KNN Algorithm Simulation Based on Quantum Information

Ning Yang
Seidenberg School of CSIS,Pace University,Pleasantville,New York 10570
Email: ny00685p@pace.edu

***Abstract*—Text classification is the process of assigning tags or categories to text according to its content. It is one of the fundamental tasks in Natural Language Processing. The KNN algorithm is a widely used text classification for it is simple, valid and non-parameters. It is one of the simplest classifications. The main idea of KNN is to calculate the similarity which is reflected by the distance between the new sample with an unknown class label and the training samples and choosing the class label of the highest k nearest neighbors as the new sample's class label. However, the real text data always contains many features. The similarity calculation will cost much time. Meanwhile, many traditional machine learning algorithms cannot compute large number of vectors in high-dimensional vectors in high dimensional space. The Quantum Computing algorithms are good at computing high-dimensional vectors in large tensor product spaces. It can provide exponential speed-up over its classical counterparts. This paper introduces a KNN algorithm based on quantum computing, which uses fidelity to compute the similarity between two quantum states. Meanwhile, we use the IBM Qubit toolkit to simulate the circle of this algorithm and take the test. The test result is satisfying the design.**

Key words: Text Classification; KNN Algorithm; Quantum Computing; IBM Qubit Toolkit.

1. Introduction

Quantum Information[1] is is the information of the state of a quantum system; it is the basic entity of study in quantum information theory and can be manipulated using quantum information processing techniques. Quantum information, like classical information, can be processed using digital computers, transmitted from one location to another, manipulated with algorithms, and analyzed with computer science mathematics. It is very different from its classical counterpart. The quantum state is in a superposition of classical states. Quantum information theory[2] is broader in scope than classical information theory, for quantum information theory includes all the static and dynamic elements of classical information theory, as well as additional static and dynamic elements. Quantum computing is good at high-dimensional vectors manipulating in large tensor product spaces, so many researchers recently focus on the problem of large scale machine learning algorithms based on quantum computing. The problems such as factoring, quantum simulation, and optimization can be speed-up by the Quantum state design. Quantum algorithms have begun to emerge that promise quantum advantage for solving problems in data processing, machine learning, and classification. It has been shown by Lloyd, Mohseni, and Rebentrost that quantum algorithms, which are good at manipulating vectors and matrices, could provide an asymptotically exponential speed-up over their classical counterparts [3]. Feynman [4] proposed that to construct a quantum computer based on quantum mechanics principle, which can solve problems more effective than classical computer. Quantum computing[5] is the use of quantum-mechanical phenomena such as superposition and entanglement to perform computation. A quantum computer is used to perform such computation, which can be implemented theoretically or physically.

Text classification is the process of assigning tags or categories to text according to its content It's one of the fundamental tasks in Natural Language Processing with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. It is the problem of automatically assigning predefined categories to free text documents. Many text classification algorithms use machine learning algorithms, which can be categorized from supervised and unsupervised learning. However, It is impossible for many machine learning or classification model to handle huge feature spaces. However, Quantum computing is good at handling large feature space. The classical data in quantum computing is expressed in the form of N-dimensional feature space which can be mapped to quantum states over log_2N qubits. The quantum machine learning algorithm would provide an efficient speed-up over their classical opponents.

In this paper, we will mainly focus on the KNN(k-Nearest Neighbor)algorithm in quantum form. The classical KNN is a very popular, simple and non-parameter text classification algorithm[5]. The main idea of the KNN algorithm is computing similarities between the unknown sample and all training samples in order to find the top k nearest neighbors of the unknown sample. The big problem of the KNN algorithm is that the text has a highdimension feature space, which will cost much time to do computation on it. However, the quantum computing can use N bits memory to create 2^N quantum states which can finish the operation at the same time This is called quantum parallelism. Thus, this paper will talk about how to adjust the KNN algorithm to an algorithm suit for the quantum mechanical to improve the efficiency.

Thanks to the IBM Faculty Award that made this research possible

2. Quantum Computing Basic

Quantum computing's basic is different from classical computing information. It involves a quantum bit(qubit), quantum register and quantum gates[6]. This section will introduce the quantum basic information based on the three parts.

