tibshirani's Gap, or indices such as Calinski & Harabasz's index.Because of space limitation, these criteria will be discussed in a full-length version of the paper. Actually, our approach relies upon Prim's algorithm for constructing MST's. It is proposed to record each iteration characteristics (namely which vertex is connected, and what is the length of the new edge), in order to get an *one-dimensional* unfolded representation of the underlying data probability density function. The method will be developed in Section 2.

Furthermore, we address the problem of designing a measure of similarity between two data points. Instead of using the popular Euclidean distance, we propose to define the similarity as a measure of spectral variability between two probability density functions, by the use of informational divergences. New metrics and motivation for resorting to information based similarity measures will be investigated in Section 2. In Section 4, we present some unsupervised clustering results obtained in the frame of two applications : the taxonomic classification of asteroids, and the classification of objects in a multi-spectral satellite image.

2. AUTOMATIC INITIALIZATION OF POPULAR CLUSTERING METHODS

2.1 Minimum Spanning Tree and Prim's Trajectory

Let $G=(V,E)$ be an undirected graph where V is the set of N vertices and E denotes the set of edges. The length of an edge measures the similarity between two vertices, and depends on the choice of the metric. The graphs considered herein are *trees*, that is, they are connected (*i.e.* every vertex is connected to at least one other vertex) and acyclic (*i.e.* there is no loop).

A spanning tree of G is a tree T passing through every vertex of G. The power-weighted length of the tree T is the sum of all edge lengths raised to a power $\gamma \in (0,L)$, denoted by : $\sum_{e\in T}|e|^\gamma$. The minimal spanning tree (MST) is the tree which has the minimal length over all spanning trees

$$\mathscr{L}(V)=\min_T \sum_{e\in T}|e|^\gamma$$

Among algorithms allowing to build a MST, one of the most popular is Prim's algorithm [10], the complexity of which is $O(N\log N)$. The Prim's algorithm connects to the partially connected graph at iteration i the closest non connected vertex (in the sense of a given chosen metric). The graph which is determined is acyclic, unique (*i.e.* independent of the initial point of the construction of the graph) and of minimal length.

Denote $g(i)=|e_i|$, the length of a new edge built at iteration i ; $[g(i), i=1\dots N]$ is referred to as *Prim's trajectory*. Function g allows us to "unfold" the probability density function of points in L dimensions into a *one-dimensional* function. The latter exhibits some valleys in the neighborhoods of high density : a set of close points is indeed connected thanks to a sequence of successive iterations yielding short segments (Fig. 1). The detection of valleys in the curve hence corres-

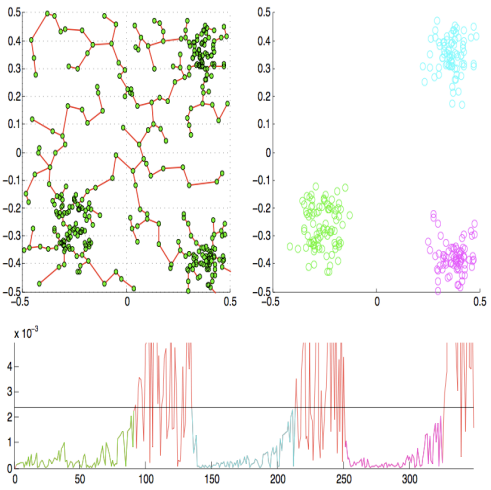

FIG. 1 – Typical example : (top left) construction of a MST, (bottom) Prim's trajectory and threshold, (top right) extraction of the clusters

ponds to the detection of clusters. It allows to identify the main modes of the probability density function, each mode being associated with a cluster.

The choice of the threshold to be applied to $g(i)$ is an important parameter, since it determines the main modes of the probability density function. In [8], Michel *et al.* have proposed to give the same cluster label to all vertices connected sequentially, as long as the function $g(i)$ stays below the threshold. An empirical solution is proposed in [8] for setting the threshold. In this paper, the number of modes and associated centroids are required in our problem, in order to initialize a classical algorithm (*e.g.* K-means). One possibility is to arbitrarily set the threshold to the standard deviation of the connections $\varepsilon = std\{e_i\}$. Because the detection of modes (or high density regions) is solely based upon finding the centroid of vertices gathered into a valley of Prim's trajectory, the estimation of the threshold does not need to be very precise. For the sake of simplicity, the threshold has been chosen to be constant over the entire Prim's trajectory, but it could vary from one cluster to another.

