A MODE-BASED CLUSTERING ALGORITHM WITHOUT MODE SEEKING

Esra Ataer-Cansizoglu, Deniz Erdogmus

Cognitive Systems Laboratory, Northeastern University, Boston, MA

ABSTRACT

Mode-based clustering approaches such as mean-shift and its variants are extremely successful. They are also computationally expensive due to their iterative hill-climbing strategy when determining cluster labels for samples. We identify the fact that mode-based cluster boundaries exhibit themselves as minor surfaces of the data distribution. Based on this observation, we develop a mode-based clustering methodology that does not involve iterative hill climbing for each sample. The method, instead, is based on searching for the presence of a minor surface on a path that connects pairs of samples. The pairwise data connections, when evaluated efficiently, may lead to a simple graph connectivity matrix based on which clusters can be identified using connected components. This search efficiency is achieved by an agglomerative clustering approach in the particular proposition presented in this paper. Illustrative experiments are carried out on synthetic datasets using Gaussian mixture models and kernel density estimates.

***Index Terms*—** Mode-based clustering, minor surface, cluster boundary

1. INTRODUCTION

Clustering is a fundamental problem in data analysis. There is a tremendous amount of work on clustering methods [1], many of which include mode seeking techniques. In mode-based clustering, each mode of the density model is considered as a cluster representative. Each sample is iterated towards its mode and the sample is assigned to the cluster which is represented by its mode.

Mean-shift [2, 3] is a popular mode-seeking algorithm which is non-parametric and iterative. In the literature many variations are proposed such as medoid-shift [4] and quick shift [5]. The advantages of using mode seeking algorithms are (i) the procedure is data driven and it is not required to know the number of clusters (ii) no need for a predefined step size for convergence to mode. However, the time complexity of the algorithm is large since a hill-climbing scheme is used for each sample. Many methods [6, 7] are proposed to accelerate mode seeking algorithms, which mostly suggest ways to decrease the number of iterations on hill climbing.

Eberly et al. [8] defined ridge and valley points in images by inspecting the relation between local gradients and Hessians. The definitions are generalized to principal and minor curves/surfaces of probability density functions (pdfs) by Ozertem and Erdogmus [9], and techniques to identify principal curves have been utilized in many domains such as signal denoising [10] and clustering [11]. Basically, a point is on minor surface if the local gradient and one or more eigenvectors of Hessian are orthogonal and corresponding eigenvalues are positive.

In this work, we are interested in clustering large amounts of data without iterating to the mode for each sample. The proposed approach exploits the fact that minor surfaces are the cluster boundaries for a mode-based clustering method. Instead of convergence to mode, a connectivity analysis is performed on pairs by searching for minor surfaces between them. Search for all pairs is avoided following an agglomerative clustering scheme.

This work is supported by NSF.

2. METHOD

In this section, we discuss the main idea by explaining the relation between minor hyper-surfaces and cluster boundaries. The details of the algorithm is presented based on this association.

2.1. Relation between Minor Surfaces and Cluster Boundaries

The proposed method is based on the fact that minor surfaces represent a natural boundary between clusters in a mode seeking clustering scheme. [1]

Assume we are given a pdf estimate $p(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{g}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x})$ be the local gradient and Hessian at $\mathbf{x}$. Let $\mathbf{L}(\mathbf{x})$ be the Hessian of the logarithm of the pdf, denoted by $\mathbf{H}_{log\ p}(\mathbf{x})$. Using the second order Taylor series expansion of $log\ p(\mathbf{x})$ we have

$$\mathbf{L}(\mathbf{x}) = -p^{-1}(\mathbf{x})\mathbf{H}(\mathbf{x}) + p^{-2}(\mathbf{x})\mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x}) \qquad (1)$$

The Hessian of the logarithm of the pdf is preferred in order for $\mathbf{L}(\mathbf{x})$ to become a quadratic when the pdf is Gaussian. This monotonic-increasing function applied to the pdf can be arbitrarily selected without a theoretical affect in the rest of this paper. Suppose $\{(\lambda_1(\mathbf{x}), \mathbf{q}_1(\mathbf{x})), \ldots, (\lambda_n(\mathbf{x}), \mathbf{q}_n(\mathbf{x}))\}$ are the eigenvalue-eigenvector pairs of $\mathbf{L}(\mathbf{x})$, where eigenvalues are sorted in ascending order: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let $\mathtt{M}^{n-1}$ denotes the $n-1$ dimensional minor surface. A point $\mathbf{x} \in \mathtt{M}^{n-1}$ iff $\mathbf{g}(\mathbf{x})^T\mathbf{q}_j(\mathbf{x}) = 0$ for some $j \in 1, ..., n$ and $\lambda_j(\mathbf{x}) > 0$. In other words, on the minor surface the local gradient and the eigenvector corresponding to an eigenvalue should be perpendicular and the associated eigenvalue should be positive. We are particularly interested in the case where $j = n$, i.e.the largest eigenvalue.

