A Symmetry Based Clustering Technique For Multi-spectral Satellite Imagery

Pinakpani Pal and Bhabatosh Chanda
Electronics and Communication Sciences Unit
Indian Statistical Institute
Kolkata 700 108
pinak@isical.ac.in and chanda@isical.ac.in

Abstract

In this paper we have proposed a clustreing technique that extracts sub- and sup-clusters based on a simple measure of circular symmetry. These sub-clusters and sup-clusters are then used as building blocks to form final clusters of any arbitry shape including concave ones through merging and splitting iteratively. The proposed method is tested on multi-spectral satellite imagery and a good result is obtained. Major advantages of this method is its simplicity and being free from initial guess about the cluster centres or the number of clusters.

1. Introduction

Cluster analysis is an unsupervised tool for exploring the underlying structure of a given data set. It is widely used in pattern recognition techniques applied to numerous domain starting from remote sensing to biomedical image processing. Clustering techniques based on similarity (or, in other words, distance) measure is a popular method [7, 2, 9, 3, 1, 4]. Based on some measure of similarity, a set of rule is specified to assign patterns to a cluster domain. A threshold is needed to define a degree of acceptable similarity in such process. Measure of symmetry instead of similarity is also used [8, 5, 6, 10] to find the clusters in the feature space because symmetry is a basic feature of shapes and objects as Nature gives great emphasis to this attribute. However, all the above mentioned techniques produces convex clusters mostly with a shape of ellipsoid. Many data sets obtained from real-life problems form clusters of any arbitrary shape including the concave ones. They may also encompass one another to some extent. Reported methods usually fail to handle such situations. In this paper we have proposed a method of extracting sub-clusters as well as sup-clusters based on a simple measure of circular symmetry. These sub-clusters are then merged and sup-clusters are split iteratively beased on pre-defined overlap criteria. When no more merging or splitting is possible, the method is terminated and from the remaining clusters required number of clusters are given out as the desired result. Note that in the proposed algorithm no initial of cluster centres is required, nor the number of clusters play any role in the clustering process.

The paper is organized as follows. Section 2 presents the proposed method. Prime objective of this work and a plausible strategy are given in Sections 2.1 and 2.2 respectively. Detail algorithm with an example is described in Section 2.3. Experimental result on multi-spectral satellite image is given in Section 3 and concluding remarks are given in Section 4.

2. Proposed Method

2.1. Objective

Result of the most popular K-means clustering algorithm depends heavily on the user-supplied parameters like number of clusters and the cluster centres. This algorithm always produces given number of clusters irrespective of actual underlying structure of data. Result can be made more data dependent and number of clusters may be relaxed to some extent by using ISODATA algorithm. Main problem with this algorithm is that it requires a lot of parameter values to be supplied by the user. Hence, the performance of these clustering algorithm is very much dependent on the parameter values, the chosen measure of similarity and the method used for identifying clusters in the data [9]. The objective of the present work is to minimise the number of external parameters supplied for cluster seeking so that it becomes more data dependent as well as robust.

2.2. Strategy

Basic assumption behind the proposed method is whatever be the shape of the actual cluster, the cluster is always composed of one or many small hyper-spherical sub-clusters. Thus each data point is assumed to be a sub-cluster centre. The largest sub-cluster at any point is the largest hypersphere formed by the neighbouring points satisfying the symmetry property.

There are various kinds of symmetry metric found in the literature. In the proposed method we use the following

simple one for the purpose. Suppose n is the dimension of the data. Let $p_j = (x_{j1}, x_{j2} \cdots, x_{jn})$ be any point with frequency of occurrence m_j. Let $c_i = (\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{in})$ be the center of the i-th cluster C_i. Let the *cluster symmetry* $S(r_k)$ of the hyper-sphere of radius r_k around c_i is defined as:

$$S(r_k) = \frac{1}{M} \left| \sum_j m_j \times \delta(c_i, p_j) \right| \qquad (1)$$

where

$$\delta(c_i, p_j) = \begin{cases} \frac{1}{n} \sum_{l=1}^{n} (\bar{x}_{il} - x_{jl}) & dist(c_i, p_j) \leq r_k \\ 0 & \text{otherwise} \end{cases}$$

