

Addendum to: An Approach to Hierarchical Clustering via Level Surfaces and Convexity

Jerome Malitz¹, Seth Malitz²

¹Department Mathematics, Emeritus, University of Colorado, Boulder, USA

²Department Research and Development, GeoEye, Thornton, USA

E-mail: jerrymalitz@yahoo.com, malitz.seth@geoeye.com

Received January 20, 2010; revised March 2, 2010; accepted April 7, 2010

Abstract

This article is an addendum to the 2001 paper [1] which investigated an approach to hierarchical clustering based on the level sets of a density function induced on data points in a d -dimensional feature space. We refer to this as the “level-sets approach” to hierarchical clustering. The density functions considered in [1] were those formed as the sum of identical radial basis functions centered at the data points, each radial basis function assumed to be continuous, monotone decreasing, convex on every ray, and rising to positive infinity at its center point. Such a framework can be investigated with respect to both the Euclidean (L_2) and Manhattan (L_1) metrics. The addendum here puts forth some observations and questions about the level-sets approach that go beyond those in [1]. In particular, we detail and ask the following questions. How does the level-sets approach compare with other related approaches? How is the resulting hierarchical clustering affected by the choice of radial basis function? What are the structural properties of a function formed as the sum of radial basis functions? Can the level-sets approach be theoretically validated? Is there an efficient algorithm to implement the level-sets approach?

Keywords: Hierarchical Clustering, Level Sets, Level Surfaces, Radial Basis Function, Convex, Heat, Gravity, Light, Cluster Validation, Ridge Path, Euclidean Distance, Manhattan Distance, Metric

1. Introduction

This article is an addendum to our 2001 paper [1] which considered algorithmic aspects of an approach to hierarchical clustering based on the level sets of a “density function” induced on data points in a d -dimensional feature space. Quotes are invoked in the previous line because the density functions considered in [1] are not necessarily integrable. They do however satisfy the property that where the data points are dense, the value of the density function is relatively high in the associated region of feature space, and where the data points are sparse, the value of the density function is relatively low in the associated region of feature space. The density functions are taken as the sum of identical radial basis functions centered at the data points, each radial basis function assumed to be continuous, monotone decreasing, convex along each ray, and rising to positive infinity at its center point. Although this framework can be investigated with respect to both the Euclidean (L_2) and Manhattan (L_1) metrics, particular attention

was paid to binary data under the Manhattan metric. That paper concluded with various open questions, which to our knowledge are still open. Among them: Is there a polynomial-time algorithm to implement the level sets approach for binary data in L_1 ? Since the publication of [1], we have encountered other works in the literature, some slightly preceding ours, that seem to bear some relationship to our particular version of a level-sets approach. It seems to us that there is still an opportunity for further development of frameworks and potential unification regarding these methods of hierarchical clustering and their validation. Toward this end, the present addendum puts forth some observations and questions beyond those in [1] about the level-sets approach. In particular, we detail and ask the following questions. How does the level-sets approach compare with other related approaches? How is the resulting hierarchical clustering affected by the choice of radial basis function? What are the structural properties of a function formed as the sum of radial basis functions? Can the level-sets approach be theoretically validated? Are there efficient algorithms to

implement the level-sets approach?

The essence of what we are calling the *level-sets* (equivalently, *level-surfaces*) approach to clustering is perhaps more commonly referred to as a *density-based* approach. The basic idea of using the level sets of a density function to create a hierarchical clustering of data points in a Euclidean space goes back at least as far as 1975 with Hartigan [2], Fukunaga and Hostler [3], and Schnell [4].

For ease of exposition, let us introduce some notation. For us, a *radial basis function* (RBF) $\psi(\mathbf{x}, \mathbf{c})$ on R^d means any real-valued function of \mathbf{x} whose value depends only on the distance (either Euclidean or Manhattan) of \mathbf{x} from a *center point* \mathbf{c} . The characters {I, N, M, S, C} will be used to denote certain properties of the RBF along any ray from its center point: (I) means the value on every ray rises to infinity at the center point, but is everywhere else defined; (N) means the value along every ray is non-increasing; (M) means the value along every ray is monotone decreasing; (S) means the value along every ray is smooth; (C) means the value along every ray is convex. The notion of “ray” here is the same whether we are talking about Euclidean space or Manhattan space – it will always refer to a straight line originating from the center of the RBF. For our purposes, all RBF’s will be assumed continuous, except possibly at the center point. Note that in [1], the term *point luminosity function* was used instead of IMC RBF, and the term *combined luminosity function* was used instead of density function or sum of RBFs. In this addendum, there will be occasion to use both terminologies. A real-valued function that still has the notion of a center or *reference point* (at which the function is either bounded or infinite) but is not necessarily radially-symmetric about that point will be called an Arbitrary Influence Function (AIF). For our purposes, all AIF’s are assumed continuous, except possibly at the reference point. The meanings of the characters {I, N, M, S, C} extend naturally to AIF’s.

