Research and evaluation of underwater acoustic signal classification capability using quantum K-mean algorithm on real dataset

Bach Nhat Hoang*, Dang Tien Sy, Doan Trung Thanh, Phan Huy Anh

Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam.
*Corresponding author: hoangbn.vdt@gmail.com
Received 10 Aug. 2025; Revised 26 Sep. 2025; Accepted 10 Oct. 2025; Published 30 Oct. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.IITE.2025.19-26

ABSTRACT

In the big data era, traditional machine learning algorithms like K-means face computational challenges in processing large, high-dimensional datasets, particularly for underwater acoustic signal classification. This study investigates the application of the quantum clustering algorithm Q-means, a quantum-enhanced variant of K-means, to real passive sonar datasets. The purpose is to evaluate Q-means' effectiveness in classifying propeller-driven ship signals under noisy conditions, leveraging quantum principles such as superposition and entanglement for exponential speedup. The research extends classical K-means to δ-k-means for robustness, implementing quantum subroutines including distance estimation, cluster assignment, and state tomography. Preprocessed features from power spectral density analysis of real datasets (from prior studies on varying ship speeds) are used as input to avoid issues with raw time-series data, such as high dimensionality and sensitivity to shifts. Simulations on 15,000 passive sonar samples demonstrate that Q-means achieves clustering quality comparable to K-means, with clear separation of three clusters and accurate centroids, while reducing complexity from O(n) to O(polylog(n)). This validates Q-means as a promising tool for large-scale, noisy acoustic data in national security applications, bridging quantum theory with practical machine learning.

Keywords: Quantum computing; Passive sonar; Q-mean; K-mean.

1. INTRODUCTION

In the era of big data, traditional machine learning algorithms are facing increasingly challenging computational demands. The ability to process and analyze massive datasets is limited by the fundamental principles of classical computing, where computational complexity often scales linearly or polynomially with the number of data points n and dimensions d. This limitation drives the search for a new computational paradigm capable of overcoming current barriers. Quantum computing emerges as a promising candidate, offering a fundamentally different approach to information processing [1]. Analysis of the limitations of classical computers reveals that machine learning algorithms such as K-means and Monte Carlo encounter significant constraints, such as stability and processing time, when applied to large-scale datasets with complex features like underwater acoustic signals [2].

Quantum computing addresses this challenge by exploiting the principles of quantum mechanics, enabling information processing in ways unattainable on classical computers. The foundation of quantum computing stems from quantum concepts [3]: (1) Qubit and Superposition: The basic unit of information in a quantum computer is the qubit. Unlike a classical bit, which can only be in state 0 or 1, a qubit can exist in a superposition of both, with complex probability amplitudes. This capability allows a register of L qubits to represent 2^L distinct values simultaneously. Consequently, a quantum computer can encode and process an entire massive dataset in a single quantum state, requiring only a number of qubits proportional to the logarithm of the number of data points, O(log n); (2) Quantum Entanglement: This is a phenomenon where qubits are intricately linked such that the state of one qubit instantaneously depends on the states

of others, regardless of physical distance. Quantum entanglement generates complex correlations, enabling parallel computations over a vastly expanded computational space.