Mode seeking algorithms assign each sample to its mode by following gradient flow. Our method is based on the fact that no gradient trajectory can pass through a minor surface. Assume a point $\mathbf{y}' \in \mathbb{R}^n$ is in the vicinity of the point $\mathbf{y} \in \mathtt{M}^{n-1}$. If we think of the eigenvectors of $\mathbf{L}(\mathbf{y})$ as a local coordinate frame, the local gradient $\mathbf{g}(\mathbf{y}')$ should have two components: one parallel to the minor surface, the other orthogonal. The orthogonal component is along the eigenvector of $\mathbf{L}(\mathbf{y})$ that is perpendicular to $\mathbf{g}(\mathbf{y})$. The parallel component is in the span of the other eigenvectors whose span also contains $\mathbf{g}(\mathbf{y})$. The parallel component of $\mathbf{g}(\mathbf{y}')$ does not make the gradient flow trajectory passing through this point approach or diverge from the minor surface $\mathtt{M}^{n-1}$. However, the orthogonal component always points away from the minor surface and causes the

[1] We will use the phrase minor surface instead of minor hyper-surface.

gradient flow trajectory to diverge. Consequently, the gradient flow trajectory passing through $\mathbf{y}'$ will move the point away from the minor surface [2]. As a broader consequence of this observation, we note that gradient flow trajectories do not intersect (pass through) minor surface. Therefore, minor surfaces represent the natural geometric boundary for mode based clusters given a pdf model.

In this study, we use this fact to construct an agglomerative clustering methodology. Instead of seeking the mode for each sample, we search for minor surfaces between samples. If there exists a path between two samples that does not pass through minor surface, then these samples climb to the same mode. Thus, they belong to the same cluster.

2.2. Minor Surface Search between Samples

Based on the definition of minor surfaces, the following measure will help us to test whether a point $\mathbf{x}$ is on $\mathtt{M}^{n-1}$ or not.

$$\beta(\mathbf{x}) = \frac{\lambda_n \mathbf{g}(\mathbf{x})^T \mathbf{q}_n}{\|\mathbf{L}(\mathbf{x})\mathbf{g}(\mathbf{x})\|\|\mathbf{g}(\mathbf{x})\|} \tag{2}$$

If $\mathbf{x} \in \mathtt{M}^{n-1}$, then $\beta(\mathbf{x})$ will be 0 and $\lambda_n > 0$. For the points around a minor surface, β will attain a positive value which is close to zero. Similarly, for points on the principal curve β will be -1, and for points close to principal curve, it will give negative values that are close to -1. Thus we will use $\beta(\mathbf{x})$ and the largest eigenvalue $\lambda_n(\mathbf{x})$ to test whether there exists a minor surface in between two samples. Basically, we perform a line search with a small step size γ and search for a point which has a β value that is smaller than a predefined threshold thr [3] and a positive λ_n. Thresholding might be problematic, since we can break off some necessary connections. For example, even though there does not exist a point which gives a β value of zero on a line, due to thresholding we can decide not to connect the samples. As our primary aim in this paper is to demonstrate the concept of using minor surfaces for clustering, we accept some potential missing edges due to this algorithmic imperfection. As a future extension, we will seek to develop improved strategies for testing for the presence minor surface intersections.

If a sample point is detected as being in close proximity of a minor surface, then that point is not connected to any of the samples in the dataset due to the test described above. Hence, we perform a post-processing step for these samples that are close to minor surfaces based on the β-test. For each such point, mean-shift iterations are performed until it gets connected to one of the current subgraphs.

In our graph representation, we add an edge between samples when there is no minor surface on the line that is connecting them. Two samples belong to same cluster if there is a piecewise linear path that does not pass through any minor surfaces. Therefore final clustering is based on connected component analysis. Two samples are assigned to the same cluster if they belong to the same subgraph on the graph representation. We follow an agglomerative clustering scheme and start with a graph having N nodes with no edges where N is the number of samples. Then we perform tests on pairs starting from the closest ones. We avoid testing for unnecessary pairs based on our cluster growing methodology. At the execution of the algorithm, if we create an edge between two samples, then their clusters are merged together. Similarly, we skip testing two samples that belong to same cluster. The details of the algorithm are presented in Algorithm 1. Note that, we can also avoid testing for pairs that are too far from each other by enforcing a tolerance value for pairwise distance.