The question remains to determine the critical number of points which could be considered as a cluster. Actually for a given realization, some vertices can be gathered onto a small neighborhood, even in the case where the theoretical density does not exhibit any local maximum. Alternatively, small clusters may correspond to noise effects and thus may not be relevant. Thus, the number of modes tends to be overestimated.

We propose to estimate automatically the minimum number of points above which one decides that a cluster is detected as a function of the threshold applied on the Prim's trajectory. This estimation is realized in the framework of a Neyman-Pearson approach. For each new connected vertex, a binary hypothesis test is performed. Under the null hypothesis, last connected vertices do not belong to some 'mode' or cluster, and they are spread all over the neighborhood according to a Poisson distribution (see below). Under the alternative, the length of the edge that connects the new vertex is too small to match with the null hypothesis.

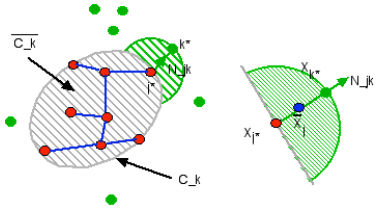

FIG. 2 – Left : C_k denotes the contour of the support where connected vertices are found. The neighborhood which is considered for finding a vertex that could be connected to v_i is shown. In the limit of large N, this latter neighborhood is the half sphere laying on the tangent (hyper)plane to C_k (right).

Let v_i be the vertex connected at iteration i. Consider an L-dimensional space and a neighborhood of v_i hereafter noted $B_{(v_i,\varepsilon)}$,with characteristic length ε. We suppose that the vertices are approximately distributed according to a Poisson distribution, with rate $\lambda\varepsilon^L$. This is justified, since the latter is the limiting distribution of the binomial. Both λ and $B_{(v_i,\varepsilon)}$ will be identified later.

We assume that under the null hypothesis H_0, the density does not exhibit any mode : the process is homogeneous over its entire support $\mathscr{V}$. Hence we can assess that λ is constant over $\mathscr{V}$. The probability that at least one vertex is found in the neighborhood $B_{(v_i,\varepsilon)}$ is given by

$$F_{v_i}(\varepsilon) = 1 - \mathrm{e}^{-\lambda\varepsilon^L}$$

In the context of the Prim construction of the MST, $F_{v_i}(\varepsilon)$ can also be considered as the probability to construct an edge of length less than ε when connecting a new vertex to v_i. Considering the asymptotic case where N is large, the neighborhood which must be considered here is the half sphere of radius ε, as illustrated on the figure 2.

Let $P_{k,\varepsilon}$ be the probability that exactly k vertices are connected with edge lengths less than ε but the next edge built is larger than ε :

$$P_{k,\varepsilon} = \left(1 - \mathrm{e}^{-\lambda\varepsilon^L}\right)^k \mathrm{e}^{-\lambda\varepsilon^L}$$

Suppose that at least k_0 successive connections of length less than the threshold value ε are required for considering that a cluster is detected. Under H_0, false alarm in the mode detection will arise for any occurrence of more than k_0 successive connections of length less than ε. Therefore, the expression of the false alarm probability is given by

$$P_{FA}(k_0,\varepsilon) = \sum_{k\geq k_0}^{\infty} P_{k,\varepsilon} = \left(1 - \mathrm{e}^{-\lambda\varepsilon^L}\right)^{k_0} \qquad (1)$$

In the case where L-dimensional Euclidean spaces are considered, the volume of the half sphere of radius ε is $B_L(\varepsilon) = \frac{1}{2}C_L\varepsilon^L$, where C_L stands for the volume of the unit ball in dimension L : $C_L = \frac{2\pi^{L/2}}{L\Gamma(L/2)}$.