$dist(.,.)$ stands for any distance metric and M is the total number of data points within the hyper-sphere of radius r_k. For ideal symmetric cluster, the value of cluster symmetry is 0, and for all other cases it is greater than 0 and less than or equal to 1. Since in real-life problem even a small cluster's symmetry value seldom becomes 0, we use a small threshold S_{th} to determine whether the cluster possesses an acceptable symmetry. Accordingly, the radius of largest acceptable symmetric cluster at a point may be defined as:

$$r_k = \max_j \{Arg[S(r_j)]\} \qquad \forall \; S(r_j) \leq S_{th} \qquad (2)$$

Usually clusters can be of any arbitrary shapes. Therefore, such symmetry measure may not be very effective in finding the actual clusters. However, it is good enough to detect sub-clusters (or sup-clusters) of valid and desired clusters. In the proposed method these symmetric clusters act as building blocks for the actual arbitrary shaped clusters. These actual clusters are formed through iterations of merging and splitting operations. A building block cluster is merged into a larger or equal size cluster if a significant portion of the former hyper-sphere overlap with the latter. It also splits (usually when the radius of the building block cluster is large) if different clusters are found placed inside the candidate one. This situation is revealed by the presence of mutually exclusive clusters in a larger cluster as its sub-clusters. The detail procedure is described in the following subsection in the form of an algorithm with an example using synthetic data.

2.3. Implementation

In this section we describe in detail the implementation of the above strategy through the following algorithm.

Algorithm:

Input

Data vectors, p_i.
The global value of threshold for symmetry, S_{th}.
Approximate number of clusters, η .

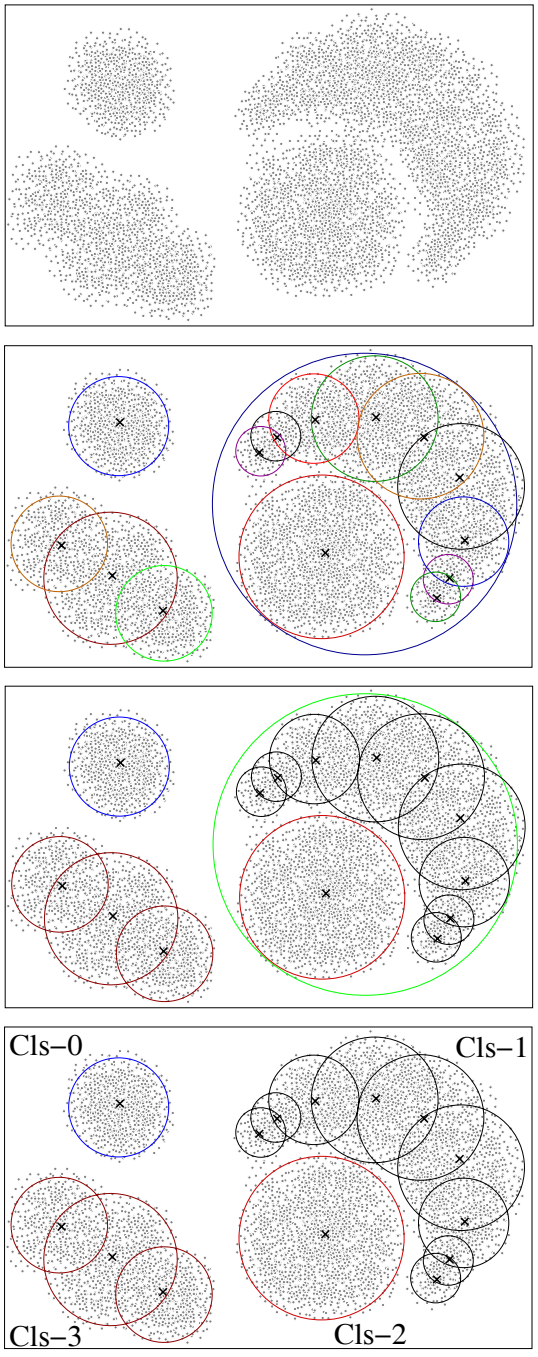


Figure 1: Illustrating proposed clustering scheme with synthetic data. For detail see the Algorithm. (a) Data points in the two-dimensional feature space. Gray points show the locations where frequency of occurrances is significant. (b) Largest circular clusters formed. Boundary of clusters are shown in different colours. (c) Result after merging. Merged clusters are shown by unique colours. (d) Result after splitting and also the final result. Note that large green cluster of figure (c) is removed.

Steps

- Compute frequency of occurances m_i of data points at each location p_i of feature space.

  [See Fig. 1(a) for an example. All p_i whose frequency of occurrances is greater than 0.001% of the total number of points are only shown by gray dots.]

- Assume that each p_i (with $m_i > 0$) is a zero-radius cluster centre c_i representing the class C_i, and set the FLAG of all such C_is to 0.

- Calculate the radius r_i of symmetry around c_i using equation (1) and (2). Let N is the number of such largest sub-clusters.

  [See Fig. 1(b) which shows only the largest sub-clusters by different coloursy. The sub-clusters which are completely inside a larger sub-cluster are not shown here for the clarity of the figure, because such sub-clusters are too many.]

- Dictionary sort the clusters in decreasing order by using the radius as primary key and the symmetry value as secondary key.