In the current global field of underwater acoustics, K-means is commonly applied for dimensionality reduction, noise removal, or supporting signal classification by combining with other techniques such as PCA or deep learning to enhance classification outcomes [4]. In passive signal classification, K-means is used to cluster data from underwater sound sources, assuming a predefined number of clusters K [5]. However, K-means is sensitive to outliers and asymmetric data; in complex underwater acoustic environments, K-means may be inefficient if the data has high noise or uneven distribution. With the increasingly large volumes of underwater acoustic data collected in applications related to national security and socioeconomic development, the problem of processing large, high-dimensional, noisy datasets demands the development of algorithms capable of accelerating computations to quickly filter anomalous signals. In previous studies [6,7], the authors detected and classified underwater acoustic signals from propeller-driven ships collected under real conditions in three scenarios: (1) Ships moving with complex, varying speeds in noisy backgrounds, (2) Ships moving at constant speeds in noisy backgrounds, and (3) Ships starting to move at slow speeds. The Q-means algorithm is the quantum version of K-means, utilizing quantum subroutines to compute distances and update centroids, providing quantum speedup for large datasets; it can handle large data with lower complexity, suitable for multi-target or high-noise underwater acoustic signal classification. Therefore, investigating the applicability of quantum algorithms to underwater acoustic data problems is essential. In this paper, the research group proposes a solution to implement quantum clustering algorithms based on extending traditional clustering algorithms in machine learning, executing the algorithm on numerical simulations to evaluate its effectiveness on real passive sonar data. From the above analyses, the paper is divided into 4 sections: Section 1: Introduction; Section 2: Literature quantum mean cluster; Section 3: Results and discussion; Section 4: Conclusions.

2. LITERATURE QUANTUM MEAN CLUSTER

2.1. Mathematical essence of the classical K-means algorithm

The K-means algorithm [8] is an iterative unsupervised clustering method that partitions a dataset into k distinct clusters. Its essence is an optimization process aimed at minimizing the total intra-cluster squared distances.

Input and initialization

The algorithm takes as input a dataset $X = \{x_i\}_{i=1}^{n}$ with $x_i \in R^d$ and the number of clusters k. The process begins by initializing k cluster centroids $c_1, c_2, ...c_k$, typically selected randomly from the data points. The algorithm iterates two steps until convergence:

- Assignment Step: Each data point x_i is assigned to the cluster with the nearest centroid. The label of x_i point at iteration t, denoted as $l(x_i)^t$, is determined by:

$$\ell(x_i)^t = \underset{j\in[k]}{\operatorname{argmin}}\, d(x_i, c_j^t)^2 = \underset{j\in[k]}{\operatorname{argmin}} \mid x_i - c_j^t \mid^2 \tag{1}$$

This partitions the dataset into k sets C_j^t.

- Update Step: Each cluster centroid is recalculated as the mean of all data points assigned to that cluster.

$$c_j^{t+1} = \frac{1}{\mid C_j^t \mid} \sum_{x_i \in C_j^t} x_i \tag{2}$$

Loss function and convergence:

The K-means algorithm iteratively minimizes the Residual Sum of Squares (RSS) loss function:

$$RSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} || x_i - c_j ||^2 \quad (3)$$

Since RSS decreases with each iteration, the algorithm is guaranteed to converge to a local minimum. Complexity analysis: The computational complexity per iteration is $O(knd)$, reflecting a linear dependence on the number of data points n. This is the primary computational bottleneck for large datasets.

2.2. Structure and subroutines of the Q-means algorithm

The q-means algorithm [8] is understood as the quantum version of a generalized variant of k-means, called (δ)-k-means. The (δ)-k-means model introduces an error margin (δ). In the assignment step, a data point x_i can be assigned to any centroid c_p as long as the squared distance difference from the nearest centroid c_j^* is no more than (δ):

$$\left| d^2(c_j^*, x_i) - d^2(c_p, x_i) \right| \leq \delta \quad (4)$$

In the update step, the new centroid only needs to lie within a ball of radius (δ/2) around the true mean. Introducing (δ) makes the algorithm more robust to small errors, a property well-suited to the probabilistic nature of quantum algorithms.

The q-means algorithm implements the steps of (δ)-k-means using a sequence of quantum subroutines:

Step 1: Quantum distance estimation: Instead of sequentially computing nk distances, q-means leverages quantum parallelism. The algorithm prepares a superposition state of all data point-centroid pairs and applies a distance estimation subroutine. The result is a single quantum state containing all squared distance estimates:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} |i\rangle \bigotimes_{j \in [k]} |j\rangle |\overline{d^2(x_i, c_j)}\rangle \quad (5)$$

Step 2: Quantum cluster assignment: To find the nearest centroid for each data point in the superposition state, a quantum minimum-finding algorithm is applied in parallel across the k distance registers for each component $|i\rangle$.

