

# Strong Statistical Correlation Revealed by Quantum Entanglement for Supervised Learning

Junwei Zhang<sup>1</sup>, Yuexian Hou<sup>2</sup>, Zhao Li<sup>3</sup>, Long Zhang<sup>4</sup> and Xia Chen<sup>5</sup>

**Abstract.** In supervised learning, the generative approach is an important one, which obtains the generative model by learning the joint probability between features and categories. In quantum mechanics, Quantum Entanglement (QE) can provide a statistical correlation between subsystems (or attributes) that is stronger than what classical systems are able to produce. It inspires us to use entangled systems (states) to characterize this strong statistical correlation between features and categories, that is, to use the joint probability derived from QE to model the correlation. Based on the separability of the density matrix of entangled systems, this paper formally clarifies the manifestation of the strong statistical correlation revealed by QE, and implements a classification algorithm (called ECA) to verify the feasibility and superiority of the correlation in specific tasks. Since QE arises from the measurement process of entangled systems, the core of ECA is quantum measurement operations. In this paper, we use the GHZ [25] and W [22] states to prepare the entangled system and use a fully connected network layer to learn the measurement operator. It can also be understood as replacing the output layer of the Multi-Layer Perceptron (MLP) with a quantum measurement operation. The experimental results show that ECA is superior to most representative classification algorithms in multiple evaluation metrics.

## 1 Introduction

In machine learning, supervised learning is an important learning method, and is also the most widely used learning method in the industrial field. Therefore, continuous in-depth research on supervised learning methods will directly promote greater progress in industrial technology. An important approach in supervised learning is the generative one, which obtains the generative model by learning the joint probability between features (or attributes) and categories (or labels). In quantum mechanics, Quantum Entanglement (QE) is an important quantum resource and has no classical counterpart and therefore has been receiving continuous theoretical attention ever since the birth of quantum mechanics. QE plays very important roles in quantum information processing, such as Quantum Teleportation [5, 28], Dense Coding [34, 32], Error Correction [8, 12], Cryptography [24, 6], and Computation [17, 7].

Quantum mechanics theory has some characteristics that are not possessed by classical theory, such as observational behavior affects the observed system [10, 40] and the observations of the system are affected by the observed context [29]. These differences (or advantages) may help us improve and perfect the current classical models and algorithms.

QE can provide a statistical correlation between subsystems (or attributes) that is stronger than what classical systems are able to produce [48]. It inspires us to use quantum entangled systems (states) to characterize this strong statistical correlation between features and categories, that is, to use the joint probability derived from QE to model the correlation. This paper uses the joint probability derived from QE to implement a classification algorithm (called ECA) to learn the strong statistical correlation between features and categories and between features. In the existing work, QE is mostly related to quantum acceleration [35] or non-local correlation [37], but these topics are not discussed in this paper. This paper constructs a classification algorithm based on pure mathematical forms of QE and does not assume that between features and categories or between features have non-locality. In other words, we focus on the superiority of the mathematical form of quantum mechanics.

QE arises from the measurement process of entangled systems (states), so the core of ECA is quantum measurement operations. In this paper, we use the GHZ [25] and W [22] states to prepare the entangled system and use a fully connected network layer to learn the parameters of the measurement operator. It can also be understood as replacing the output layer of the Multi-Layer Perceptron (MLP) with a quantum measurement operation. Since ECA is simulated on classical computers, it is not able to verify the algorithm under the large datasets. The main difficulty in its existence is that the resources required by the algorithm will increase exponentially due to the tensor operation [35]. Experiments are performed on three classic machine learning datasets. The experimental results show that ECA is superior to most representative classification algorithms in multiple evaluation metrics.

The main contributions are listed as follows: 1) Based on the separability of the density matrix of entangled systems, this paper decomposes the joint probability derived from QE into two parts, the classical probability part and the quantum probability part, so that we can clarify the essential reason of the strong statistical correlation revealed by QE, that is, a quantum interference term for classical probability is added. 2) According to the analysis conclusions, this paper uses the joint probability derived from QE to implement a classification algorithm, and fully validates it on three machine learning datasets.

