Agglomerative Mean Shift Clustering

Swami Shivputra Guruling
Department of Computer Science and Engineering
TPCT College of Engineering,Osmanabad,Maharashtra,India

Dr.Holambe A.N.
Department of Computer Science and Engineering
TPCT College of Engineering,Osmanabad,Maharashtra,India

Abstract –Clustering is an integral part of data mining and has attracted much attention recently. In this paper, we focus our discussion on Agglomerative Mean Shift clustering with what is clustering, why to clustering, Mystery of clustering VS Classification, conceptual Clustering procedures.

Keywords: Agglomerative Mean Shift,KDE.

I. INTRODUCTION

A.Definition-

A group of same or similar Elements (objects) gathered or occurring closely together i.e. let us define some basic concepts of clusters in a mathematical way.

Let X be a set of data, that is

$$X = \{x_1, x_2,\ldots,x_n\} \qquad (1)$$

A so called m-clustering of X is its partition into m parts (clusters) $C_1,\ldots, C_m$, so that

1. None of the clusters is empty; $C_i \neq \emptyset$
2. Every sample belongs to a cluster
3. Every sample belongs to a single cluster (crisp clustering); $C_i \cap C_j = \emptyset$, $i \neq j$

Naturally, it is assumed that vectors in cluster C_i are in some way "more similar" to each other than to the vectors in other clusters. Fig. 1 illustrates a couple of different kind of clusters; compact, linear and circular.

Figure 1. Couple of different kind of clusters

B. Why Clustering-

i)Organizing data into clusters shows internal structure of the data

Ex. Clusty and clustering genes above.

ii) Sometimes the partitioning is the goal

Ex. Market segmentation

iii) Prepare for other AI techniques

Ex. Summarize news (cluster and then find centroid)

iv) Techniques for clustering is useful in knowledge discovery in data

Ex. Underlying rules, reoccurring patterns, topics, etc.

C. Mystery of clustering VS Classification-

Classification: the task is to learn to assign instances to predefined classes.

Clustering: no predefined classification is required. The task is to learn a classification from the data.

II. CLUSTERING PROCEDURES

This method works on both bottom up and top down approaches .The agglomerative hierarchical technique works on bottom up approach. The general approach of hierarchical clustering is in using an appropriate metric which measures distance between 2 tuples and linkage criteria which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

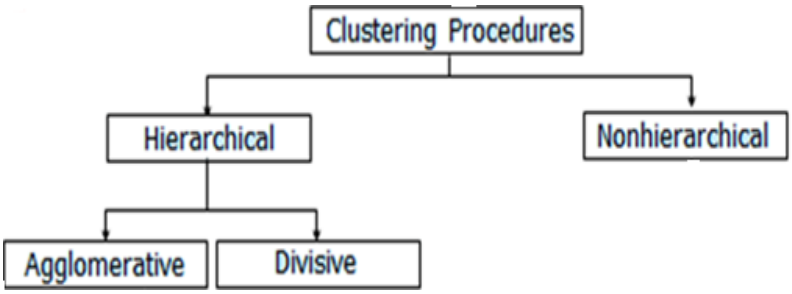

Figure 2. Clustering Procedure

A. Hierarchical clustering-

It is characterized by the development of a hierarchy or tree-like structure.
Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level).
Hierarchical methods can be agglomerative or divisive.

B Divisive clustering-

Starts with all the objects grouped in a single cluster. Clusters are divided or split until each object is in a separate cluster. Agglomerative methods are commonly used in marketing research. They consist of linkage methods, error sums of squares or variance methods, and centroid methods.

C.Agglomerative clustering -

Starts with each object in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters. This process is continued until all objects are members of a single cluster.

D. Nonhierarchical -

The nonhierarchical clustering methods are frequently referred to as k-means clustering. These methods include sequential threshold, parallel threshold, and optimizing partitioning.
In the sequential threshold method, a cluster center is selected and all objects within a prespecified threshold value from the center are grouped together. Then a new cluster center or seed is selected, and the process is repeated for the unclustered points. Once an object is clustered with a seed, it is no longer considered for clustering with subsequent seeds.
The parallel threshold method operates similarly, except that several cluster centers are selected simultaneously and objects within the threshold level are grouped with the nearest center.
The optimizing partitioning method differs from the two threshold procedures in that objects can later be reassigned to clusters to optimize an overall criterion, such as average within cluster distance for a given number of clusters.