After uncomputing the distance registers, the final state is a superposition of data point indices and their assigned cluster labels:

$$|\psi_t\rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |i\rangle |l_t(x_i)\rangle \quad (6)$$

Step 3: Centroid state recreation: Updating centroids equates to matrix multiplication. By measuring the label register of $|\psi_t\rangle$, the algorithm can prepare probabilistic states $|\chi_j^t\rangle$ (feature vectors of clusters). Quantum matrix multiplication is then used to generate quantum states $|c_j^{t+1}\rangle$ corresponding to the new centroids.

Step 4: Centroid update: To obtain classical descriptions of the new centroids for the next iteration, Quantum State Tomography is applied to each state $|c_j^{t+1}\rangle$. This process extracts classical centroid vectors with controllable accuracy.

3. RESULTS AND DISCUSSION

3.1. Comparison of complexity and advantages of Q-means

The fundamental architectural differences lead to significant performance disparities, particularly for large datasets.

Complexity analysis: The complexity of one q-means iteration [8] is given by:

$$\tilde{O}(kd\frac{\eta}{\delta^2}\kappa(X)(\mu(X)+kd\frac{\eta}{\delta})+...) \quad (7)$$

The most critical aspect of this expression is the absence of n as a multiplicative factor. The dependence on the number of data points has been reduced to $O\left(polylog\left(n\right)\right)$. This is the source of the exponential speedup.

Table 1. *The comparison between the K-means and Q-means algorithms.*

Criterion	Classical K-means	Quantum Q-means
Basic Model	Exact RSS Optimization	Robust (δ)-k-means Execution
Data Access	Sequential	Parallel
Distance Computation	Classical Arithmetic Operations	Quantum Distance Estimation
Complexity	$O(n)$	$O(polylog(n)$
New Parameters	None	$\kappa(X)$, ($\mu(X)$, (δ), (η)
Output	Deterministic Centroids	Probabilistic Centroids with Error

The most profound difference between K-means and Q-means lies not only in speed but also in the nature of the solutions they provide. K-means is a deterministic algorithm aimed at precisely optimizing the RSS loss function. In contrast, q-means does not solve the exact K-means problem. Instead, it addresses an approximate and more robust problem, (δ)-k-means. The introduction of the (δ)-k-means model in the literature is not incidental; it is a conceptual preparation.

Quantum subroutines like distance estimation and state tomography are inherently probabilistic and error-prone. Thus, a quantum algorithm cannot perfectly replicate a deterministic process like K-means. By redefining the classical problem to include an error margin (δ), the authors create a target that quantum algorithms can provably achieve. The parameter (δ) is not a flaw but an integral part of the model, reflecting the probabilistic nature of quantum computing. This illustrates a broader trend in Quantum Machine Learning (QML): to achieve quantum speedup, we often must redefine the problem in a more robust way, accepting approximations. For many real-world machine learning applications, where data is inherently noisy, a robust and approximate solution may be more valuable than a precise but sluggish one.

3.2. Algorithm implementation and evaluation results

Directly using raw underwater acoustic signals in the form of time series of amplitude values as input for Q-means is infeasible and inefficient. A raw underwater acoustic signal segment contains tens or hundreds of thousands of data points (samples). If each data point is treated as a dimension, the input vector x_i would have an extremely large dimensionality d. Although Q-means offers exponential speedup with respect to the number of data points (n), its complexity still scales polynomially with the number of dimensions d. An excessively large d would render the algorithm slow and resource-intensive, negating the quantum advantage.

Moreover, Q-means, like K-means, operates by minimizing Euclidean distances between data points and centroids. For raw time-series signals, Euclidean distance is highly sensitive to phase or time shifts. Two underwater acoustic signals may originate from the same source (e.g., the same type of ship) but be slightly time-shifted. Although essentially similar, the Euclidean distance

between their raw signal vectors could be large, leading to misclassification. Raw signals contain substantial noise and irrelevant information. The Q-means algorithm would attempt to cluster based on this noisy information, resulting in inaccurate and unstable outcomes.