To give proper context to some of the questions alluded to earlier, we wish to draw attention to two other approaches in the literature that are related to the level-sets approach in that all three can be viewed as *physics-based* approaches to hierarchical clustering.

2. Physics-Based Approaches to Hierarchical Clustering

Although the level-sets approach to hierarchical clustering is considered a *density-based* approach, it may also be considered in a natural way a *physics-based* approach. The two perspectives are explained below. We assume the data points lie in Euclidean d -space unless otherwise stated.

Density-based approaches to hierarchical clustering typically regard clusters as dense regions of data points

in the feature space separated by regions of low density. Formalizing this requires the notion of a non-negative real-valued *density function*, which is defined over the feature space. Where the data points are dense, the value of the density function is relatively high, and where the data points are sparse, the value of the density function is relatively low. Such a density function might be obtained, for example, by evaluating the number of data points within a sliding volumetric window, or by evaluating the sum of identical, non-increasing RBF’s centered at the data points.

One of several density-based algorithms frequently cited in the literature [5,6] is DENCLUE, due to Hinneburg and Keim [7]. In one version of DENCLUE, the density function is taken as the sum of identical Gaussian RBF’s centered at the data points. The associated algorithm provides an efficient, approximate way to compute which data points are enclosed within the same level surface component of the density function where the level surface components are induced by a threshold value ξ . A hierarchy of clusters may be obtained by varying the standard deviation σ of the Gaussian RBF’s while adjusting the value of ξ in concert. The reference [7] considers other RBF’s too, in addition to the Gaussian.

The level-sets approach is density-based as well, but unlike the approach in [7], it builds only one density function. The hierarchy of clusters is obtained by varying the threshold alone and not the density function.

Physics-based approaches to hierarchical clustering are related to real-world physical processes involving matter or energy. Level-sets hierarchical clustering (one version thereof) qualifies as such because the underlying density function may be viewed as the sum of point luminosity functions centered at the data points, where the density function models the total light flux impinging on each point of the feature space. A hierarchy of clusters is obtained from the level sets of this density function. Two other algorithms that could be called physics-based are GRIN [8,9], and again, DENCLUE from [7].

In GRIN, the clustering algorithm simulates how point masses, placed initially at the data points of the Euclidean feature space, cluster together hierarchically under the influence of a gravitational force.

In DENCLUE, expanding on the discussion above, a hierarchical clustering of the data, induced by a nesting of level surfaces, can be obtained by varying the standard deviation σ of the underlying Gaussian RBF’s while varying the threshold ξ in concert. Although not explicitly mentioned in [7], the nesting of level surfaces they touch upon may be obtained from a different yet equivalent perspective, namely one involving the propagation of heat through a convective medium as follows.

Imagine that the Euclidean feature space is filled with a convective medium. At time $T = 0$, let us assume that throughout the medium the temperature is 0 everywhere except for the data points, each of which is modeled as a

Dirac delta function “hotspot” with integrated temperature value equal to 1. Call this the *initial heat configuration*. The temperature at any point in the medium is given by the time-dependent heat Equation. It is well-known that at any fixed time T , the solution of the heat equation may be obtained as (see for instance, Strang [10]) the convolution of the initial heat configuration by a Gaussian kernel with σ^2 proportional to T . But it is easily seen that for an initial heat configuration of finitely many delta function hotspots, convolution by a Gaussian kernel is equivalent to summing copies of that Gaussian kernel centered at the hot spots.