<sup>1</sup> School of Computer Science and Technology, Tianjin University, China, email: junwei@tju.edu.cn

<sup>2</sup> School of Computer Science and Technology, Tianjin University, China, email: yxhou@tju.edu.cn (Corresponding Author)

<sup>3</sup> Search Division, Alibaba Group, China, email: lizhao.lz@alibaba-inc.com (Corresponding Author)

<sup>4</sup> Search Division, Alibaba Group, China, email: james.zl@alibaba-inc.com

<sup>5</sup> Search Division, Alibaba Group, China, email: xia.cx@alibaba-inc.com

## 2 Related Work

Since the birth of statistics, new classification algorithms and their various improvement methods have emerged, which are widely used in all areas of life. We briefly mention some important classical classification algorithms, and mainly focus on the learning methods inspired by quantum mechanics.

In 1936, Fisher [23] proposed the famous Fisher discriminant analysis; in 1957, Rosenblatt [39] proposed the perceptron method; in 1958, Cox [14] proposed the logistic regression algorithm; in 1963, Vapnik et al. [47] proposed the support vector machine algorithm; in 1967, Cover et al. [13] proposed the nearest neighbor classification algorithm. See also [27, 41, 45].

Due to the limitation of the computing power and popularity of quantum computers, a small number of researchers can implement quantum algorithms on quantum computers, but most of them use the simulation method on classical computers to complete the research. Considering the advantages of quantum computer processing high-dimensional vectors in large tensor product space, Lloyd et al. [33] proposed a supervised and unsupervised quantum machine learning algorithm for cluster assignment and cluster finding. This algorithm reduces the time complexity of the classical algorithm from the polynomial-time to the logarithmic-time. Similarly, Rebentrost et al. [38] demonstrated that support vector machines implemented on quantum computers can reduce the time complexity from the polynomial-time to the exponential-time for classical sampling tasks. From the perspective of quantum information, Dunjko et al. [21] proposed an approach for the systematic treatment of machine learning. See also [46, 9, 36, 2, 16].

Levine et al. [31] established a contemporary deep learning architecture that effectively represents highly quantum entangled systems in the form of deep convolutional and recursive networks. By constructing tensor network equivalents of these architectures, they identify the inherent reuse of information in the network operation as a key trait that distinguishes them from standard tensor network-based representations, and that enhances their entanglement capacity. Schuld et al. [43] interpreted the process of encoding inputs in a quantum state as a nonlinear feature map, mapping the data into a quantum Hilbert space. According to the theory, two quantum classification models are established, which use the variational quantum circuit as a linear model to explicitly classify the data in Hilbert space. The quantum device estimates inner products of quantum states to compute a classically intractable kernel. See also [11, 20, 19, 1].

## 3 Theoretical Analysis and Verification

In order to have an intuitive understanding of the manifestation and action mechanism of the strong statistical correlation revealed by QE, we use the separability of the density matrix of entangled systems to decompose the joint probability derived from QE into classical probability part and quantum probability part (called the quantum interference term) to clarify the essential reason for the strong statistical correlation.

In 1927, von Neumann and Landau independently proposed the concept of density matrix, and finally applied this concept to realize the quantization of statistical thermodynamics. Due to the superposition of states in quantum mechanics, non-diagonal elements of the density matrix appear, and it describes the coherent correlation between states. In fact, the decoherence process of the entangled system is the process of gradually diagonalizing the density matrix. Because

the diagonal elements of the density matrix represent classic probability statistics between states, the off-diagonal elements represent coherent correlation between states [15, 49]. The idea of treating the diagonal part of the density matrix as a classical statistical description of states and the non-diagonal part as a coherence description between states has been used in Ref. [44, 50].

Let us begin with an arbitrary two-qubit entangled state in the bases  $\sigma_z|\pm\rangle = \pm|\pm\rangle$ <sup>6</sup> that

$$|\psi\rangle = \alpha|++\rangle + \beta|--\rangle \quad (1)$$

where  $\alpha$  and  $\beta$  are the normalization condition  $|\alpha|^2 + |\beta|^2 = 1$  but  $\alpha, \beta \neq 0$ . Without losing generality,  $\alpha$  and  $\beta$  can be parameterized as  $\alpha = e^{i\eta} \sin \xi$ ,  $\beta = e^{-i\eta} \cos \xi$ , but  $\sin \xi, \cos \xi \neq 0$ , where  $\eta$  and  $\xi$  are two real parameters and  $i$  is the imaginary number with  $i^2 = -1$ .