In this framework, under H_0, consider the radius ε_0 of a sphere covering the set of all vertices ; λ is identified by the set of equations

$$\left\{ \begin{array}{ll} \mathscr{V} & = C_L\varepsilon_0^L \\ \lambda\varepsilon_0^L & = N \end{array} \right. \quad \Longrightarrow \quad \lambda = C_L\frac{N}{\mathscr{V}} \qquad (2)$$

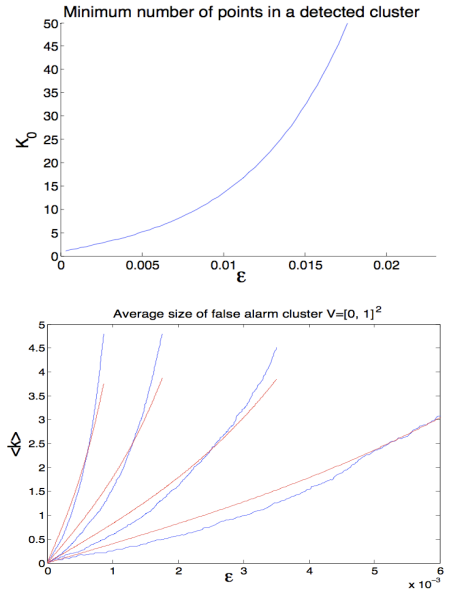

FIG. 3 – Upper plot : $k_0 = f(\varepsilon)$ for $P_{FA} = 0.05$; Lower plot : Average size of false alarm detected cluster from sets of uniformly distributed vertices over $[0,1]^2$, and $N = 128, 256, 512, 1024$ vertices respectively. Theoretical curve (red) and numerical simulation (blue).

Finally, we get from equations (1) and (2)

$$P_{FA}(k_0, \varepsilon) = \left(1 - \mathrm{e}^{-C_L \varepsilon^L \frac{N}{\mathcal{V}}}\right)^{k_0}$$

This formula allows to determine the relationship between k_0 (minimum number of vertices that form a cluster) and the threshold value ε in the framework of a Neyman-Pearson approach for cluster detection. Though these calculations were derived under some rough simplifying assumptions, the obtained results are close to what is obtained by numerical simulations, as illustrated in Figure 3[1].

Though the next sections will introduce the usefulness of informational divergences for the clustering problems, and consequently formulate the Prim's trajectory construction in some evidently non-Euclidean space, the previous results will still be applied as approximations. So far, we have no evidence but the quality of our results for accepting these approximations. This will be the matter of a future work.

2.2 Choice of metrics

As already pointed out, the distance measure between points plays a key role to characterize their similarity or dissimilarity. In this paper, the physical data that are considered are non-negative (they are homogeneous to a spectral measure), and may therefore be easily understood as behaving like probability densities, up to some scale factor.

Let $X = \{x_1, \ldots, x_L\}$ and $Y = \{y_1, \ldots, y_L\}$ two feature vectors, *e.g.* corresponding to a pixel in the imagery domain or to a reflectance spectrum in astrophysics. The most popular distance used to characterize similarities between two points is the Euclidean distance. Though this metric enjoys usefull properties (symmetry, non-negativity, triangular inequality), it turns out not to be the perfect measure of distance that one can think of to calculate similarities. This distance indeed has the following drawbacks : (i) it increases when the dimension of the data (*e.g.* number of wavelengths) increases ; (ii) it does not handle cases when spectra contain missing values at some wavelengths, (iii) it gives essentially a spatial distance, and does not take into account the positivity of data. For these reasons, following the works of Chang [4], we prefer information divergences as measures of similarity.

First, at a given wavelength λ_i, each data point is associated with a (positive) normalized quantity : $\tilde{x}_i = x_i / \sum_{j=1}^{L} x_j$. Let $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_L\}$ and $\tilde{Y}$ be defined accordingly. $\tilde{X}$ (resp. $\tilde{Y}$) can be interpreted as the probability distribution that a certain amount of information has been captured (measured) at the wavelength λ_i.

Next, our goal is now to measure the similarity between these two probability density functions. Using the popular symmetrized Kullback-Leibler divergence [5] leads to :

$$d_{KL}(X||Y) = \sum_{i=1}^{L} (\tilde{x}_i - \tilde{y}_i) \log \frac{\tilde{x}_i}{\tilde{y}_i} \tag{3}$$

It corresponds to the relative entropy of $\tilde{Y}$ with respect to $\tilde{X}$. This information divergence is widely used in Information Theory, and is strongly related to the Shannon theory (see *e.g.* [5]).

