International Journal of Theoretical Physics (2020) 59:3134–3144
https://doi.org/10.1007/s10773-020-04567-1

Quantum Algorithms for Similarity Measurement Based on Euclidean Distance

Kai Yu[1] · Gong-De Guo[1] · Jing Li[1] · Song Lin[1]

Received: 26 February 2020 / Accepted: 4 August 2020 / Published online: 20 August 2020

Abstract
Similarity measurement is a fundamental problem that arise both on its own and as a key subroutine in more complex tasks, such as machine learning. However, in classical algorithms, the time used to similarity measurement usually increases exponentially as the amount of data and the number of data dimensions increase. In this paper, we presented three quantum algorithms based on Euclidean distance to measure the similarity between data sets. In the proposed algorithms, some special unitary operations are utilized to construct imperative quantum states from quantum random access memory. Then, a badly needed result for estimating the similarity between data sets, can be got by performing projective measurements. Furthermore, it is shown that these algorithms can achieve the exponential acceleration of the classical algorithm in the quantity or the dimension of data.

Keywords Quantum algorithm · Quantum machine learning · Similarity measurement · Euclidean distance

1 Introduction

Dating from the 80's of last century, it is shown that quantum computing allows for more efficient solutions of some problems than classical computing. The prime-factorization problem is one of this. In 1994, by exploiting quantum mechanical properties, Shor proposed the famous Shor algorithm [1], which is exponentially faster than the best-known classical factoring algorithm. The other one is Grover algorithm for solving unsorted database search problem [2], which makes quadratic speedup over the classical search algorithm. Inspired by these algorithms, people try to exploit quantum computing to solve other problems in the field of information technology [3–20], e.g., machine learning. The most representative one is HHL algorithm [3]. In this algorithm, Harrow et al. utilized Hamiltonian simulation [4] and phase estimation technology to solve linear equations. Furthermore, it has exponential speedup compared with classical counterpart, and can be widely applied in machine learning numerical calculation or other scenarios. After that, more and

✉ Song Lin
lins95@gmail.com

[1] College of Mathematics and Informatics, Fujian Normal University, Fuzhou, 350007, China

more attention has been paid to this new multi-knowledge crossed research field, quantum machine learning. As for now, some subtle quantum machine learning algorithms have been proposed for various machine learning tasks, such as quantum principal component analysis [5], quantum supervised learning [6–10], quantum unsupervised learning [7, 11], quantum regression [12–14], quantum search engine ranking [15–18], quantum neural network [19], and so on. More excitingly, all these algorithms achieve desirable speedup effect.

Similarity measurement is a primitive machine learning task, which plays an indispensable role in data clustering [21, 22] and classification [23, 24]. It is a method to evaluate the degree of similarity between individuals. There exist many ways to depict similarity, the most common of which is to use Euclidean distance. That is, the closer the Euclidean distance between two data points (data vectors) is, the more similar they are; and vice versa. With the advent of the era of big data, the number of the data set and the dimension of the data points grow larger. As noted by Aaronson [25], sampling and estimating distances and inner products between post-processed vectors on a classical computer is apparently exponentially hard. Therefore, how to effectively similarity measurement has become a new challenge. Quantum computing may be a good solution to this problem. In 2014, Lloyd exploited quantum parallelism to accelerate the estimation of Euclidean distance from a training data vector to a sample cluster [7]. However, there are many other similarity measurements based on Euclidean distance in classical machine learning algorithms, such as group average distance [26]. Hence, in this work, we examine the issue further and propose three quantum algorithms based on Euclidean distance similarity measurement. In these algorithms, we use the characteristic of quantum random access memory (QRAM) to access data in parallel [27] and the amplitude estimation technology [28] to estimate the Euclidean distance between data sets. And the analysis shows that these algorithms are more efficient than the corresponding classical algorithms.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the essential preliminaries. Then, the quantum algorithms of three distance measurement methods respectively are described in Section 3. A brief analysis of three proposed quantum algorithms is presented in Section 4. Finally, a short conclusion is provided in Section 5.

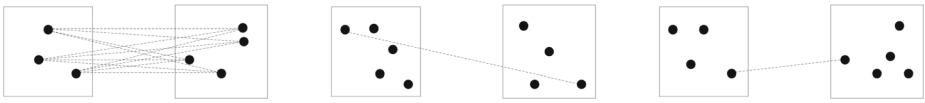
2 Preliminaries

2.1 Similarity Measurement Based on Euclidean Distance

As the key of clustering algorithm, similarity measurement can be expressed in many ways. The most common way is Euclidean distance. Given two N-dimensional data vectors $\mathbf{u}$ and $\mathbf{v}$, the similarity of them expressed in squared Euclidean distance is $d^2(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{N}(u_k - v_k)^2$. However, we need not only to estimate the similarity between the points, but also to calculate the distance of two sets in practice. For example, there are three different Euclidean distance measurements in hierarchical clustering, which will lead to different results. Therefore, different methods are needed for different types of data sets. It makes sense to design quantum algorithms for these distances respectively. Before describing quantum algorithms, we review three classical methods, briefly.

Considering the task of similarity measurement for two data sets U and V. And p data points $\overrightarrow{u_i} \in U$ and q data points $\overrightarrow{v_j} \in V$ are given.

i *Average Linkage* (average distance). As shown in Fig. 1a, the squared Euclidean distance between each data point in one set and that in the other set is calculated firstly.