[2] Detailed proof is omitted due to lack of space, but this proof sketch clearly illustrates the concept. Proofs will be included in future journal publications.

[3] Ideally, we need to identify if a point that achieves 0 exactly exists in this interval. Iterative root searching is possible but still there is no guarantee that we will get exactly the point of interest if it exists. In this initial design, we use a threshold value which is positive and close to 0.

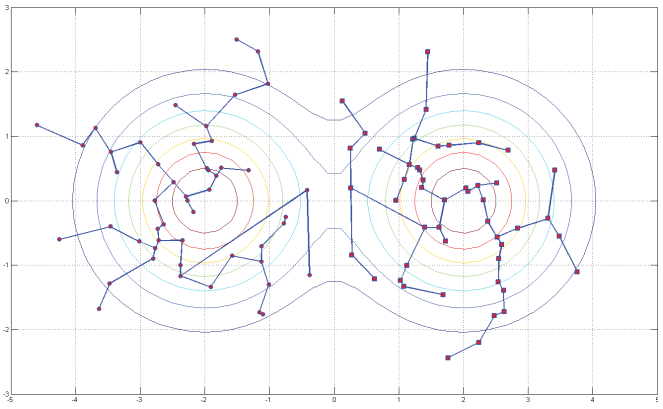

Fig. 1. The results on a mixture of two Gaussians with equal weights and identity covariances. The means of two mixtures are at $(-2, 0)$ and $(2, 0)$.

3. EXPERIMENTS AND RESULTS

We have simulated different datasets to run our experiments on. The first set of experiments are performed on data generated from Gaussian Mixture models (GMMs). Our goal is to better visualize the results on a known density where we know the location of the minor curve - the cluster boundary in 2-dimensional examples. Figure 1 displays the results for a set of samples that are generated from a mixture of two Gaussians with equal weights and identity covariances. The generated samples are successfully clustered into two groups. As seen in the figure, connectivity test outputs two sparse subgraphs and there is no connection passing through the minor curve lying at $x = 0$. Figure 2 displays the result for a mixture of three Gaussians with equal weights and identity covariances, organized into an equilateral triangle. Here there are minor curve segments between each pair of clusters designating the separation boundary. The technique gives accurate results.

In the second set of experiments, Kernel Density Estimation (KDE) using Gaussian kernels is exploited. We used leave one out log-likelihood cross validation to find the optimal kernel width, where isotropic fixed-bandwidth kernels are used. Some samples that are close to a minor curve cannot be connected to any other samples during their initial connectivity test. These samples are connected successfully, after a few mean-shift iterations. Figure 3 displays two circular clusters, where no connectivity is assigned to samples of pairs corresponding to different clusters. The number of clusters is correctly found as two. Clustering on another set of samples is seen on figure 4. Notice that since we start connectivity test on samples that are close to each other, we do not perform any test on samples that are at the tips of the same crescent. Moreover, the samples at the tips of the same cluster are not considered connected due to the minor curve passing through the line connecting them. Although they are not directly connected by an edge in the graph, based on the agglomerative strategy, they are successfully put into the same cluster through connecting paths on the graph.

Algorithm 1 Clustering algorithm for given set of samples $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, a pdf, a step size γ for line search and a threshold value thr for testing β. The outputs are the cluster labels $c_1, c_2, \ldots c_N$ and $N \times N$ edge matrice E for graph representation. $MinorSurfaceTest(\mathbf{x}_a, \mathbf{x}_b)$ function returns true when there is no minor surface on the line connecting $\mathbf{x}_a$ and $\mathbf{x}_b$. It returns false otherwise.

