Quantum K-Means Clustering Method For Detecting Heart Disease Using Quantum Circuit Approach

Kavitha S S (✉ kavithaece@nie.ac.in)

The National Institute of Engineering <https://orcid.org/0000-0002-0843-5143>

Narasimha Kaulgud

The National Institute of Engineering

Research Article

Keywords: Quantum, computing, clustering, classical

Posted Date: October 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-954132/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Quantum K-means Clustering method for detecting heart disease using Quantum circuit approach

Kavitha S S¹ and Dr. Narasimha Kaulgud²

¹ The National Institute of Engineering, Mysuru

² The National Institute of Engineering, Mysuru

kavithaece@nie.ac.in

narasimhakaulgud@nie.ac.in

Abstract. The development of Noisy Intermediate Scale Quantum computers is expected to signify potential advantages of quantum computing over classical computing. This paper focuses on quantum paradigm usage to speed up unsupervised machine learning algorithms particularly the K-means clustering method. The main approach is to build a quantum circuit that performs the distance calculation required for the clustering process. This proposed technique is a collaboration of data mining techniques with quantum computation. Initially extracted heart disease dataset is preprocessed and classical K-means performance is evaluated. Later, the quantum concept is applied to the classical approach of the clustering algorithm. The comparative analysis is performed between quantum and classical processing to check performance metrics.

1 Introduction

Machine learning plays a vital role in solving various highly challenging tasks due to its ability to handle computational improvements. Although computing power has grown rapidly in recent decades and new algorithms have hit the market regularly, the growth in data has outpaced the growth in computing power. As a result, in the field of artificial intelligence, which in many cases relies on massive data, a lack of computational capacity eventually becomes a shortcoming. Here the quantum computer comes into the picture which has a high speed of computing due to its parallelism feature which can solve specific algorithms with specific problems [1][2]. As the dataset becomes huge, it requires fast and efficient algorithms to process. Polynomial-time algorithms aren't always thought to be the most efficient as it takes more time to process it [3]. Therefore it is necessary to preprocess the data by applying PCA and standard minimizer before applying ML algorithms. Developing healthcare systems based on patient data models and technologies can significantly improve the decision-making process. However, existing ML models are generally characterized by a sophisticated quantum of data that leads to process overhead [4] [5]. This high dimensionality of the data makes it difficult to visualize even in 3D, which is

known as the curse of dimensionality [6]. As a result, performing operations on this data requires a lot of memory, and the data can grow exponentially at times, leading to over fitting. The weighting features are utilized to eliminate redundancy in the dataset, which helps to speed up the execution time [7] [8]. While dealing with high dimensional data, PCA is more likely used dimensionality reduction technique. However, processing overhead still exists taking much more time to exhibit the result. Executing these ML algorithms on a quantum system reduces processing overhead and also improves/matches with performance metrics of classical systems. To perform this, classical algorithms are re-designed to match to the principles of quantum mechanics [9]. This manuscript presents a quantum K-means approach using a quantum circuit for unsupervised learning with the goal of performing heart disease prediction. The major contributions are as follows:

1. Preprocess the heart disease prediction dataset by applying Principal component analysis(PCA) and standard minimizer, then and the K value in the K-means algorithm using the Elbow method.
2. Calculate the distance using Euclidean distance calculation equations in classical and convert them into quantum states using quantum circuit analogy.
3. Comparison of classical and quantum approach for k-means clustering.

Methodology carried out to predict heart disease in classical and quantum context using k-means clustering method is shown as 1 In this article, quantum K-means clustering method is used, which provides a great foundation to cluster unlabeled data with less time complexity of $O(LNK)$ compared to a classical version that takes $O(LNpK)$ time. The empirical evaluations show the benefits of this work. First, the heart disease prediction dataset preprocessing stage is summarized (section II). Then the classical k-means algorithm to cluster the dataset is explained briefly (section III). Implementation and analysis of Quantum K-means using quantum circuit are provided in section IV. Followed by this results obtained by the classical and quantum approach are tabulated in section IV and finally, section V presents the conclusion.