There are two major issues in thinking about clustering:

1) What is a good inter-cluster distance?

Agglomerative clustering uses an inter-cluster distance to fuse nearby clusters; divisive clustering uses it to split in-sufficiently coherent clusters. Even if a natural distance between data pointsis available (which might not be the case for vision problems), there is no canonical inter-cluster distance. Generally, one chooses a distance that seems appropriate for the data set.
For example, one might choose the distancebetween the closest elements as the inter-cluster distance, which tends to yield extended clusters (statisticians call this method single-link cluster-ing). Another natural choice is the maximum distance between an element of the first cluster and one of the second, which tends to yield rounded clusters (statisticians call this method complete-link clustering). Finally, one could use an average of distances between elements in the cluster, which also tends to yield "rounded" clusters (statisticians call this method group averageclustering).

2) How many clusters are there?
This is an intrinsically difficult task if there is no model for the process that generated the clusters. The algorithms we have described generate a hierarchy of clusters.

E. How do we Clustering –

a. Agglomerative clustering-

- How to define cluster similarity-
  - Average distance between points, maximum distance, minimum distance
  - Distance between means or medoids.
- How many clusters-
  - Clustering creates a dendrogram (a tree)
  - Threshold based on max number of clusters or based on distance between merges.

F. Steps of Agglomerative clustering -

The algorithm forms clusters in a bottom-up manner, as follows:

1. Initially, put each article in its own cluster.

Figure 3. Step-1

2. Among all current clusters, pick the two clusters with the smallest distance.

Figure 4. Step-2

3. Replace these two clusters with a new cluster, formed by merging the two original ones.

Figure 5. Step-3

4. Repeat the above two steps until there is only one remaining cluster in the pool.

G. Merits and Demerits -

- Merits -
  - Simple to implement, widespread application.
  - Clusters have adaptive shapes.
  - Provides a hierarchy of clusters.
- Demerits -
  - May have imbalanced clusters.
  - Still have to choose number of clusters or threshold.
  - Need to use an "ultrametric" to get a meaningful hierarchy.

III.MEAN SHIFT ALGORITHM

Mean Shift is a powerful and versatile non parametric iterative algorithm that can be used for lot of purposes like finding modes, clustering etc. Mean Shift was introduced in Fukunaga and Hostetler and has been extended to be applicable in other fields like Computer Vision. This document will provide a discussion of Mean Shift, prove its convergence and slightly discuss its important applications.

The most important application is using Mean Shift for clustering. The fact that Mean Shift does not make assumptions about the number of clusters or the shape of the cluster makes it ideal for handling clusters of arbitrary shape and number.

Although, Mean Shift is primarily a mode finding algorithm, we can find clusters using it. The stationary points obtained via gradient ascent represent the modes of the density function. All points associated with the same stationary point belong to the same cluster.

An alternate way is to use the concept of Basin of Attraction. Informally, the set of points that converge to the same mode forms the basin of attraction for that mode. All the points in the same basin of attraction are associated with the same cluster. The number of clusters is obtained by the number of modes.

This is a versatile technique for clustering-based segmentation.The Mean-Shift algorithm is a powerful local optimization algorithm for maximizing Kernel Density Estimator (KDE).

The most important application of this algorithm is the nonparametric mode-seeking and clustering.

A. Kernel density estimation -

- Kernel density estimation function -

  Kernel density estimation is a non parametric way to estimate the density function of a random variable. This is usually called as the Parzen window technique. Given a kernel K, bandwidth parameter h , Kernel density estimator for a given set of d-dimensional points is:

$$\hat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (2)$$

  where $K(\bullet)$ is the kernel — a symmetric but not necessarily positive function that integrates to one — and $h > 0$ is a smoothing parameter called the *bandwidth*.

  A kernel with subscript h is called the *scaled kernel* and defined as $K_h(x) = 1/h\ K(x/h)$. Intuitively one wants to choose h as small as the data allow, however there is always a trade-off between the bias of the estimator and its variance; more on the choice of bandwidth later.