2.1. Quantum Bits

Quantum bits are the fundamental units of information in quantum computing, just as bits are fundamental units in classical computing. Many ways can be used for the qubits' realization such as polarized photons. In general, the qubit is always described as a vector in a two-dimensional Hilbert space.

Qubit is the unit of quantum information[7]. The qubit also has two basic states as the classical one(0 and 1), it described as $|0\rangle$ and $|1\rangle$:

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{1}$$

Any superposition of qubit$|\phi\rangle$ can be represented with these two basic states:

$$|\phi\rangle = a|0\rangle + b|1\rangle \tag{2}$$

The a and b are some complex numbers which mean the probability amplitudes of this qubit in Hilbert spaces. It means the probability of$|0\rangle$ and $|1\rangle$ is $|a|^2$ and $|b|^2$

2.2. Quantum Register

A quantum register is the quantum mechanical analog of a classical processor register. A mathematical description of a quantum register is achieved by using tensor products of qubit bra or ket vectors. For example, a 2 qubit quantum register is described by the element

$$|\phi\rangle = |\phi 1\rangle \otimes |\phi 2\rangle = |\phi 1\rangle|\phi 2\rangle = |\phi 1 \phi 2\rangle \tag{3}$$

Thus a quantum register comprises a number of qubits, the number of the qubit is its size. For example,5 a 2-qubit quantum register can store individual numbers such as 2 or 3:

$$|1\rangle \otimes |0\rangle = |10\rangle = |2\rangle, |1\rangle \otimes |1\rangle = |11\rangle = |3\rangle \tag{4}$$

Thus the 2-bits quantum register contains 4 bais states:

$$|\phi\rangle = a_0|00\rangle + a_1|01\rangle + a_2|10\rangle + a_3|11\rangle = a_0|0\rangle + a_1|1\rangle + a_2|2\rangle + a_3|3\rangle \tag{5}$$

The probability of measurement for each basis should equal to 1:

$$|a_0|^2 + |a_1|^2 + |a_2|^2 + |a_3|^2 = 1 \tag{6}$$

Thus, a n-bit quantum register should have 2^n qubit states and can be represented as

$$|\phi\rangle = a_0|00...0\rangle + a_1|00...1\rangle + ... + a_{2^n-1}|11...1\rangle \tag{7}$$

and the formula can be represented with replacing the binary states with decimal state:

$$|\phi\rangle = \sum_{i=0}^{2^n-1} a_i|i\rangle \tag{8}$$

2.3. Quantum Gate

In quantum computing and specifically the quantum circuit model of computation, a quantum logic gate[8] (or simply quantum gate) is a basic quantum circuit operating on a small number of qubits. They are the building blocks of quantum circuits like classical logic gates are for conventional digital circuits. Unlike many classical logic gates, quantum logic gates are reversible. However, it is possible to perform classical computing using only reversible gates. Quantum gates are unitary operators of the Hilbert space. It changes the quantum state of qubits or quantum registers. We always use a unitary matrix to indicate quantum gates. Quantum gate map quantum states onto other quantum states[9]. In order to keep the orthonormal eigenstate of the qubit states in Hilbert space, the transformer mapping quantum state is a unitary matrix. The definition of the unitary matrix is as follows:

$$UU^{\dagger} = I \tag{9}$$

The basic one-bit quantum gates are shown in figure1: The first gate is called Hadamard gate. It acts on a single

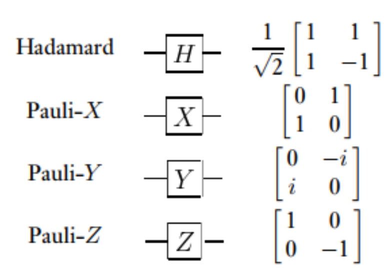

Figure 1. The Basic Quantum Gate

qubit. It maps the basis state$|0\rangle$ to $\frac{|0\rangle+|1\rangle}{2}$ and $|1\rangle$to $\frac{|0\rangle-|1\rangle}{2}$, which means that measurement will have equal probabilities to become 1 or 0. The Hadamard gate is the one-qubit version of the quantum Fourier transform and it is a unitary matrix.

The second one is called Pauli-X gate. The Pauli-X gate acts on a single qubit. It is the quantum equivalent of the NOT gate for classical computers (with respect to the standard basis $|0\rangle$ and $|1\rangle$ which distinguishes the Z-direction; in the sense that measurement of the eigenvalue +1 corresponds to classical 1/true and -1 to 0/false).

