Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090202
ISSN 2371-8838 Vol. 9 Num. 2

# *Improvement of K-means Clustering Algorithm Based on Quantum State Similarity Measurement*

**Hongfei Zhang, Mingwei Li***

*Northeastern University at Qinhuangdao, Qinhuangdao, 066004, China*
**Corresponding author*

***Keywords:*** Quantum K-means clustering; Quantum state encoding; Similarity metrics for quantum-classical mixing

***Abstract:*** The classical K-means clustering algorithm is widely used in various fields due to its simple implementation and efficient computation, but the classical K-means clustering algorithm relies on the random selection of the initial center of mass, which is prone to fall into the deadlock of local optimality. In order to break through this limitation, the quantum K-means clustering algorithm is introduced, which is able to explore multiple potential clustering center combinations at the same time through the parallelism of quantum computation, so as to have a greater probability of converging to the globally optimal solution. Quantum K-means clustering algorithms typically employ fidelity as a similarity measure between quantum states, and similarity is assessed by calculating the probability of overlap between quantum states. However, the fidelity only quantizes the pure state information of the quantum states and ignores the classical statistical features of the data itself, which may lead to unreasonable clustering boundaries in mixed state or noise interference scenarios. In response to the above problems, this paper proposes an improved quantum-classical hybrid similarity metric, whose core idea is to incorporate the dual constraints of quantum information and classical features.

## 1. Introduction

In today's data-driven scientific and industrial applications, although classical machine learning algorithms have made remarkable achievements, they face computational challenges when the amount of data increases. In recent years, with the accelerated development of quantum technology, quantum computing has demonstrated superior performance in dealing with classical problems, e.g., Shor[1] proposed the Las Vegas algorithm for finding discrete logarithms and factorizing integers, proving the exponential acceleration advantage of quantum computers for these two types of number theoretic problems, revealing the potential of quantum computing to subvert the classical computing paradigm. In 1996, Grover's quantum search algorithm[2] was used to solve the search problem of unstructured databases, proving the advantage of quantum computing in search problems and becoming an important part of quantum algorithms. Based on the proposal of these algorithms, some scholars have widely applied them as subroutines in various quantum machine learning algorithms. Anguita[3] et al. used Grover's quantum search algorithm in support vector machine to improve the training efficiency to optimize SVM. Ruan[4] et al. proposed a quantum principal component analysis method applied to face recognition, which utilizes the quantum state to face feature encoding and using Grover quantum search algorithm for face recognition with secondary acceleration to improve the efficiency of the algorithm.

Quantum computing has also shown great potential in tasks such as clustering and classification. Kerenidis[5] et al. extended the work of quantum machine learning by proposing an end-to-end quantum algorithm to perform spectral clustering. The algorithm is capable of completing the clustering task with high accuracy and a more efficient runtime, which provides a new way of thinking about other machine learning and optimization algorithms based on graph structures. Wiebe[6] et al. proposed a quantum algorithm for nearest neighbor classification can effectively handle the task of classifying datasets with high-dimensional feature space and large-scale training samples, and is robust to noise interference, and the algorithm outperforms classical algorithms in terms of time performance and classification accuracy.

The classical K-means clustering algorithm is widely used in unsupervised learning tasks due to its simple operation and high efficiency, while quantum computing provides a potential solution for clustering big data, the quantum K-means clustering algorithm is proposed in this context, which aims to optimize the classical K-means clustering algorithm by taking advantage of the advantages of quantum computing. Khan[7] et al. proposed quantum K-means clustering analysis using shallow quantum circuits, which not only reduced the number of quantum operations but also significantly improved the accuracy of the algorithm. Arthur[8] et al. proposed a quantum method for training balanced K-means clustering models, which is a quantum scheme that can more efficiently approximate the globally optimal solution of the training problem than the classical method, and shows better scalability when dealing with large-scale datasets. Ohno[9] proposed a quantum subroutine for the quantum-enhanced K-means algorithm, which achieves algorithm optimization by eliminating the traditional center-of-mass computation step. Based on the principle of quantum entanglement, the quantum subroutine can output an estimate of the Euclidean distances between the data points and the clustering center-of-mass for a given set of clusters, and it is capable of achieving exponential speedup for large-scale datasets. The current status of these studies shows the potential of quantum computing and lays a solid foundation for the future development of quantum technology.