Now for some fixed time $T_0 > 0$, choose a threshold $\xi(T_0)$ such that when the sum $S_0(\mathbf{x})$ of time-dependent Gaussians (with variance $\sigma^2(T_0) = T_0$) centered at the data points submits to this threshold, each data point is encapsulated within its own level surface component. Here the value of each Gaussian at its respective data point is proportional to $1/\sqrt{2\pi\sigma^2(T_0)}$. Now consider a time $T_1 > T_0$. The value of each new Gaussian at its respective data point is $1/\sqrt{2\pi\sigma^2(T_1)}$, where standard deviation $\sigma(T_1) > \sigma(T_0)$. Let $S_1(\mathbf{x})$ denote the sum of these new Gaussians. If we multiply each new Gaussian by $\sigma(T_1)/\sigma(T_0)$, then the resulting sum $S_{11}(\mathbf{x})$ is greater than or equal to $S_0(\mathbf{x})$ everywhere. Thus the level sets of $S_{11}(\mathbf{x})$ at threshold $\xi(T_0)$ encapsulate those of $S_0(\mathbf{x})$ at threshold $\xi(T_0)$. The equivalent effect can be achieved without multiplying the Gaussians at time T_1 by $\sigma(T_1)/\sigma(T_0)$, but rather instead by introducing a new threshold $\xi(T_1) = \xi(T_0)\sigma(T_0)/\sigma(T_1)$. That is to say, the level sets of $S_1(\mathbf{x})$ at threshold $\xi(T_1)$ are identical to the level sets of $S_{11}(\mathbf{x})$ at threshold $\xi(T_0)$, and thus encapsulate the level sets of $S_0(\mathbf{x})$ at threshold $\xi(T_0)$. More generally, if we set the time-dependent threshold $\xi(T) = \xi(T_0)\sigma(T_0)/\sigma(T)$, then the sum of the time-dependent Gaussians yields a hierarchy of level surface components under encapsulation. Thus the approach in [7], given its connection to heat, may be considered a physics-based approach to hierarchical clustering, or more precisely, a *heat-based* approach.

3. Open Questions

The questions below are distinct from those mentioned at the end of [1], and are open to the best of our knowledge. We think they are of both mathematical and practical interest.

3.1. Effect of Different Radial Basis Functions

Given two different RBF's, $f(\mathbf{x})$ and $g(\mathbf{x})$, what is the difference between the level set hierarchy associated with $F(\mathbf{z})$, the sum of RBF's $f(\mathbf{x})$ centered at the data points, versus $G(\mathbf{z})$, the sum of RBF's $g(\mathbf{x})$ centered at the data points? What is the difference in the hierarchical clustering of the data points associated with $F(\mathbf{z})$

versus $G(\mathbf{z})$? For example, what is the difference when $f(\mathbf{x})$ and $g(\mathbf{x})$ are two IM RBFs such that $f(\mathbf{x}) > g(\mathbf{x})$ for all \mathbf{x} ?

In a related vein, suppose $g(t, \mathbf{x})$ is a continuous family of RBF's parameterized by the real-valued parameter $t \geq 0$. Let us assume that some natural notion of “radius” of $g(t, \mathbf{x})$ grows monotonically with t . For example, when $g(t, \mathbf{x})$ is the Gaussian heat kernel, this radius can be taken as the standard deviation σ , where $\sigma^2 = t$. Suppose that $\xi(t)$ is a (possibly) time-dependent threshold, such that the $\xi(t)$ -level sets of $g(t, \mathbf{x})$ form a hierarchy under encapsulation, where the $\xi(0)$ -level set encapsulates each data point in its own component, and at some time $t > 0$, the $\xi(t)$ -level set encapsulates all the data points within the same component. Thus $g(t, \mathbf{x})$ and $\xi(t)$ induce a hierarchy of level sets, which in turn induces a hierarchical clustering of the data points. Call this the generalized heat-based approach to hierarchical clustering utilizing the generalized heat kernel $g(t, \mathbf{x})$. What is the difference between the level set hierarchy in the generalized heat-based approach using $g(t, \mathbf{x})$ versus some other parameterized family of RBF's $f(t, \mathbf{x})$? What is the difference between the hierarchical clustering of the data points in the generalized heat-based approach using $g(t, \mathbf{x})$ versus some other parameterized family of RBF's $f(t, \mathbf{x})$?

Given two different hierarchical clusterings of the same data points (as for example might be generated by two runs of the level sets approach, each using a different RBF), is there another hierarchical clustering that in some sense represents the hybrid of the two, in other words is in some sense consistent with both?

