K-means Clustering Over Time
CSCI-471-02

Author: Jeffery B. Russell

Reviewer: Daniel Moore

Submitted: April 23, 2020

Computer Science at RIT

Overview

With the ubiquity of data in today's age, machine learning has become a driving force in research and innovation in both the public and private sectors. Due to the vast quantity of data generated by the internet, unsupervised learning has been at the forefront of data science over the past few decades. Clustering has consistently been an essential way of extracting data from large sets of information.

This project aims to break down and analyze two papers that use k-means clustering. In 1967 J. MacQueen wrote the historical article that introduced [5] the original k-means clustering. MacQueen's paper initially coined the term k-means [5] and is going to be reviewed because it provides useful historical insight into where this field of research started and what influenced it. Since 1967 the topics produced in the paper continues to be an exciting area of investigation because we can compare a relatively older research paper on clustering with a new one. Project Euclid(hosts 1.8 million pages of open-access content) has this paper on its website[1].

John Paparrizos and Luis Gravano in 2016 wrote a paper that covers how we can cluster time series data using an algorithm they call k-shape– the k-means algorithm [6] inspired this work. The University of Colombia's website[2] hosts this paper.

For the scope of this project in CSCI-471, up till section 3 is covered on the k-shape algorithm, and we excluded section two from the original k-means paper.

[1] `https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992`

[2] `http://web2.cs.columbia.edu/~gravano/Papers/2015/sigmod2015.pdf`

Summary

This section provides a summary of each article in depth.

K-Means

The general idea of clustering is to group data with similar traits. The main benefit of this is the ability to extract information from new data because you know what it is most similar to, thus giving you valuable insight. In the field of machine learning, this is considered as unsupervised learning because it requires no labels on the data – the algorithm auto assigns clusters, and you infer behavior off of those clusters.

Clustering has many applications such as image segmentation, preference predictions, compression, model fitting.

Although you can trace the idea of k-means clustering back to 1967 with a paper by Hugo Steinhaus [7], James MacQueen was the first to coin the term k-means in 1956 [5]. MacQueen's paper title "Some Methods For Classification and Analysis of Multivariate Observations" goes over the k-means process that segments an N-dimensional population into k sets. Note: when we refer to k in the algorithm, that is the number of sets that we are dividing the population.

A great deal of this article discusses optimality for the k-means algorithm, which is an important area to discuss, especially when considering the time at which the article got published. Back in 1967, computers were very slow and expensive. Although we had proofs that can guarantee that we could find an optimal solution, they were a NP-Hard problem [3]. This is critical because NP-Hard problems are problems that are exponential to solve.

Although the k-means algorithm did not guarantee the optimal solution, there was a subset of problems that it did guarantee an optimal solution– the specifics of these problems got discussed later in the article. Nerveless, since this algorithm wasn't computationally expensive and generally gave good results, it was a huge breakthrough at the time.

In section three, the paper examines specific applications of the k-means algorithm. The paper ran these experiments with an IBM 7094. Section 3.1 looked at clustering student documents data in high dimensions to find syntactical differences. Section 3.2 looked at a more theoretical and mathmatical ways to test the k-means algorithm. They created four-dimensional data and clustered them into two groups. After running the algorithm, it was able to identify the correct class 87 percent of the times correctly. Section 3.5 poses a unique approach that looks at the lexicographical analysis of papers written.

K-Shape

Time series data is a series of data points taken at time intervals. Each data point measured could have multiple dimensions. When doing k-means clustering on images, you would just ignore the relative x-y position in the image and treat everything as if they were data-points to feed into the algorithm. You could do the same thing with time-series data; however, you would just be throwing away the time information. Time-series clustering is frequently used for anomaly detection and pattern identification for use later on in forecasting.

Due to its immediate applications in fields of finance for stock prediction and the medical field, time series analysis is as old as computing itself.

In 2015 Paparrizos J and Gravano L introduced the K-shape clustering algorithm in their paper title "k-Shape: Efficient and Accurate Clustering of Time Series" [6].

Cyclical patterns that repeat over time is unique to time series analysis and requires special treatment. Unlike a typical k-means clustering algorithm, k-shape tries to find these shapes in the time series data that repeat itself. One cluster, for example, could be a gradual increase in values and then a sudden drop. Identification and categorization of these shapes is no trivial task. Compared to standard clustering, clustering on shape is far more time-consuming. Misalighments and magnitude differences have to get accounted for in any time-series clustering algorithm. Aligning two sequences using dynamic time warping is much more time consuming than a typical comparison of two sequences.

The three major things discussed in this paper are the distance measure used, the clustering algorithm, and performance experiments. The k-shape algorithm used a centroid based clustering technique (different from other methods like hierarchically and bottom-up/topdown) that uses a shape based distance measure.