The main contributions of this paper are:

First, different quantum state encoding approaches are taken and the effects of these encoding approaches on the quantum K-means clustering algorithm are compared.

Second, a quantum-classical hybrid distance is proposed considering the influence of classical data itself on the clustering effect.

Finally, three UCI datasets are selected to compare the quantum K-means clustering algorithm with the classical K-means clustering algorithm to verify the effectiveness of the proposed method.

## 2. Quantum K-means clustering based on improved quantum state similarity metrics

Quantum K-means clustering enhances classical K-means through quantum computing capabilities, leveraging quantum parallelism and state superposition for efficient processing of large-scale and high-dimensional datasets. The algorithm encodes data points and centroids as quantum states manipulated via quantum circuits, employing core quantum techniques including state encoding and quantum distance measurement.

### 2.1 Quantum state encoding approach

Quantum computation demands specialized encoding to bridge classical-quantum data compatibility, achieved through mathematical mappings into Hilbert space via quantum gates. This section analyzes two primary encoding schemes: amplitude and angle encoding, which maintain critical data properties while enabling quantum superposition and entanglement operations.

### 2.1.1 Amplitude encoding

Amplitude encoding achieves exponential quantum resource efficiency by embedding high-dimensional vector data into quantum state amplitudes. Specifically, for a normalized real vector $\vec{x}=(x_1,x_2,\cdots x_N)$ of length $N$, it can be mapped onto a superposition state of $n=\log_2 N$ quantum bits by amplitude encoding, which is formally represented as:

$$|\psi\rangle=\sum_{i=1}^{N}x_i|i\rangle \tag{1}$$

Where $|i\rangle$ is the computational ground state of the $n$ quantum bit system. Amplitude encoding's logarithmic qubit scaling enables efficient handling of complex data, remaining vital for quantum ML and chemistry despite stringent normalization requirements.

### 2.1.2 Angle encoding

Angle encoding converts classical data into quantum states by mapping input values to rotation angles of qubit gates, enabling parameter embedding through single-qubit rotations. In terms of implementation, a single quantum bit can encode multiple classical parameters by means of a generic single quantum bit gate $U_3(\theta,\varphi,\lambda)$. For $n$ dimensional classical data $x=(x_1,x_2,\cdots,x_n)$, an angle encoding method corresponding to one quantum bit per dimension can also be used, and if the $R_y(\theta)$ gate is used, the classical data can be encoded into the following quantum state:

$$|\psi\rangle=\bigotimes_{i=1}^{n}R_y(x_i)|0\rangle^{\otimes n} \tag{2}$$

This encoding creates tensor product states through gate rotation parameters, achieving hardware-efficient implementation ideal for dynamic parameter optimization in variational quantum classifiers.

## 2.2 Parameter selection for the quantum state encoding method

Quantum encoding parameters serve as critical interfaces between classical data and quantum feature spaces, governing model representational capacity and convergence. This work implements quantum K-means clustering with amplitude and angle encoding, and will analyze their parameter configurations.

(1) Amplitude encoding

When using amplitude encoding, the single quantum bit gate usually selected is the $R_y(\theta)$ gate, and the calculation of $\theta$ is mainly performed in the following way: for any data point $(a,b)$ in the data set, which needs to be normalized first in order to satisfy the normalization condition of the quantum state, there are

$$\bar{a}=\frac{a}{\sqrt{a^2+b^2}}\ ,\ \bar{b}=\frac{b}{\sqrt{a^2+b^2}} \tag{3}$$