(a) Average Linkage (b) Complete Linkage (c) Single Linkage

Fig. 1 The illustration of distance

Then, by taking the mean as the distance between the two data sets. The average distance between U and V is obtained, which can be written as:

$$D(U,V) = \frac{\sum_{\overrightarrow{u_i}\in U,\overrightarrow{v_j}\in V} d^2(\overrightarrow{u_i},\overrightarrow{v_j})}{pq}. \tag{1}$$

ii *Complete Linkage* (furthest distance). As shown in Fig. 1b, this method is to take the distance between the two furthest data points in the two data sets as the distance between two sets. The furthest distance between U and V can be written as:

$$D'(U,V) = \max_{\overrightarrow{u_i}\in U,\overrightarrow{v_j}\in V} d^2(\overrightarrow{u_i},\overrightarrow{v_j}). \tag{2}$$

iii *Single Linkage* (nearest distance). As shown in Fig. 1c, the single linkage, contrary to the complete linkage, takes the distance between the nearest data points in the two data sets as the distance between the sets. The nearest distance between U and V can be written as:

$$D''(U,V) = \min_{\overrightarrow{u_i}\in U,\overrightarrow{v_j}\in V} d^2(\overrightarrow{u_i},\overrightarrow{v_j}). \tag{3}$$

These methods complement each other, satisfying common data types. For example, the average linkage is suitable for ellipsoid data sets, and the single linkage is better for strip data sets [29]. Hence, the three methods could be widely applied to classification and clustering of data sets.

2.2 Data Preparation and Pre-processing

In quantum machine learning, we generally assume that data sets consist of arrays of vectors stored in QRAM. Given a $N = 2^n$ dimensional complex vector $\overrightarrow{u}$, we can express it as the composition of N components of $\overrightarrow{u} = (u_1, u_2, \cdots, u_N)$, where $\{u_k = |u_k|e^{\mathrm{i}\phi_k},\ \mathrm{i} = \sqrt{-1}\}$. Moreover, we assume that $\{|u_k|, \phi_k\}$ are stored as floating point numbers in quantum random access memory, and stored as "binary trees" data structure that is shown in Fig. 2. The details can be found in its initial proposal for designing quantum recommendation system [30], as well as its successful application in designing the quantum linear system algorithm for dense matrices [31]. In short, N dimensional vector $\overrightarrow{u}$ can be described by a quantum state $|u\rangle = |\overrightarrow{u}|^{-1}\overrightarrow{u}$ of n qubits in time $O\left(\log_2 N\right)$. Once the exponentially compressed quantum versions of the vectors have been created, we can apply these vectors to the proposed quantum algorithms.

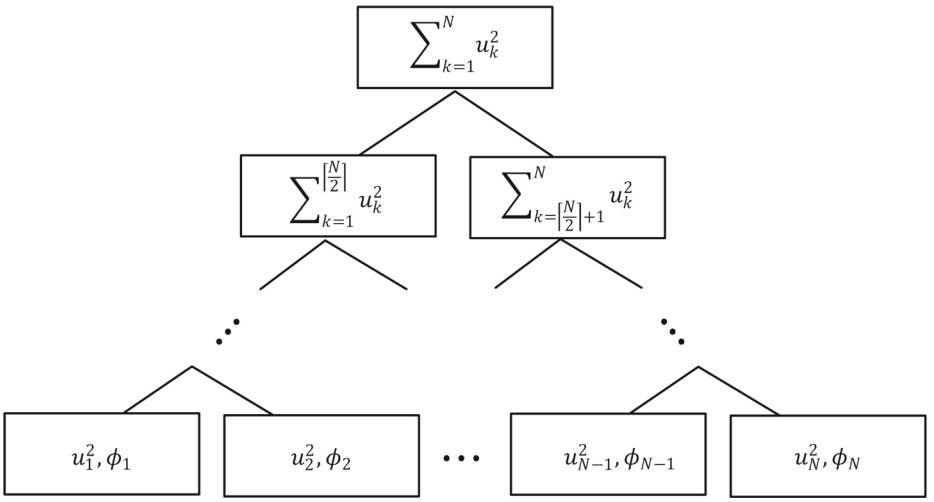


Fig. 2 The binary tree storage a data point $\mathbf{u}$, store its u_k^2 in binary form. Furthermore, this kind of storage structure build N dimensional vector corresponding quantum state simply by $O\left(\log_2 N\right)$

3 Quantum Algorithms

In this section, we present quantum algorithms for similarity measurement. It consists of three ways to compute similarity of two data sets based on Euclidean distance: quantum algorithms for estimating the average linkage, the complete linkage and the single linkage.