The density matrix of the entangled state,  $\rho_{all} = |\psi\rangle\langle\psi|$ , can be separated to two parts, i.e.,

$$\rho_{all} = \rho_c + \rho_q, \quad (2)$$

in order to see the quantum correlation explicitly. The first part is

$$\rho_c = \sin^2 \xi |++\rangle\langle ++| + \cos^2 \xi |--\rangle\langle --|, \quad (3)$$

which describes two qubits obeying the classical (or local) statistics. The other part is

$$\rho_q = \sin \xi \cos \xi \left( e^{i2\eta} |++\rangle\langle --| + e^{-i2\eta} |--\rangle\langle ++| \right), \quad (4)$$

which describes the quantum (or non-local) correlation between two qubits.

### 3.1 Measuring Density Matrix

From the definition of projective measurement in quantum mechanics theory [35], a projective measurement is described by an observable,  $M$ , a Hermitian operator on the state space of the system being observed. The observable has a spectral decomposition,

$$M = \sum_m m M^m \quad (5)$$

where  $M^m$  is the measurement operator (or projector) onto the eigenspace of  $M$  with eigenvalue  $m$ . The possible outcomes of the measurement correspond to the eigenvalues,  $m$ , of the observable. Upon measuring the state  $|\psi\rangle$ , the probability of getting result  $m$  is given by

$$P(m) = \langle\psi|M^m|\psi\rangle = Tr[M^m(|\psi\rangle\langle\psi|)] \quad (6)$$

where  $Tr$  denotes a matrix function, that is, the trace of the matrix.

Suppose that the observables of the two subsystems of the entangled system  $|\psi\rangle$  are  $a$  and  $b$ , respectively, and they are defined as:

$$M_r = \begin{bmatrix} \cos \theta_r & e^{-i\phi_r} \sin \theta_r \\ e^{i\phi_r} \sin \theta_r & -\cos \theta_r \end{bmatrix} \quad (7)$$

where  $\theta$  and  $\phi$  are two arbitrary real parameters, and  $r \in \{a, b\}$ . Its spectral decomposition is  $M_r = M_r^+ - M_r^-$ . The probability of the combination is obtained as:

$$P_{all}(+a, +b) = Tr[(M_a^+ \otimes M_b^+) \rho_{all}], \quad (8)$$

<sup>6</sup> We here adopt the widely used Dirac notations, in which a unit vector  $\vec{v}$  and its transpose  $\vec{v}^T$  are denoted as a ket  $|v\rangle$  and a bra  $\langle v|$ , respectively.  $\sigma_z$  denotes Pauli matrix, and Pauli matrix refers to four common matrices, which are  $2 \times 2$  matrix, each with its own mark, namely  $\sigma_x \equiv X$ ,  $\sigma_y \equiv Y$ ,  $\sigma_z \equiv Z$  and  $I$ .

which can be separated to classical (local) probability part and quantum (non-local) probability part (called a quantum interference term)

$$P_{all}(+a, +b) = Tr[(M_a^+ \otimes M_b^+)(\rho_c + \rho_q)] \quad (9)$$

$$= Tr[(M_a^+ \otimes M_b^+)\rho_c] + Tr[(M_a^+ \otimes M_b^+)\rho_q] \quad (10)$$

$$= P_c(+a, +b) + P_q(+a, +b) \quad (11)$$

where  $\otimes$  denotes the tensor (Kronecker) product.

Accordingly, the probability of other combinations, i.e.,  $P_c(\pm a, \pm b)$ ,  $P_c(\mp a, \pm b)$ ,  $P_q(\pm a, \pm b)$  and  $P_q(\mp a, \pm b)$ , can also be obtained. Then under the outcome-independent base vectors  $|+a, +b\rangle$ ,  $|+a, -b\rangle$ ,  $|-a, +b\rangle$  and  $|-a, -b\rangle$ , the average values of  $a$  and  $b$  in the classical and quantum cases are

$$\langle ab \rangle_c = P_c(+a, +b) - P_c(+a, -b) - P_c(-a, +b) + P_c(-a, -b) \quad (12)$$

$$= \cos \theta_a \cos \theta_b \quad (13)$$

and

$$\langle ab \rangle_q = P_q(+a, +b) - P_q(+a, -b) - P_q(-a, +b) + P_q(-a, -b) \quad (14)$$

$$= \sin \theta_a \sin \theta_b \sin 2\xi \cos(\phi_a + \phi_b + 2\eta), \quad (15)$$

respectively, and  $\langle \cdot \rangle$  is the quantum mechanical symbol of the average value.

