K-Means Clustering

Preston Engstrom

***Abstract*—The K-Means clustering algorithm has been a mainstay of unsupervised machine learning for decades. The algorithm is explored in this report using known datasets of 2D points, using several cluster counts and distance measures, with results compared using the Dunn Index for cluster analysis.**

***Index Terms*—Clustering, K-means, Machine Learning**

I. Introduction

THE K-Means Clustering Algortihm has continued to be the workhorse of clustering tasks since its development in 1967 by James MacQueen. In this implementation, the data consists of points on a 2D plain. The clusters are simply the collection of points closest to the center of the cluster based on some distance measure. K-means, and clustering algorithms in general, suffer greatly from the curse of dimensionallity due to their reliance on relating data vectors to each other in a multidimensional space. Presenting data in a 2D space is the best case scenario for exploring the behavior of the algorithm.

A. Problem Definition

Clustering is the task of grouping a set of unlabled data vectors into groups, or clusters. Items in a cluster should be more alike items in the same cluster, and much less like items in other clusters. Clustering is one of the most common tasks of exploratory data mining for extracting preliminary observations about datasets. This paper begins with a description of the K-Means Algorithm and its implementation using the Euclidean, Chebyshev and Minkowski distance measures. The results are then presented and compared using the Dunn Index.

II. The K-Means Algorithm

K-Means is a method of vector quantization taken from signal processing. Vector quantization itself is the process of taking a set of complex vectors and compressing them into broader group. With its origins for lossy data compression and lossy error correction, the same processes have been applied for data exploration, pattern recognition and clustering, as in K-Means.

The K-Means alogrithm can be divided into 2 distinct steps. The assignment step and the update step, where the centroids are moved and member data points updated until convergence is reached.

A. Description

K-Means begins with an initial set of K centroids, wich serve as the centers of each cluster. The assignment step can be expressed as follows[7]:

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \ \forall j, 1 \leq j \leq k\}$$

Fig. 1. Simple Backprop Skeleton

```
given training example and label
for each training example:

for o in outputs:
calculate output_delta

for h in hidden:
calculate hidden delta

update hidden weights
update input weights
```

Where each data vector x_p is assigned to exactly one cluster S_i based on, in the case of this representation, the smallest euclidean distance to the closest mean $m_i....m_j$. For each group, the mean of the member data vectors is then calculated, and the clusters centroid is then moved to that position:

$$m_i^{t+1} = \frac{1}{|S_i^{(t)}|}\Sigma x_j$$

This is run for each centroid x_j. These steps, assignment and update, are repeated until no change occurs between two assignment steps. This convergence generally happens quite quickly, but can scale exponentially with the number of dimensions.

B. Centroid Intialization

C. Pseudocode Implementation

D. Distance Measures

E. Euclidean Distance

F. Chebyshev Distance

G. Minkowski Distance

III. Cluster Analysis Using The Dunn Index

IV. Results and Discussion

TABLE I
Comparison of S-Set Dunn Index, Euclidian Distance

Dataset	10 Centroids	15 Centroids	20 Centroids	Optimal Value
S1	0.004797	0.012200	0.005278	0.0367893
S2	0.002901	0.005611	0.006178	0.020947
S3	0.006027	0.008156	0.008058	0.004394
S4	0.005708	0.006873	0.005099	0.007474

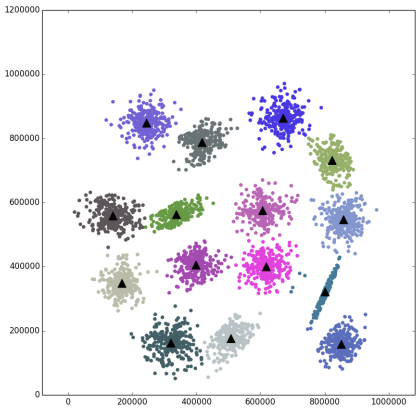

Fig. 2. Optimal Clusters for the S1 dataset.

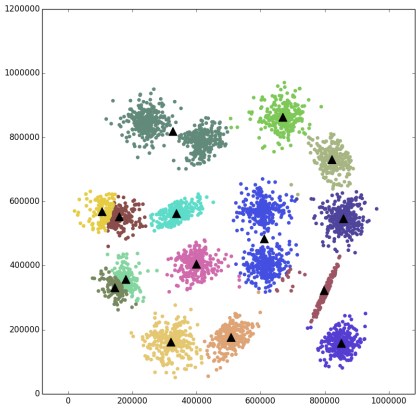

Fig. 3. Results using 15 randomly initialized centroids on the S1 Dataset.

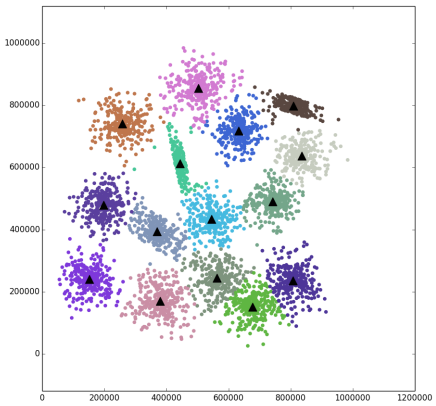

Fig. 4. Optimal Clusters for the S2 dataset.

V. Conclusion

The conclusion goes here.

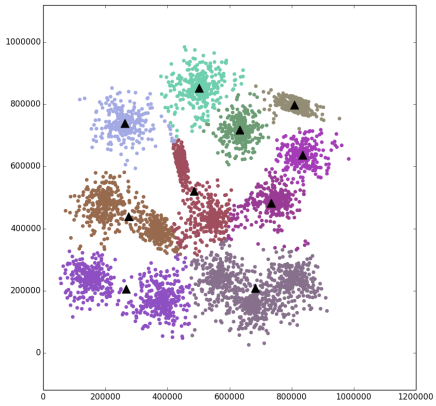

Fig. 5. Results using 10 randomly initialized centroids on the S2 Dataset.

References

[1] [1]D. MacKay, Information theory, inference, and learning algorithms. Cambridge, UK: Cambridge University Press, 2003.