Fig. 1: Process flow for clustering

2 Heart disease prediction data description

The data content of the heart disease prediction consists collection of records, representing the diagnosis of the patient heart disease [10]. The extracted dataset was the Public Health dataset [11]. It contains 13 features including the predicting attribute. Since, the chosen dataset contains 8 categorical data, it is handled by converting to integer during preprocessing stage as shown in Table 1. data The "target" field indicates whether or not the patient has the cardiac disease. Target = '0' refers to no disease and Target = '1' refers presence of disease. A total of 1025 patients record are summarized for the prediction of cardiac disease. The first two columns tabulate the 'age' and 'sex' of the patient. The rest of the columns describes the type of chest pain, resting blood pressure value, cholesterol level, fasting blood sugar value, resting electrocardiographic results, maximum heart rate, exercise-induced angina, Oldpeak referring to ST depression, the slope of ST-segment [12].

Table 1: Preprocessing dataset by converting Category to Integer

Attributes	Type	Data Encoding (Category to Integer)
sex	Category	Male:1, Female:2
chest_pain_type	Category	Typical angina:1, Non-anginal pain:2, Atypical angina:3, Asymptomatic:4
fasting_blood_sugar	Category	Lower_than_120:1, Greater_than_120:2
rest_ecg	Category	ST_T_wave_abnormality:1, Normal:2, Left_ventricular_hypertrophy:3
exercise_induced_angina	Category	No:1, Yes:2
slope	Category	Flat:0, Downsloping:1, Upsloping:2
vessels_colored_by_flourosopy	Category	Zero:1, One:2, Two:3, Three:4, Four:5
thalassemia	Category	Fixed Defect:0, Reversable Defect:1, Normal:2, No:3

Fig. 2: Outlier visualization

2.1 Heart disease data preprocessing

As the dataset is not properly distributed and consists many outliers, data has to undergo preprocessing. Initially, the amount of outliers present in the dataset is detected. These outliers are removed from data using the mathematical function IQR(Inter Quartile Range). Using the quantile method provided by python find Q1 and Q3 values, which are then subtracted ($Q3 - Q1$) to obtain IQR. Lower bound($Q1 - 1.5 * IQR$) and Upper bound($Q3 + 1.5 * IQR$) values are used to check outlier and this outlier can be easily removed from data as shown in fig 2.

2.2 Data visualization

During preprocessing, the Pairs Plot visualization technique is used to see the distribution of single variables and the relationship between two variables. This helps to find the most separated clusters. Based on the dataset analysis oldpeak, vessels_colored_by_flourosopy, resting_blood_pressure, rest_ecg features are dropped. Fig 3 shows the correlation among features and hence chosen for subsequent analysis.

Fig. 3: Pair Plot of 13 features

2.3 Dimensionality reduction technique and feature selection

Increasing dimensionality results in computationally expensive and poor performance of clustering algorithms. The Principal Component Analysis (PCA) and standard MinMax scaler are used to compress higher dimensional data for dimensionality reduction of the dataset. Setting the dimensionality to 2 has the added benefit that clusters can be visualized appropriately. MinMax scaling often called Normalization is used to scale the data to a fixed range from +1 to -1. This results in having smaller standard deviations and thereby suppressing the effect of outliers.

3 Classical K-means clustering method

K-means clustering [13] is an unsupervised clustering method first proposed by Lloyd [14]. It is one of the most extensively used clustering methods. Let $X = x_1, x_2, \dots, x_N$ represent the datapoints, where each row element $x_n \in \mathbb{R}^M, 1 \leq n \leq N$ consists of 'M' number of features. Clustering of datapoints is driven by a set of centroids $W = w_1, w_2, \dots, w_K, w_n \in \mathbb{R}^M, 1 \leq n \leq K$ generated from the

dataset X. Thus set of datapoints is divided into K clusters $C = C_1, C_2, \dots, C_K$ by minimizing sum of squared errors provided using "Elbow method". As shown in fig 4 the k value is selected as 2.

$$E = \sum_{k=1}^K \sum_{n \in C_k} \|x_n - w_k\|^2 \quad (1)$$

Fig. 4: Elbow plot for selecting k value in cluster

The classical K-means algorithm described below iterates between two major steps: 1) Updating of the clusters 2) Updating of the centroids.