Then the data point $(a,b)$ amplitude can be encoded as

$$|\psi\rangle=\bar{a}|0\rangle+\bar{b}|1\rangle \tag{4}$$

Because any single quantum bit state $|\psi\rangle$ , can be expressed as:

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + \sin\left(\frac{\theta}{2}\right)|1\rangle \tag{5}$$

Where $\theta$ is a polar angle and $\theta \in [0,\pi]$ , so the coefficients correspond to have

$$\theta = 2\arccos\frac{a}{\sqrt{a^2+b^2}} \tag{6}$$

Then the data point $(a,b)$ can be encoded as

$$|\psi\rangle = R_y(\theta)|0\rangle \tag{7}$$

(2) Angle encoding

In the SWAP test circuit utilizing angle encoding, the single quantum bit gate usually selected is the $U_3(\theta,\varphi,\lambda)$ gate, which will cause information redundancy if the three parameters are selected simultaneously, so in this paper, we make $\lambda = 0$. The parameters $\theta$ and $\varphi$ are computed mainly in the following way: for any data point $(a,b)$ in the dataset, it is normalized and transformed to[10]

$$\theta = \frac{\pi}{2}\left(\frac{x}{x^2+y^2}+1\right) \tag{8}$$

$$\varphi = \frac{\pi}{2}\left(\frac{y}{x^2+y^2}+1\right) \tag{9}$$

Then the data point $(a,b)$ can be encoded as

$$|\psi\rangle = U_3(\theta,\varphi,0)|0\rangle \tag{10}$$

## 2.3 Improved quantum state similarity metrics

For two pure states $|\varphi\rangle$ and $|\psi\rangle$, fidelity as the fundamental distance metric between data representations is defined as the squared overlap:

$$F(|\varphi\rangle,|\psi\rangle) = |\langle\varphi|\psi\rangle|^2 \tag{11}$$

the range from 0 to 1, with higher values indicating greater proximity in Hilbert space. The SWAP test quantifies quantum state similarity by entangling states through controlled SWAP gates and measuring ancilla qubit probabilities. The test circuit and its core steps are as shown in Figure 1:

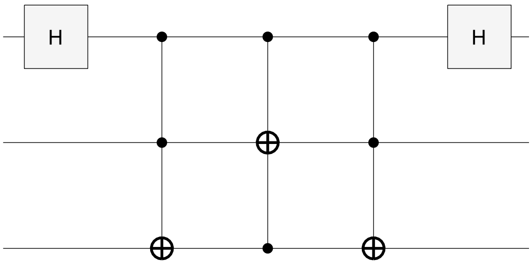

Fig. 1 SWAP test circuit diagram

First, the auxiliary quantum bit is initialized to $|0\rangle$ , and the quantum states to be compared $|\varphi\rangle$ and $|\psi\rangle$ are stored in two registers, respectively, so that the system input states are

$$|\psi_1\rangle = |0\rangle \otimes |\varphi\rangle \otimes |\psi\rangle \tag{12}$$

Applying a Hadamard gate to the auxiliary quantum bit yields

$$|\psi_2\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right)|\varphi\rangle|\psi\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle|\varphi\rangle|\psi\rangle + |1\rangle|\varphi\rangle|\psi\rangle\right) \tag{13}$$

Applying SWAP gate yields

$$|\psi_3\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle|\varphi\rangle|\psi\rangle + |1\rangle|\psi\rangle|\varphi\rangle\right) \tag{14}$$

Finally, applying another Hadamard gate to manipulate the auxiliary quantum bits yields

$$|\psi_4\rangle = H|\psi_3\rangle = \frac{1}{2}\left[|0\rangle\left(|\varphi\rangle|\psi\rangle + |\psi\rangle|\varphi\rangle\right) + |1\rangle\left(|\varphi\rangle|\psi\rangle - |\psi\rangle|\varphi\rangle\right)\right] \tag{15}$$

Then the probability of measuring auxiliary quantum bits to get $|0\rangle$ is:

$$P(0) = \langle\psi_4|0\rangle\langle 0|\psi_4\rangle = \frac{1}{2} + \frac{1}{2}\langle\psi|\varphi\rangle^2 \tag{16}$$