### 3.2 Main Conclusions and Critical Value

The theoretical tool for verifying QE is the Bell inequality [3]. Violating (or Destroying) Bell inequality is a sufficient condition for the existence of QE, that is, it belongs to the category of quantum probability. Bell inequality has many well-known promotion forms, the first of which is the Clauser-Horne-Shimony-Holt (CHSH) inequality [30]. The form of the CHSH inequality is simpler and more symmetrical than many other Bell inequalities that are later proposed. The specific form of the CHSH inequality is

$$|E(QS) + E(RS) + E(RT) - E(QT)| \leq 2 \quad (16)$$

where  $E(\cdot)$  denotes the average value, and  $Q$ ,  $R$ ,  $S$  and  $T$  denote observables.

Based on the CHSH inequality, we can confirm the category of each part of the joint probability derived from QE, and we can draw the following conclusion:

**Conclusion 1.**  $P_c(\cdot)$  belongs to the category of classical probability.

**Conclusion 2.**  $P_q(\cdot)$  belongs to the category of quantum probability, so it can be called a quantum interference term for classical probability.

**Conclusion 3.**  $P_{all}(\cdot) = P_c(\cdot) + P_q(\cdot)$  belongs to the category of quantum probability.

Moreover, based on the above analysis, we can also obtain the critical value of the correlation between the subsystems of the entangled system when the CHSH inequality is broken, that is, the values of the off-diagonal elements of the density matrix when the CHSH inequality is broken. From this, we can define a density matrix  $\rho$  with off-diagonal elements as the independent variable  $0 \leq x \leq \frac{1}{2}$ , i.e.,

$$\rho = \rho_c + \rho_q \quad (17)$$

with

$$\rho_c = \frac{1}{2}(|++\rangle\langle++| + |--\rangle\langle--|) \quad (18)$$

and

$$\rho_q = x(|++\rangle\langle--| + |--\rangle\langle++|). \quad (19)$$

Calculated to get

$$\langle ab \rangle_c = \cos \theta_a \cos \theta_b \quad (20)$$

and

$$\langle ab \rangle_q = 2x \sin \theta_a \sin \theta_b \cos(\phi_a + \phi_b). \quad (21)$$

From the numerical analysis, when  $\theta_Q = 0$ ,  $\theta_r = \frac{\pi}{2}$ ,  $\theta_S = \frac{5\pi}{4}$ ,  $\theta_T = \frac{7\pi}{4}$  and  $\phi_Q = \phi_R = \phi_S = \phi_T = 0$ ,

$$|\langle QS \rangle_{all} + \langle RS \rangle_{all} + \langle RT \rangle_{all} - \langle QT \rangle_{all}| = \sqrt{2}(2x + 1). \quad (22)$$

Therefore, when  $x \approx 0.207$ , it is the critical value of off-diagonal elements when the CHSH inequality is broken.

### 3.3 Discussions

The above analysis shows that the joint probability derived from QE can violate the Bell inequality, that is, it belongs to the category of quantum probability. Moreover, the joint probability can be divided into the classical probability part and the quantum probability part (called the quantum interference term), i.e.,  $P_{all}(\cdot) = P_c(\cdot) + P_q(\cdot)$ , and the quantum interference term can also violate the bell inequality.

From Eqs. 13 and 15, we can get the following conclusions: 1) The phase of the entangled state, namely  $e^{\pm i\eta}$ , is independent of the classical probability part, but together with the azimuth angles  $\phi$  of the observables determine the quantum interference term. 2) When the quantum interference term is not zero, the azimuth angles  $\phi$  of the observables again describe the correlation between observables in a manner different from the polar angles  $\theta$ . It can be seen that because quantum mechanics uses complex numbers, it is possible to describe a strong statistical correlation between observables. 3) It can be seen from the quantum interference term that when the quantum interference term is not zero, the interference term contains all parameters including the observables and the entangled system. The form of quantum interference terms can explain many (strange) phenomena in the quantum field, such as the observables of the system are affected by the observed context [29].

