Symbiosis

Research Article | Journal of Computer Science Applications and Information Technology | Open Access

Received: February 17, 2017
Accepted: February 28, 2017
Published: March 13, 2017

Graph Based Hybrid Clustering With Unbounded Regions

Hongjun Su and Hong Zhang*

Department of Computer Science and Information Technology, Armstrong State University, Savannah, Georgia, USA

***Corresponding author:** Hong Zhang, Department of Computer Science and Information Technology, Armstrong State University, Savannah, Georgia, USA, E-mail: hong.zhang@armstrong.edu

1 Abstract

An extension of the hybrid clustering approach is proposed for partitioning data with possibly unbounded polygon regions. Clustering or partitioning data into relatively homogeneous and coherent subpopulations can be an effective pre-processing method to achieve data analysis tasks such as pattern recognition and classification. Our method uses a graph to model the initial manual partition of the dataset. Based on the graph model, an algorithm is developed for automatic detection of the regions defined by the partition. A clustering algorithm using Markov Chain Monte Carlo method is developed for finding optimal adjustments to the partition automatically. The regions are generalized polygons which may include points at infinity. Homogeneous coordinates are used to represent the points at infinity and to derive algorithms in a unified fashion.

2 Keywords:

Partition; clustering; Planar graph; Markov Chain Monte Carlo method; Homogeneous coordinates;

3 Introduction

In applications involving large volume, high dimensional data analysis, it is common to present the data in a series of graphical plots that resemble digital images. For example, in flow cytometry, a typical analysis would involve many such data plots [1,2]. Even though such a 2D plot only represents a projection of the data into a 2D space, the graphical view provides convenience in practical applications. In such an application, a "divide and conquer" approach to partition the data is often useful in reducing the complexity of the system for further analysis.

A standard, general purpose clustering algorithm such as K-means may not perform well in this type of applications, because of the structural complexity [3]. The resulting partition may not correspond to the desired clustering patterns. User input on the partition can provide useful guidance on the overall pattern, but the manual operations may not easily achieve the necessary accuracy and consistency.

In a previous work, we proposed a hybrid approach to facilitate the partitioning and clustering of the data in an intuitive and convenient way. A user will be able to draw an initial partition on a finite image through a graphical user interface [4]. The partition will be modeled as a graph and an algorithm is developed to automatically determine all the polygon regions in the partition. The partition scheme is then optimized to best fit the data using a special clustering algorithm based on the MCMC paradigm.

The method in [4] is limited to a finite bounded region such as an image and all the partitions containing data points are bounded polygons. Such a limitation may not be appropriate for applications in which the dataset does not have a natural bound. In this paper, we propose an extension to the method to allow unbounded regions in the partition. Each partition region is defined as an extended version of polygon with vertices possibly located at infinity. Homogeneous coordinates are used to represent points at infinity. The partitioning and clustering algorithm can be carried out in a similar fashion as the bounded case.

4 Partitioning

The 2D projection of a dataset provides an intuitive, image like view for a specific cross section of the data, which is convenient for subdividing the dataset and creating partitions. We will consider partitions of the 2D plane formed by interconnected line segments as shown in Figure 1. Each region of the partition is formed by a set of line segments. The region is a generalized polygon because it can be unbounded. The initial partition is drawn manually through a GUI system. The dataset is divided into a collection of subsets according to the generalized polygon regions in the partition.

Each subset of the partition represents a relatively homogeneous group of elements with potentially reduced complexity for further analysis and recognition tasks. This process can be continued recursively to produce a hierarchical structure of partitions.

Polygons are simple convenient shapes to represent the partitioned regions. In our previous work, only finite polygons are considered [4]. We will try to extend the results to unbounded polygon regions with vertices possibly located at infinity.

Citation: Hongjun Su, Hong Zhang (2017) Graph Based Hybrid Clustering With Unbounded Regions. J Comp Sci Appl Inform Technol. 2(2):1-5