Suppose that two N dimensional data sets $U : \left\{\overrightarrow{u_i} | i = 1, 2, \cdots, p\right\}$ and $V : \left\{\overrightarrow{v_j} | j = 1, 2, \cdots, q\right\}$ are given. From Section 2.2, the corresponding normalization $\left|\overrightarrow{u_i}\right|$ and $\left|\overrightarrow{v_j}\right|$ can be respectively obtained from quantum random access memory. Then, we can read data from QRAM efficiently and build their corresponding form of quantum state $|u_i\rangle$ and $\left|v_j\right\rangle$ through the following two unitary operations

$$U_{\mathcal{F}} : \frac{1}{\sqrt{p}} \sum_{i=1}^{p} |i\rangle|0\rangle \mapsto \frac{1}{\sqrt{A}} \sum_{i=1}^{p} |i\rangle \otimes \overrightarrow{u_i} = \frac{\sum_{i=1}^{p} \left|\overrightarrow{u_i}\right\| i\rangle |u_i\rangle}{\sqrt{\sum_{i=1}^{p} \left|\overrightarrow{u_i}\right|^2}}, \tag{4}$$

$$U_{\mathcal{T}} : \frac{1}{\sqrt{q}} \sum_{j=1}^{q} |0\rangle|j\rangle \mapsto \frac{1}{\sqrt{B}} \sum_{j=1}^{q} \overrightarrow{v_j} \otimes |j\rangle = \frac{\sum_{j=1}^{q} \left|\overrightarrow{v_j}\right\| v_j\rangle |j\rangle}{\sqrt{\sum_{j=1}^{q} \left|\overrightarrow{v_j}\right|^2}}. \tag{5}$$

Here, $A = \sum_{i=1}^{p} \left|\overrightarrow{u}_i\right|^2$, $B = \sum_{j=1}^{q} \left|\overrightarrow{v_j}\right|^2$. That is, the operations $U_{\mathcal{F}}$, $U_{\mathcal{T}}$ allow preparation of quantum state encoding an original data point of data sets $\left\{\overrightarrow{u_i}\right\}$ and $\left\{\overrightarrow{v_j}\right\}$ respectively. Moreover, based on the conclusions of Ref. [32], two controlled operators $C(U_{\mathcal{F}})$, $C(U_{\mathcal{T}})$ may be constructed with the circuits, if the operators $U_{\mathcal{F}}$ and $U_{\mathcal{T}}$ are implemented efficiently. In the following, the implementations of quantum algorithms for three computing methods are put forward.

3.1 Algorithm 1: a Quantum Algorithm for Estimating the Average Linkage

From (1), the average distance between two data sets can be expressed as quantum form

$$\begin{aligned} D(U,V) &= \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}|\overrightarrow{u_i}-\overrightarrow{v_j}|^2}{pq} \\ &= \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\left(|\overrightarrow{u_i}|^2+|\overrightarrow{v_j}|^2-2|\overrightarrow{u_i}||\overrightarrow{v_j}|\langle u_i|v_j\rangle\right)}{pq}. \end{aligned} \tag{6}$$

Obviously, we are able to use Lloyd algorithm [7] to accomplish this task. Namely, by Lloyd algorithm, the distances between all data points in one data set and the other data set are got, then the average distance is directly derived from the average value of these distances. However, it needs time $O(pq\log N)$ and is not efficient. To achieve the internal quantum efficiency, we reexamine this task and propose a new quantum algorithm, which is described as follows.

(A1.1) Let us start with constructing a quantum system that consists of four registers, in the initial state, $|\phi\rangle = |0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4$, where the subscript indicates the location of register.

(A1.2) An operation $H_1 \otimes F_2^p \otimes I_3 \otimes F_4^q$ is applied on these four registers. Here, $H = \frac{1}{\sqrt{2}}(|0\rangle\langle 0| + |0\rangle\langle 1| + |1\rangle\langle 0| - |1\rangle\langle 1|)$ is a Hadamard operator, and F^p (F^q) is a $p(q)$-level Fourier transform operator depicted as follows,

$$F^p = \sum_{i=1}^{p}\omega^{il}|i\rangle\langle l|, \tag{7}$$

where, $\omega = e^{\frac{2\pi \mathrm{i}}{p}}$. After this operation, the whole system is in the state,

$$\begin{aligned} |\varphi\rangle &= H_1 \otimes F_2^p \otimes I_3 \otimes F_4^q|\phi\rangle_{1234} \\ &= \tfrac{1}{\sqrt{2pq}}\textstyle\sum_{i=1}^{p}\sum_{j=1}^{q}(|0\rangle|i\rangle|0\rangle|j\rangle + |1\rangle|i\rangle|0\rangle|j\rangle)_{1234}, \end{aligned} \tag{8}$$

(A1.3) In this step, a controlled operator $C(U_{\mathcal{F}})_{123}$ ($C(U_{\mathcal{T}})_{134}$) is performed on $|\varphi\rangle$, where the first register is the control qubit, and the second and third (second and fourth) registers are the target. That is, if the control qubit is in state $|0\rangle$ then apply $U_{\mathcal{F}}$ to the target registers, otherwise apply $U_{\mathcal{T}}$, as shown in Fig. 3. After performing the operations $C(U_{\mathcal{F}})_{123}$ and $C(U_{\mathcal{T}})_{134}$, the desired state can be obtained,