Therefore, in this paper, we utilize feature maps obtained after power spectral density analysis of the data collected in previous publications by the research group [6, 7]. The data used in the paper are real hydroacoustic data, used in published papers [6, 7], collected from the real ShipEars dataset (collected off the coast of Spain).

The dataset X is treated as a classical matrix containing real numbers, denoted as $X \in R^{nxd}$ with n being the number of data points; d being the number of features/ dimensions per data point. Each row of this matrix x_i is a classical vector representing a unique data point. The Q-means algorithm operates directly on this data matrix.

The model focuses on the outer loop for convergence checking and the inner loop for assigning clusters to each data point, with the quantum component centered on the Quantum Interference Circuit (QIC) for distance computation to centroids. Data is encoded directly into quantum circuits for each point-centroid pair individually using U gates (no quantum memory needed to store or access the entire dataset). The implementation steps are described below:

Step 1:

Outer Loop: Check convergence by computing the average distance between old and new centroids $\frac{1}{K}\sum_{j=1}^{K} d(c_j^t, c_j^{t-1}) \le \tau$, with τ being the threshold.

Step 2:

Inner Loop: For each data point x_i, use a tournament-style approach to compare pairwise centroids via QIC, eliminating until one winner remains (nearest centroid). To encode the classical vector $v = (v_0, v_1, \ldots, v_{N-1})$ into a quantum state, use amplitude encoding $|\psi\rangle = \frac{1}{M}\sum_{i=0}^{N-1} v_i |i\rangle$ with $M = \sqrt{\sum_{i=0}^{N-1} v_i^2}$ being the norm of the vector.

Step 3:

Use QIC to compute interference probability, from which the distance is inferred. This implementation relies on estimating Euclidean distance via controlled swap (CSWAP) and interference probability. In QIC, for two vectors t and c (data point and centroid), after applying the Hadamard gate to the state $|\psi\rangle = \begin{pmatrix} t'_x \\ t'_y \\ c'_x \\ c'_y \end{pmatrix}$, the probability of qubit in $|1\rangle$ being is $P(|1\rangle) = \frac{1}{2}[(t'_x - c'_x)^2 + (t'_y - c'_y)^2]$, leading to distance $d(t,c) = \text{Norm} \times \sqrt{2} \times \sqrt{P(|1\rangle)}$ with $\text{Norm} = \sqrt{t_x^2 + t_y^2 + c_x^2 + c_y^2}$.

Step 4:

Convergence stops when cluster assignments no longer change. Use circuit U_{min} to find argmin among (k) distances: $(\otimes_{j\in[k]} |a_j\rangle)|0\rangle \to (\otimes_{j\in[k]} |a_j\rangle)|\text{argmin}(a_j)\rangle$, timing $O(k\ log\ p)$.

Step 5:

Evaluation and return results to the classical vector space domain. Update new centroids $c_j^{t+1} = \frac{1}{|C_j^{t+1}|} \sum_{x_i \in C_j^{t+1}} x_i$ using quantum matrix multiplication and tomography to convert from quantum state to classical vector, ensuring convergence similar to the classical K-means algorithm.

Results when implemented on the preprocessed dataset from [6, 7] in figure 1 show that the K-means algorithm does not perform overly effectively. The three data clusters are clearly separated, but the centroids are not accurately determined at the center of each cluster, reflecting the limited capability of K-means in finding optimal structures on low-noise but high-dimensional datasets.