The significance of the section is as follows:

- The joint probability derived from QE is given;
- The classical probability and quantum probability parts are clearly divided, and the quantum probability part can be called the quantum interference term for classical probability;
- The quantum interference term is formalized, and the components and their relationships are clarified.

### 4 The Classification Algorithm with QE

It can be known from quantum mechanics theory that QE occurs in the measurement process of entangled systems, so the overall structure of ECA will unfold around the measurement process of entangled systems. For the entangled system to be measured, it can be known from the analysis of the theoretical section that both the probability amplitude and phase information of the entangled system act on the quantum interference term, and its effect can be covered by the azimuth angle of the observables. Therefore, we use certain entangled states to represent the measured system, such as the GHZ and W states. For the observable of the subsystems, it can be known from the analysis of the theoretical section that the polar angle of the observable is related to both the classical probability term and the

quantum interference term, and the azimuth angle is only related to the quantum interference term. Therefore, for targeted learning the strong statistical correlation, We will mainly learn the azimuth angle related only to the quantum interference term, that is, the azimuth angle is learned from the features and categories, and the polar angle is expressed as a learnable weight variable, that is, the degree of freedom.

The following sections will describe in detail the construction method of both the entangled system and its observable, the model structure of ECA, and the parameter learning process of ECA.

#### 4.1 Define Entangled System

We first give formal definitions of the GHZ and W states and use them to construct the entangled system of ECA.

**Definition 1.** *The Greenberger-Horne-Zeilinger state (GHZ state) is a entangled state that involves at least three subsystems (particles). It was first studied by Daniel Greenberger, Michael Horne and Anton Zeilinger in 1989 [25]. In the case of each of the subsystems being two-dimensional, i.e., qubits, it is*

$$|GHZ\rangle = \frac{1}{\sqrt{2}} (|0\rangle^{\otimes N} + |1\rangle^{\otimes N}) \quad (23)$$

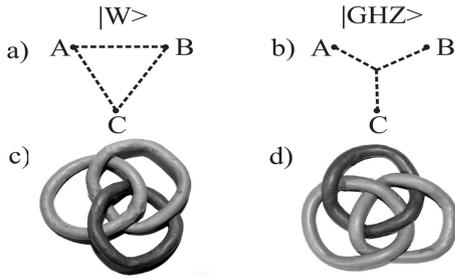
where  $N$  is the number of qubits. The GHZ state is a maximally quantum entangled state.

**Definition 2.** *The W state [22] is an entangled state of three qubits, which has the form*

$$|W\rangle = \frac{1}{\sqrt{3}} (|001\rangle + |010\rangle + |100\rangle). \quad (24)$$

The notion of the W state has been generalized for  $N$  qubits, which has the form

$$|W\rangle = \frac{1}{\sqrt{N}} (|10\dots 0\rangle + |01\dots 0\rangle + \dots + |00\dots 1\rangle). \quad (25)$$



**Figure 1.** Graphical description of the entanglement relationship of the W and GHZ states. This figure is referenced from Ref. [4]. a) and b) represent the dot-line structure diagrams of the W and GHZ states, respectively, while c) and d) are visualized to give an example diagram.

There has been a lot of discussion about the entanglement relationship and strength of the W and GHZ states, and we will not repeat it here. In order to enable the reader to have a preliminary understanding of the W and GHZ states, Fig. 1 given by Ref. [4] is cited here.

#### 4.2 Define Observable of Entangled System

In quantum mechanics, the properties of the system are called observables and are represented by Hermitian operators (matrices), e.g., Eq. 7. Therefore, for the classification task, we can respectively represent features and categories with observables, e.g.,  $M_{fea}$  and  $M_{cat}$ . Because of this, ECA can only be used to solve the two-class (binary) classification task. Of course, the method of adding the number of qubits of the categories can be used to solve the multi-class classification task, but this is not the focus of this paper. Spectral decomposition of the observables  $M_r$  yields two measurement operators,

$$M_r(\theta_r, \phi_r) = M_r^+(\theta_r, \phi_r) - M_r^-(\theta_r, \phi_r), \quad (26)$$

that is,  $M_r^+(\theta_r, \phi_r)$  and  $M_r^-(\theta_r, \phi_r)$ .