{initialize cluster labels, start with N clusters}
$c_i \leftarrow i \; \forall i \in \{1, 2, \ldots N\}$
{Initialize edge matrice}
$E(i, j) \leftarrow 0 \; \forall i, j \in \{1, 2, \ldots N\}$
{Compute pairwise distance matrix D}
$\mathbf{D}(i, j) \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\| \; \forall i, j \in \{1, 2, \ldots, N\}$
{pairs sorted according to distances in ascending order}
$\{f_1, f_2, \ldots f_{N^2}\} = \{\{1, 2, \ldots N\} \times \{1, 2, \ldots N\}$ and $\mathbf{D}(f_i) \leq \mathbf{D}(f_j)$ iff $i \leq j\}$
{Keep a set S containing the points detected on the minor surface}
$S \leftarrow \{\}$
for $(a, b) = f_1 \rightarrow f_{N^2}$ **do**
  if $a < b$ AND $c_a \neq c_b$ AND $\{a, b\} \notin S$ **then**
    if $MinorSurfaceTest(\mathbf{x}_a, \mathbf{x}_b)$ **then**
      {Merge clusters and connect the nodes.}
      $c_i \leftarrow c_b \; \forall i : c_i = c_a$
      $E(a, b) \leftarrow 1$
      $E(b, a) \leftarrow 1$
    else
      if $0 \leq \beta(\mathbf{x}_a) \leq thr$ AND $\lambda_n(\mathbf{x}_a) > 0$ **then**
        $S \leftarrow S \cup \{a\}$
      end if
      if $0 \leq \beta(\mathbf{x}_b) \leq thr$ AND $\lambda_n(\mathbf{x}_b) > 0$ **then**
        $S \leftarrow S \cup \{b\}$
      end if
    end if
  end if
end for
{Post-processing on samples in S}
for $\mathbf{k} \in S$ **do**
  $\mathbf{x} \leftarrow \mathbf{x}_k$
  while $\nexists j \in \{1, 2 \ldots, N\} : MinorSurfaceTest(\mathbf{x}, \mathbf{x_j})$ **do**
    {iterate with the mean-shift update $msupdate(\mathbf{x})$}
    $\mathbf{x} \leftarrow \mathbf{x} + msupdate(\mathbf{x})$
  end while
  $c_k \leftarrow c_j \mid MinorSurfaceTest(\mathbf{x}, \mathbf{x}_j)$
end for

procedure: MinorSurfaceTest($\mathbf{x}_a, \mathbf{x}_b$)
$\mathbf{e} \leftarrow \frac{\gamma(\mathbf{x}_b - \mathbf{x}_a)}{\|\mathbf{x}_b - \mathbf{x}_a\|}$
$\mathbf{x} \leftarrow \mathbf{x}_a + \mathbf{e}$
while $\|\mathbf{x} - \mathbf{x}_b\| \geq \gamma$ **do**
  if $0 \leq \beta(\mathbf{x}) \leq thr$ AND $\lambda_n(\mathbf{x}) > 0$ **then**
    return false
  end if
  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{e}$
end while
return true

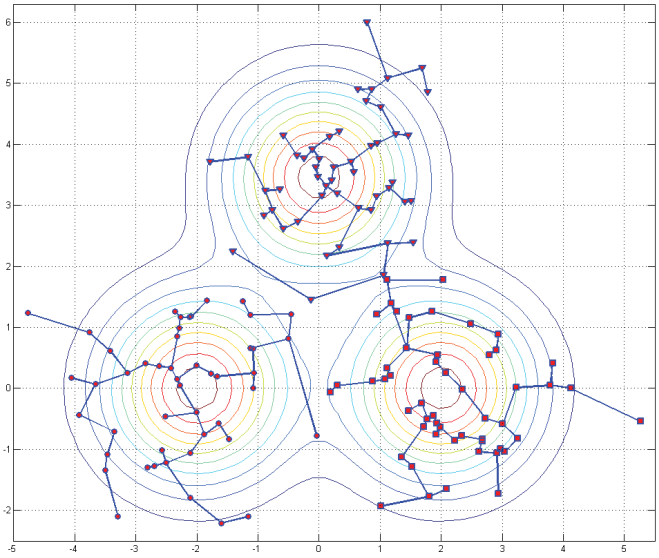

Fig. 2. The results on a mixture of three Gaussians with equal weights and identity covariances. The means of the mixtures construct an equilateral triangle with an edge length of 4. Resulting three clusters are illustrated with circle, square and triangle nodes.