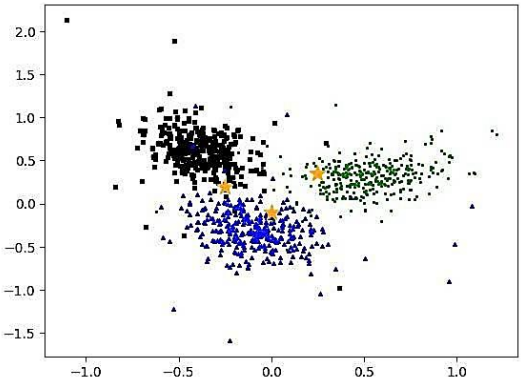

Figure 1. *K-means algorithm results with 15,000 data samples.*

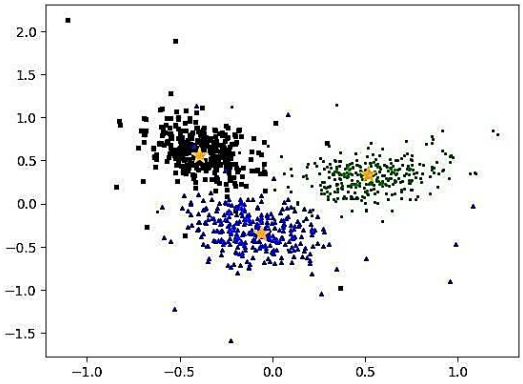

Figure 2. *Q-means algorithm results with 15,000 data samples.*

Results when implemented on the preprocessed dataset from [6, 7] in figure 2 show that the Q-means algorithm performs relatively effectively. The distribution of clusters, positions of final centroids, and data point assignments are visually similar. This indicates that Q-means, despite operating on quantum principles and the error-tolerant (δ)-k-means model, successfully reproduces high-quality clustering results comparable to the classical algorithm.

It demonstrates that the approximation steps and inherent probabilistic nature in quantum subroutines (such as distance estimation and state tomography) do not degrade the final result quality. Q-means still converges to a clustering solution of equivalent quality to the well-established classical method. This also validates that the (δ)-k-means model, a robust and fault-tolerant version of K-means, is a suitable alternative target for achieving quantum speedup without sacrificing practical accuracy. When representing hydroacoustic data in 2D, if the randomly initialized centroids fall into poor positions (near outliers), the algorithm may converge to local optima (non-optimal local solutions), leading to skewed centroids. Therefore, in subsequent studies, the research team will use 3D models to verify the effectiveness of K-means and Q-means. Overlap clusters between black-blue-green, leading to point labeling changes over iterations, causing centroids to fluctuate and eventually deviate from the "ideal" position based on qualitative observations.

In summary, from theoretical foundations to real simulation results, the Q-means algorithm has proven itself as a breakthrough advancement, successfully bridging the theoretical potential of quantum computing with practical machine learning applications. Theoretically, the core and most appealing advantage of Q-means is its exponential speedup relative to the number of data points (n). By replacing the linear dependence $O(n)$ of classical K-means with a polylogarithmic dependence $O\left(polylog\left(n\right)\right)$, Q-means opens a new approach for analyzing large, complex, high-dimensional datasets previously considered computationally challenging.

Table 2. *The performance comparison results between K-means and Q-means.*

Metric	K-mean	Q-mean	Comparison
Silhouette Score	0.4632	0.4578	K-means is slightly better (1.2% higher separation), suitable for spherical data. Q-means is poor due to approx distance (like L1 or noisy estimation). Both are average (0.4-0.5), indicating good clusters, but have small overlap.
Davies-Bouldin Index	0.7683	0.7571	Q-means is superior (1.5% lower, lower cluster similarity), indicating more compact clusters. Value <0.8 confirms both are usable.
Calinski-Harabasz	632.5	638.2	Q-means is 0.9% higher, with better dispersion between clusters
Purity	0.8850	0.9138	Q-means is 1.8% more accurate than true labels, with fewer misassignments (outliers)
Iteration	8	11	K-means converges faster

However, this speedup comes with a trade-off: the algorithm's complexity no longer depends on (n) but instead on data-specific parameters like the condition number (κ) and other matrix properties (μ). This indicates that Q-means is not a universal solution but a specialized tool, most effective on "quantum-friendly" or "quantum-transformable" datasets.