In this paper, we use the azimuth angle  $\phi$  to represent the feature  $k$  of the instance, and the polar angle  $\theta$  to represent the degrees of freedom of  $k$ , that is, a learnable weight variable, so the measurement operators of  $k$  are  $M_k^+(\theta_k, \phi_k)$  and  $M_k^-(\theta_k, \phi_k)$ , respectively. In this paper, we define the polar angle as a fixed value, that is,  $\theta = \frac{\pi}{2}$ , in order to reduce the number of learning parameters. Of course, there are advantages to using learnable variables, but the learning difficulty increases accordingly. We use the determined measurement operators to represent the categories of the instance,

$$M_{cat}^+ = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, M_{cat}^- = \frac{1}{2} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad (27)$$

that is,  $\theta = \frac{\pi}{2}$  and  $\phi = 0$ . In fact, any set of eigenstates of the observable can be used to represent the categories, as long as they meet the orthogonality. The reason why we need to select a set of orthogonal bases to represent the categories is because the positive and negative examples of the two-class classification task are often binary opposition.

Now we can formally define the measurement operator for the entangled system. Assuming that each instance has  $N$  features and one category, the measurement operator of the entangled system can be defined as

$$\mathcal{M}^+(\theta, \phi) = \bigotimes_{k=1}^N M_k^+(\theta_k, \phi_k) \otimes M_{cat}^+ \quad (28)$$

and

$$\mathcal{M}^-(\theta, \phi) = \bigotimes_{k=1}^N M_k^+(\theta_k, \phi_k) \otimes M_{cat}^-. \quad (29)$$

Here we can also use  $M_k^-$  to instead of the  $M_k^+$  in the above formula, Eqs. 28 and 29, they have the same effect.

Applying  $\mathcal{M}^+$  and  $\mathcal{M}^-$  separately to the entangled system, e.g., the GHZ state, yields probability values that are the positive and negative examples,

$$P^+ = \langle GHZ | \mathcal{M}^+ | GHZ \rangle \quad (30)$$

and

$$P^- = \langle GHZ | \mathcal{M}^- | GHZ \rangle. \quad (31)$$

#### 4.3 Model Structure of ECA

The above sections have given the core parts of ECA, namely Eqs. 30 and 31. For the feature-related parameter  $\phi_r$ , we use a fully connected network layer to learn. Formally, we can also think of ECA

as the MLP that replaces its output layer with the quantum measurement process. Moreover, in order to reproduce the scene required to generate quantum entanglement, we train three sets of measurement operators at the same time to obtain the results of different measurement situations. In order to make the reader understand the overall structure of ECA more clearly, here we use the pseudo-code to display, that is, Algorithm 1.

---

**Algorithm 1** Predicts Data by ECA

---

**Input:**  $x \in \mathbb{R}^N$   
**Output:**  $y \in \{+1, -1\}$   
1: Define  $w_s \in \mathbb{R}^{N \times N}$  and  $b_s, v_s, \theta_s \in \mathbb{R}^N$ ,  $s = 1, 2, 3$ , the value of which is determined by the training process;  
2:  $|GHZ\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes(N+1)} + |1\rangle^{\otimes(N+1)})$ ;  
3:  $\phi_s = ReLU(w_s * x + b_s) * v_s$ ,  $s = 1, 2, 3$ ;  
4:  $P_s^+ = \langle GHZ | \mathcal{M}^+(\theta_s, \phi_s) | GHZ \rangle$ ,  $s = 1, 2, 3$ ;  
5:  $P_s^- = \langle GHZ | \mathcal{M}^-(\theta_s, \phi_s) | GHZ \rangle$ ,  $s = 1, 2, 3$ ;  
6: **if**  $\frac{1}{3} \sum_s P_s^+ > \frac{1}{3} \sum_s P_s^-$  **then**  
7:     return +1;  
8: **else**  
9:     return -1;  
10: **end if**

---

#### 4.4 Parameter Learning Process of ECA

Machine learning often uses the loss function to evaluate the pros and cons of the model, and also uses the loss function to improve the performance of the model. This process of improvement is called optimization. ECA uses the classical cross-entropy loss function,

$$H(p, q) = - \sum_x p(x) \log q(x), \quad (32)$$

to act on its loss function. Since Adam (adaptive moment estimation) is a clear range of learning rates per iteration, making the parameters change smoothly, we use Adam as the optimizer for ECA. In fact, Adam is also the most commonly used optimizer in mainstream learning algorithms.

