Mini Project: Implementing K-Means algorithm using Quantum Computing

Rokin Maharjan & Nishan Karki

August 10, 2022

1 Project Summary

1.1 Overview:

Cluster Analysis is an unsupervised machine learning data analysis technique that is used to find hidden patterns within a given dataset [2]. One such unsupervised machine learning algorithm is the K-Means algorithm [3]. K-Means clusters data points based on the pairwise distance between partitions to minimize overall intra-cluster variance [5]. With the adoption of quantum K-Means, we can achieve speed-up in distance calculation if we allow the input and output vectors to be quantum states [5]. In this project, we are going to implement the classic K-means Clustering algorithm using quantum computers to cluster unlabeled primate (monkey) data. Our data set has 115 primates' data points with 11 features. We plan to reduce the data dimension using PCA (Principal Component Analyses) and perform K-Means clustering on the resulting data. The distance calculating part of the K-Means algorithm will be implemented using a quantum circuit in Qiskit.

1.2 Intellectual Merit:

K-means clustering algorithm is a very time and resource-heavy algorithm and can suffer with huge data sets. With the implementation of K-means in quantum computers, we can drastically speed-up the training and prediction time [3]. This project looks forward to adapting classical algorithms to quantum algorithms and dwells upon possibilities of adapting other similar algorithms.

1.3 Broader Impacts of the Proposed Work:

K-Means is a widely popular unsupervised machine learning algorithm. Its uses in the field of biology, big data, chemistry, etc. are well documented. Various techniques have been applied to speed up the run-time of similar algorithms. With the adaption of the K-Means algorithm in quantum computing, we might be able to generate faster classical algorithms.

2 Project Description

2.1 Introduction

The MATRR (Monkey Alcohol Tissue Research Resource) repository has a well-defined categorization of non-human primates (NHP) based on alcohol self-administration studies. Previous work categorized individuals into four categories based on alcohol consumption patterns: low drinking (LD), binge drinking (BD), heavy drinking (HD), and very heavy drinking (VHD) [4]. Further research carried out has asserted the drinking categories based on the 90-day alcohol intake induction period [1]. Though most categorization has been established based on the amount of alcohol consumed, it is unknown whether hormone and immune biomarkers measured during different time points during the experimental period reflect the pre-established categorization, or if a new NHP categorization can be established. Using the MATRR data, we investigate if such NHP categorization based on biomarkers is possible. We plan to use the NHP induction data to generate clusters and group the primates. Our experiment will process and cluster the data using the K-Means algorithm. We further wish to expand the clustering to process through Quantum K-Means. In this study, we plan to discuss NHP categorization and the output analysis between classical K-Means and Quantum K-Means algorithms.

2.2 Methodology

2.2.1 Data Collection

For our experiment, Hormone data from 110 macaques (Rhesus and Cynomolgus macaques), ranging in age from 4 years (adolescent) to 10 years (adult) at the induction time point were measured. To generate precise clusters, we used macaques with a complete hormone biomarkers profile that were also previously classified as LD, BD, HD, and VHD based on their alcohol consumption patterns. We used 'Age', 'Cortisol', 'ACTH', 'Testosterone', 'Deoxycorticosterone', 'Aldosterone', 'DHEAS', 'Osteocalcin' and 'CTX' as primary features for clustering. For our experiments, the complete hormone dataset contained 31 NHPs at the induction timepoint.

2.2.2 Principal Component Analysis (PCA)

After data preprocessing, we generated a dataset containing 31 NHPs with 9 features that were relevant for cluster generation. We normalized the resulting data using *StandardScalar*. We choose standard scalar as it will transform each value in the column to range about the mean 0 and standard deviation 1. After standardization, we performed PCA on the data and reduced the data dimension to 2. The resulting dataset contained 31 primates with their features described by the 2-dimension output data achieved through PCA.