Algorithm 1 Classical K-means algorithm

Input Datapoints $X = x_1, x_2, \dots, x_N, x_n \in \mathbb{R}^M, 1 \leq n \leq N$ consists of 'M' number of features, set initial centroids $W = w_1, w_2, \dots, w_K, w_n \in \mathbb{R}^M, 1 \leq n \leq K$.

Output Number of K clusters consisting of vectors C_k

repeat

Assignment step(Clustering updation): New datapoint is assigned to the cluster C_{k^*} , k^* is computed by:

$$k^* = \operatorname{argmin}_{k=1,2,\dots,K} \|x_n - w_k\|^2$$

Centroid Updation: Centroid w_k is updated for each cluster C_k by:

$$w_k = \frac{1}{\sqrt{|C_k|}} \sum_{n=1}^N g_{nk} x_n$$

until a stopping condition is reached

The time-complexity of the algorithm is specified by $O(LNpK)$, where L = number of iterations taken to form clusters, N = number of datapoints, p = dataset dimension, K = centroids.

4 Quantum K-means clustering

The initial step to achieve the Quantum machine learning process is to convert the data points into quantum states [15]. Following that, the distance between each state's centroids is computed. In terms of quantum learning, estimating the distance between quantum states is completely different from classical learning. The distance computation in the classical version is performed using coordinates, whereas in quantum learning the distance calculation is exhibited in terms of the probabilistic nature of qubits. Although quantum learning allows for the measurement and quantification of notions such as phase differences and amplitudes of different probabilities, the calculation of the distance between two states is unfortunately not achievable due to quantum noise and instability [16].

Set $|\phi\rangle$ to be the centroid cluster in the chosen dataset and $|\psi\rangle$ to be the new datapoint used to find for which cluster it belongs to. To find the distance between these two vectors, one more ancillary qubit in state $|0\rangle$ is used. A Hadamard gate is appended to qubit $|0\rangle$ to bring it to a superposed state. Using swap gate-controlled on the auxiliary qubit, the distance between the state $|\phi\rangle$ and the state $|\psi\rangle$ entangled with the ancillary qubit is computed. The ancillary qubit is then subjected to another Hadamard gate before being measured. Each state is assigned to its nearest cluster after the distance between each state and centroid is determined using a swap test circuit. Iterative steps are carried out by updating new centroids of each cluster till the stop criterion is satisfied. Quantum K-means clustering is concluded in three subroutines: 1) Distance calculation using swap-test-circuit 2) Cluster updation 3) Centroid updation.

4.1 Distance calculation using swap-test-circuit

Prior to computing the distance between two vectors, it is necessary to encode the data from classical to Quantum world. This is carried out by Amplitude encoding, which can be briefly summarized as

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{2^n} \end{bmatrix} = \sum_{i=1}^{2^n} x_i |i\rangle \quad (2)$$

In a quantum context, datapoint assignment to the relative cluster is achieved by the overlap function, which measures the state of overlap between two vectors [17]. Outputting '1' indicates, two vectors are identical, and '0' indicates they are

Fig. 5: Quantum-swap-test-circuit

orthogonal or no overlap found. Turning this analogy to quantum part requires a swap test that results in a squared probability of the overlap fig ??