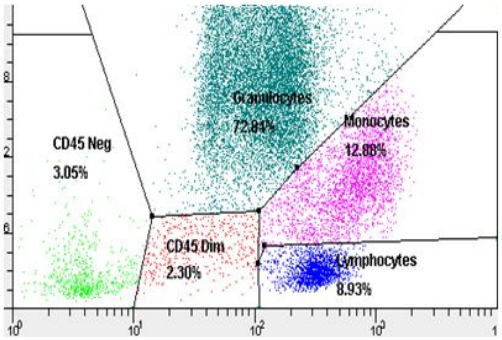

Figure 1 Partition of data

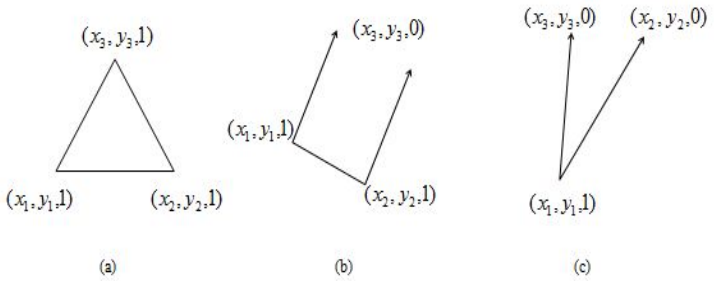

Figure 2 Generalized polygons

The projective plane is a 2D space that incorporates the points at infinity in a natural way [5]. The modern definition of the projective plane is the set of all lines through the origin in the 3D affine space. Homogeneous coordinates provide a natural and unified representation of all the points in the projective plane. Topologically, the projective plane is a semi-sphere with a Mobius band attached. The non-orientable surface presents some inconveniences for our partition algorithm [5]. For our purposes in this paper, we will take the view of the extended plane as an affine plane augmented with a line at infinity without identifying the antipodal points. Using this model the plane will remain orientable even though some geometric properties of the projective plane will be lost. We use homogeneous coordinates to represent the points at infinity as in projective geometry. However, concepts of Euclidean space such as angles will still be used.

A point in the extended plane can be represented as (x,y,w) where not all coordinates are 0. An ordinary point has a no-zero w and the corresponding Euclidean coordinates (x/w,y/w). A point at infinity will have w=0 with (x,y) indicating the vector pointing in the direction of the point at infinity. Most of the analytic geometry tools can be extended to homogeneous coordinates. For example, the equation of the line through the ordinary point $(x_1, y_1, 1)$ and the point at infinity $(x_2, y_2, 0)$ is:

$$y - x_1 = (y_2/x_2)(x - x_1)$$

The angle between the vector from $(x_1, y_1, 1)$ to $(x_2, y_2, 1)$ and the direction $(x_3, y_3, 0)$ can be found using the inner product:

$$\cos\theta = \frac{(x_2 - x_1, y_2 - y_1, 0) \cdot (x_3, y_3, 0)}{\|(x_2 - x_1, y_2 - y_1, 0)\| \cdot \|(x_3, y_3, 0)\|}$$

We define a generalized polygon as a sequence of vertices $< v_1, v_2, \ldots, v_n >$ in the extended plane. Each vertex is represented using homogeneous coordinates. For example, Figure 2 shows (a) a triangle with three ordinary points, (b) a triangle with two ordinary points and a point at infinity, (c) a triangle with one ordinary point and two points at infinity.

5 Region Detection

The partition scheme can be naturally represented as a graph. The vertices of the graph are the end points of the line segments and the edges are the lines. Because of our assumption on the extended plane, the surface is orientable with genus 0 and the graph is planar. If the graph is connected with v vertices, e edges, and f faces, then by Euler's formula [6], v-e+f=2.

To perform the partitioning on the underlying dataset, it is necessary to determine the polygon regions explicitly. If this task is done manually, it will be extra steps for users to complete which will be rather tedious and repetitive since the information for defining the regions is already contained in the graph. We present an algorithm to obtain the generalized polygons from the given graph automatically.

The algorithm presented in [4] can be extended to the unbounded polygon regions. The basic idea of the algorithm is to construct a polygon by tracing the interior sides of the boundary edges as illustrated in Figure 3. Each edge containing an ordinary point is considered to have two sides, each of that belongs to one polygon region.

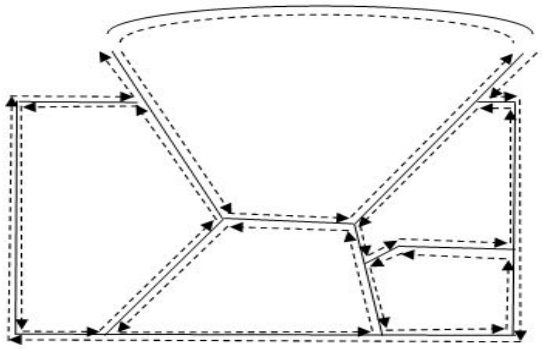

Figure 3 Region detection algorithm

The pseudo code of the algorithm is given below. Each edge of the graph constructed from two vertices of ordinary points or points at infinity is modeled with two directed edges in opposite directions. The graph is assumed to be connected and planar. The graph model produced by the GUI will be a proper planar embedding.

