A Novel Clustering Algorithm Based on Bayesian Sequential Partition and Its Application in Image Segmentation

Ruijie Zhou RUIJIE@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Abstract

In this work, we propose a novel clustering algorithm based on Bayesian Sequential Partition (BSP) and the spectral clustering algorithm. Since the BSP is capable of providing much more accurate density estimates when the sample space is of moderate to high dimension, the proposed clustering algorithm is believed to be superior to available ones when dealing with high dimensional problems. To demonstrate the effectiveness and efficiency of our novel clustering method, we have compared our new algorithm with the original spectral clustering algorithm and the mean-shirt method in several image segmentation problems. Based on the results from our experimental studies, it can be shown that our new algorithm is indeed much more powerful.

1. Introduction

Clustering is a very powerful class of algorithm that divides data into groups for the purposes of improving understanding. While a large number of clustering techniques have been developed, significant challenges still remain, especially for high dimensional data. The principal challenge in extending cluster analysis to high dimensional data is to overcome the curse of dimensionality and the ways in which high dimensional data are different from low dimensional data (such as the correlation of attributes), and how these differences might affect the process of cluster analysis. Clustering in high dimension depends on reliable and accurate density estimate of high dimensional data and one very promising methods for this purpose is the Bayesian Sequential Partition (BSP). The theoretical foundations of BSP was first proposed in the seminal work by Luo et al.(L. Luo & Wong, 2013 Published Online). Densities estimation based BSP exploits sequential importance sampling to explore the space of simple functions based on binary partitions. BSP first relies on is a closed form expression for the posterior probability of a binary partition and then a computational efficient procedure is introduced to maximize this posterior probability, which is equivalent to minimizing the KL divergence between the true density and the histogram built by BSP. Compared to traditional approaches such as the kernel method or the histogram, the BSP is more capable of providing accurate estimates when the sample space is of moderate to high dimension.

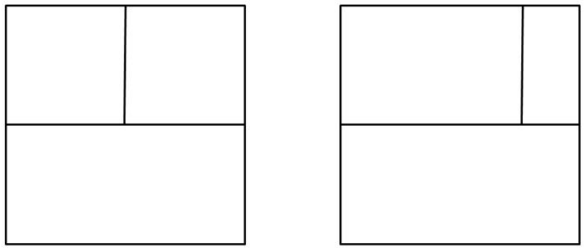

Figure 1. BP Example (The left partition is a binary partition, but the right partition is not)

The theory of BSP was established in the paper by Luo and coauthors. However, to the best knowledge of the author, there has been no direct applications of BSP in clustering analysis. In this work, we will combine the BSP with the spectral matting algorithm (A. Levin & Lischinski, 2008) to design a novel clustering method. We expect that our new algorithm performs better than available algorithms in dealing with high dimensional data. To describe and demonstrate our algorithm, the rest of this paper is organized as follows. Our novel BSP clustering algorithm is first introduced in Section 2. In Section 3, three examples on image segmentation are studied to demonstrate the

CS229 Final Project

validity and advantages of the proposed algorithm. Finally, the conclusion of this work is provided and some possible future research topics are suggested in Section 4.

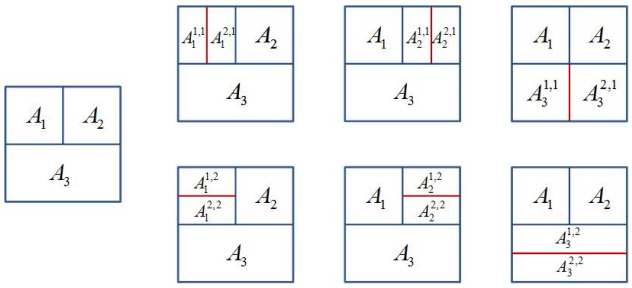

Figure 2. An example of partition construction with d=2 and t=3. The next partition will be determined by maximizing the conditional probability.

2. Algorithm and Methodology

The BSP-based clustering has two consecutive steps. In the first place, BSP will be applied to estimate the density of the feature space and make appropriate partitions to the sample space. Then, a graph will be built based on all the partitions we have and spectral clustering will be applied. We will describe each step in the sequel.