3.2. Comparing the Physics-Based Approaches

Is there a formal sense in which the level-sets approach to hierarchical clustering is approximately equivalent to the generalized heat-based approach? In other words, given a parameterized RBF $g(t, \mathbf{x})$ and threshold $\xi(t)$ for the generalized heat-based approach, is there a corresponding RBF $f(\mathbf{x})$ for the level sets approach such that the level set hierarchies generated by the two approaches are in some sense close, or the hierarchical clusterings that they induce on the data points are in some sense close?

Given a parameterized RBF $g(t, \mathbf{x})$ and threshold $\xi(t)$ for the generalized heat-based approach, is there a corresponding function $H(t, \mathbf{x})$ satisfying $H(t, \mathbf{x}) = H(0, \mathbf{x}) + t$ for all t , and such that the 0-level sets of $H(t, \mathbf{x})$ indexed by t are identical to the level sets of the generalized heat-based approach indexed by t ? (The question as formulated is reminiscent of the perspective taken in the “Level Set Method” by S. Osher and J. Sethian [11], which may be of relevance here.) Assuming the answer is yes, can $H(0, \mathbf{x})$ be expressed as, or well-approximated by, a sum of RBF's centered at the data points?

Is either the level-sets approach or heat-based approach to hierarchical clustering approximately equivalent to the gravity-based approach of [8] and [9]?

3.3. Validating the Physics-Based Approaches

Here we are concerned with the possible theoretical validation of the physics-based approaches, rather than their empirical validation against selected data sets. To this end we ask several questions.

Are there natural stochastic birth (or birth/death) processes to generate hierarchical data in Euclidean or Manhattan d -space, such that the descendents of a sample point of the process can form arbitrary-shaped clusters, not just spherical or ellipsoidal ones? Are there such processes that can simulate natural physical processes (e.g., galactic cluster formation) and natural biological processes (e.g., biological evolution)? Can such stochastic processes be designed as variants of the *Hierarchical Dirichlet Process* [12]?

Is it possible to prove, under general assumptions about the stochastic process, that one or more of the three physics-based approaches to hierarchical clustering is inherently good, with high probability, at recovering the approximate phylogenetic structure of a sample run of the process? Inherent in this question, is another question that asks how do we measure the distance between two hierarchical clusterings of the data points, in this case the hierarchy constructed by the algorithm versus the true hierarchy of the data generation?

3.4. Ghost Clusters

Does the sum of identical, IMSC RBF's centered at the data points in Euclidean d -space ever exhibit a finite local maximum? Notice that if the answer is yes, then there is a threshold that yields a level surface component whose interior is devoid of data points. Such a component could rightly be called a *ghost cluster*. Note that if there is reason to believe that a particular data set under investigation has an underlying phylogenetic structure, a finite local maximum might have a natural interpretation as encapsulating a one-time progenitor that has since gone extinct.

What are necessary and sufficient conditions against an IMSC RBF such that the sum of identical such RBF's centered at the data points does not exhibit a finite local maximum? A sufficient condition is that the RBF be harmonic in R^d – for if a finite local maximum were to occur, it would violate the Maximum Principle for Harmonic Functions. (It is well-known that up to multiplicative and additive constant, there is a unique positive radially-symmetric harmonic function in Euclidean R^d that goes to positive infinity at the origin and is elsewhere defined throughout the space. This fact is a corollary of

Bôcher's Theorem [13]. If necessary and sufficient conditions do exist for RBF's, do they also exist for IMSC AIF's? Can the RBF's or AIF's in the sum be non-identical?

Does the heat-based approach to hierarchical clustering, using the time-dependent Gaussian heat kernel with $\sigma^2 = T$ and time-dependent threshold $\xi(T) = \xi(T_0)\sigma(T_0)/\sigma(T)$, ever produce a ghost cluster?