The third one is called Pauli-Y gate. It also acts on a single qubit. It equates to a rotation around the Y-axis of the Bloch sphere by π radians. It maps $|0\rangle$ to $i|1\rangle$ and $|1\rangle$ to $-i|0\rangle$.

The last one is called Pauli-Z gate. It equates to a rotation around the Z-axis of the Bloch sphere by π radians. Thus, it is a special case of a phase shift gate with $\phi = \pi$. It leaves the basis state $|0\rangle$ unchanged and maps $|1\rangle$ to $-|1\rangle$. Due to this nature, it is sometimes called phase-flip.

The most important two-qubit quantum register is The Controlled-NOT(CNOT) gate, it is a 4-dimensional unitary matrix which can be described as: The first qubit is the

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} .$$

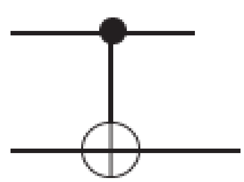

Figure 2. The CNOT Gate

control qubit, and the second one is the target bit. The target will reverse its state if and only if the control qubit is in state$|1\rangle$. Hence, we can find that:

$$CNOT|00\rangle = |00\rangle, CNOT|01\rangle = |01\rangle \tag{10}$$

$$CNOT|10\rangle = |10\rangle, CNOT|11\rangle = |10\rangle \tag{11}$$

Another important state for 2-qubit gate is the Bell State. The Bell states[10] are four specific maximally entangled quantum states of two qubits. They are in a superposition of 0 and 1–that is, a linear combination of the two states. Four specific two-qubit states with the maximal value of $2\sqrt{2}$ are designated as "Bell states". They are known as the four maximally entangled two-qubit Bell states, and they form a maximally entangled basis, known as the Bell basis, of the four-dimensional Hilbert space for two qubits:

$$|\psi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |0\rangle_B + |1\rangle_A \otimes |1\rangle_B) \tag{12}$$

$$|\psi^-\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |0\rangle_B - |1\rangle_A \otimes |1\rangle_B) \tag{13}$$

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |1\rangle_B + |1\rangle_A \otimes |0\rangle_B) \tag{14}$$

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |1\rangle_B - |1\rangle_A \otimes |0\rangle_B) \tag{15}$$

When we measure one of the qubit states of the Bell state, the Bell state will collapse to a certain state.

3. Quantum KNN algorithm

3.1. Data Structure and Description

The text classification is automatically assigning predefined categories to new samples. Many traditional machine learning algorithms are used in this field. These algorithms learn the features from the original data, build a classification model for new data samples. The traditional data sets in classical computing can be described as:

$$S_n = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)...(x_i, y_i)...(x_n, y_n)\} \tag{16}$$

The S_n is the description of the training data set, the (x_i,y_i) is i^th training data sample. x_i is the input parameter of the data while y_i is the class label which reflects the input's value or category.

Quantum information has a different structure of computing. However, the training dataset for Quantum machine learning is still necessary. The data set in Quantum computing contains quantum states. A training dataset can be described as follows:

$$S_n = \{(|x_1\rangle, y_1), (|x_2\rangle, y_2), (|x_3\rangle, y_3)...(|x_i\rangle, y_i)...(|x_n\rangle, y_n)\} \tag{17}$$

The $|x_i\rangle$ is the i^th quantum state of the training set and y_i is the class label which can reflect the input's state value and category. In classical computing, the binary string is always be used to represent the data value such as: (x_1,x_2,x_3...x_i..x_n) where x_i∈(0,1),*i*=1,2,3...n. Thus, the structure in Quantum computing can be the same as in classical computing, the binary string value can be directly translated to n-qubit quantum state $|x_1...x_i...x_n\rangle$ with basis $\{|0...00\rangle$,$|0...01\rangle$...$|1...11\rangle\}$. Seth L. and his co-workers[1] have been proposed alternative data representation. the classical information can be encoded into the norm of a quantum state: $\langle$x|x$\rangle$=$|\overline{x}|^{-1}\times\overline{x}^2$. Thus, the quantum state$|x\rangle$ can be described as follows:

$$|x\rangle = |\overline{x}|^{-1/2} \times \overline{x} \tag{18}$$

where $\overline{x}$is the vector of $|x\rangle$ and $|\overline{x}|$ is the nom of x. A pure qubit state is a coherent superposition[2] of the basis states. With superposition state, a quantum bit can be represented as $|\psi\rangle$=$\alpha|0\rangle$+$\beta|1\rangle$. The α and β are probability amplitude, and $|\alpha|^2$+ $|\beta|^2$=1. Therefore, an n-dimensional quantum register can be encoded as:

$$\begin{bmatrix} \alpha_1 & \alpha_2 & ... & \alpha_n \\ \beta_1 & \beta_2 & ... & \beta_n \end{bmatrix} \tag{19}$$

where $|\alpha_i|^2$+$|\beta_i|^2$=1,i=1,2,3.... . For example, a 3-qubit state can be described as:

$$\begin{bmatrix} \sqrt{2}/2 & 1 & 1/2 \\ \sqrt{2}/2 & 0 & \sqrt{3}/2 \end{bmatrix} \tag{20}$$

The formula is a superstation of 3-qubit which can be described as:

$$Q = \frac{1}{2\sqrt{2}}|000\rangle + \frac{\sqrt{3}}{2\sqrt{2}}|001\rangle + \frac{1}{2\sqrt{2}}|100\rangle + \frac{\sqrt{3}}{2\sqrt{2}}|101\rangle \tag{21}$$

There are four superstation states $|000\rangle$,$|001\rangle$,$|100\rangle$,$|101\rangle$. with the probability amplitude are $\frac{1}{8}$,$\frac{3}{8}$,$\frac{1}{8}$,$\frac{3}{8}$.

3.2. Quantum KNN algorithm

KNN is a very popular, simple and non-parameter text classification algorithm. The idea of the KNN algorithm is to choose the class label for the new input that appears most often amongst its k near neighbors. This is based on the assumption that the class label of vector is the same as the nearest neighbors'.

KNN algorithm is obviously based on a distance metric to evaluate the similarity of two feature vectors. The two vector $\overline{v}_A$ and $\overline{v}_B$ are labeled as A and B. To classify the new sample which is represented by the vector $\overline{u}$. One common method is to calculate and compare the distance between the three vectors. The new sample is assigned to the vector's class label which the distance is smaller.

In order to translate this algorithm into a quantum version, we focus on the efficient evaluation of classical distance through a quantum algorithm. One idea is to use similarity of two quantum states $|\psi\rangle$and $|\phi\rangle$ as the similarity measure, which is known as fidelity.

$$Fid(|\psi\rangle, |\phi\rangle) = |\langle\phi|\psi\rangle|^2 \tag{22}$$

The fidelity is similar to cosine similarity used in classical information. The value of fidelity ranges from 0 if the two quantum states are orthogonal to 1 if the states are identical. The fidelity can be obtained through a simple quantum routine sometimes referred to as a control swap test. The circle design is shown in Figure.3: The C-Swap gate is a

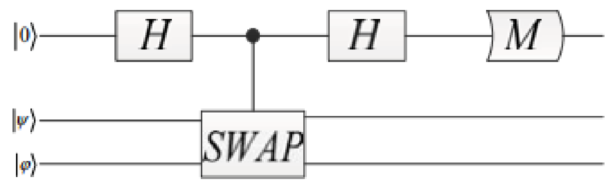

Figure 3. The Quantum Circle Design

3-bit quantum register that performs a controlled swap. The C-Swap matrix as follows, it is also a unitary matrix.

$$\begin{bmatrix} 1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&1 \end{bmatrix} \tag{23}$$

Given a quantum register $|\phi, \psi\rangle$, as well as an ancilla register initially set to $|0\rangle$, the input quantum states is $|0, \phi, \psi\rangle$. A Hadamard gate transform the ancilla into a superposition:$1/\sqrt{2}(|0\rangle + |1\rangle)$. Followed SWAP gate on $|\psi\rangle$ and $|\phi\rangle$ which swaps the two states under the condition that the ancilla is in state $|1\rangle$. Therefore, the quantum state evolves to:

$$\frac{1}{\sqrt{2}}|\phi\rangle|\psi\rangle + \frac{1}{\sqrt{2}}|\psi\rangle|\phi\rangle \tag{24}$$