2.1. Data-driven Partitions

The idea of BSP is to use the class of simple function to approximate the densities in high dimension. However, this class of function is too large to serve as the building blocks for density estimates. To derive computationally tractable algorithm, we only consider the simple function that can be realized by binary partitions. Specifically, a binary partition of size $k+1$ is a partition derived from a BP of size k by further performing a bisection of one of its regions along one of the coordinates. Figure 1 demonstrates a binary partition versus a partition which is not a binary partition in a two dimensional rectangle.

Let $T = \bigcup_{i=0}^{t} A_i$ be a rectangle in $\mathcal{R}^d$ and let f be a simple function defined over T. To construct a prior for f, we assume that the partition of size t has a prior density proportional to $\exp(-t)$ and all partitions that have the same size have the same prior densities. Assume the probability of each region be $\mathbf{p} = \{\theta_1, ..., \theta_t\}$, which has a Dirichlet distribution with parameters $(\alpha, ..., \alpha)$. Let $|A_k|$ and n_k be the volume and number of data points in region k, the posterior distribution of partition given data points can be written as follows:

$$
\begin{aligned}
&P(T|D) \propto P(T)P(D|T) \\
&= P(T)\int_{\mathbf{p}} P(D|\mathbf{p},T)\times P(\mathbf{p}|T)d\mathbf{p} \\
&\propto e^{-\beta t}\int \prod_{k=1}^{t}(\frac{\theta_k}{|A_k|})^{n_k}(\frac{1}{D(\alpha,...,\alpha)}\prod_{j=1}^{t}\theta_j^{\alpha-1})d\theta_1...d\theta_t \\
&\propto e^{-t}\frac{D(n_1+\alpha,...,n_t+\alpha)}{D(\alpha,...,\alpha)}\prod_{k=1}^{t}(\frac{1}{|A_k|})^{n_k}
\end{aligned}
$$

where $D(\delta_1, ..., \delta_t) = \frac{\prod_{j=1}^{t}\Gamma(\delta_j)}{\Gamma(\sum_{j=1}^{t}\delta_j)}$. It should be noted that the above posterior probability of partition is valid for both discrete and continuous data. For discrete data, we just need to replace $|A_k|$ with the number of possible data points in A_k. For each specific partition T, we then define the partition score $s(T)$ as the logarithm of the above posterior probability. Specifically, we have $s(T) = \log \pi(T)$, where

$$
\pi(T) = Ce^{-t}\frac{D(n_1+\alpha,...,n_t+\alpha)}{D(\alpha,...,\alpha)}\prod_{k=1}^{t}(\frac{1}{|A_k|})^{n_k}
$$

where C is the normalization constant. The optimal partition is obtained by maximizing the above partition score and the optimal number of partitions will be determined automatically by the algorithm. From the above expression, it can be found that $\exp(-t)$ can be viewed as a penalizing factor and this can actually prevent our model from overfitting.

2.2. Partition Constructions

In this part, we will introduce how the partitions are constructed. The idea is to use the sequential important sampling to build up the partitions from low dimension to high dimension. Let $x_1, ..., x_t$ denote the cut of each step, then we have:

$$
\begin{aligned}
\pi(x_1,...,x_t) &= \pi(x_1)\pi(x_2|x_1)...\pi(x_t|x_{t-1}) \\
&= \pi(x_1)\frac{\pi(x_1,x_2)}{\pi(x_1)}...\frac{\pi(x_1,...,x_t)}{\pi(x_1,...,x_{t-1})}
\end{aligned}
$$

Regarding the choice of x_i at step i, we generate partitions randomly and find the one with large posterior probability which is used to construct the histogram. This idea is the illustrated in Figure 2. We will always find partition x_3 that maximize the conditional probability of x_3 given x_2. It can be shown that we actually need to maximize:

$$
P(x_3|x_2) = C(x_2)\frac{\Gamma(\frac{1}{2}+n_j^{1,d})\Gamma(\frac{1}{2}+n_j^{2,d})}{\Gamma(\frac{1}{2}+n_j)}\frac{|A_j|^{n_j}}{|A_j^{1,d}|^{n_j^{1,d}}|A_j^{2,d}|^{n_j^{2,d}}}
$$

where $C(x_2)$ is a constant term depending on previous partition x_2. Also, it is worthy of note that most of the operations needed to find x_{i+1} given x_i is to count the number of data points in the subregion of each possible cut.