3.5. Ridge Paths in Euclidean Space

Suppose identical IMSC RBF's are centered at the data points $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$ in Euclidean d -space. Consider the sum F of these RBF's. Say a continuous path P from \mathbf{p}_i to \mathbf{p}_j ($i \neq j$) in R^d is a *ridge path* if for all but finitely many points \mathbf{x} along the path, all the following conditions hold: 1) the path tangent at \mathbf{x} is well-defined; 2) the path tangent at \mathbf{x} is aligned with the gradient of F at \mathbf{x} ; 3) the restriction of F to the $(d-1)$ -plane normal to the path tangent at \mathbf{x} assumes a local maximum at \mathbf{x} . A ridge path is said to be a *direct ridge path* if it additionally satisfies: 4) no other data point \mathbf{p}_k lies on the path. Thus a direct ridge path from \mathbf{p}_i to \mathbf{p}_j is in a sense locally maximum with respect to F all along its trajectory. Define the *height* of path P from \mathbf{p}_i to \mathbf{p}_j to be the minimum F -value assumed along its trajectory. Is there a highest path (one with maximum height as defined in the previous line) from \mathbf{p}_i to \mathbf{p}_j that is also a ridge path? If so, then there is the following connection between ridge paths and level-sets hierarchical clustering. If for some threshold, T_1 , the points \mathbf{p}_i and \mathbf{p}_j lie within the interiors of distinct level surface components C_i and C_j , respectively, and at threshold T_2 those components merge, then there is a direct ridge path between some data point in C_i and some data point in C_j whose lowest F -value is T_2 .

Many other natural questions suggest themselves in relation to these ridge paths. How many direct ridge paths are there as a function of the number of data points and the dimension of the space in which they reside? How different is the connectivity structure of the direct ridge paths from say the Delaunay triangulation of the points? Can there ever be two distinct direct ridge paths between a pair of data points? If direct ridge path P is parameterized by t in $[0, 1]$, is the function $G(t) = F(P(t))$ a convex function of t ? As a function of dimension, what is the largest possible ratio between the arc length of a direct ridge path and the Euclidean distance between the same pair of end points?

With an eye toward generating an efficient algorithm to compute the level-sets hierarchical clustering, is there an efficient algorithm (that avoids the curse of dimensionality, *i.e.*, exponential run-time dependence on problem-space dimension) to find all the direct ridge paths induced by $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$ in Euclidean d -space? In particular, is there an efficient algorithm based on the fol-

lowing two-step paradigm: 1) Using an easy-to-evaluate criterion, determine which pairs of points $\{\mathbf{p}_i, \mathbf{p}_j\}$ are candidates for a direct ridge path; 2) find the direct ridge path from \mathbf{p}_i to \mathbf{p}_j numerically, starting with a straight line of beads from \mathbf{p}_i to \mathbf{p}_j and then iteratively optimizing the bead positions in a manner similar to the variational active contour method of *Snakes* [14]. To reduce the number of calculations, if a bead B is “far” from a data point \mathbf{p}_k , then one might consider excluding the contribution of \mathbf{p}_k ’s influence function while optimizing the position of B .

It also seems reasonable to ask for efficient approximation algorithms to recover the level-sets hierarchical clustering, especially given that many other problems related to clustering are already known to be NP-hard. These include: Optimal Binary Decision Tree [15], Optimal Rectilinear Steiner Tree [16], Steiner Problem in Phylogeny [17], Maximum Likelihood Tree [18], and various versions of the Parsimony Problem in Phylogenetics [19].

Finally, in this section, we ask if the discussion on ridge paths can be generalized to the function $H(t, \mathbf{x})$ in Subsection 3.2, when evaluated a time $t = 0$. So doing might lead to an efficient algorithm for recovering the heat-based hierarchical clustering introduced in [7].

3.6. Approximate Level-Set Hierarchical Clustering

An approximate level-set hierarchical clustering that does not involve finding direct ridge paths might be based, at least notionally, on the following two-step paradigm. First, using an easy-to-evaluate criterion, determine which pairs of points $\{\mathbf{p}_i, \mathbf{p}_j\}$ are candidates for a direct ridge path. Second, connect the two points of each candidate pair by a straight line segment and evaluate the density function along the line segments. This configuration may be viewed as having *restricted* the density function, originally defined over the entirety of the feature space, to a restricted space now consisting only of the introduced line segments. Within this restricted space, it still makes sense to talk about level sets, and thus it still makes sense to speak of the level-sets approach to hierarchical clustering in this space. Can one design an efficient implementation of such an approach? How does the resulting hierarchical clustering compare with the unrestricted level-sets hierarchical clustering?