4. CONCLUSIONS

This study has presented a comprehensive comparative analysis between the classical K-means clustering algorithm and the quantum Q-means. Experimental simulation results have convincingly demonstrated that Q-means can produce clustering outcomes of equivalent quality to classical K-means, even on large and complex datasets. These results highlight a combination of theoretical efficiency and practical quality: Q-means not only promises exponential speedup relative to the number of data points, a key factor in the big data era, but also delivers a reliable and accurate clustering solution. This affirms Q-means as a viable and promising alternative for large-scale clustering problems, where classical algorithms become inefficient due to computational barriers.

REFERENCES

[1]. Montanaro, Ashley. "*Quantum algorithms: An overview*", Quantum Information Nature, 2.1, 1–8, (2016).

[2]. Theocharidis, Theocharis, and Ergina Kavallieratou. "*Underwater communication technologies: A review*", Telecommunication Systems, 88, 2, 54, (2025).

[3]. Bohm, David. "*Quantum theory*", Courier Corporation, (1989).

[4]. Ahmad, Izhar. "*K-mean and K-prototype algorithms performance analysis*", International Journal of Computer and Information Technology, 3, 4, 823–828, (2014).

[5]. Liu, Mingqian, et al. "*Intelligent passive detection of aerial target in space-air-ground integrated networks*", China Communications, 19, 1, 52–63, (2022).

[6]. Bach, Hoang Nhat, Duc Van Nguyen, and Ha Le Vu. "*Enhancing the capacity of detecting and classifying cavitation noise generated from propeller using the convolution neural network*", International Conference on Industrial Networks and Intelligent Systems, Cham: Springer International Publishing, (2021).

[7]. Bach, Nhat Hoang, et al. "*Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational auto-encoder*", Scientific Reports, 13, 1, 19984, (2023).

[8]. Dalzell, Alexander M., et al. "*Quantum algorithms: A survey of applications and end-to-end complexities*", arXiv preprint arXiv:2310.03011, (2023).

TÓM TẮT

Nghiên cứu và đánh giá khả năng phân loại tín hiệu âm thanh dưới nước bằng thuật toán Quantum K-mean trên tập dữ liệu thực

Trong kỷ nguyên dữ liệu lớn, các thuật toán học máy truyền thống như K-means gặp thách thức tính toán khi xử lý các bộ dữ liệu lớn, nhiều chiều, đặc biệt là phân loại tín hiệu thủy âm. Nghiên cứu này khảo sát việc áp dụng thuật toán phân cụm lượng tử Q-means, phiên bản nâng cao lượng tử của K-means, trên bộ dữ liệu sonar thụ động thực tế. Mục đích là đánh giá hiệu quả của Q-means trong việc phân loại tín hiệu tàu chân vịt trong điều kiện nhiễu, khai thác các nguyên lý lượng tử như chồng chập và rối để đạt tăng tốc theo cấp số mũ. Nghiên cứu mở rộng K-means cổ điển sang δ-k-means để tăng tính mạnh mẽ, thực hiện các chương trình con lượng tử bao gồm ước tính khoảng cách, gán cụm và chụp cắt lớp trạng thái. Các đặc trưng được tiền xử lý từ phân tích mật độ phổ công suất của bộ dữ liệu thực tế (từ các nghiên cứu trước về tốc độ tàu thay đổi) được sử dụng làm đầu vào để tránh vấn đề với dữ liệu chuỗi thời gian thô, như chiều cao và nhạy cảm với dịch chuyển. Mô phỏng trên 15.000 mẫu cho thấy Q-means đạt chất lượng phân cụm tương đương K-means, với sự phân tách rõ ràng ba cụm và tâm cụm chính xác, đồng thời giảm độ phức tạp từ O(n) sang O(polylog(n)). Điều này xác nhận Q-means là công cụ đầy hứa hẹn cho dữ liệu thủy âm lớn, nhiễu trong các ứng dụng an ninh quốc gia, kết nối lý thuyết lượng tử với học máy thực tiễn.

Từ khoá: Điện toán lượng tử; Sonar thụ động; Q-mean; K-mean.