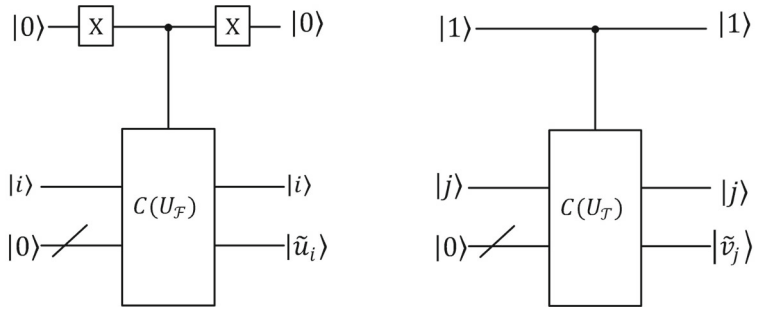
$$\begin{aligned} |\psi\rangle &= C(U_{\mathcal{F}})_{123}C(U_{\mathcal{T}})_{134}|\varphi\rangle_{1234} \\ &= \tfrac{1}{\sqrt{qA+pB}}\textstyle\sum_{i=1}^{p}\sum_{j=1}^{q}\left(|0\rangle|i\rangle\left|\overrightarrow{u_i}\right||u_i\rangle|j\rangle + |1\rangle|i\rangle\left|\overrightarrow{v_j}\right\|v_j\rangle|j\rangle\right)_{1234}, \end{aligned} \tag{9}$$

where, $A = \sum_{i=1}^{p}\left|\overrightarrow{u_i}\right|^2$ and $B = \sum_{j=1}^{q}\left|\overrightarrow{v_j}\right|^2$.

(A1.4) A projective measurement is performed on the first register of state $|\psi\rangle$ in the basis $\{|+\rangle, |-\rangle \mid |\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)\}$. A direct and simple calculation shows that

$$Pr(-) = \frac{\left(q\sum_{i=1}^{p}\left|\overrightarrow{u_i}\right|^2 + p\sum_{j=1}^{q}\left|\overrightarrow{v_j}\right|^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\left|\overrightarrow{u_i}\right|\left|\overrightarrow{v_j}\right|\langle u_i|v_j\rangle\right)}{2(qA+pB)}. \tag{10}$$

which is the probability of that the measurement outcome is $|-\rangle$.



(a) Quantum $C\,(U_{\mathcal{F}})$ circuit (b) Quantum $C\,(U_{\mathcal{T}})$ circuit

Fig. 3 Schematic diagram of controlled operations $C\,(U_{\mathcal{F}})$ and $C\,(U_{\mathcal{T}})$, where $|\tilde{u}_i\rangle$, $|\tilde{v}_j\rangle$ represent vectors which may not be normalized

(A1.5) Now, we can straightforward to calculate that square of desired distance,

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\left|\overrightarrow{u_i}-\overrightarrow{v_j}\right|^2 = 2(qA+pB)\times Pr(-). \tag{11}$$

After that, the similarity can be easy to deduce, is equal to $\frac{2(qA+pB)\times Pr(-)}{pq}$.

Through the above steps, we can obtain the average Euclidean distance between two data sets, which can be used to describe the similarity of these. The schematic quantum circuit of **Algorithm 1** is shown in Fig. 4.

3.2 Algorithm 2: a Quantum Algorithm for Estimating the Complete Linkage

In classical machine learning, because of the complexity of the average linkage calculation, the complete linkage is used as the distance between data sets sometimes. From (2), it is evident that the complete linkage is equal to the task of finding the maximum, from the distances between all data points in set $U = \left\{\overrightarrow{u_i}\right\}$ and each data point of set $V = \left\{\overrightarrow{v_j}\right\}$, i.e., $D'(U,V) = \max_i d_i^2 = \max_i (\max_j d^2(\overrightarrow{u_i}, \overrightarrow{v_j}))$.

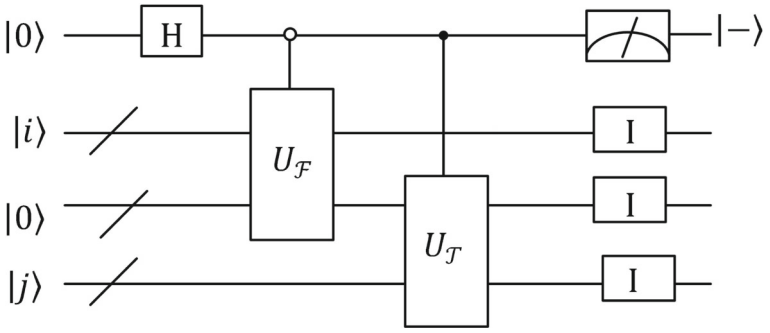


Fig. 4 Quantum circuit of **Algorithm 1**

Next, the second quantum algorithm for estimating the complete linkage, **Algorithm 2**, is presented. In this algorithm, we take the same assumption as that in Section 2, the data points are stored as quantum states in QRAM. So, the quantum counterpart of the distance between two points $\overrightarrow{u_i}$ and $\overrightarrow{v_j}$ is represented as

$$d_{i,j}^2 = \left|\left|\overrightarrow{u_i}\right||u_i\rangle - \left|\overrightarrow{v_j}\right||v_j\rangle\right|^2 = \left|\overrightarrow{u_i}\right|^2 + \left|\overrightarrow{v_j}\right|^2 - 2\left|\overrightarrow{u_i}\right|\left|\overrightarrow{v_j}\right|\langle u_i|v_j\rangle. \tag{12}$$

Besides, without loss of generality, we can assume $p \leq q$. In this case, we reckon the maximum distance between the sample vector $\overrightarrow{u_i}$ $(i = 1, 2, \cdots, p)$ and each data point of data set $\left\{\overrightarrow{v_j}\right\}$ in turn, and get set $\left\{d_1^2, d_2^2, \cdots, d_p^2\right\}$ $(d_i^2 = \max_j d^2(\overrightarrow{u_i}, \overrightarrow{v_j}),\ i = 1, 2, \cdots, p)$. Then, the desired distance can be find based on it. The details of the second algorithm are described in the following steps.