Specifically, when the amplitude encoding in Section 3.2 is used, The test circuit is refined as shown in Figure 2:

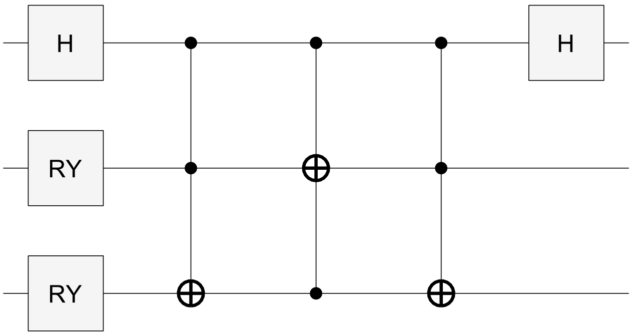

Fig. 2 SWAP test circuit for amplitude encoding

When the angle encoding in Section 3.2 is used, the test circuit is refined as follows in Figure 3:

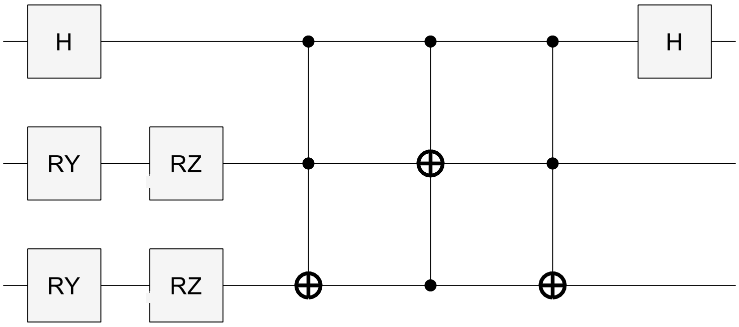

Fig. 3 SWAP test circuit for angle encoding

While quantum state fidelity effectively measures similarity in Hilbert space, traditional quantum

K-means clustering overlooks the original data's geometric topology. Furthermore, dimensional compression during quantum encoding induces local manifold distortion, adversely affecting clustering performance. To overcome the above limitations, this paper innovatively proposes a hybrid quantum-classical distance metric. Define the hybrid quantum-classical distance metric formula as:

$$dis=\sqrt{\max_{i}\left(\left|x_{1i}-x_{2i}\right|\right)}\mathrm{P}(0) \tag{17}$$

where $A(x_{11},x_{12},\cdots x_{1n})$ and $B(x_{21},x_{22},\cdots x_{2n})$ are any two points in the $n$-dimensional space before quantum encoding, and $\mathrm{P}(0)$ denotes the probability that the auxiliary quantum bit is in the ground state $|0\rangle$ as measured by the SWAP test circuit.

## 2.4 Implementation process of quantum K-means clustering algorithm

Quantum K-means clustering algorithm is obtained by introducing quantum computation on the basis of classical K-means clustering algorithm and is based on the quantum mixing distance proposed in Section **2.3** as a quantum state similarity metric. The specific steps of quantum K-means clustering algorithm are as follows in Table 1:

Table 1 Quantum K-means algorithm

```
Algorithm 1 Quantum K-means algorithm
Input: Dataset $D=\{x_1,x_2,\cdots x_n\}$, where $x_i \in R^d$
          Number of clusters $k$
          Maximum iterations $T_{\max}$
          Convergence threshold $\varepsilon$
Output: Cluster assignments $C=\{C_1,C_2,\cdots C_k\}$
          Cluster centroids $c=\{c_1,c_2,\cdots c_k\}$
Procedure:
  1.Initialize Centroids
    Randomly select $k$ distinct data points from $D$ as initial centroids:
    $c_j^{(0)} \leftarrow \{x_{i_1},x_{i2},\cdots x_{ik}\}$ for $j=1,2,\cdots k$
  2. Iterate until convergence
      a. Cluster Assignment (Quantum Step):
          For each data point $x_i \in D$:
            i. Quantum Encoding:
                Prepare quantum state $|\psi_i\rangle$ using angle or amplitude encoding
            ii. Distance Calculation:
                For each centroid $c_j^{(t)}$:
                  1. Encode centroid as $|\varphi_j\rangle$ using same method
                  2. Construct quantum circuit with $|\psi_i\rangle, |\varphi_j\rangle$, ancilla qubit
                  3. Apply controlled-SWAP operations
                  4. Measure ancilla M times, record $\mathrm{P}(0)$
                  5. Compute hybrid distance:
                        $\mathrm{d}_{ij}=\sqrt{\max_m\left(\left|x_{im}-c_{jm}^t\right|\right)}\mathrm{P}(0)$
            iii. Assign $x_i$ to nearest cluster:
                $l=\arg\min_j \mathrm{d}_{ij}$
            $C_l \leftarrow C_l \cup \{x_i\}$
    **b. Centroid Update (Classical Step):**
        For each cluster $j = 1, \cdots, K$:
            $c_j^{t+1} = \frac{1}{|C_j|} \times \sum_{x \in C_j} x$
    **c. Convergence Check:**
        Compute $\Delta = \max_j \| c_j^{t+1} - c_j^t \|$
        **Until** $\Delta < \varepsilon$ **or** $t \geq T_{\max}$
**3. Return final clusters** $C = \{C_1, C_2, \cdots C_k\}$ **and centroids** $c = \{c_1, c_2, \cdots c_k\}$
```

## 3. Experimental simulation and analysis

### 3.1 Dataset Selection and Preprocessing

In order to evaluate the performance of the quantum K-means clustering algorithm based on the improved quantum state similarity measure, three real UCI datasets are selected in this paper, namely, the Iris, Seeds, and Wine datasets, which are available for download at http://archive.ics.uci.edu/ml/datasets.php, and Table 2 lists the basic information of these three datasets.

Table 2 UCI dataset information table

| number | data set | sample size | dimensionality | Number of categories |
|---|---|---|---|---|
| 1 | Iris. | 150 | 4 | 3 |
| 2 | Seeds | 210 | 7 | 3 |
| 3 | Wine | 178 | 13 | 3 |

To address the computational efficiency and interpretability challenges of quantum K-means clustering for high-dimensional data, this paper proposes a dimensionality reduction optimization strategy. By compressing UCI standard datasets into two-dimensional space through Principal Component Analysis while retaining the core features of the data, this approach effectively reduces quantum bit consumption and mitigates quantum noise interference.

### 3.2 Experimental Simulation and Analysis

This section compares the performance of the proposed quantum K-means-Amplitude, quantum K-means-Angle with the classical K-means clustering algorithm on the Iris, Seeds and Wine datasets through three sets of comparative experiments, Table 3 – 5 Quantitatively presents the comparative data for the five evaluation metrics.

Table 3 Iris dataset comparison of clustering effects using different methods

| Evaluation indicators | K-means | quantum K-means-Amplitude | quantum K-means-Angle |
|---|---|---|---|
| ARI | 0.6146 | 0.6205 | **0.6312** |
| NMI | 0.6956 | 0.6893 | **0.7042** |
| SC | **0.8674** | 0.8551 | 0.8596 |
| CH | **681.0889** | 670.0490 | 674.7586 |
| DB | **0.4361** | 0.4661 | 0.4566 |

Table 3 shows quantum K-means-Angle achieves superior ARI/NMI scores in Iris classification, effectively capturing categorical features. Although amplitude encoding reaches 0.6205 ARI, its weaker SC/DB metrics reveal sensitivity to specific traits. The triangular PCA distribution naturally complements angle encoding's directional separation, while amplitude encoding's dimension reduction compromises feature representation through information loss.