This algorithm got tested on the ECGFiveDays dataset. Similar to the MNIST dataset for neural networks, the ECGFiveDays dataset commonly gets used in time-series analysis. This dataset looks at an electrocardiograph, which demonstrates the electrical activity of the heart over time. This dataset and implications have immediate usage since the rapid detection of anomalies in the electrocardiograph can help prevent fatal health issues. The experiment was able to identify the two different classes in the dataset correctly. Additional datasets were used in the experiments section later on.

Critique

This section critiques each article discussed in the prior section.

K-Means

The paper was well constructed and contained little to no grammatical errors. Most of the paper focused on the algorithm itself and all of the math that was associated with it. Rather than having separate sections for experiments, results, and applications, MacQueen lumped them all into a single application section that looked at different distance measures you could use with the algorithm. Compared to newer papers on the subject, this paper had relatively few computer simulations reflecting the cost and scarcity of computing at the time.

Other than the math provided for proofs, no figures, tables, or other diagrams got provided in this paper. Compared to other papers at the time, this was common. Computer graphics were relatively new and very expensive to create.

Due to the k-means algorithm not always converting on an optimum answer and being profoundly affected by outlines, it is rarely used by itself. However, it frequently used to bootstrap and influence other algorithms, as we saw in the K-shape algorithm. As the author stated initially in his article, "there is no feasible, general method which always yield an optimal partition." Due to the nature of NP-Hard problems, this fact is unlikely to change anytime soon. More recently, people have rebooted k-means to include a beam search to avoid converging on local maxima– this process is called "k-means++" and Vassilvitskii outlines it in his 2007 paper titled "K-means++: the advantages of careful seeding" [2].

Another major question posed after this research is: how do we choose k? Although this paper outlines how we can cluster data into k clusters, it did not mention how we should select k. Newer papers are extending the work with k-means and investigating how we can automatically select k using the elbow technique or the goodness value fit (GVF) in Jenks clustering.

This paper has had a lasting effect on the field of machine learning. Most textbooks and AI classes cover k-means clustering as a starting point when teaching people about unsupervised learning. Moreover, algorithms to this day are still using k-means as a tool behind the scenes to pre-process data before it gets fed to the next step in the data pipeline.

K-Shape

As a newer paper k-shape has recent papers looking into it. In 2017 a group of researchers looked into how they could use k-shape to monitor metrics on a distributed system [4]. The K-shape algorithm also has an open-source [3] implementation in python that has a considerable audience. The scalability and performance of the k-shape algorithm, like the original k-mean clustering algorithm, has made this very popular.

The writing style of the k-shape paper was very formal and had no noticeable grammatical issues. The overall structure of the paper is pretty standard for the field. It had an abstract followed by a section going over the background and then a section describing the paper and then experiments, results, and conclusions section. One thing that was unique about this paper is that it contained a whopping 90 references. This abundance of references makes it hard to check all the references used thoroughly. The author tended to pile on a ton of references mentioning the same piece of information. For example: in the introduction to state that time-series data gets used in many fields, the author cited 8 papers. Although it is not a bad thing, it is typically seen as impulsive to include multiple references when citing something that is common knowledge in the field.

The experimental trials that the paper did were very appropriate for the scope of the work that it was doing. Rather than testing the algorithm on one or two datasets, the researchers tested their algorithm on 48 different datasets. The experiments looked at a combination of different clustering algorithms with different distance measures for time-series clustering and compared it against the k-shape algorithm presented in the paper.

This project released the source code – something that many projects don't do. The Matlab source code and their datasets are published on the University of Colombia's website[4]. Disclosing source code is a notable thing to mention since it builds credibility on their paper because it makes it easier to reproduce the results of the experiments ran. However, it is worth noting that running all the experiments took two months to run on a 10

[3] https://github.com/johnpaparrizos/kshape

[4] http://www.cs.columbia.edu/~jopa/kshape.html

server cluster with Intel Xeon processors.

Similar to many research articles, the verbiage of the k-shape paper makes it hard for people outside of this sphere of research to digest. Acronyms like Dynamic Time Warping(DTW) and Euclidean Distance(ED) were frequently used – which is typical for research papers. Some terms, like ECG (Electrocardiograph), were never even defined in the paper. An appendix with acronyms would help less versed readers understand the research paper.

Summary

Although being first introduced in 1967, k-means continues to be a flourishing facet of research in computer science. As we continue to gather and produce more data on the internet, state of the art clustering research has focused on time series and imagery to extract critical information [1].

References

[1] A. Ahmad and S. S. Khan. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902, 2019.

[2] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. volume 8, pages 1027–1035, 01 2007.

[3] M. R. Garey and D. S. Johnson. “ strong ’ ’ np-completeness results: Motivation, examples, and implications. *J. ACM*, 25(3):499–508, July 1978.

[4] Istemi Ekin Akkus Pramod Bhatotia Ruichuan Chen Bimal Viswanath Lei Jiao Christof Fetzer Jörg Thalheim, Antonio Rodrigues. Sieve: Actionable insights from monitored metrics in distributed systems.

[5] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967)., 1967.

[6] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. *ACM SIGMOD Record*, 45:69–76, 06 2016.

[7] Hugo Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*, 4:801–804, 1957.