Alternatively, the symmetrized Rényi divergence of order α ($0 < \alpha < 1$) can similarly be used as a spectral measure :

$$d_\alpha(X||Y) = \frac{1}{\alpha - 1} \left(\log \sum_{i=1}^{L} \tilde{x}_i{}^{\alpha} \tilde{y}_i{}^{1-\alpha} + \log \sum_{i=1}^{L} \tilde{y}_i{}^{\alpha} \tilde{x}_i{}^{1-\alpha} \right) \tag{4}$$

Properties and advantages of the Rényi α-divergence have been detailed by Hero *et al.* [7]. Note that when α tends to 1, the α-divergence (4) converges to the Kullback-Leibler divergence (3). When $\alpha = 1/2$, the α-divergence matches with the Hellinger affinity $d_{1/2}(X\|Y)$ or Hellinger distance which is often used to assess how close a probability density is to a reference one. As the Bayes optimal exponential rate of decay of the decision error in a binary test (*i.e.*'is this spectrum almost identical to this other one ?') involves the α-divergence of order $1/2$ [7], only the value $\alpha = 1/2$ will be considered in the following.

These metrics have been used to determine similarity between points in order to form the MST. In the clustering part of our algorithm, we use the K-means algorithm (minimization of the squared error function between the points and the centroids). This classical partitioning algorithm can also be adapted to match the metric used in the construction of the tree [1]. Nevertheless, in our approach, we have only focused our interest on the initialization of K-means.

3. IMPLEMENTATION OF MST'S FOR LARGE SETS OF DATA

Suppose that the pairwise distance or affinity between vertices is known. The construction of an MST requires to

[1] The average size of a false alarm cluster can be shown (this will be published elsewhere) to be expressed as :

$$< k >= 2 \sinh \left(\frac{C_L}{2} \frac{N}{\mathcal{V}} \varepsilon^L \right)$$

sort the lengths of all possible edges. This operation requires $Comp.sort = O(N\log(N))$ logical operations (*e.g.* by using the "quick-sort" algorithm). Thus the overall computational cost is mostly due to the computation of all the distances. This becomes prohibitive for large datasets, since it would lead to a computational burden of $Comp.dist = O(LN^2/2)$ flops.

This is avoided by using a pre-conditionning data driven hierarchical classification tree, which could be learned from a small randomly chosen subset of R data[2]. The latter classification tree allows to identify neighborhoods of each vertex by comparing the vertex coordinates along each of the L dimensions to sets of thresholds. The size of the neighborhood can be set up in such a way that in the average, each neighborhood counts around M vertices. Consequently, the number of pairwise distances that need to be evaluated is of the order of $M^2/2$. This algorithm will be discussed in an extended version of this paper, and leads to a maximal computational burden expressed by $Comp.dist = O(NLM^2/2)$ flops and $Comp.sort = O(NM\log M)$. Assuming that M is set such that $M^2 \ll N$, the computational load is thus significantly lowered. In the next sections, construction of MST using this algorithm will be referred to as 'Nearest-Neighbor MST'.

4. RESULTS

The first application deals with the taxonomic classification of asteroids, by using reflectance measures at different wavelengths ; the second application concerns the segmentation of a multi-spectral image of Paris surroundings.

4.1 Astrophysical data

Here we report the results obtained on astrophysical data, by comparing several clustering methods. Different similarity measures have been used for constructing Prim's trajectories. In order to provide some comparisons with an alternative existing method, clustering results obtained with the so-called *spectral clustering* [11] approach are given. The spectral clustering algorithm used is that of Ng *et al.* [9] based on an eigen-decomposition of the normalized Laplacian of the graph. The affinity which is considered in the latter case relies upon a new metric introduced recently by Grishkat *et al.* [6], which uses hitting time of Prim's trajectories rooted at each vertex. Some tests have been realized with symmetrical divergences as similarity measures in the spectral clustering algorithm, and are not reported here. On one hand, they have not been proved to be able to handle non-euclidean cases ; on the other hand, the results obtained were not better than those reported in table 1. Note that this algorithm will not be detailed here, as it is only mentioned for providing a comparison to the one proposed in this paper.

The asteroid data are reflectances measured at different wavelengths, from which a mineralogic classification of the asteroid is seeked. More details on the physics underlying the classification problems are in [8].

