DETERMINING THE NUMBER OF CLUSTERS FOR A K-MEANS CLUSTERING ALGORTIHM

Abhijit Kane
Department of Computer Science
Birla Institute of Technology and Science, Pilani
Jawahar Nagar, Shamirpet (M)
Hyderabad, Andhra Pradesh 500078, India
abhijitkane@gmail.com

Abstract

Clustering is a process used to divide data into a number of groups. All data points have some mathematical parameter according to which grouping can be done. For instance, if we have a number of points on a two-dimensional grid, the x and y coordinates of the points are the parameters according to which clustering is done. If the k-means algorithm is run with k=3, the data points will be split into 3 groups such that the sum of the variance for each group is minimized. The problem here, of course, is the choice of the parameter k. We may get a much better modeling of the data if we split the data points into 2 or 4 groups. Determining the ‘best’ value of k is a broad problem – there is no obvious parameter according to which this can be done. This paper looks at a new, efficient approach to determine the number of clusters.

Keywords: data, k-means, clustering, variance, data-mining

1. Introduction

Determining the number of clusters to divide a dataset into is a common problem in data clustering, and is a different problem altogether from actually making the clusters. The number of clusters ‘k’, is ambiguous, as the data may have its own inherent meaning. Eg. The speeds of different cars on the road and the number of stars visible in each constellation are two very different types of datasets, and have to be interpreted differently. There are multiple algorithms available to cluster data. The paper will be focusing on the centroid-based approach, called k-means clustering, which groups data points based on spatial extent. The algorithm is limited to datasets with very few outliers.

2. The volume metric

2.1. *Definition*

For a single cluster, we define a property called ‘volume’. It is valid for any cluster of n-dimensional data, where each point is a vector of the form ($P_{1,}$ $P_{2,}$ $P_{3....}$,$P_{n)}$ .P_i represents the value for the nth dimension. The volume of a cluster is defined as the product of the range of the cluster in each dimension. For a two-dimensional dataset, like points in a 2-D Cartesian plane, the ‘volume’ of each cluster is equal to the area of the smallest square (aligned with the axes) that can completely contain all points in the cluster. For a three-dimensional dataset, the ‘volume’ of a cluster is equal to the volume of the smallest cube (aligned with the axes) that can contain all points in the cluster.

2.2. *Determination*

The algorithm to find the ‘volume’ of a cluster is as follows:

1. For every dimension of data, repeat steps 2-4.
2. Find the maximum value for this dimension among all points in the dataset.
3. Find the minimum value for this dimension among all points in the dataset.
4. Subtract the minimum value from the maximum value to get the ‘range’ for this dimension.
5. Calculate the ‘volume’ of the data cluster as the product of the ranges of all the dimensions.

3. Finding the optimal number of clusters

The algorithm is as follows:

1. Start with one cluster (k=1)
2. Calculate the volume of this cluster. Store the volume as V_1 (Volume of clusters with k=1)

3. Increment k by 1.
4. Perform k-means clustering on the dataset to generate k clusters.
5. Calculate the sum of the volumes of all k clusters. Store this value as V_k.
6. Calculate the ratio (V_k/V_{k-1}).
7. If this ratio is less than a fixed threshold, repeat from step 3. Else, stop clustering, go to step 8.
8. The value (k-1) is the optimal number of clusters.

The algorithm relies on the fact that if k clusters are optimal, creating k+1 clusters will not lead to a significant reduction in total volume. A few illustrations are shown, where the number of clusters is being increased from k=1 to k=4. The total volume of all the clusters is also mentioned.

The process is shown below:

Table 1: Illustration of the algorithm for 4 clusters

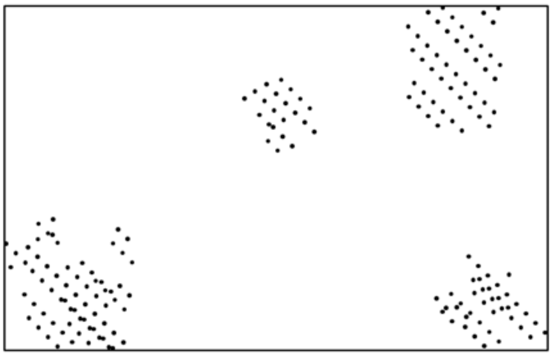
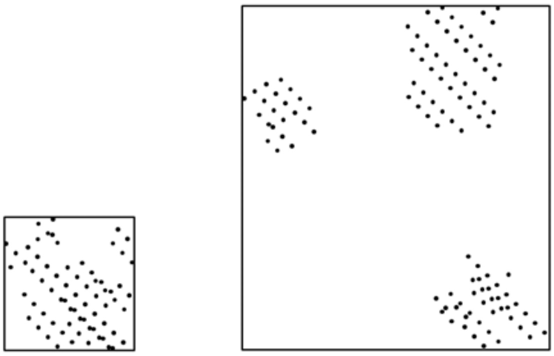
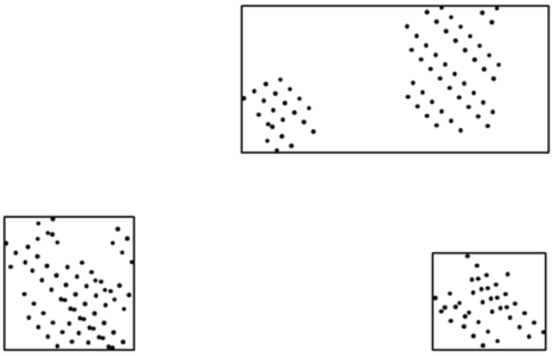
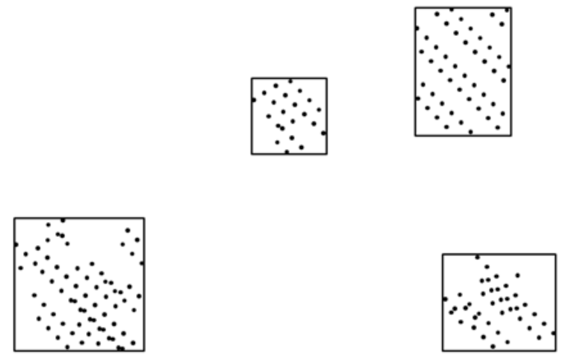
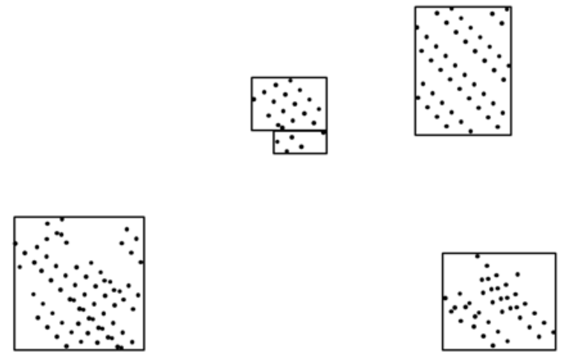
No. of clusters = k	Diagram	Cluster Volumes	Total 'Volume' in clusters = V_k	Ratio (V_k/V_{k-1})
1		368	368	N/A
2		34 + 206	240	0.652
3		34 + 88 + 21	143	0.596

Table 1 (Continued)

4		34 + 11 + 21 + 24	90	0.629
5		34+24+21+7+2	88	0.977

As can be seen, the total area under the clusters (column 5) decreases with the number of clusters. For k=2,3 and 4, the reduction in area is significant. The ratio V_k/V_{k-1} is around 0.6, meaning that there is a 40% reduction in area in each step. For the fifth cluster (k=5), however, the reduction in area is less than 3% (from 90 to 88). The reduction ratio is 0.977, greater than our empirical threshold of 0.80. Therefore, we say that k=4 is the optimal number of clusters.

4. Comparison with other methods

This method is a variation of the ‘elbow method’ [Thorndike (1953)], which calculates the variance of the clusters formed at each step. Our algorithm relies on a metric that’s more visually intuitive, leading to better results for datasets where the data is spatial. This algorithm will be less effective when the data points are ‘non-spatial’, text, for example.

5. Conclusion

This paper highlights a simple way to automatically find the number of clusters in a dataset. Though every step requires re-clustering of the dataset, calculating the ratios after that is an O(n) operation. This method works well for clusters that are distinctly separated. This method is also density-independent, making it useful for clustering algorithms like the Expectation-maximization algorithm [Dempster et al. (1977)].

References

[1] Dempster A.P., Laird N.M., Rubin D.B (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society

[2] Hartigan J.A. , Wong MA. (1979): A K-Means Clustering Algorithm. Journal of the Royal Statistical Society

[3] Kanungo Tapas et al. (2002): An Efficient K-Means clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence

[4] Sugar C.A. and James G.M. (2003). Finding the number of clusters in a data set: An information theoretic approach. Journal of the American Statistical Association

[5] Thorndike, Robert (1953) “Who Belongs in the family?”