for each vertex v with outdegree > 0
  add v to polygon
  find first egde (v,u) in counter-clockwise order
  remove (v,u)
  while u != v

Citation: Hongjun Su, Hong Zhang (2017) Graph Based Hybrid Clustering With Unbounded Regions. J Comp Sci Appl Inform Technol. 2(2):1-5

v = u
add v to polygon
find first egde (v,u) in counter-clockwise order
remove (v, u)

The algorithm will find both bounded and unbounded regions. Unlike in [4], unbounded regions are modelled as generalized polygons and included in the partition. In [4], the signed area of a polygon is introduced to determine the orientation and to exclude the unbounded region.

$$A = \frac{1}{2}\sum_{i=1}^{M}(x_i y_{i+1} - x_{i+1} y_i)$$

This formula is no longer valid since the area of an unbounded polygon may not be finite. However, we do not need the signed area because unbounded regions are no longer excluded. If necessary, the orientation of a polygon can be determined with the sign of the winding number.

6 Clustering

A popular clustering method is the K-means algorithm. Given a set of data points and an initial set of k means, the algorithm proceeds by alternating between two steps: [3]

1. Assign each point to the cluster with the "nearest" mean.
2. Update the new means of the new clusters.

The algorithm converges when the clusters no longer change. Although simple to implement, the K-means algorithm is only guaranteed to converge to a local optimum. In the K-means method, it is also not easy to incorporate prior information about the clustering information.

Our partitioning approach provides a convenient way to introduce prior knowledge about the clustering. To optimize the fitting of the partition to the data, we propose a clustering algorithm that will seek the optimal adjustments automatically similar to the approach in [4]. The partition defined by the generalized polygons will guide the clustering process. The clustering algorithm will make the structure of the partition (the graph model) invariant, but will change the vertex locations of the partition to achieve an optimal fit. The measure for the quality of the clustering is defined as

$$E = \sum_{k=1}^{M} \frac{1}{|D_k|} \sum_{x_i \in D_k} |x_i - m_k|^2$$

Where D_k denotes the subset of 2D projections of elements in the k-th cluster and m_k the mean (centroid) of the cluster. The corresponding Boltzmann weight is given by

$$e^{-E/kT}$$

Where k is the Boltzmann constant and T the temperature.

The only special consideration required to handle the unbounded regions is the determination of the cluster containing a data point. For a regular polygon, the common method to determine whether a point is inside the region is to use the winding number [7]. Intuitively, the winding number of a polygon with respect to a point is the number of turns around the point made by traversing along the polygon. A non-zero winding number indicates that the point is inside the polygon. A straightforward method to calculate the winding number is to take the sum of the subtended angles over all vertices of the polygon. This is method can be generalized to polygons with vertices at infinity, because the angles can still be calculated using homogeneous coordinates. A more efficient method to calculate the winding number is to count the signed crossings along a ray in a fixed direction [8]. This method can also be adapted to the generalized polygons.

The proposed algorithm employs a MCMC (Markov Chain Monte Carlo) approach [9,10]. The value E defined above serves as the energy function for the MCMC method. The algorithm will seek a configuration with minimal energy. The pseudo code of the algorithm is given below.

loop until convergence
propose a perturbation of vertices of
partition through random walk
if proposed vertex locations form a
valid partition
compute the energy E' for new partition
if E' < E
accept the new partition
else
accept the new partition with
probability $e^{-}(E' - E)/kT$

The algorithm attempts to search for an optimal partition by perturbing the locations of the vertices. It follows the general Metropolis-Hastings style acceptance/rejection on the proposed movements.

7 Experimental Results

The region detection algorithm is seen to be robust on legitimate input graphs. Figure.4 illustrates the automatic region detection. The partition graph is drawn manually. The algorithm detects all the regions and displays the data points in the regions with different colors.

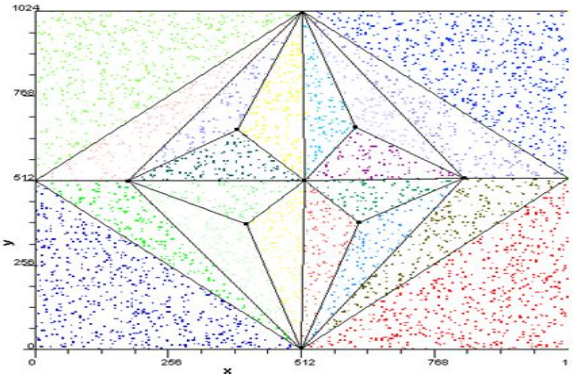

Figure 4 Automatic region detection

If the graph is not properly formed (e.g., disconnected graph, crossing edges, overlapping regions), the algorithm may not produce a correct output. Certain exceptional cases can be corrected automatically. For example, our implementation will connect a disconnected graph by adding a minimum numbers of edges.