Figure 3. Illustration of feature space partition by BSP. In the figure below, it can be seen that boundary parts are clearly separated.

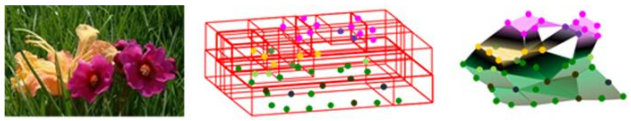

Figure 4. Graph built after applying BSP. It can be found that components with high degree of affinity are connected.

2.3. Spectral Clustering

Given a d dimensional feature space, we first apply the BSP based on Section 2.1 and 2.2 to partition the region into N subregions. Let $\mathbf{c_i} \in R^d$ for $i = 1, ..., N$ denote the center point of each subregion. Then, we can use the method proposed by Levin to build an undirected graph, where each edge represents the affinity of the partition, and apply spectral clustering. Note that the spectral clustering method and the way to build the undirected graph are the same as the method used by Levin (A. Levin & Lischinski, 2008). However, with the aid of BSP, it is expected better performance can always be obtained. We will analysis the result from some experimental studies in the next section.

3. Experimental Studies

In this section, we will demonstrate our novel clustering on several image segmentation problems. Image segmentation is the process of partitioning a digital image into multiple segments, which simplify and change the representation of an image into something that is more meaningful and easier to analyze. More precisely,

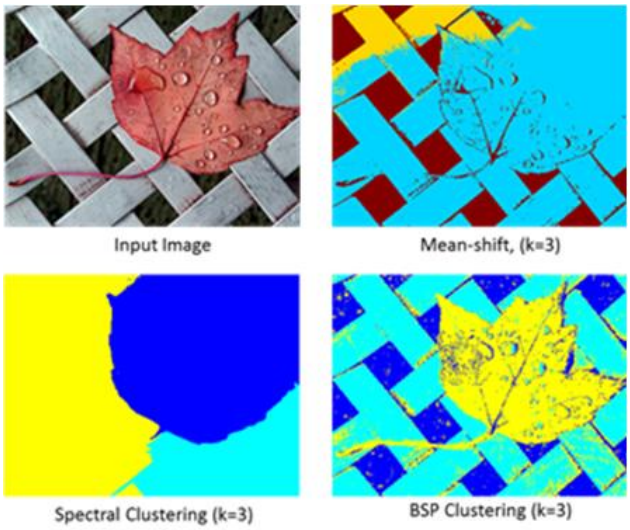

Figure 5. Experimental Result with $k = 3$: We can see that the matting Laplacian failed to capture the essential information in the original image, while the mean-shift method missed many important features. BSP generates 644 partitions for this case and its performance is better.

image segmentation is to assign a label to every pixel in an image such that pixels with the same label share certain visual characteristics.

Now, given a digital image and we define its feature vector as (r, g, b, x, y), where (r, g, b) represents the color and (x, y) denotes the spatial information. We first apply BSP to perform partitions on this feature space and this process is illustrated in Figure 3. Next, we apply Levin's method to construct the undirected graph and this procedure is illustrated in Figure 4. To illustrate the merit of our proposed clustering algorithm, we apply our novel clustering algorithm and compare its performance with the spectral matting algorithm (A. Levin & Lischinski, 2008) and the mean-shift method (Comaniciu & Meer, 2002).

Figures 5-7 show the original input images and the corresponding segmentation result from three different algorithms. We have tested our method on three different level of k values ($k = 3, k = 4$ and $k = 12$). In the results, we labeled the pixels in the same group with the same color. It can be observed that our proposed clustering method is much more better than the other two existing approaches. In particular, for spectral matting algorithm, it is very difficult to determine the optimal local window size and by defining affinity the algorithm will inevitably collapse the high dimensional features into one similarity value. By using BSP, we can get avoid of the problem of selecting the size

of widow as BSP affords us with a highly reliable and accurate way for pre-grouping based on density estimation. Moreover, the kernel-based mean-shift is also less effective than BSP and we can conclude it is not suitable for clustering even with a five dimensiona feature space.