The most popular applications of PCA, an unsupervised linear transformation technique, are feature extraction and dimensionality reduction.. PCA helps us to identify patterns in data based on the correlation between features. It aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

2.2.3 K-means Algorithm

K-means is a distance-based clustering algorithm that groups the data into different clusters[3]. The K-means algorithm is a popular unsupervised iterative clustering technique that partitions the given data into several predetermined clusters based on the similarities exhibited by the data points [2]. Here, k is the number of clusters. The algorithm for k-means clustering is defined below:

1. Select k - the number of clusters
2. Select k random centroids from the data points
3. Assign each data point to its closest centroid. After this, k clusters will be formed.
4. Calculate the mean of each of the clusters and re-assign the data points to the mean. If any re-assignment is done, repeat this step.

We used K-means to initialize k centroids and use the centroids to allocate the data points to the nearest centroid. In our experiment, we chose two centroids (k = 2). Our choice of k was informed by preliminary work showing significant differences between low and heavy drinking animals, and experiments using different numbers of cluster centroids. We applied K-means clustering to generate two clusters for primate categorization. From our experiments, we anticipated generating clusters that distinguish between the NHP categorized as LD/BD and HD/VHD.

2.2.4 Quantum K-Means

K-means algorithm is based on the distance calculation function to find similarities using Euclidean distance. K-means lies in the realm of problems that are NP-hard with a runtime complexity of O(t*k*n*d) [3]. Here, t is the number of iterations before convergence, d is the dimension of the dataset, n is the number of data points, and k is the number of clusters. In general, for a huge amount of data k means is faster than other clustering algorithms but as the size of the data increases the algorithm suffers computationally. This has motivated us to come up with ways to make K-Means computationally efficient. This problem has opened spaces for research for a Quantum K-Means algorithm. With effective implementation, a polylogarithmic runtime of O(N) can be achieved with the implementation of quantum K-Means [3].

In this experiment, we implemented the Euclidean distance calculation between the data points and centers using the Quantum circuit. The quantum circuit is implemented using Qiskit.

2.2.5 Quantum Circuit Design

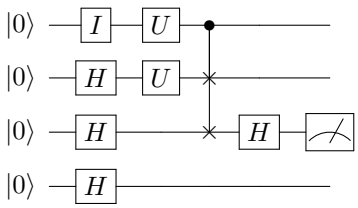
We implemented the quantum circuit using 4 quantum registers and 1 classical register. This implementation is for the Euclidean distance calculation. We use the swap test as a primary methodology to calculate the Euclidean distance. We randomly initialize centroids and then recalculate the distance between the points and new centroids. All data points are assigned to their closest centroids and a cluster label is generated. We set an error rate and iterate through until the new error rate is lower than the threshold error rate. Finally, in the end, we observe the datapoint classified into two cluster labels. We use this cluster label to calculate the accuracy of the Quantum K-Means algorithm.

The circuit design has 4 quantum registers and 1 classical register.

Circuit pseudocode:

```
qc.h(qr[1])
qc.h(qr[2])
qc.h(qr[3])
qc.u(thetalist[0], philist[0], 0, qr[0])
qc.u(thetalist[i], philist[i], 0, qr[1])
qc.cswap(qr[2], qr[0], qr[1])
qc.h(qr[2])
qc.measure(qr[2], cr[0])
qc.reset(qr)
```

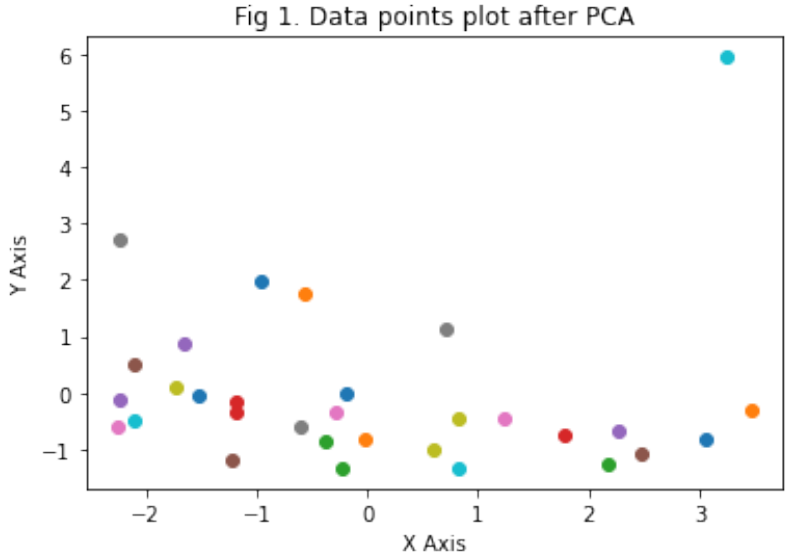
Our circuit design is displayed below:

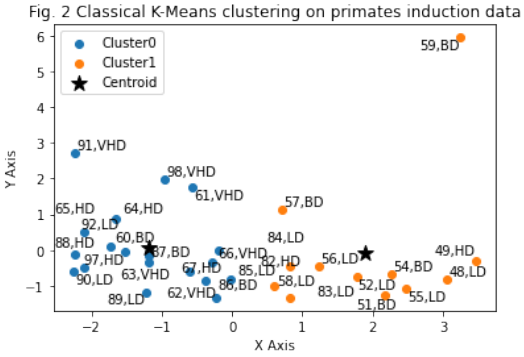


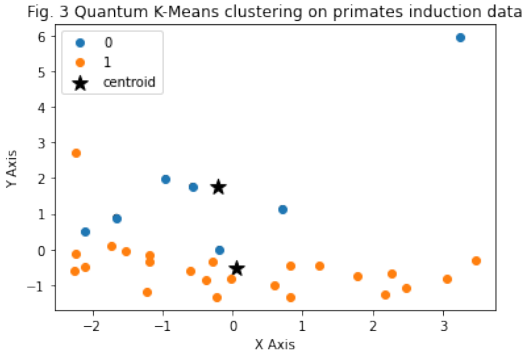
2.3 Results

Our initial task was to implement PCA to reduce the dimension of our dataset. Figure 1. represents the data plots of the output of PCA. This was followed by the implementation of the classical K-Means algorithm. After the implementation of the classical K-Means algorithm, we generated two clusters. The clusters generated has two distinct categorizations. There was no consistent separation between the BD and HD, but a clear separation has been achieved

between LD and VHD. Figure 1. shows the cluster generated after the application of the classical K-Means algorithm. We can see two clusters i.e., Cluster 0 and Cluster 1. From our experiment, we achieved an accuracy of 67.75% for the primate categorization. Figure 3. represents the output of the Quantum K-Means algorithm. The figure shows two clusters (0 and 1) formed. Like classical K-Means the Quantum K-means show no concrete separation between the drinking categories. We achieved an accuracy of 61.3 %.







2.4 Conclusion and Future Works

These findings indicate the hormone data derived from the NHP did not fully replicate the initial categorization based on alcohol consumption patterns. These findings might be due to the low numbers of data for our primary dataset. The accuracy of our model dropped from 67.75% to 61.3% with the implementation of the Quantum K-Means algorithm. This is due to the quantum circuit inefficiency.

For our future work, we wish to implement Quantum and classical K-Means with a larger dataset with more features. We also look forward to testing different circuit designs for quantum K-Means.

References

[1] Erich Baker, Nicole Walter, Alex Salo, Pablo Rivas, Sharon Moore, Steven Gonzales, and Kathleen Grant. Identifying future drinkers: Behavioral analysis of monkeys initiating drinking to intoxication is predictive of future drinking classification. *Alcoholism, clinical and experimental research*, 41, 01 2017.

[2] Stephen DiAdamo, Corey O'Meara, Giorgio Cortiana, and Juan Bernabé-Moreno. Practical quantum k-means clustering: Performance analysis and applications in energy grid classification, Dec 2021.

[3] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum algorithms for supervised and unsupervised machine learning, Nov 2013.

[4] Sharon Moore, Ami Radunskaya, Elizabeth Zollinger, Kathleen Grant, Steven Gonzales, and Erich Baker. Time for a drink? a mathematical model of non-human primate alcohol consumption. *Frontiers in Applied Mathematics and Statistics*, 5, 02 2019.

[5] Peter Wittek. *Quantum machine learning what quantum computing means to data mining*. Elsevier, AP, 2016.