Enhanced Fuzzy based Affinity Propagation for Multispectral Images

Venkata Ramana K & Sailaja Vishnubhatla
Computer Science and System Engineering
Andhra University, India

Abstract: Affinity Propagation (AP) clustering algorithm that is a new clustering algorithm based on message passing. The standard affinity propagation clustering algorithm suffers from two limitations. One that it is hard to know the value of the parameter “preference” which can yield an optimal clustering solution. Two Euclidean similarity measure is used which can give only ellipsoidal clusters. To overcome the first limitation, in this paper an adaptive affinity propagation method. The method first finds out the range of “preference”, then searches the space of “preference” to find a good value which can optimize the clustering result. The second limitation is overcome by using Fuzzy Statistics Similarity Measure.

1. Introduction

Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super-pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

To partition multispectral and hyperspectral feature spaces for extracting clusters of patterns that can be associated with land-cover classes clustering techniques can be used. A common approach is to use data to learn a set of centers such that the sum of squared errors between data points and their nearest centers is small. When the centers are selected from actual data points, they are called “exemplars”. In order to mathematically identify clusters in a dataset, it is usually necessary to first define a measure of similarity, which establishes a rule for assigning patterns to the domain of a particular cluster center (or exemplar). This similarity measure places similar data close to one another to form a group, thus generating different clusters.

1.1. Methods of Clustering Multispectral Images

K-Means discovered as one of most well-known methods for clustering, is developed by Mac Queen in 1967. The simplicity of K-means made this algorithm could be implemented in various cases that has a big size data. K-means is a partition clustering method that separates data into k groups. K-means algorithm has many advantages such as simplicity, low computational complexity. Therefore, it is widely used in remote sensing. However, K-means is sensitive to initialization, and to the choice of the number of clusters, which usually is a critical issue. Different random initializations of the cluster centers result in significantly different clusters at the convergence. Thus, the algorithm is usually run many times with different initializations in an attempt to find a good solution. In addition, K-means is prone to find clusters with spherical shape, and it is sensitive to noisy data. Fuzzy K-means algorithm is the extension of crisp K-means. It has been shown to have a better performance than K-means due to its ability to deal with uncertain situations. One of the most significant advantages of fuzzy K-means is that it more naturally handles situations in which subclasses are formed by mixing extreme examples. Fuzzy K-means has been widely used in remote sensing, and many algorithms are derived from it. However, these foregoing algorithms have similar drawbacks when used in remote-sensing imagery analysis.

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means. It can find arbitrarily
shaped clusters. DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed. DBSCAN cannot cluster data sets well with large differences in densities.

Mean Shift does not require the number of clusters to be specified as an input. Mean shift cleverly exploits the density of the points in an attempt to generate a reasonable number of clusters. The kernel bandwidth value can often times be chosen based on some domain-specific knowledge. Due to its edge preserving filtering property the salient features of the overall image are retained. This property is important for segmenting remotely sensed images in which several distinct regions are used to represent the whole scene. The most glaring disadvantage is its slowness. More specifically, it is an $N^2$ algorithm. For problems with many points, it can take a long time to execute. The one silver lining is that, while it is slow, it is also embarrassingly parallelizable, as each point could be shifted in parallel with every other point.

A new clustering approach named affinity propagation [1] (AP) has been proposed recently. AP clustering method performs well in many applications such as image categorization [2], gene expressions and text summarization [3] and speaker clustering [4]. This paper applies it to document clustering. Unlike other methods, AP simultaneously considers all data points as potential exemplars. AP recursively transmits real-valued messages along edges of the network by viewing each data point as a node in a network until a good set of centers is determined. Rather than requiring that the number of clusters be pre-specified, affinity propagation takes as input a real number $s(k, k)$ for each data point $k$ so that data points with larger values of $s(k, k)$ are more likely to be chosen as exemplars. These values are referred to as "preferences," which is a kind of the self-similarity. The number of identified exemplars is influenced by the values of the input preferences. Frey suggested preference be set as the median of the input similarities $(p_m)$ without any prior knowledge. But in most cases, $p_m$ can’t lead to optimal clustering solutions. The use of Euclidean distance as a measure of similarity is not very suitable for remote-sensing clustering, because the scatter diagram of multispectral remote-sensing data tends to hyperellipsoid distributions in the feature space, owing to uncertainty and existence of mixed pixels.

### 2. Proposed Clustering Method

Multispectral and hyperspectral remote-sensing images often have extensive interband correlations. As a result, the images may contain similar information and have similar spatial structures [4]. At the same time, multispectral and hyper spectral images have their own special characteristics, namely, the spatial variability of the spectral signature. According to this, a statistical characteristics is introduced which are based on fuzzy statistics.