3.7. Ridge Paths in Manhattan Space

In certain hierarchical clustering applications, the L_1 metric may be preferred to the L_2 metric for computing inter-point distances. For example, quoting from the abstract of Aggarwal, Hinneburg, Keim [20], “...we spe-

cifically examine the behavior of the commonly used L_k norm and show that the problem of meaningfulness in high dimensionality is sensitive to the value of k . For example, this means that the Manhattan distance metric L_1 is consistently more preferable than the Euclidean distance metric L_2 for high dimensional data mining applications.” Hence there can be motivation to consider the same basic questions of Subsection 3.5, but now with respect to the L_1 metric; for example, even the special case of the L_1 metric on binary data.

Suppose identical IMSC RBF’s are centered at the data points $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$ which are located at some corners of the closed binary cube $[0, 1]^d$ within *Manhattan* d -space. Consider the sum F of these RBF’s. Say a continuous path P from \mathbf{p}_i to \mathbf{p}_j ($i \neq j$) in $[0, 1]^d$ is a ridge path if for all but finitely many points \mathbf{x} along the path, all the following conditions hold: 1) the path tangent at \mathbf{x} is well-defined; 2) the path tangent at \mathbf{x} is aligned with the limiting value of $\mathbf{grad} F(\mathbf{y})$ (for \mathbf{y} in the open binary cube $(0,1)^d$ arbitrarily near \mathbf{x}) projected into the closed binary cube $[0, 1]^d$; 3) the restriction of F to the $(d-1)$ -plane normal to the path tangent at \mathbf{x} assumes a local maximum at \mathbf{x} within the closed binary cube $[0, 1]^d$. A ridge path is said to be a direct ridge path if it additionally satisfies: 4) no other data point \mathbf{p}_k lies on the path.

In view of Theorem 3.2 of [1], if there is a direct ridge path from \mathbf{p}_i to \mathbf{p}_j , is there a direct ridge path from \mathbf{p}_i to \mathbf{p}_j that is also an edge-path of the closed cube, *i.e.*, that consists solely of edges of the closed cube $[0, 1]^d$? If so, call such a path a direct ridge edge-path. Again various questions arise. How many direct ridge edge-paths are there as a function of the number of data points and the dimension of the space in which they reside? When there are at least three data points in the data set, can there ever be two distinct direct ridge edge-paths between a pair of data points? If direct ridge path P is parameterized by t in $[0, 1]$, then is the function $G(t) = F(P(t))$ a convex function of t ? As a function of dimension, what is the largest possible ratio between the Manhattan length of a direct ridge path and the Manhattan distance between its end points?

With an eye toward generating an efficient algorithm to compute the level-sets hierarchical clustering, is there an efficient algorithm, that avoids the curse of dimensionality, to find all the direct ridge edge-paths induced by $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$, in the closed binary cube $[0, 1]^d$? In particular, is there an efficient algorithm based on the following two-step paradigm: 1) Using an easy-to-evaluate criterion, determine which pairs of points $\{\mathbf{p}_i, \mathbf{p}_j\}$ are candidates for a direct ridge edge-path; 2) find a direct ridge edge-path from \mathbf{p}_i to \mathbf{p}_j by starting with an arbitrary shortest edge path from \mathbf{p}_i to \mathbf{p}_j and then iteratively generating a finite sequence of edge paths from \mathbf{p}_i to \mathbf{p}_j , each a local and improved modification of the previous, until finally a direct ridge edge-path is obtained from \mathbf{p}_i to \mathbf{p}_j ?

3.8. Highest Paths of a Convex Function over a Convex Body

Theorems 3.1 and 3.2 from [1] state, respectively, that given two points $\mathbf{p}_1, \mathbf{p}_2$ on the surface of a closed convex body (convex polytope) B in R^d and a convex function F defined throughout B , there is a highest B -path from \mathbf{p}_1 to \mathbf{p}_2 with respect to F that resides wholly within the surface ∂B of B (that manifests itself as a sequence of edges within the surface ∂B of B .) A highest B -path with respect to F from \mathbf{p}_1 to \mathbf{p}_2 is defined to be a path within B connecting \mathbf{p}_1 to \mathbf{p}_2 whose minimum F -value is maximized. Can these theorems be generalized? For example, can their conclusions be maintained in certain circumstances under relaxed assumptions that would require B or F to be only star-convex rather than convex?

