

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 Algorithm 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 dimensionality 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 unlabeled 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 algorithm 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.

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

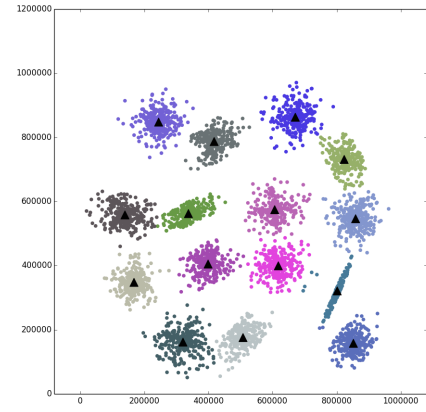


Fig. 1. Optimal Clusters for the S1 dataset.

A. Description

B. Centroid Initialization

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

V. CONCLUSION

The conclusion goes here.

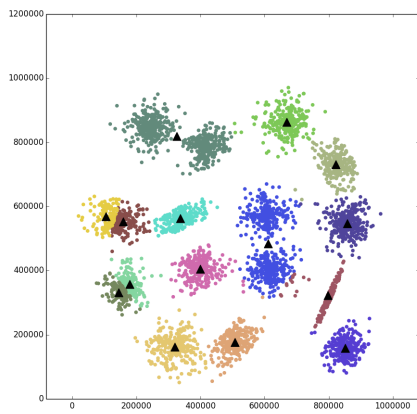


Fig. 2. Results using 15 randomly initialized centroids on the S1 Dataset