The fuzzy set is defined by

$$ F = \{x_1, z_1; \mu_1(1)\} \cup \{x_1, z_2; \mu_1(2)\} \cup \{x_n, z_m; \mu_m(m)\} $$

Fuzzy mean distance is computed using the membership degree. It represents the distance between two pixels and is defined as

$$ dis_i(j) = \frac{\sum_{k=1}^{p} \mu_i(k) \mu_j(k)}{\sum_{k=1}^{p} \mu_i(k)} $$

where $\mu_i(k)$ is the $i$th pixel vector $\mu_i = [x_i, z_i]$ and $\mu_j(k)$ is the $j$th pixel vector $\mu_j = [x_j, z_j]$. The mean distance deviation of the $i$th pixel vector is defined as

$$ \text{Dev}_i = \frac{\sum_{j=1}^{p} \mu_i(k) \mu_j(k)}{\sum_{k=1}^{p} \mu_i(k)} $$

The membership degree of the fuzzy mean distance is computed using

$$ \mu_i(j) = \exp(- \text{dev}_i(j)^\beta) $$

where $\beta$ is a parameter determining the scalar of $\mu_i$ and ranges from $0, +\infty$. It determines the degree of fuzziness of the final solution, which is the degree of overlapping between groups. If $\beta = 0$, the solution is a hard partition. As $\beta$ becomes close to infinity, the solution approaches its highest degree of fuzziness. $\beta$ is aimed to accommodate the outliers in a special class to decrease their effect on clustering. The membership degree depends on distance deviation in the spectral space.

The FSS between pixel vector $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ and clustering exemplar $z_k = [z_{k1}, z_{k2}, \ldots, z_{kp}]$ can be computed using the following

$$ \text{FSS} = s(i, k) = -\text{dis}(i, k) $$

Affinity propagation tries to maximize the net similarity [6]. Net similarity is a score for
explaining the data, and it represents how appropriate the exemplars are. The score sums up all similarities between data points and their exemplar (The similarity between exemplar to itself is the preference of the exemplar). Affinity propagation aims at maximizing Net Similarity and tests each data point whether it is an exemplar. To solve the problem of computing the preference, this paper proposes an Enhanced affinity propagation clustering algorithm to determine the range of preference, searches for the optimal value and then applies it to clustering image. Searches the space of preferences $[-\infty, p_{max}]$, to maximize the value of Silhouette [5].

**Algorithm 1 Preference Range Computing**

**Input:** $s(i, k)$: the similarity between point I and point k ($i \neq k$)

**Output:** the maximal value and minimal value of preferences: $p_{max}, p_{min}$

Step 1. Initialize $s(k, k)$ to zero:

$$s(k, k) = 0$$

Step 2: Compute the maximal value of preferences:

$$p_{max} = \max\{s(i, k)\}$$

Step 3: Compute the minimal value of preferences:

Step 3.1 Compute the net similarity when the number of clusters is 1:

$$dpsim_1 = \max\{\sum_j s(i, j)\}$$

Step 3.2 Compute the net similarity when the number of clusters is 2:

$$dpsim_2 = \max\{\max_i \sum_k \{s(i, k), s(j, k)\}\}$$

Step 3.3 Compute the minimal value of preferences $p_{min} = dpsim_1 - dpsim_2$

The maximum preference ($\max p$) in the range is the value which clusters the N datapoints into N clusters, and this is equal to the maximum similarity, since a preference lower than that would make the object better to have the data point associated with that maximum similarity assigned to be a cluster member rather than an exemplar. The derivation for $p_{max}$ is similar to $p_{max}$. After computing the range of preferences, we can scan through preferences space to find the optimal clustering result. Different preferences would lead to different cluster results. Cluster validation techniques are used to evaluate which clustering result is optimal for the datasets. In order to sample the whole space, we set the base of scanning step as $p_{max} - p_{min}/N$. However, this fixed increasing step cannot meet the different requirement of different cases such as more clusters and less clusters. Because more-clusters case is more sensitive than that of less-cluster case. We adopt the adaptive step method similar to Wang's [8], an adaptive coefficient is introduced. The preference step can be given as

$$q = 1.01 \times \sqrt{k + 50}$$

In this way, we set the value of $p_{step}$ with the count of clusters ($k$) dynamically. When $k$ is large, $p_{step}$ will be small and vice versa. The largest global silhouette index ($sil$) indicates the best clustering quality and the optimal number of clusters. A series of $sil$ values corresponding to clustering result with different number of cluster are calculated. The optimal clustering result is found when is $sil$ largest.

In this way, we set the value of $p_{step}$ with the count of clusters ($k$) dynamically. When $k$ is large, $p_{step}$ will be small and vice versa. The largest global silhouette index ($sil$) indicates the best clustering quality and the optimal number of clusters. A series of $sil$ values corresponding to clustering result with different number of cluster are calculated. The optimal clustering result is found when is $sil$ largest.

![Figure 1. FlowChart For EAP-FSM](image-url)
3. Conclusion

In this paper, a novel EAP-FSM clustering method has been presented and implemented. The key concepts contained in the EAP-FSM include fuzzy mean deviation and FSS. FSS allows the algorithm to assign proper memberships to uncertain data and can get an accurate and objective estimate of how closely two pixel vectors resemble each other. This ensures high robustness against noise and involves accurate clustering results in the case of mixed (or complex)

4. References