- **Gaussian kernel :**

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}} \quad (3)$$

  Where h is the window width and x_i are the values of the independent variable in the data, and x is the value of the independent variable for which one seeks an estimate. Unlike most kernel functions this one is unbounded on x.

B. Computing the Mean Shift -

Mean shift treats the points the feature space as an probability density function. Dense regions in feature space correspond to local maxima or modes. So for each data point, we perform gradient ascent on the local estimated density until convergence. The stationary points obtained via gradient ascent represent the modes of the density function. All points associated with the same stationary point belong to the same cluster.

$$m(x) = \frac{\sum_{i=1}^{n} x_i g\left(\frac{\|x-x_i\|^2}{h}\right)}{\sum_{i=1}^{n} g\left(\frac{\|x-x_i\|^2}{h}\right)} - x \quad (4)$$

The quantity m(x)is called as the mean shift

C. Mean shift clustering -

The mean shift algorithm seeks *modes* of the given set of points.

Choose kernel and bandwidth.

For each point:

- Center a window on that point.
- Compute the mean of the data in the search window.
- Center the search window at the new mean location.
- Repeat (above two steps) until convergence.
- Assign points that lead to nearby modes to the same cluster.

D. Merits & Demerits -

Merits -

- Application independent tool.
- Suitable for real data analysis.
- Does not assume any prior shape.(e.g. elliptical) on data clusters
- Can handle arbitrary feature spaces.
- Only ONE parameter to choose.
- *h* (window size) has a physical meaning, unlike K-Means.

Demerits -

- The window size (bandwidth selection) is not trivial.
- Inappropriate window size can cause modes to be merged, or generate additional shadows.

Example :consider 1,2,3,4,5 & 6 are articles with final Agglomerative MS output.

IV.CONCLUSION

It is the goal that most parts of the paper can be appreciated by a new comer to the field of Agglomerative Mean Shift clustering with what is clustering, why to clustering, Mystery of clustering VS Classification, conceptual Clustering procedures.

Theoretically it is analysed that Agglo-MS algorithm is a fast, stable, and accurate MS clustering algorithms that can achieve competitive solutions.

REFERENCES

[1] M. Carreira-Perpinan, "Acceleration Strategies for Gaussian Mean-Shift Image Segmentation,"Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 1160-1167, 2006.

[2] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 7, pp. 790-799, July 1995.

[3] Y. Zho and G. Karypis, "Hierarchical Clustering Algorithms for Document Datasets,"Data Mining and Knowledge Discovery,vol. 10, no. 2, pp. 141-168, Mar. 2005.

[4] M. Allain, J. Idier, and Y. Goussard, "On Global and Local Convergence of Half-quadratic Algorithms," IEEE Trans. Image Processing, vol. 15, no. 5, pp. 1130-1142, May 2006.

[5] M. Carreira-Perpinan."Fast nonparametric clustering with gaussian blurring mean-shift". In International Conference on Machine Learning, pages 153–160, 2006.
M. Carreira-Perpinan."Gaussian mean-shift is an emalgorithm".IEEE TPAMI, 29(5):767–776, 2007.

[6] Y. Cheng. "Mean shift, mode seeking, and clustering".IEEE TPAMI, 17(7):790–799, 1995.

[7] D. Comaniciu and P. Meer. "Mean shift: A robust approach toward feature space analysis". IEEE TPAMI, 24(5):603–619, May 2002.

[8] I. Davidson and S. Ravi. "Clustering with constraints: Feasibility issues and the k-means algorithm". In International Conference on Data Mining. SIAM, 2005.

[9] I. Davidson and S. Ravi. "Hierarchical clustering with constraints: theory and practice".In principles and Practice of Knowledge Discovery in Databases(PKDD), 2005.

[10] Henry Lin "Clustering" 15-381 Artificial Intelligence.

[11] JukkaKainulainen "Clustering Algorithms: Basics and Visualization" 47942F.

[12] Frank Keller " Clustering" Connectionist and Statistical Language ProcessingComputerlinguistikUniversität des Saarlandes.