K-Means clustering variants and improvements

Stefan Sebastian, 242

November 11, 2018

Abstract

TODO make abstract

Contents

1 Introduction

1.1 Motivation

In recent times there has been an explosive growth of data. This new, digital resource has become so important that a recent article from The Economist magazine has called it 'the new oil'[1]. Some important sources of data, gathered by Bernard Marr[2], a business consultant in big data and AI technologies, are: Social Media (every minute 456000 tweets are sent on Twitter, 46740 photos are uploaded on Instagram, 4146600 YouTube videos are being watched), Communication (16 million text messages and 156 million emails are sent every minute), Services(45788 trips are made through Uber per minute), Internet of Things(8 million people use voice control every month). The amount of data in the world is predicted to reach 163 zettabytes by 2025[3], where one zettabyte represents a trillion gigabytes.

The main reason for choosing the K-means method for this paper is because I believe that the development of a computationally efficient unsupervised learning method will be necessary to keep up with future's data, which will continue to grow and sometimes lack meaningful labels.

1.2 Clustering

Clustering is a field of data analysis whose purpose is to organize points, or objects into natural clusters. In other words given N objects described by a set of features, find a number of groups, such that objects in the same cluster are similar to each other and different from objects in other clusters[4]. Given that there is no rigorous definition of similarity and that clusters in real data can overlap there are multiple algorithms and difference measures proposed in research.

1.3 K-means

The K-means algorithm was discovered independently in multiple scientific fields around the same time and its name comes from a paper published by MacQueen in 1967[4]. Despite being so old, this method is still widely used in clustering tasks because of its efficiency, simplicity and proven success.

The main goal of this algorithm is to minimize the distance between cluster centers and points assigned to them, which is a NP-hard problem. For this reason K-means is a greedy algorithm and has a chance to get stuck in local minimum instead of the best solution[4].

A basic version of the K-means algorithm is[5]:

1. Select an initial partition with K clusters and repeat 2 and 3 until the point membership no longer changes
2. Assign each point to the closest center
3. Update cluster centers based on the new assignment

2 Improvements

2.1 Execution time optimizations

TODO parallel stuff

2.2 Seed selection

TODO seed selection stuff

3 Variants

3.1 Bisecting K-means

bisecting kmeans

3.2 Evolutionary K-means

genetic kmeans

4 Conclusions

TODO something conclusive

References

[1] "The most valuable resource." https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. Accessed: 2018-11-11.

[2] "How much data do we create every day? the mind-blowing stats everyone should read." https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7f96d4a060ba. Accessed: 2018-11-11.

[3] "What will we do when the world's data hits 163 zettabytes in 2025?." https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#4ff934f6349a. Accessed: 2018-11-11.

[4] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[5] R. C. D. Anil K. Jain, *Algorithms for clustering data*, ch. Clustering Methods and Algorithms, pp. 96–97. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.