(A2.1) In a similar way to **Algorithm 1**, p initial states $|\phi_i'\rangle$ $(i = 1, 2, \cdots, p)$ are prepared firstly. Each state $|\phi_i'\rangle$ consists of four registers and in the initial state, $|\phi_i'\rangle = |0\rangle_1|i\rangle_2|0\rangle_3|0\rangle_4$. It will be utilized to calculate distance between the i^{th} data point $\overrightarrow{u_i}$ in data set $\left\{\overrightarrow{u_i}\right\}$ and each point of data set $\left\{\overrightarrow{v_j}\right\}$.

(A2.2) An operation $H_1 \otimes I_2 \otimes I_3 \otimes F_4^q$ is applied on the state $|\phi_i'\rangle$. After this operation, it is in the state,

$$\begin{aligned}\left|\varphi_i'\right\rangle &= H_1 \otimes I_2 \otimes I_3 \otimes F_4^q|\phi_i'\rangle_{1234} \\ &= \tfrac{1}{\sqrt{2q}}\textstyle\sum_{j=1}^{q}(|0\rangle|i\rangle|0\rangle|j\rangle + |1\rangle|i\rangle|0\rangle|j\rangle)_{1234}.\end{aligned} \tag{13}$$

(A2.3) Two controlled operations $C(U_{\mathcal{F}})_{123}$, $C(U_{\mathcal{T}})_{134}$ are performed on the state $\left|\varphi_i'\right\rangle$. And the details are the same as that in step (A1.3) of **Algorithm 1**. Then, we obtain the state

$$\begin{aligned}\left|\psi_i'\right\rangle &= C(U_{\mathcal{F}})_{123}C(U_{\mathcal{T}})_{134}|\varphi_i'\rangle_{1234} \\ &= \frac{\sum_{j=1}^{q}(|0\rangle|i\rangle|\overrightarrow{u_i}||u_i\rangle|j\rangle + |1\rangle|i\rangle|\overrightarrow{v_j}||v_j\rangle|j\rangle)_{1234}}{\sqrt{q|\overrightarrow{u_i}|^2 + \Sigma_j|\overrightarrow{v_j}|^2}}.\end{aligned} \tag{14}$$

(A2.4) In this step, we measure these q desired states $\left|\psi_i'\right\rangle$. Concretely, one projective measurement is performed on the first register of this state in the $\{|+\rangle, |-\rangle\}$ basis. At the same time, the other projective measurement is on the fourth register. This projective measurement is described by an observable, $\mathrm{M} = \Sigma_j|j\rangle\langle j|$, a Hermitian operator on the state space of the system being observed. After that, these two registers will project into $|-\rangle\langle -|$ and $|j\rangle\langle j|$ space respectively, with probability

$$Pr_{ij} = \frac{\left|\overrightarrow{u_i}\right|^2 + \left|\overrightarrow{v_j}\right|^2 - 2\left|\overrightarrow{u_i}\right|\left|\overrightarrow{v_j}\right|\langle u_i|v_j\rangle}{2\left(q\left|\overrightarrow{u_i}\right|^2 + \sum_{j=1}^{q}\left|\overrightarrow{v_j}\right|^2\right)}. \tag{15}$$

Thus, according to (12) and (15), the auxiliary qubit projects to the space spanned by the index $|j\rangle$ of set $\left\{\left|v_j\right\rangle\right\}$, that is the farthest from the i^{th} data point in set $\{\overrightarrow{u_i}\}$, with maximum probability mp_i. Finally, it is straightforward to calculate the square of maximum distance between the i^{th} sample vector in the data set $\{\overrightarrow{u_i}\}$ and the data point each of data set $\{\overrightarrow{v_j}\}$,

$$d_i^2 = mp_i \times 2\left(q\left|\overrightarrow{u_i}\right|^2 + \sum_{j=1}^{q}\left|\overrightarrow{v_j}\right|^2\right). \tag{16}$$

(A2.5) Repeating steps (A2.2)-(A2.4) for each $|\phi_i'\rangle$ $(i = 1, 2, \cdots, p)$ in turn. It allows us to get the distance set $\Omega = \left\{d_1^2, d_2^2, \cdots, d_p^2\right\}$. Then, it costs time $O(p)$ to check

every d_i^2 and find the maximum by using classical methods. Hence, we retrieve the farthest distance $d_{\max}^2$ of data point in the data sets $\{\overrightarrow{u_i}\}$ and $\{\overrightarrow{v_j}\}$.

Through the above steps, we can obtain the squared Euclidean distance between the furthest data vectors in two sets. Its execution operations are similar to **Algorithm 1**, so the circuit diagram can refer to Fig. 4.

3.3 Algorithm 3: a Quantum Algorithm for Estimating the Single Linkage

The single linkage method is more suitable for estimating the similarity of s-shaped and strip-shaped data sets in the clustering algorithm [29]. As shown in Section 2, the single linkage is measuring distance of the nearest pair of data vector in data sets $\{\overrightarrow{u_i}\}$ and $\{\overrightarrow{v_j}\}$, i.e., $D''(U,V) = \min d_i^2 = \min_i (\min_j d^2(\overrightarrow{u_i}, \overrightarrow{v_j}))$. Obviously, the method of estimating the complete linkage can be generalized to that of the single linkage. Next, we will describe the third quantum algorithm for estimating the single linkage. Before describing the steps of it, we first assume that the data sets have been processed which makes the different sets are in different spatial quadrants. Then, we introduce two quantum oracles.