Initially, two quantum states $|\psi\rangle$ and $|\phi\rangle$ and one ancillary qubit $|0\rangle$ is chosen and its outer product is given as

$$|S_0\rangle = |0, |\psi\rangle, |\phi\rangle\rangle = |0\rangle \otimes |\psi\rangle \otimes |\phi\rangle \quad (3)$$

Applying Hadamard gate to the ancillary qubit, an superposition state is obtained as:

$$|S_1\rangle = \frac{1}{\sqrt{2}}(|0, |\psi\rangle, |\phi\rangle\rangle + |1, |\psi\rangle, |\phi\rangle\rangle) \quad (4)$$

The controlled SWAP-gate is then applied to both $|\psi\rangle$ and $|\phi\rangle$ in order to swap $|\psi\rangle$ and $|\phi\rangle$ with the target qubit $|1\rangle$. This results in

$$|S_2\rangle = \frac{1}{\sqrt{2}}(|0, |\psi\rangle, |\phi\rangle\rangle + |1, |\phi\rangle, |\psi\rangle\rangle) \quad (5)$$

Another Hadamard gate is placed in the ancillary qubit at the end, resulting in:

$$|S_3\rangle = \frac{1}{\sqrt{2}} \left(\left| \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \psi, \phi \right\rangle + \left| \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \phi, \psi \right\rangle \right) \quad (6)$$

Expanding the above equation results in:

$$|S_3\rangle = \frac{1}{2}|0\rangle(|\psi, \phi\rangle + |\phi, \psi\rangle) + \frac{1}{2}|1\rangle(|\psi, \phi\rangle - |\phi, \psi\rangle) \quad (7)$$

Finally, the measurement is applied to the ancillary qubit [15]. The probability of getting state $|0\rangle$ is given by,

$$\begin{aligned}
P(|0\rangle) &= \left| \frac{1}{2} \langle 0|0 \rangle (|\psi, \phi\rangle + |\phi, \psi\rangle) + \frac{1}{2} \langle 0|1 \rangle (|\psi, \phi\rangle - |\phi, \psi\rangle) \right|^2 \\
&= \frac{1}{4} |(|\psi, \phi\rangle + |\phi, \psi\rangle)|^2 \\
&= \frac{1}{4} (\langle \phi, \psi | + \langle \psi, \phi |) (|\psi, \phi\rangle + |\phi, \psi\rangle) \\
&= \frac{1}{4} (\langle \phi | \psi \rangle \langle \psi | \phi \rangle + \langle \phi | \phi \rangle \langle \psi | \psi \rangle + \langle \psi | \psi \rangle \langle \phi | \phi \rangle + \langle \psi | \phi \rangle \langle \phi | \psi \rangle) \\
&= \frac{1}{4} (|\langle \psi | \phi \rangle|^2 + 1 + 1 + |\langle \psi | \phi \rangle|^2) \\
&= \frac{1}{2} + \frac{1}{2} |\langle \psi | \phi \rangle|^2
\end{aligned} \tag{8}$$

If the probability value $P(|0\rangle) = 0.5$, the two states ϕ and ψ are orthogonal or have no overlap, and if $P(|0\rangle) = 1$, the two states are identical. Similarly, $P(|1\rangle)$ is specified by

$$P(|1\rangle) = \frac{1}{2} - \frac{1}{2} |\langle \psi | \phi \rangle|^2 \tag{9}$$

Futhermore, distance between 2 states are computed by,

$$Dist = d^2 r_{max}^4 (2P(|0\rangle) - 1) \tag{10}$$

Where, d = input non zero datapoints, r_{max} = maximum limit of any feature in the data. Thus the overlap function leads a foundation to perform clustering tasks.

4.2 Cluster updation

After calculating the distance between the new data-point and each cluster-centroid, the nearest-neighborfunction is used to allocate the new data-point to the closest cluster [18]. Cluster assignment is given by,

$$C_k = Cluster(X_n) = \arg \min \sum_{k=1}^K \sum_{x_n \in C_k} g_{nk} \|x_n - w_k\|^2 \tag{11}$$

$$g_{nk} = \begin{cases} 1 & \text{if } x_n \in C_k \\ 0 & \text{Otherwise} \end{cases} \tag{12}$$

4.3 Centroid updation

Centroids are recomputed by taking the mean of all the datapoints assigned to a particular cluster. Centroid updation is given by,

$$w_k = \frac{1}{\sqrt{|C_k|}} \sum_{n=1}^N g_{nk} x_n \quad (13)$$

These three phases are performed until the cluster centroid's position does not change, indicating convergence. The algorithm for Q K-means is summarized below:

Algorithm 2 Quantum K-means algorithm

Input: Convert input datapoints $X = x_1, x_2, \dots, x_N, x_n \in \mathbb{R}^M, 1 \leq n \leq N$ into Quantum states $|\psi\rangle |\phi\rangle$ using Amplitude encoding. Initial cluster centroid: $|w_1\rangle, |w_2\rangle, \dots, |w_K\rangle$

Output: K clusters consisting of vectors C_k **Distance Calculation using swap-test:**

$$Dist = d^2 r_{max}^4 (2P(|0\rangle) - 1)$$

repeat

Clustering Updation): Each data is assigned to the nearest cluster using nearest- neighbor function:

$$C_k = Cluster(X_n) = \arg \min \sum_{k=1}^K \sum_{x_n \in C_k} g_{nk} \|x_n - w_k\|^2$$

Centroid Updation: Centroid w_k of each cluster is updated by-

$$w_k = \frac{1}{\sqrt{|C_k|}} \sum_{n=1}^N g_{nk} x_n$$

until Convergence is reached

5 Result and analysis

The number of gates that can be executed and the number of qubits that can be used in a given algorithm are both limited in IBM Q Experience. This dataset is preprocessed before distance calculation for clustering. By using principal component analysis(PCA), the heart disease dataset consisting of 1025 features are reduced to 2 features. Quantum computers could also be used for dimensionality reduction algorithms such as [19] [20]. Implementation and execution of the algorithm is carried out in classical and IBM's open-source quantum software

development kit QISKIT [21]. Qiskit allows an access to several quantum simulators both locally and in the cloud. A state vector simulator can simulate circuits with up to 25 qubits, whereas a unitary simulator can simulate circuits with up to 12 qubits. For the chosen Heart disease prediction the original data scatter and processed cluster steps are shown in fig 6, similarly clustering process carried out in Quantum system is shown in fig. 7. The results from the classical k-means are compared with the quantum k-means algorithm. Table 2 summarizes the results obtained on simulation on classical and quantum systems.

Table 2: Comparison between classical and quantum clustering method

Methods	Accuracy	Execution Time (Seconds)	Precision	Recall	F1-score
Classical K-means	83.3	1.49	83	82	83
Quantum K-means (Qasm Simulator)	82.9	1.26	83	82	82
Quantum K-means (statevector Simulator)	82.4	0.73	82	81	82

Fig 6: Classical K-means clustering, fig(a):Original scattered dataset, fig(b):Clustering process iteration 1, fig(c):Clustering process iteration 2, fig(d):Final clustered data

Clustering achieved between classical and quantum shows relatively comparable results. Accuracy of quantum method is decline by a small amount, this is due to quantum noise present in the system. Other parameters such as precision, recall, F1-score shows comparable results. Quantum k-means executed in statevector simulator exhibits less execution time compared to classical system. This proves the supremacy of quantum computer in terms of speed. As the dataset becomes huge, Quantum systems performs better than classical in terms of speed. Along with this, confusion matrix for classical and quantum gives the measure of performance of clustering fig 8. The quantum circuit used to perform clustering is shown in fig 9.

Fig. 7: Quantum K-means clustering, fig(a):Original scattered dataset, fig(b):Clustering process iteration 1, fig(c):Clustering process iteration 2, fig(d):Final clustered data

Fig. 8: Confusion matrix

Fig. 9: Quantum Circuits for Clustering

6 Conclusion

The possibility of quantum computers to give exponential speedups has prompted substantial research into quantum algorithms. Quantum computers currently have some constraints in terms of quantum noise and qubit coherence times. These impediments limit their ability to solve problems with reasonable precision. In this article, heart disease prediction is presented using the classical and Quantum k-means method. Quantum k-means clustering is performed utilizing a set of quantum-gates consisting less number of quantum operations and thereby resulting in significant improvement in terms of accuracy and execution time. In contrast to traditional K-means clustering, which takes $O(LNpK)$ time, quantum mechanical K-means clustering takes only $O(LNK)$ time. This might save a lot of time in data mining applications, where data dimensions are typically in the hundreds.