Finally, it is worthy noting that computational complexity of BSP is not high. It has been proved that computational complexity of the partitions is linear to the sample size. For all three images we have, the partitions can be obtained within less than 6 seconds in a laptop with a 2.90 GHz Intel Core i7-3520M Processor in Visual Studio 2012 environment.

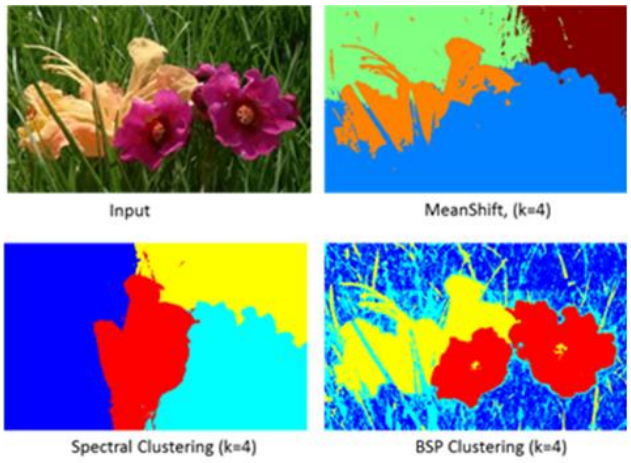

Figure 6. Experimental Result with $k = 4$: As before, the spectral clustering did very bad and mean-shift was better. BSP in this case generated 681 regions and was significantly better the other two methods.

4. Conclusions and Future work

In our work, we have presented a novel clustering algorithm based on the BSP and spectral clustering. Compared with other available methods, the proposed method performs significantly better for medium or high dimensional problems. Several image segmentation problems have been studied to demonstrated the advantages of our clustering method.

The clustering technique introduced in this paper is a general tool for more challenging high dimensional problem. Also, the method is not restricted to the image segmentation discussed here. In our experimental studies, we only use our clustering algorithm for low level segmentation. In other words, our clustering algorithm is performed based only on the colors and the spatial information of an image. A very obvious extension is to work on semantic segmentation by including more features, such as textures. Since BSP is an accurate and efficient algorithm for high dimensional density estimation and feature space partition, it is expected that the proposed clustering algorithm will yield better performance than others.

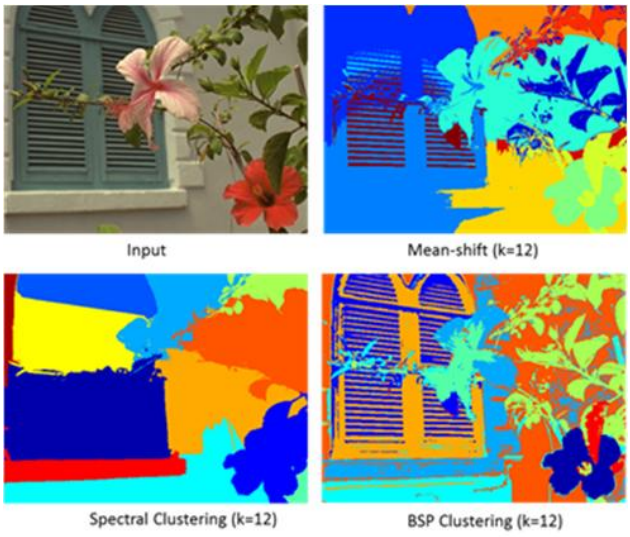

Figure 7. Experimental Result with $k = 12$: For large k, spectral clustering still worked poorly. Mean-shift was much more better and it captured many important features for part of the image. BSP generated 748 partitions and again outperformed the other two.

Acknowledgments

The author of this report would like to acknowledge the help from Tung-yu Wu for understanding BSP. Also, the author would like to acknowledge Prof. Andrew Ng for inciting his interest in machine learning.

References

A. Levin, A. Rav-Acha and Lischinski, D. Spectral matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1699–1712, 2008.

Comaniciu, D. and Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

L. Luo, H. Jiang and Wong, W. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association*, 2013 Published Online.