Suppose that the following two quantum oracles O_u, O_v exist to obtain the reciprocal of each component of $\overrightarrow{u_i}$ and $\overrightarrow{v_j}$ respectively,

$$O_u : |i\rangle|k\rangle \mapsto \frac{1}{u_{ik}}|i\rangle|k\rangle = |i\rangle\left|\tilde{u}'_{ik}\right\rangle, \tag{17}$$

$$O_v : |k\rangle|j\rangle \mapsto \frac{1}{v_{jk}}|k\rangle|j\rangle = \left|\tilde{v}'_{jk}\right\rangle|j\rangle. \tag{18}$$

Here, the data vectors' components are stored as floating point numbers in quantum random access memory, and the states $\left|\tilde{u}'_{ik}\right\rangle$ and $|\tilde{v}'_{jk}\rangle$ representing vectors may not be normalized. Meanwhile, the two oracles can be used to efficiently prepare the vector states $\left|\tilde{u}'_i\right\rangle$, $\left|\tilde{v}'_j\right\rangle$ in time $O(\log N)$.

$$O_u : \sum_{k=1}^{N}|i\rangle|k\rangle \mapsto \sum_{k=1}^{N}\frac{1}{u_{ik}}|i\rangle|k\rangle = \sum_{k=1}^{N}\left|\tilde{u}'_{ik}\right\rangle|i\rangle \mapsto \overrightarrow{u_i}' = \left|\overrightarrow{u_i}'\right|\left|u'_i\right\rangle, \tag{19}$$

$$O_v : \sum_{k=1}^{N}|k\rangle|j\rangle \mapsto \sum_{k=1}^{N}\frac{1}{v_{jk}}|k\rangle|j\rangle = \sum_{k=1}^{N}\left|\tilde{v}'_{jk}\right\rangle|j\rangle \mapsto \overrightarrow{v_j}' = \left|\overrightarrow{v_j}'\right|\left|v'_j\right\rangle. \tag{20}$$

The essential steps of this algorithm is depicted as follows.

(A3.1) At first, we construct p initial states $|\phi_i''\rangle = |0\rangle_1|i\rangle_2|0\rangle_3|0\rangle_4$, and each one consists of four registers. As similar to **Algorithm 2**, the state $|\phi_i''\rangle$ means we estimate the distance between the i^{th} vector $\overrightarrow{u_i}$ and each point of data set $\left\{\overrightarrow{v_j}\right\}$, in this time.

(A3.2) We apply an operation $H_1 \otimes I_2 \otimes F_3^N \otimes F_4^q$ on the state $|\phi_i''\rangle$. Here, F^N (F^q) is a $N(q)$-level Fourier transform. In this way, we can obtain state

$$\begin{aligned}\left|\varphi_i''\right\rangle &= H_1 \otimes I_2 \otimes F_3^N \otimes F_4^q|\phi_i''\rangle_{1234} \\ &= \tfrac{1}{\sqrt{2q}}\textstyle\sum_{j=1}^{q}\sum_{k=1}^{N}(|0\rangle|i\rangle|k\rangle|j\rangle + |1\rangle|i\rangle|k\rangle|j\rangle)_{1234}.\end{aligned} \tag{21}$$

(A3.3) Two controlled operations are performed on the state $\left|\varphi_i''\right\rangle$ based on the oracles. That is, the oracle O_u is applied on the second and the third register if the first

qubit of $\left|\varphi_i''\right\rangle$ is in the state $|0\rangle$; otherwise oracle O_v is applied on the third and fourth register. In this way, we will have the specific state

$$\begin{aligned}\left|\psi_i''\right\rangle &= \frac{1}{\sqrt{2q}}\sum_{j=1}^{q}\sum_{k=1}^{N}\left(|0\rangle O_u|i\rangle|k\rangle|j\rangle + |1\rangle|i\rangle O_v|k\rangle|j\rangle\right)_{1234} \\ &= \frac{\sum_{j=1}^{q}\left(|0\rangle|i\rangle|\overrightarrow{u_i}'|\left|u_i'\right\rangle|j\rangle+|1\rangle|i\rangle|\overrightarrow{v_j}'|\left|v_j'\right\rangle|j\rangle\right)_{1234}}{\sqrt{q|\overrightarrow{u_i}'|^2+\sum_{j=1}^{q}|\overrightarrow{v_j}'|^2}}.\end{aligned} \tag{22}$$

(A3.4) In this step, the state $\left|\psi_i''\right\rangle$ instead of state $\left|\psi_i'\right\rangle$ in the second quantum algorithm. Then, we repeat steps (A2.2)-(A2.4) of **Algorithm 2**, it can easily adapts to compute the maximum $d_i'^2$.

(A3.5) We get the distance set $\Omega' = \left\{d_1'^2, d_2'^2, \cdots, d_p'^2\right\}$ after we repeat steps (A3.2)-(A3.4) for each state $|\phi_i''\rangle$ $(i = 1, 2, \cdots, p)$. Hence, we retrieve a distance between the farthest pair of data sets (these correspond to the nearest pair of sets in the original data sets) in the way of step (A2.5) from **Algorithm 2**.

Although the nearest distance between the original data sets is not calculated through the above steps, our algorithm can map the Euclidean distance between the same closest data point pairs. Then, we can depict the similarity between the two sets by it. This result has the same effect on other machine learning tasks, such as hierarchical clustering.