Table 4 Seeds dataset comparison of clustering effects using different methods

| Evaluation indicators | K-means | quantum K-means-Amplitude | quantum K-means-Angle |
|---|---|---|---|
| ARI | 0.6142 | 0.6192 | **0.6506** |
| NMI | **0.6395** | 0.6047 | 0.6355 |
| SC | 0.7958 | 0.7694 | **0.8103** |
| CH | 592.3293 | 564.1185 | **643.6940** |
| DB | 0.5454 | 0.5410 | **0.5341** |

Table 4 highlights angle encoding's superiority in seed classification, with quantum K-means-Angle showing 8.7% higher CH index through directional pattern capture in PCA space via quantum phase modulation. Amplitude encoding's inferior NMI performance stems from lost high-dimensional correlations during dimensionality reduction and measurement-induced state perturbations causing label inconsistencies.

Table 5 Wine dataset comparison of clustering effects using different methods

| Evaluation indicators | K-means | quantum K-means-Amplitude | quantum K-means-Angle |
|---|---|---|---|
| ARI | 0.7587 | 0.7726 | **0.7742** |
| NMI | 0.7624 | 0.7728 | **0.7740** |
| SC | **0.8186** | 0.8167 | 0.8171 |
| CH | **497.7627** | 495.5086 | 497.3176 |
| DB | **0.4723** | 0.4739 | 0.4749 |

Table 5 shows quantum K-means-Angle's modest 2.04% ARI and 1.52% NMI gains over classical methods in PCA-reduced Wine data. Although amplitude encoding achieves comparable ARI, its compromised DB scores reveal separation deficiencies. Minimal SC and CH variations across methods indicate that 13D→2D compression constrains quantum advantages by eliminating high-dimensional patterns essential for quantum feature representation.

## 4. Conclusion

In this paper, a new quantum-classical hybrid distance is proposed and clustering performance comparison experiments are carried out based on three UCI standard datasets. By comparing and analyzing the classical K-means algorithm with two quantum K-means clustering algorithms with different encoding methods, it is found that quantum K-means with angle encoding exhibits relatively superior clustering performance. The experimental results show that with the continuous development of quantum computing hardware, quantum-enhanced clustering algorithms are expected to break through the computational bottleneck of traditional machine learning in dealing with high-dimensional and large-scale datasets, and provide a new technological path for intelligent data analysis.

## References

*[1] Shor P. Algorithms for quantum computation: discrete logarithms and factoring[J]. In Proceedings of 35th Annual Symposium on the Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1994:124-134.*

*[2] Grover L K. A fast quantum mechanical algorithm for database search[C]. Proceedings of the twenty-eighth annual ACM symposium on Theory of computing, 1996: 212-219.*

*[3] Anguita D, Ridella S , Rivieccio F ,et al. Quantum optimization for training support vector machines[J]. Neural*

*Networks, 2003, 16(5-6):763-770.*

*[4] Ruan Y, Chen H W, Liu Z H, et al. Quantum Principal Component Analysis Algorithm[J]. Chinese Journal of Computers, 2014.*

*[5] Kerenidis I, Landman J. Quantum spectral clustering[J]. Physical Review A, 2021, 103(4): 042415.*

*[6] Wiebe N, Kapoor A, Svore K. Quantum Nearest-Neighbor Algorithms for Machine Learning[J]. Quantum Information & Computation, 2014, 15:0318-0358.*

*[7] Khan S U, Awan A J, Vall-Llosera G. K-Means Clustering on Noisy Intermediate Scale Quantum Computers[J]. arXiv preprint arXiv:1909.12183, 2019.*

*[8] Arthur D, Date P. Balanced k-means clustering on an adiabatic quantum computer[J]. Quantum Information Processing, 2021, 20(9): 294.*

*[9] Ohno H. A quantum algorithm of K-means toward practical use[J]. Quantum Information Processing, 2022, 21(4).*

*[10] DiAdamo S, O'Meara C, Cortiana G, et al. Practical quantum k-means clustering: Performance analysis and applications in energy grid classification[J]. IEEE Transactions on Quantum Engineering, 2022, 3: 1-16.*