- for $i = 1, \ldots, N$

  If the FLAG of C_i is 0, then

  assign a new label to it and set the FLAG to 1.

  for $j = i + 1, \ldots, N$

  If the FLAG of C_j is 1 then

  do nothing;

  otherwise if $c_j \in C_i$ then

  assign label of C_i to C_j and set the FLAG to 1.

  [This is a merging step and the result is shown in Fig. 1(c). Merged clusters are shown in same colour.]

- Suppose in the previous step C_{ik} and C_{il} are two clusters that got label of C_i. If $C_{ik} \cap C_{ij} = \emptyset$, then C_i is a invalid cluster and do the followings:

  Set the FLAGs of all the sub-clusters of C_i to 0;

  Remove C_i from the list of clusters; and

  Go to previous Step.

  [This is a splitting step and the result is shown in Fig. 1(d).]

- Remove all the clusters of readius zero.

- Accept the first η clusters, and STOP.

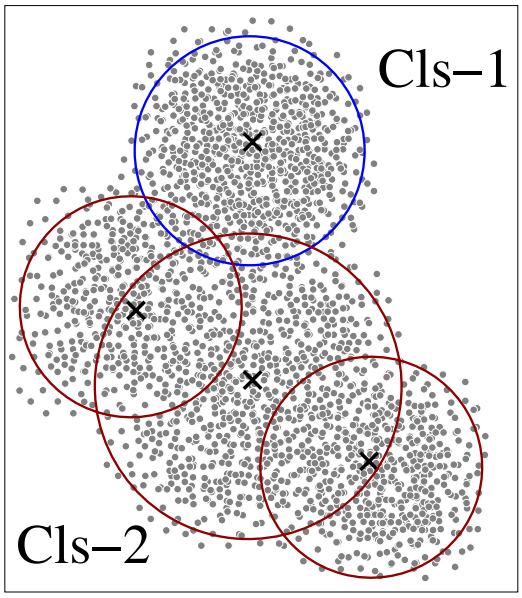


Figure 2: Illustrating overlapping of clusters. Some points are common to both the cluster-1 and cluster-2.

Output

Clustered output where each cluster at the output can be a set of circularly symmetric clusters represented by: $\{(c_{i1}, r_{i1}), ((c_{i2}, r_{i2}), \cdots, (c_{ik}, r_{ik})\}$. Note that the number of clusters produced is less than or equal to the desired number.

It may be noted that some points may belong to different clusters as they fall within the encloser of different hyper-spheres which are not merged. An example situation is shown in Fig 2. This happens when hyper-spheres do not have significant overlap as defined in the merging step.

3. Experimental Result

Here the data set used is multi-spectral satellite images, which is suitable for applying clustering techniques. Since we are interested to investigate the capability of the proposed clustering technique, we have deliberately avoided any kind of supervision even in the selection of training set. Ususally block of pixels are picked up from the image area where ground-cover is known apriori to form the training data set. In this experiment we have just picked up data points from the said image at every 8th row and every 8th column to generate the training set. Size of the image is 2500x2213 (=5532500 pixels) and the size of the training set is 87035 pixels. We run the clustering technique with the desired number of clusters 8, 10 and 20. Once the clusters are formed we use them for classifying the entire image data. A p_j pixel is classified to belong to the Class C_i if $dist(c_i, p_j) \leq r_i$ for all i. If a point belongs to more than

one clusters, that means if p_j satisfies the previous criterion for more than one C_i, then it is classified based on the minimum of distances weighted by the *apriori* propability that a point belongs to a given cluster. The pixels that cannot satisfy this criterion are marked as unclassified pixels. The percentage of unclssified pixels for 8, 10 and 20 classes are 7.76%, 5.64% and 1.72% respectively. Original and the classified (to 8 classes) images are shown in Fig. 3(a) and (b) respectively.

4. Conclusion

Common distance based algorithms (e.g. Maximum-Distance algorithm, K-means algorithm, Isodata algorithm etc.) require users to specify initial cluster centers and the number of clusters, and the results largely depend on these intial guesses. On the other hand, the proposed algorithm initially considers each data point as a cluster center. Then based on a simple symmetry measure defined as the sum of difference of coordinate values, it forms largest hyper-spherical clusters at each point. These hyper-spherical clusters are then merged and split to arrive at the final stable clusters which are accepted as the final clusters. Unlike the common agglomerative-and-divisible or split-and-merge techniques, the proposed scheme does not require any intra-cluster and inter-cluster criteria to be satisfied except simple overlapping criteria. However, this simple symmetry based approach produces extremely encouraging result and strongly suggest further investigation in this direction.

References

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[2] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

[3] J. Hartigan. *Clustering Algorithms*. John Wiley, New York, 1988.

[4] A. K. Jain and R. C. Dubes. *Algorithm for clustering*. Prentice Hall, Englewood Cliffs, N. J, 1988.

[5] K. Kanatani. Comments on symetry as a continuous feature. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):246–247, 1997.