4 Runtime Analysis

In this paper, we proposed three algorithms' main ideas are similarly, that they all utilize special measurement results to estimate the Euclidean distance between data sets. However, their implementation details are different, so there are some differences in running time. In this section, the performance of them is analyzed briefly.

From Section 3.1, it is known that the time cost is mainly for two main steps in **Algorithm 1**, i.e., steps (A1.3) and (A1.4). Evidently, what step (A1.3) generates the state of (9) takes time $O(\log(pqN^2))$ [31] based on two controlled operations $C(U_{\mathcal{F}})$ and $C(U_{\mathcal{T}})$. In step (A1.4), the time cost is evidently dominated by the amplitude estimation. According to [28], taking $O\left(\sqrt{\frac{1-Pr(-)}{Pr(-)}}\cdot\frac{1}{\epsilon}\right) = O\left(\sqrt{\frac{1}{Pr(-)}}\cdot\frac{1}{\epsilon}\right)$ times to perform projective measurement ensures that desired result $Pr(-)$ being within relative error ϵ. Finally, we can be easy to calculate the average distance by the special measure result. Hence, the time cost of the total **Algorithm 1** is $O\left(\frac{1}{\epsilon\cdot\sqrt{Pr(-)}}\log\left(pqN^2\right)\right)$. Compared with time $O(ploy(pqN))$ for the best-known classical algorithm, in this case the exponential speedup can be achieved.

As for **Algorithm 2**, it is shown that there are two main differences with the last algorithm. One is preparation of state $\left|\psi_i'\right\rangle$ as shown in step (A2.3). It takes time $O(\log(N) + \log(qN))$. Another is that **Algorithm 2** gets the maximum distance set $\Omega = \left\{d_1^2, d_2^2, \cdots, d_p^2\right\}$. It is described in step (A2.5), and each d_i^2 for $i = 1, 2, \cdots, p$ costs $O\left(\frac{1-mp_i}{mp_i}\cdot\frac{1}{\epsilon_d}\right) = O\left(\frac{1}{mp_i\cdot\epsilon_d}\right)$ times within relative error ϵ_d of mp_i. Therefore, getting the distance set Ω takes time $O\left(\sum_{i=1}^{p}\frac{1}{mp_i\cdot\epsilon_d}\log(qN^2)\right)$. Then, we find the maximum of set $\left\{d_1^2, d_2^2, \cdots, d_p^2\right\}$ by using the simple classical method in time $O(p)$. Through above brief analysis, the total time of **Algorithm 2** is $O\left(\sum_{i=1}^{p}\frac{1}{mp_i\cdot\epsilon_d}\log(qN^2) + p\right)$. It means

that when the number of data space and the dimension of data are large, **Algorithm 2** can achieve exponential speedup over the classical similarity measurement algorithms.

In **Algorithm 3**, the method adopted is the generalization of the second algorithm. The only difference is that in step (A3.3) of the third quantum algorithm, we preprocess the original data and generate a specific state $\left|\psi_i''\right\rangle$ with our quantum oracles O_u and O_v. Thus, the nearest distance is converted to the farthest distance to solve the problem. Due to the parallel processing of the quantum algorithm, the time complexity of **Algorithm 3** is similar to that of **Algorithm 2**. Therefore, it is also exponentially accelerated relative to the steps that produce the same effect on the classical algorithm.

Through the brief analysis, we presented algorithms are accelerated to a certain extent compared with the classical algorithm. It is worth emphasizing that the quantum algorithm implementation of the proposed three programs allow us to call different subroutines according to different needs, just like the classical clustering algorithm.

5 Conclusion

In this paper, we made a further study on the quantum similarity measurement. Instead of date points, the distance between data sets is considered, which usually acts as a basic component of some machine learning algorithms. First, we discussed three common kinds of Euclidean distance methods. Then, the characteristics of parallel access data of quantum random access memory and amplitude estimation technology are utilized to propose the corresponding three quantum algorithms. At last, it is shown that the proposed algorithms can achieve exponential speedup over the classical counterparts for data sets with a mass of high-dimensional vector, via a brief complexity analysis. These results demonstrate that measuring the similarity between data sets can implement other interesting tasks with significantly less quantum resource, e.g., document clustering [33, 34], medical analysis [35, 36], and so on.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grants No. 61976053 and No. 61772134), Fujian Province Natural Science Foundation (Grant No. 2018J01776), and Program for New Century Excellent Talents in Fujian Province University.

References

1. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring[C]. In: Proceedings 35th Annual Symposium on Foundations of Computer Science, pp. 124-134. IEEE Press, New York (1994)
2. Grover, L.K.: Quantum mechanics helps in searching for a needle in a haystack[J]. Phys. Rev. Lett. **79**, 325 (1997)
3. Harrow, A.W., Hassidim, A., Lloyd, S.: Quantum algorithm for linear systems of equations[J]. Phys. Rev. Lett. **103**, 150502 (2009)
4. Berry, D.W., Ahokas, G., Cleve, R., et al.: Efficient quantum algorithms for simulating sparse Hamiltonians[J]. Commun. Math. Phys. **270**, 359–371 (2007)
5. Lloyd, S., Mohseni, M., Rebentrost, P.: Quantum principal component analysis[J]. Nat. Phys. **10**, 631–633 (2014)
6. Anguita, D., Ridella, S., Rivieccio, F., et al.: Quantum optimization for training support vector machines[J]. Neural Netw. **16**, 763–770 (2003)
7. Lloyd, S., Mohseni, M., Rebentrost, P.: Quantum algorithms for supervised and unsupervised machine learning[J]. arXiv:1307.0411 (2013)