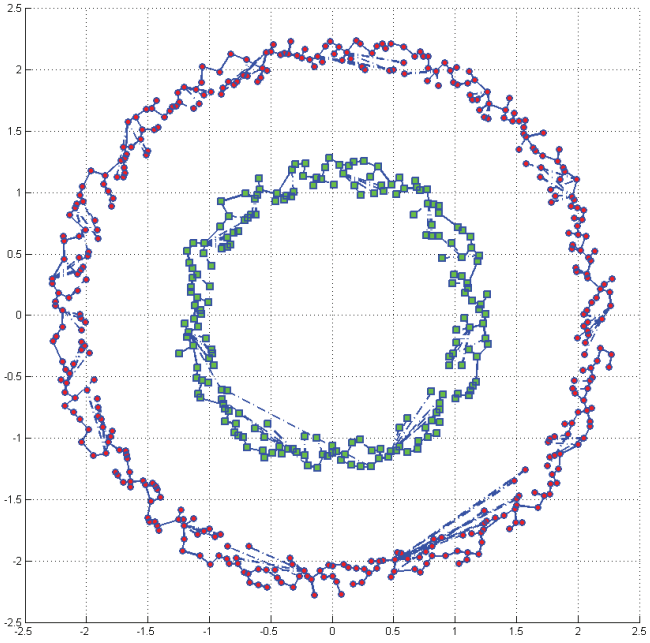

Fig. 3. The results on circles data. Dashed lines indicate the connections that are made after some mean-shift iterations. Two clusters are displayed with red circles and green squares.

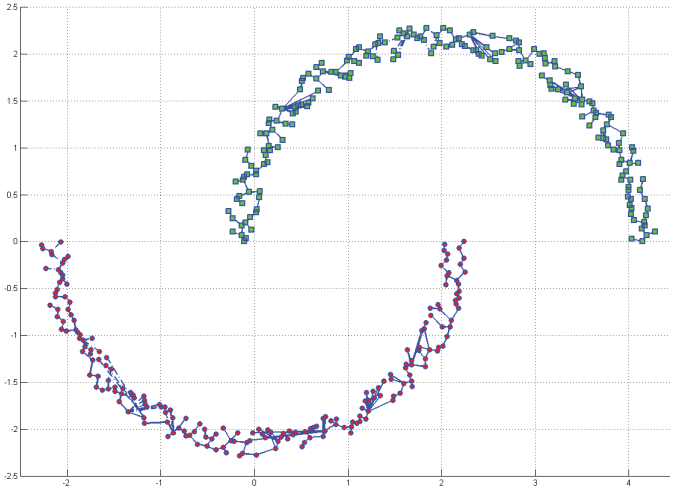

Fig. 4. The results on the crescents dataset. There are two connected components illustrated with red circles and green squares. Dashed lines are the connections made after some mean-shift iterations.

4. CONCLUSION AND DISCUSSION

In this study, we proposed a mode-based clustering algorithm which does not involve mode seeking. The proposed technique is based on minor surfaces forming the boundary between clusters. Instead of following a hill climbing strategy for each sample, we search for minor surface between samples. We propose a metric which discriminates minor surface points from the others. The pairs which do not contain any minor curve on the line connecting them are considered to be in the same cluster. Following an agglomerative clustering scheme, clusters are merged when a minor surface test between them fails. The experiments are carried out on datasets where low-dimensional GMM and KDE models are used for illustrative purposes. The results show that the method is successful on clustering samples based on minor surface test. In the future, we would like to extend our work by using an improved search strategy for the minor surface intersection test. Further improvements to handle large number of samples and high dimensional data will also be considered.

5. REFERENCES

[1] Rui Xu and II Wunsch, D., "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645 –678, 2005.

[2] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32 – 40, 1975.

[3] Yizong Cheng, "Mean shift, mode seeking, and clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 790 –799, 1995.

[4] Takeo Kanade Yaser Ajmal Sheikh, Erum Arif Khan, "Mode-seeking via medoidshifts," *IEEE International Conference on Computer Vision*, 2007.

[5] Andrea Vedaldi and Stefano Soatto, "Quick shift and kernel methods for mode seeking," in *Computer Vision ECCV 2008*, vol. 5305 of *Lecture Notes in Computer Science*, pp. 705–718. 2008.

[6] Miguel A. Carreira-Perpinan, "Acceleration strategies for gaussian mean-shift image segmentation," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, pp. 1160–1167, 2006.

[7] Kai Zhang, Jamesk Kwok, and Ming Tang, "Accelerated convergence using dynamic mean shift," in *Computer Vision ECCV 2006*, vol. 3952 of *Lecture Notes in Computer Science*, pp. 257–268. 2006.

[8] D. Eberly, R. Gardner, B. Morse, S. Pizer, and C. Scharlach, "Ridges for image analysis," *J. Math. Imaging Vis.*, vol. 4, pp. 353–373, 1994.

[9] Umut Ozertem and Deniz Erdogmus, "Locally Defined Principal Curves and Surfaces," *Journal of Machine Learning Research*, pp. 1249–1286, 2011.

[10] Umut Ozertem and Deniz Erdogmus, "Signal denoising using principal curves: Application to timewarping," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3709–3712.

[11] Erhan Bas and Deniz Erdogmus, "Sampling on locally defined principal manifolds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.