Citation: Hongjun Su, Hong Zhang (2017) Graph Based Hybrid Clustering With Unbounded Regions. J Comp Sci Appl Inform Technol. 2(2):1-5

When the graph cannot be properly corrected, the condition can usually be detected in the algorithm. For example, a violation of Euler's formula can be easily detected.

Artificial datasets are generated to test the effectiveness of the clustering algorithm. Figure 5 shows a sample from a 2-component Gaussian mixture model. The initial partition is clearly not optimal. Figure 6 shows the result of applying the MCMC clustering algorithm. The adjusted partition is much closer to the optimal separation of the 2 components.

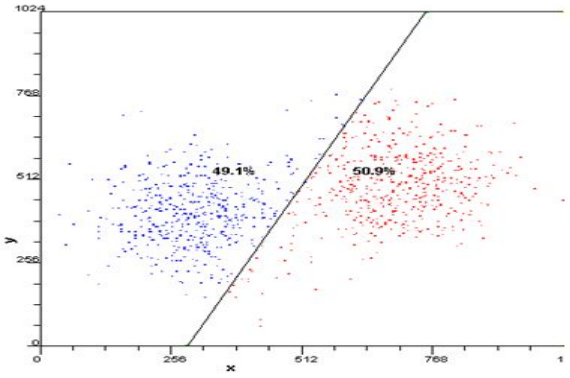

Figure 5 Sample of a 2-component Gaussian mixture

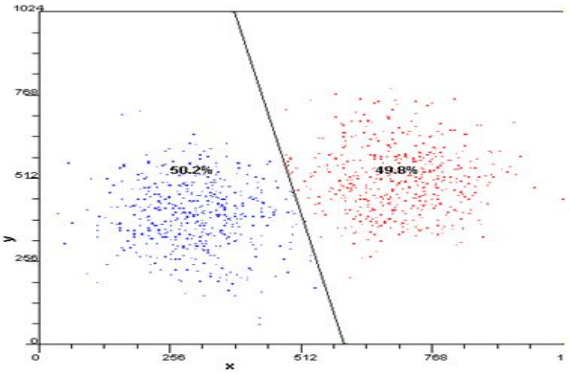

Figure 6 Optimized partition with clustering algorithm

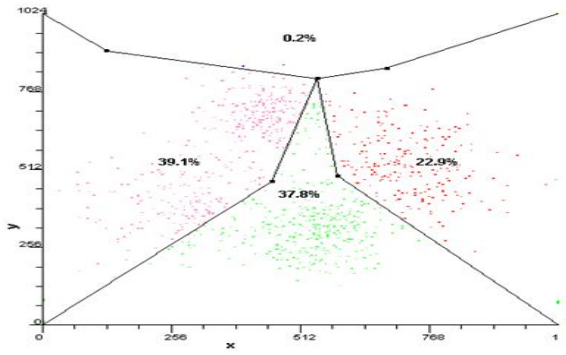

Figure 7 Sample of a 4-component Gaussian mixture

Similarly, a more complex example is shown in Figure 7 and Figure 8, Figure 7 shows a sample drawn from a 4-component

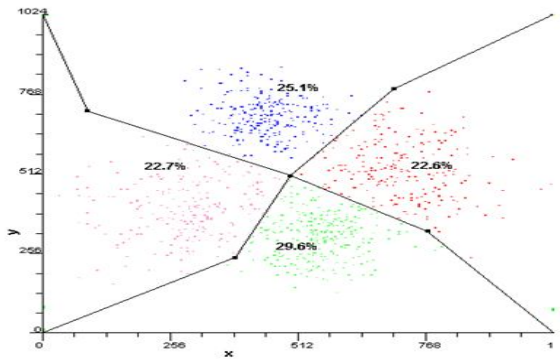

Figure 8 Optimized partition with clustering algorithm

Gaussian mixture model and an initial partition. Figure 8 shows the effects of the MCMC clustering algorithm.

These examples demonstrated the efficacy of the clustering algorithm in optimizing the clustering while maintaining the overall structure of the original partition.