8. Rebentrost, P., Mohseni, M., Lloyd, S.: Quantum support vector machine for big data classification[J]. Phys. Rev. Lett. **113**, 130503 (2014)
9. Cong, I., Duan, L.: Quantum discriminant analysis for dimensionality reduction and classification[J]. New J. Phys. **18**, 073011 (2016)
10. Ruan, Y., Xue, X., Liu, H., et al.: Quantum algorithm for k-nearest neighbors classification based on the metric of hamming distance[J]. Int. J. Theor. Phys. **56**, 3496–3507 (2017)
11. Aïmeur, E., Brassard, G., Gambs, S.: Quantum speed-up for unsupervised learning[J]. Mach. Learn. **90**, 261-287 (2013)
12. Wiebe, N., Braun, D., Lloyd, S.: Quantum algorithm for data fitting[J]. Phys. Rev. Lett. **109**, 050505 (2012)
13. Schuld, M., Sinayskiy, I., Petruccione, F.: Prediction by linear regression on a quantum computer[J]. Phys. Rev. A **94**, 022342 (2016)
14. Yu, C.H., Gao, F., Wen, Q.Y.: Quantum algorithm for ridge regression[J]. arXiv:1707.09524 (2017)
15. Garnerone, S., Zanardi, P., Lidar, D.A.: Adiabatic quantum algorithm for search engine ranking[J]. Phys. Rev. Lett. **108**, 230506 (2012)
16. Paparo, G.D., Martin-Delgado, M.A.: Google in a quantum network[J]. Sci. Rep. **2**, 444 (2012)
17. Sánchez-Burillo, E., Duch, J., Gómez-Gardenes, J., et al.: Quantum navigation and ranking in complex networks[J]. Sci. Rep. **2**, 605 (2012)
18. Paparo, G.D., Müller, M., Comellas, F., et al.: Quantum google in a complex network[J]. Sci. Rep. **3**, 2773 (2013)
19. Schuld, M., Sinayskiy, I., Petruccione, F.: The quest for a quantum neural network[J]. Quantum Inf. Process **13**, 2567–2586 (2014)
20. Jiang, D.H., Wang, J., Liang, X.Q., et al.: Quantum voting scheme based on locally indistinguishable orthogonal product states[J]. Int. J. Theor. Phys. **59**, 436–444 (2020)
21. Cohen-Addad, V., Klein, P.N., Mathieu, C.: Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics[J]. SIAM J. Comput. **48**, 644–667 (2019)
22. Ahmadian, S., Norouzi-Fard, A., Svensson, O., et al.: Better guarantees for k-means and euclidean k-median by primal-dual algorithms[J]. SIAM Journal on Computing, 2019, FOCS17-97 (2019)
23. Soucy, P., Mineau, G.W.: A simple KNN algorithm for text categorization[C]. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 647-648. IEEE, California (2001)
24. Xing, W., Bei, Y.: Medical health big data classification based on KNN classification algorithm[J] IEEE access (2019)
25. Aaronson, S.: BQP And the polynomial hierarchy[C]. In: Proceedings of the Forty-second ACM Symposium on Theory Of Computing, 141-150, California (2010)
26. Seifoddini, H.K.: Single linkage versus average linkage clustering in machine cells formation applications[J]. Comput. Ind. Eng. **16**, 419–426 (1989)
27. Giovannetti, V., Lloyd, S., Maccone, L.: Quantum random access memory[J]. Phys. Rev. Lett. **100**, 160501 (2008)
28. Brassard, G., Hoyer, P., Mosca, M., et al.: Quantum amplitude amplification and estimation[J]. Contemp. Math. **305**, 53–74 (2002)
29. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation[J]. ACM Trans. Knowl. Discov. Data **1**, 4-es (2007)
30. Kerenidis, I., Prakash, A.: Quantum recommendation systems[J]. arXiv:1603.08675 (2016)
31. Wossnig, L., Zhao, Z., Prakash, A.: Quantum linear system algorithm for dense matrices[J]. Phys. Rev. Lett. **120**, 050502 (2018)
32. Araújo, M., Feix, A., Costa, F., et al.: Quantum circuits cannot control unknown operations[J]. New J. Phys. **16**, 093026 (2014)
33. Sardar, T.H., Ansari, Z.: An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm[J]. Future Comput. Inform. J. **3**, 200–209 (2018)
34. Yang, R., Qu, D., Qian, Y., et al.: An online log template extraction method based on hierarchical clustering[J]. EURASIP J. Wirel. Commun. Netw. **2019**, 1–12 (2019)
35. Lange, J.K., DiSegna, S.T., Yang, W., et al.: Using cluster analysis to identify patient factors linked to differential functional gains after total knee Arthroplasty[J]. J. Arthroplasty **35**, 121–126 (2020)
36. Zhang, Y., Liu, Y., Li, Y., et al.: Hierarchical and complex system entropy clustering analysis based validation for traditional Chinese medicine syndrome patterns of chronic atrophic gastritis[J]. J. Altern. Complement. Med. **25**, 1130–1139 (2019)