Afterwards, the second Hadamard gate transforms formula to:

$$\frac{1}{\sqrt{2}}|0\rangle(|\phi\rangle|\psi\rangle + |\psi\rangle|\phi\rangle) + \frac{1}{\sqrt{2}}|1\rangle(|\phi\rangle|\psi\rangle - |\psi\rangle|\phi\rangle) \tag{25}$$

Finally, measurement is on the second Hadamard gate output, which the probability of measuring the ground state is given by:

$$P(|0\rangle_anc) = \frac{1}{2} + \frac{1}{2}|\langle\phi|\psi\rangle|^2 \tag{26}$$

Hence, if $|\phi\rangle$ and $|\psi\rangle$ are identical, the probability is 1. If $|\phi\rangle$ and $|\psi\rangle$ are orthogonal the probability is 1/2. Therefore, the C-Swap test is an estimator for the fidelity between $|\phi\rangle$ and $|\psi\rangle$.

4. Simulation and Result Analysis

4.1. Tools and Environment

The quantum computing system is different from classical computer. In order to stimulate the environment, we use the Qiskit[10] with python to build the Quantum circle. The Qiskit is an open-source framework for working with noisy quantum computers at the level of pulses, circuits, and algorithms. Its goal is to build a software stack that makes it easy for anyone to use quantum computers. However, Qiskit also aims to facilitate research on the most important open issues facing quantum computation today.

Qiskit consists of four foundational elements: Terra (the code foundation, for composing quantum programs at the level of circuits and pulses), Aqua (for building algorithms and applications), Ignis (for addressing noise and errors), and Aer (for accelerating development via simulators, emulators and debuggers). We mainly use the Qiskit Terra for the building of quantum register and circle stimulation.

4.2. Simulation and Result

Firstly we build the Quantum circle which is shown in figure4: In this circle, we simulate the process we introduced

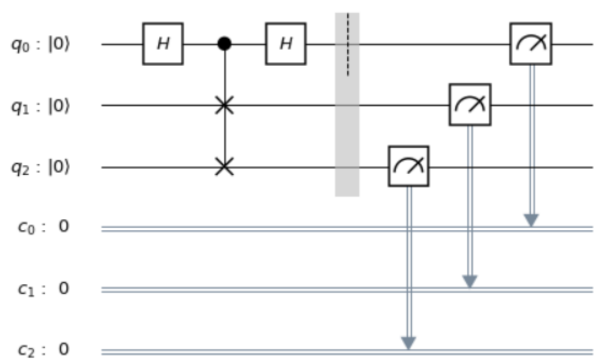

Figure 4. The Quantum Circle

in last section. The first qubit p0's signal pass the first Hadamard gate to transfer to the superposition state. The q1 and q2 register store the state which needs to be compared. The signal in one Qubit is $|0\rangle$ and $|1\rangle$. The gate between the three registers is called Fredkin gate whos function is the same as CSWAP gate which is mentioned in part 3. The final signal will pass the second Hadamard gate and be measured at last. If the two states are orthogonal, the measure is 0.5 while if they are identical, the probability is 1.

We use two backends to simulate the circle. First we use the unitary simulator that works provided all the elements in the circuit are unitary operations. The result is shown in figure5: The unitary result is a 2^3 x 2^3 matrix representing the gates in this circle. Most of the result is 0.5 until the Qubit q1 and q2 is the same.

```
[[ 1. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  1. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j  0.5+0.j  0.5+0.j  0.5+0.j -0.5+0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j  0.5+0.j  0.5+0.j -0.5+0.j  0.5+0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j  0.5+0.j -0.5+0.j  0.5+0.j  0.5+0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j -0.5+0.j  0.5+0.j  0.5+0.j  0.5+0.j  0. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  1. +0.j  0. +0.j]
 [ 0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  0. +0.j  1. +0.j]]
```

Figure 5. The Result for Unitary Backend

We also use the State-vector backend to analyze the result. The Statevector simulator is the most common backend in Qiskit. It will return the quantum states which is a complex vector of dimensions 2^3. The process and result are shown in figure6: In this simulation, we only use one state

```
backend = BasicAer.get_backend('statevector_simulator')
# Create a Quantum Program for execution
job = execute(circ, backend)
result = job.result()
outputstate = result.get_statevector(circ, decimals=3)
print(outputstate)
```

```
[1.+0.j 0.+0.j 0.+0.j 0.+0.j 0.+0.j 0.+0.j 0.+0.j 0.+0.j]
```