The pseudo-code of the training process of ECA is shown in Algorithm 2.

## 5 Experiments

To evaluate the effect of ECA, we conduct two groups of comparative experiments which are compared with standard MLP and compared with most representative classification algorithms. Moreover, to verify the impact of different entangled states on ECA, e.g., the GHZ and W states, we also conduct the comparative experiment.

### 5.1 Datasets and Evaluation Metrics

The experiments are conducted on three most frequently used and most popular machine learning datasets from UCI [18], i.e., Abalone<sup>7</sup>, Wine Quality<sup>8</sup> (Red) and Wine Quality<sup>8</sup> (White). The statistics of each dataset are given in Tab. 1. Since the performance of ECA will be verified under the two-class classification task, the above multi-class datasets need to be adjusted to meet the requirements of the task.

<sup>7</sup> <http://archive.ics.uci.edu/ml/datasets/Abalone>

<sup>8</sup> <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

---

**Algorithm 2** Training Process of ECA

---

**Input:** Training Set  $\mathcal{D}$   
 $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^N, y_i \in \{[0, 1], [1, 0]\}\}_{i=1}^M$ .  
1: Initialise  $w_s \in \mathbb{R}^{N \times N}$  and  $b_s, v_s, \theta_s \in \mathbb{R}^N$ ,  $s = 1, 2, 3$ , the parameters that need to be learned;  
2:  $|GHZ\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes(N+1)} + |1\rangle^{\otimes(N+1)})$ ;  
3: **repeat**  
4:     **for each**  $(x_i, y_i)$  in  $\mathcal{D}$  **do**  
5:          $\phi_s = ReLU(w_s * x_i + b_s) * v_s$ ,  $s = 1, 2, 3$ ;  
6:          $P_s^+ = \langle GHZ | \mathcal{M}^+(\theta_s, \phi_s) | GHZ \rangle$ ,  $s = 1, 2, 3$ ;  
7:          $P_s^- = \langle GHZ | \mathcal{M}^-(\theta_s, \phi_s) | GHZ \rangle$ ,  $s = 1, 2, 3$ ;  
8:         **if**  $\frac{1}{3} \sum_s P_s^+ > \frac{1}{3} \sum_s P_s^-$  **then**  
9:              $P_i = [0, 1]$ ;  
10:         **else**  
11:              $P_i = [1, 0]$ ;  
12:         **end if**  
13:         Calculate the cross-entropy of  $P_i$  and  $y_i$ , i.e. Eq. 32; Use the optimizer *Adam* to minimize  $H(P_i, y_i)$  and update parameters.  
14:     **end for**  
15: **until** Epochs

---

**Abalone** is a dataset that predicts the age of abalone. We divide the age less than 10 into one class, and the others into another. The reason for dividing the datasets in this way is that the number of instances (samples) in two categories can be as close as possible, making the data unbiased.

**Wine Quality (WQ)** is a dataset that scores on wine quality. The score is between 0 and 10. We divide the scores less than or equal to 5 into one class, and the others into another.

All experiments use the 5-fold Cross-Validation method to divide the training set and test set. The experimental evaluation metrics, *F1-score*, *Accuracy* and *AUC* (Area Under Curve), are taken as the average of 5 results.

**Table 1.** Dataset Statistics: For each dataset, the number of instances (samples) and the number of attributes are given in the table. The number of positive and negative samples is shown in parentheses.