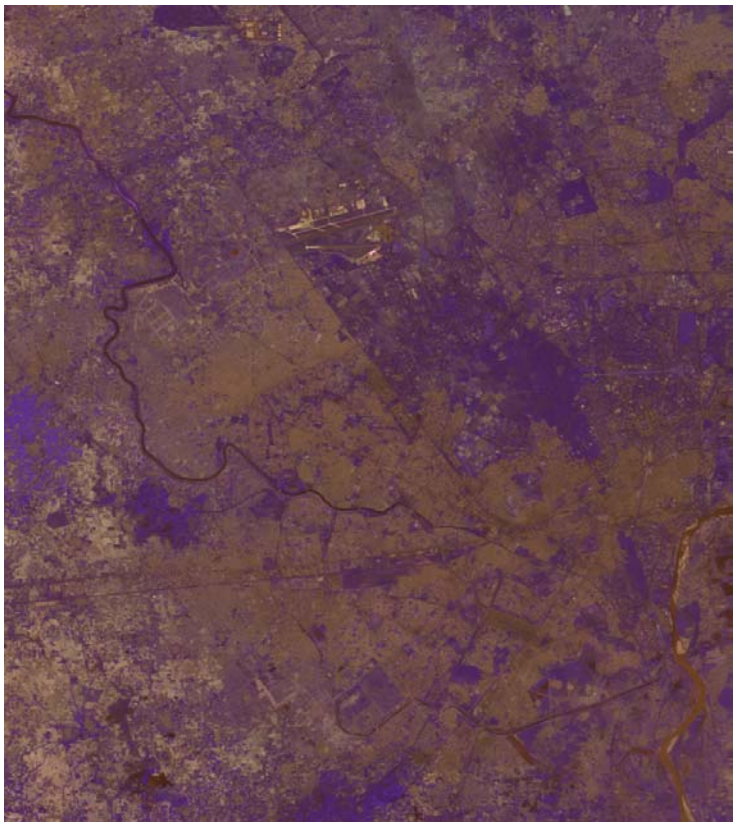
[6] D. Reisfled, H. Wolfsow, and Y. Yeshurun. Context-free attentional operators: the generalized symetry transform. *International Journal of Computer Vision*, 14:119–130, 1995.

[7] E. Ruspini. A new approach to clustering. *Information Control*, 15(1):22–32, 1969.

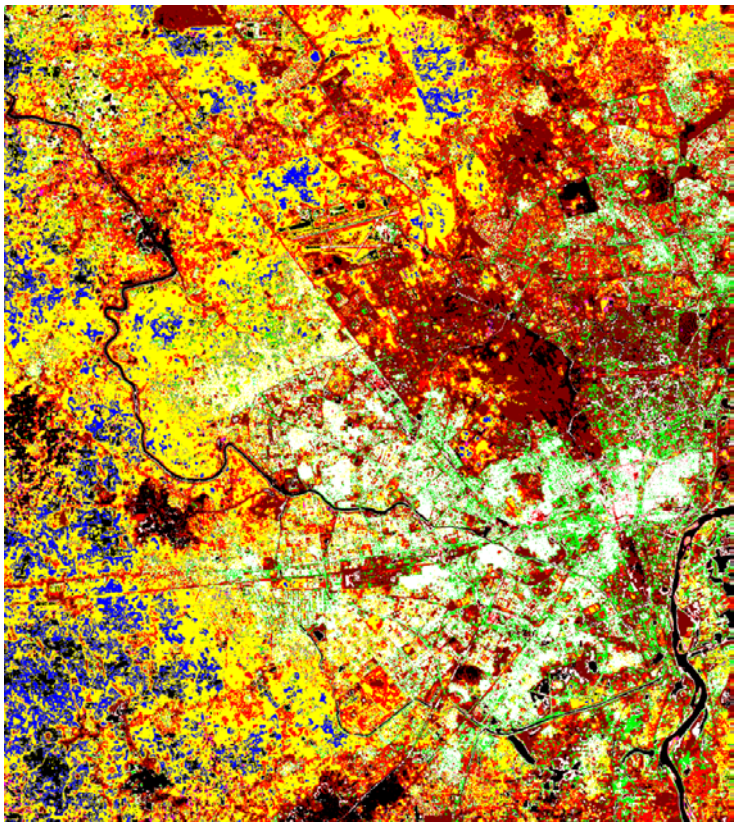
[8] M.-C. Su and C.-H. Chou. A modified version of the k-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):674–680, 2002.

[9] J. T. Tou and R. C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley Publishing Company, Reading, MA, 1974.

[10] D. Zabrodsky, S. Peleg, and D. Avnir. Symetry as a continuous feature. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1154–1166, 1995.



(a)



(b)

Figure 3: (a) Original Satellite image, (b) Clustered image (8 classes).