4. Conclusions

In this addendum, we considered a level-sets approach to hierarchical clustering where the underlying density function is formed as the sum of identical radial basis functions centered at the data points in either Euclidean or Manhattan Cartesian space. Each radial basis function was assumed to be continuous, monotone decreasing, convex on every ray, and rising to positive infinity at its center. We indicated how this setup can be viewed as a physics-based approach to hierarchical clustering modeled on light flux. Two other physics-based approaches were identified in the literature, one based on heat, the other based on gravity. This addendum puts forth basic and natural questions concerning the level sets approach. In particular, we asked the following questions. How does the level-sets approach compare with the other physics-based approaches? How is the resulting hierarchical clustering affected by the choice of radial basis function? What are the structural properties of a function formed as the sum of radial basis functions, each with convex profile? Can the levels-sets approach be theoretically validated? Is there an efficient algorithm to implement the level-sets approach in L_1 and L_2 ? We hope these questions will stir interest within the scientific community and be resolved.

5. References

- [1] R. Holley, J. Malitz and S. Malitz, "An Approach to Hierarchical Clustering via Level Surfaces and Convexity," *Discrete and Computational Geometry*, Vol. 25, No. 2, 2001, pp. 221-233.
- [2] J. Hartigan, "Clustering Algorithms," Wiley, 1975.
- [3] K. Fukunaga and L. Hostler, "The Estimation of the Gradient of a Density Function with Application in Pattern Recognition," *IEEE Transactions on Information Theory*, Vol. 21, No. 1, 1975, pp. 32-40.
- [4] P. Schnell, "A Method to Find Point-Groups," *Biometrika*, Vol. 6, 1964, pp. 47-48.
- [5] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques," *Journal of Intelligent Information Systems*, Academic Publishers, Vol. 17, No. 2-3, 2001, pp. 107-145.
- [6] S. Kotsiantis and P. Pintelas, "Recent Advances in Clustering: A Brief Survey," *WSEAS Transactions on Information Science and Applications*, Vol. 1, No. 1, 2004, pp. 73-81.
- [7] A. Hinneburg and D. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1998, pp. 58-65.
- [8] Y. J. Oyang, C. Y. Chen and T. W. Yang, "A Study on the Hierarchical Data Clustering Algorithm Based on Gravity Theory," *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, 2001, pp. 350-361.
- [9] C. Y. Chen, S. C. Hwang and Y. J. Oyang, "An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory," *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, 2002, pp. 237-250.
- [10] G. Strang, "Introduction to Applied Mathematics," Wellesley-Cambridge Press, Wellesley, 1985.
- [11] J. Sethian, "Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science," Cambridge University Press, Cambridge, 2002.
- [12] Y. W. Teh, M. Jordan, M. Beal and D. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, Vol. 101, No. 476, 2006, pp. 1566-1581.
- [13] S. Axler, P. Bourdon and W. Ramey, "Harmonic Function Theory," Springer-Verlag, New York, 2001.
- [14] M. Kass, A. Witkin and D. Terzopolous, "Snakes: Active Contour Models," *International Journal of Computer Vision*, Kluwer Academic Publishers, Norwell, 1988, pp. 321-331.
- [15] L. Hyafil and R. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete," *Information Processing Letters*, Vol. 5, No. 1, 1976, pp. 15-17.
- [16] M. Garey and D. Johnson, "The Rectilinear Steiner Tree Problem is NP-Complete," *SIAM Journal on Applied Mathematics*, Vol. 32, No. 4, 1977, pp. 826-834.
- [17] L. Foulds and R. Graham, "The Steiner Problem in Phylogeny is NP-Complete," *Advances in Applied Mathematics*, Vol. 3, No. 1, 1982, pp. 43-49.
- [18] B. Chor and T. Tuller, "Finding a Maximum Likelihood Tree is Hard," *Journal of the ACM*, Vol. 53, No. 5, September 2006, pp. 722-744.
- [19] W. Day, "Computationally Difficult Parsimony Problems in Phylogenetics Systematics," *Journal of Theoretical Biology*, Vol. 103, No. 3, 1983, pp. 429-438.

- [20] C. Aggarwal, A. Hinneburg and D. Keim, “On the Surprising Behavior of Distance Metrics in High Dimensional Space,” *Database Theory—ICDT 2001, Lecture*

Notes in Computer Science, Springer, Berlin/Heidelberg, Vol. 1973, 2001, pp. 420-434.