Figure 6. The Result for Statevector Backend

as the input state. In order to check whether the state can be extended, we design a new quantum circle for 2 qubit state comparison which is shown in figure7.: We extended

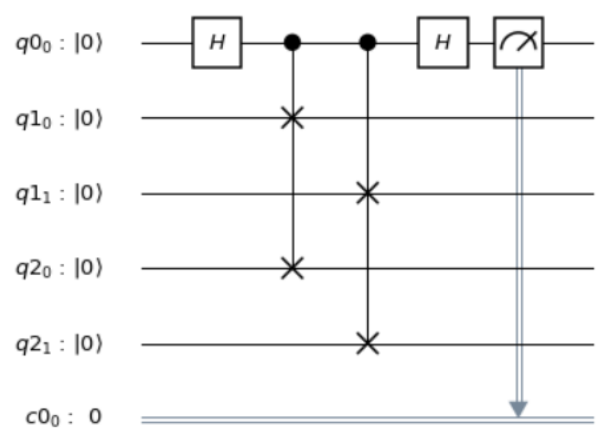

Figure 7. The Extended Quantum Circle

the state in this circle, the state is changed from $|00\rangle$ to $|11\rangle$. The measured result will be 1 at the same state while it will be 0.5 if it is different. The Unitary backend result is shown in figure 8:

5. Conclusion and Future work

In our work, we analyze the difference between classical computing and Quantum computing and introduce a way to realize the traditional KNN algorithm in Quantum computing environment. In the simulation, the designed quantum

```
[[1. +0.j 0. +0.j 0. +0.j ... 0. +0.j 0. +0.j 0. +0.j]
 [0. +0.j 1. +0.j 0. +0.j ... 0. +0.j 0. +0.j 0. +0.j]
 [0. +0.j 0. +0.j 0.5+0.j ... 0. +0.j 0. +0.j 0. +0.j]
 ...
 [0. +0.j 0. +0.j 0. +0.j ... 0.5+0.j 0. +0.j 0. +0.j]
 [0. +0.j 0. +0.j 0. +0.j ... 0. +0.j 1. +0.j 0. +0.j]
 [0. +0.j 0. +0.j 0. +0.j ... 0. +0.j 0. +0.j 1. +0.j]]
```

Figure 8. The Extended Quantum Circle Unitary Result

circle can judge whether two stated is the same as the Fredkin Quantum gate. Mostly simulations work well and can present what we thought. However, it can only show whether the two states are same or not. It cannot show a detailed distance between two states. In future work, we will focus on the states extension and the distance calculation to let this Quantum Circle more efficient and powerful.

References

[1] https://en.wikipedia.org/wiki/Quantum information

[2] Esma A., Gilles B., and Sebastien G., " Machine Learning in a Quantum World", L. Lamontagne and M. Marchand (Eds.): Canadian AI 2006, LNAI 4013, pp 431-442, 2006.

[3] Feynman R. P."Simulating Physics with Computers",International Journal of Theoretical Physics, Vol.21(6&7) pp. 467-488,1982.

[4] Grover, L. K. "Quantum Mechanism Helps in Searching for a Needle in a Haystack", Physical Revies Letters, 79, 325-8, 1997.

[5] https://en.wikipedia.org/wiki/Quantum computing

[6] Lijuan G., and Xingyi L., "Improvements of KNN algorithm for Big Data Classification", Applicatiion Research of Computers, Vol. 31(5), pp. 1342-1344,1373, 2014.

[7] Paul I. H., and Ioannis G. K., "Quantum Computers: Registers, Gates and Algorithms", Proc. 28th International Conference on Microelectronics (MIEL 2012), pp. 15-21, 2012.

[8] https://en.wikipedia.org/wiki/Quantum logic gate#Universal quantum gates

[9] Yi Z., Kai L. and Yinghui G., "Quantum Algorithms and QuantumInspired Algorithms", Chinese Journal of Computers, Vol. 36(9), pp.1835-1842, 2013.

[10] https://en.wikipedia.org/wiki/Bell state

[11] https://qiskit.org/documentation/overview.html