The Small Main Belt Asteroid Spectroscopic Survey phase II (SMASSII) contains spectra of 1341 asteroids recorded in the band 0.44 and 0.92 μm. It has been used as a reference for the Bus and Binzel taxonomy [3]. To make a fair comparison with the supervised classification method proposed on this survey [12], we kept only spectra which do not contain missing values ; hence the survey reduces to 1329 asteroids spectra.

A cluster C will be associated with the taxonomic class Tax (defined by Bus and Binzel) which has the largest overlapp with C. Let us define some variables : $\mathcal{N}_C$ represents the cardinal of C, $\mathcal{N}_{Tax}$ the cardinal of Tax and $\mathcal{N}_{inter}$ the cardinal of the intersection of C and Tax.

A clustering validity index is defined as $Score = \mathcal{N}_{inter}/\mathcal{N}_{Tax}$. $Score$ characterizes the ratio of asteroids belonging to a taxonomic class and that are correctly labeled.

Clustering Methods	Score
Nearest Neighbor MST (Euclidean) + K-means	815/1329 = 61,32%
Nearest Neighbor MST (Kullback-Leibler) + K-means	976/1329 = 73,44%
Nearest Neighbor MST (Rènyi) + K-means	972/1329 = 73,14%
Spectral Clustering [9]	878/1329 = 66,06%
Spectral Clustering (Dual Rooted Hitting Time) [6]	913/1329 = 68,69 %
K-Nearest Neighbor [12]	777/1329 = 58,46 %
K-means (randomly initialized)	773/1329 = 58,16%

TAB. 1 – Synthesis of results obtained on SMASSII

From the table 1, we see that properly initialized K-means used simultaneously with informational divergence based affinity measures outperforms previously proposed approaches.

4.2 Multi-spectral image of Paris

In this experiment, 4 ($512\times$ 512) images of the same scene are available ; each image is recorded from a device operating at a different wavelength (more precisely, around a different wavelength). The affinity measure is based upon the Kullback-Leibler divergence between the (4-points) spectra associated to each pixel. Image registration problems are not tackled here, and it is supposed that the images are perfectly registered. Figure 4 illustrates this. The proposed algorithm (Prim based initializing of K-means clustering method) is tested on this multi-spectral image, where each vertex is nothing but a 4-points spectrum. For avoiding to deal with 512^2 vertices, the Nearest Neighbor MST algorithm depicted in section 3 is applied to an image which has been subsampled by a factor of 4. Figure 5 shows the obtained results :

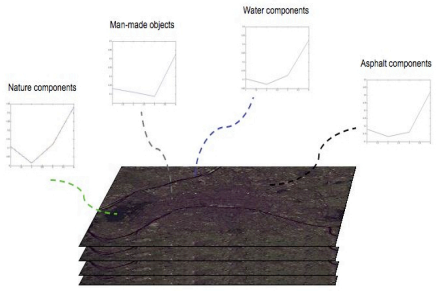

FIG. 4 – Multi-spectral image of Paris composed by multi-components

8 clusters are identified, from which 3 are easy to understand. Cluster 2 contains the pixels that are characteristic from trees and grass regions. Therefore, one can recognize recreation areas and natural parks in Paris surroundings (Boulogne, Vincennes). Cluster 3 exhibits the 'water areas' in Paris, and

[2]The hierarchical tree requires $O(LR\log R)$ logical operations.

the Seine river together with some known ponds is easily extracted. Cluster 4 is clearly associated with roads, asphalt and concrete. Other clusters cannot be fairly interpreted without cross analysing our results with *e.g.* pollution imaging or gas detection systems. This very rough unsupervised approach leads to think that the proposed method is very promising. Note that a similar approach using Euclidean distances was tested and led to gather 'water pixels' and 'green pixels' as belonging to the same cluster.