A real world example involving flow cytometry data analysis is shown in Figure 9 and Figure 10. In current practice, a typical

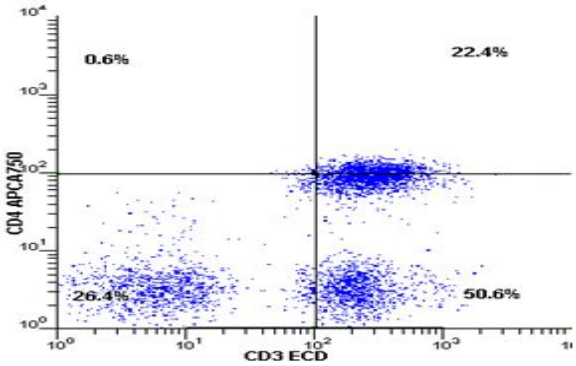

Figure 9 Example of flow cytometry data plot

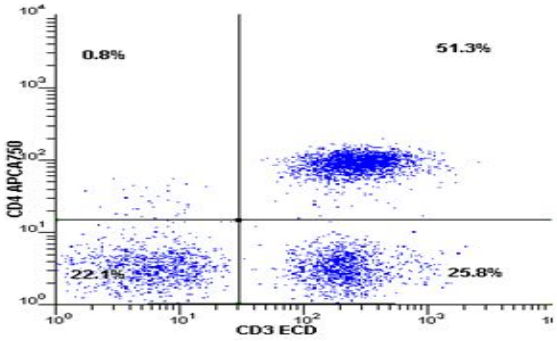

Figure 10 Automatic adjustment by clustering algorithm

flow cytometry analysis is performed with manual gating on 2D plots [1]. The manual process can often introduce inconsistency and inaccuracy. There have been proposals of automatic gating with statistical models [2]. However, the fully automatic process

Citation: Hongjun Su, Hong Zhang (2017) Graph Based Hybrid Clustering With Unbounded Regions. J Comp Sci Appl Inform Technol. 2(2):1-5

may not incorporate the prior information on the partition and can lead to improper clustering structures. Our proposed hybrid method attempts to provide a balanced solution. The partitioning algorithm offers a convenient way to specify the overall gating structure. The automatic clustering algorithm provides accurate and consistent results within the structural constraints of the gating partition. A plot of flow cytometry data with a manual gating specification is given in Figure 9. The clustering algorithm automatically determines a proper separation of the sub populations as shown in Figure 10.

Both of our proposed algorithms are computationally efficient. The complexity of the region finding algorithm is clearly bounded by $O(n^3)$ for a graph with n vertices.

The convergence of the MCMC algorithm is not easily determined theoretically. Since our algorithm is only acting on the sparse vertex set, the search space is much smaller than that of a typical clustering algorithm. Our experimental examples showed that typically the convergence occurred within a thousand iterations.

8 Conclusion

In this paper, we consider the extension of the hybrid system for partitioning and clustering to general unbounded regions. We use homogeneous coordinates to represent points at infinity and to define generalized polygons for representing possibly unbounded regions of a partition.

Two algorithms are extended to generalized polygons for automatic region detection and for optimal partitioning by clustering. The method provides an easy to use, intuitive interface for manual drawing of initial partitioning, and automatic algorithms for generalized polygon region identification and adjustment. The polygon detection algorithm is based on a planar graph model. The clustering algorithm uses an MCMC approach.

9 References

1. Howard M. Practical flow cytometry. John Wiley & Sons. 2005. doi: 10.1002/0471722731

2. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A. 2008;73(4):321-332. doi: 10.1002/cyto.a.20531

3. Jain AK , Dubes RC. Algorithms for clustering data. Prentice-Hall. 1988.

4. Su H and Zhang H. Hybrid clustering based on a graph model. Proceedings of Nineth International Symposium on Computational Intelligence and Design. 2016:242-245. doi:10.1109/ISCID.2016.1062

5. Stolfi J. Oriented Projective Geometry. Academic Press. 1991.

6. Bondy JA, Murty USR. Graph Theory with Applications. New York: Elsevier;1976.

7. Basener WF. Topology and its applications. John Wiley & Sons:Hoboken;2006.

8. Rourke JO. Computational Geometry in C. 2nd ed. Cambridge University Press:Canbridge;1998.

9. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machine. Journal of Chemical Physics 21. 1953:1087-1092.

10. Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 1970;57(1):97-109. doi: 10.2307/2334940

Citation: Hongjun Su, Hong Zhang (2017) Graph Based Hybrid Clustering With Unbounded Regions. J Comp Sci Appl Inform Technol. 2(2):1-5