- Funding: The work presented came from the PhD research undertaken at Department of Electronics and Communication Engineering, The National Institute of Engineering, Mysuru, Karnataka, India.
- Conflict: There is no conflict of interests that are directly or indirectly related to the work submitted for publication.
- Availability of data and material: Yes, data and material for transparency will be made available.
- Code availability: N/A.
- Author’s contributions: All authors contributed to the study conception and design.

References

- [1] Iordanis Kerenidis et al. “q-means: A quantum algorithm for unsupervised machine learning”. In: *arXiv preprint arXiv:1812.03584* (2018).
- [2] Abhijat Sarma et al. “Quantum unsupervised and supervised learning on superconducting processors”. In: *arXiv preprint arXiv:1909.04226* (2019).
- [3] Dongkuan Xu and Yingjie Tian. “A comprehensive survey of clustering algorithms”. In: *Annals of Data Science* 2.2 (2015), pp. 165–193.
- [4] Amin Ul Haq et al. “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms”. In: *Mobile Information Systems* 2018 (2018).
- [5] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. “Effective heart disease prediction using hybrid machine learning techniques”. In: *IEEE access* 7 (2019), pp. 81542–81554.
- [6] Eamonn J Keogh and Abdullah Mueen. “Curse of dimensionality.” In: *Encyclopedia of machine learning and data mining* 2017 (2017), pp. 314–315.
- [7] Miin-Shen Yang and Yessica Nataliani. “A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy”. In: *IEEE Transactions on Fuzzy Systems* 26.2 (2017), pp. 817–835.

- [8] Dietrich Wetttschereck, David W Aha, and Takao Mohri. “A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms”. In: *Artificial Intelligence Review* 11.1 (1997), pp. 273–314.
- [9] Sumsam Ullah Khan, Ahsan Javed Awan, and Gemma Vall-Llosera. “K-means clustering on noisy intermediate scale quantum computers”. In: *arXiv preprint arXiv:1909.12183* (2019).
- [10] Fuad Ali Mohammed Al-Yarimi et al. “Feature optimization by discrete weights for heart disease prediction using supervised learning”. In: *Soft Computing* 25.3 (2021), pp. 1821–1831.
- [11] Ketan Gangal. *Heart disease dataset uci*. Aug. 2021. URL: <https://www.kaggle.com/ketangangal/heart-disease-dataset-uci?select=HeartDiseaseTrain-Test.csv>.
- [12] Rohit Bharti et al. “Prediction of heart disease using a combination of machine learning and deep learning”. In: *Computational Intelligence and Neuroscience* 2021 (2021).
- [13] Shi Na, Liu Xumin, and Guan Yong. “Research on k-means clustering algorithm: An improved k-means clustering algorithm”. In: *2010 Third International Symposium on intelligent information technology and security informatics*. Ieee. 2010, pp. 63–67.
- [14] Stuart Lloyd. “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.
- [15] Kaoutar Benlamine et al. “Quantum Collaborative K-means”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.
- [16] Changqing Gong et al. “Quantum k-means algorithm based on Trusted server in Quantum Cloud Computing”. In: *Quantum Information Processing* 20.4 (2021), pp. 1–22.
- [17] Yao Zhang and Qiang Ni. “Recent advances in quantum machine learning”. In: *Quantum Engineering* 2.1 (2020), e34.
- [18] Giuseppe Sergioli et al. “A quantum-inspired version of the nearest mean classifier”. In: *Soft Computing* 22.3 (2018), pp. 691–705.
- [19] Iris Cong and Luming Duan. “Quantum discriminant analysis for dimensionality reduction and classification”. In: *New Journal of Physics* 18.7 (2016), p. 073011.
- [20] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. “Quantum principal component analysis”. In: *Nature Physics* 10.9 (2014), pp. 631–633.
- [21] Gadi Aleksandrowicz et al. “Qiskit: An open-source framework for quantum computing”. In: *Accessed on: Mar 16* (2019).