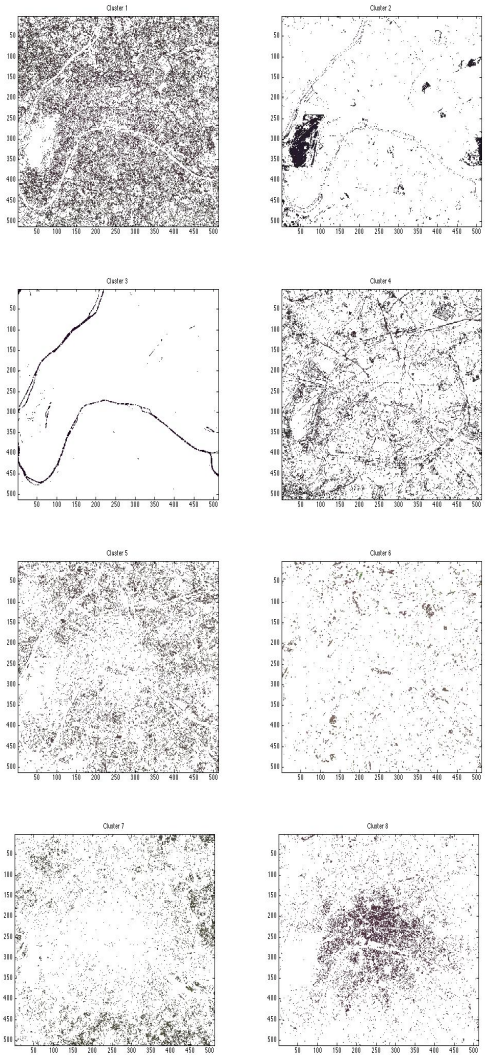

FIG. 5 – Clusters obtained on the multi-spectral image of Paris

5. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed an original approach for clustering multi-dimensional data. The method is based on the estimation of the number of clusters from the construction of a minimum spanning tree, in order to provide the initialization parameters of classical K-means algorithm. New criteria are derived for setting the false alarm rate (power) of a test over the Prim's trajectory associated with a MST built over the set of data. We assumed that the vertices are distributed according to a Poisson distribution, in the absence of additional information. Should prior information be available, this reasoning could be extended to other distributions. The usefulness of the information divergence based affinity measure is illustrated throughout many examples taken from astrophysical field or multi-spectral image analysis. In this paper, the threshold value is constant along the Prim's trajectory. We can think of setting up a variable threshold as a function of connected segments.

Some improvement in the understanding of the behavior of Prim's trajectory for vertex distributions exhibiting different modes are under study, and will allow to define clusters and labels directly from the MST, without resorting to *e.g.* K-means. In the case of hyper-spectral images, the proposed method will also require to be developed onto some lower dimensional subspace. Dimension reduction and its relationship to spectral clustering methods applied to graphs using information divergence or MST-based distances must be investigated [2].

REFERENCES

[1] A. Banerjee, S. Merugu, I. S. Dhillon and J. Gosh, "Clustering with Bregman Divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705 – 1749, 2005.

[2] M. Belkin and P. Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural Computation,* vol. 15, no. 6, June 2003, pp. 1373-1396.

[3] S. J. Bus and R. P. Binzel. "Phase II of the Small Main-Belt Asteroid Spectroscopic Survey : The Observations", *Icarus*, vol. 58, no. 1, Jul. 2002, pp. 146–177.

[4] C. -I. Chang. "An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis", *IEEE Transactions on information theory*, vol. 46, no. 5, Aug 2000.

[5] T. Cover and J. Thomas, *Elements of Information Theory* : John Wiley and Sons, Inc., 1991.

[6] S. Griskschat, J. A. Costa, A. O. Hero and O. J. J. Michel. "Dual rooted-diffusions for clustering and classification on manifolds, " in *Proc. ICASSP 2006*, Toulouse, France, May 14-19. 2006

[7] A. Hero, B. Ma, O. Michel and J. Gorman, "Applications of Entropic Spanning Graphs", *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, 2002.

[8] O. J. J. Michel, P. Bendjoya and P. RojoGuer, "Unsupervised clustering with MST : Application to asteroïd data, " in *Proc. Physics in Signal and Image Processing (PSIP) 2005*, Toulouse, France, January 31-February 2. 2005.

[9] A. Y. Ng, M. I. Jordan and Y. Weiss, "On Spectral Clustering : Analysis and an algorithm, " in *Proc. Neural Information Processing Systems (NIPS) 2001*, Vancouver, Canada, December 3-8. 2001.

[10] R. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. J.*, vol. 36, pp. 1389–1401, 1957.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[12] J. Warell and C.-I. Lagerkvist, "Asteroid taxonomic classification in the Gaia photometric system," *Astronomy and Astrophysics*, vol. 467, no. 2, pp.749–752, May 2007.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers whose helpful comments contributed to improve the paper.