Dataset	Instances	Attributes
Abalone	4177 (2096+2081)	8
Wine Quality (Red)	1599 (855 + 744)	11
Wine Quality (White)	4898 (3258+1640)	11

### 5.2 Compared with Standard MLP

Since ECA uses a fully connected network layer to learn the parameters of the measurement operator, which is similar in form to a standard MLP, we first compare the effects of ECA and MLP under the same settings. Both ECA and MLP use *Adam* as an optimizer for the model and set up similar structures, that is, they have the same number of parameters.

#### 5.2.1 Baselines

The MLP has an input layer, a hidden layer and an output layer. The input layer has  $N$  input neurons (nodes), where  $N$  is the number of features (attributes) of the instance; the hidden layer also has  $N$  neurons and its activation function is *ReLU* (Rectified Linear Unit); the output layer has one neuron and its activation function is *Sigmoid*.

Like ECA, the optimizer *Adam* is used to learn the parameters of the MLP.

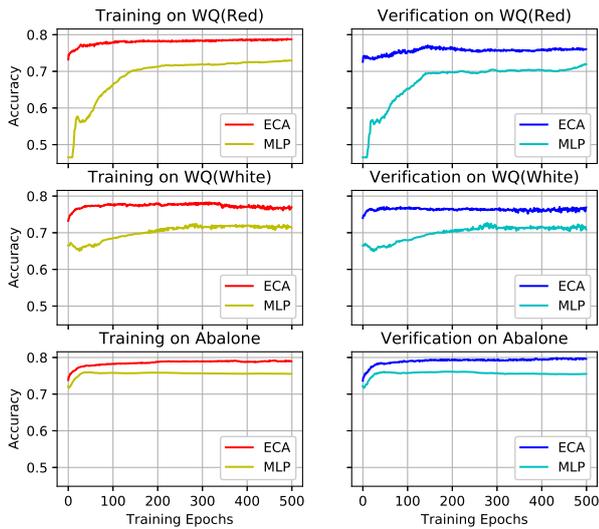
### 5.2.2 Hyper-Parameter Settings

ECA has three hyper-parameters, which are *learning rate*, *mini-batch* and *training epoch*, respectively. ECA uses the same settings on three datasets, i.e., the *learning rate* is 0.0001, the *mini-batch* is 1 and the *training epoch* is 500. The entangled system uses the GHZ state.

The MLP also has three hyper-parameters, which are identical to ECA. The MLP uses the same settings on three datasets, i.e., the *learning rate* is 0.0001, the *mini-batch* is 1 and the *training epoch* is 500. Their *weights* are initialized to a truncated positive distribution, and the *biases* are initialized to 0.01. In this section, we use the test set divided by the Cross-Validation method as the validation set.

### 5.2.3 Experiment Results

Fig. 2 exhibits the training and verification accuracy of both ECA and MLP on datasets Abalone, WQ (Red) and WQ (White). The experimental results basically verifies the effectiveness of ECA, and it also shows that ECA has a significant improvement in the experimental effect compared to its auxiliary model, namely MLP.



**Figure 2.** Under the three datasets Abalone, WQ (Red) and WQ (White), the accuracy curves of ECA and the MLP on the training set (left column) and verification set (right column), respectively.

## 5.3 Compared with Classical Algorithms

### 5.3.1 Baselines

We have selected some representative classification algorithms for comparison experiments, including Logistic Regressive (LR), Naive Bayesian Model (NBM), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Multi-Layer Perceptron (MLP).

### 5.3.2 Hyper-Parameter Settings

The hyper-parameters of ECA are exactly the same as those of Experiment 5.2. The entangled system uses the GHZ state.

The hyper-parameters in the baselines are set to: in LR, ‘penalty’ is *L2*; in SVM, ‘C’ is 1.0 and ‘kernel’ is *rbf*; in KNN, ‘n-neighbors’ is 5; in LDA, ‘solver’ is *svd*; in MLP, ‘activation’ is *relu*, ‘solver’ is *adam* and ‘alpha’ is 0.0001. Other hyper-parameters not listed use the default value of the framework scikit-learn<sup>9</sup>.

### 5.3.3 Experiment Results

Tab. 2 exhibits the experiment results on datasets Abalone, WQ (Red) and WQ (White) respectively, where bold values are the best performances out of all algorithms. From the experimental results, three metrics of ECA on three datasets are significantly better than the majority of other algorithms. It verifies the effectiveness of ECA from a broader perspective. The experimental results in this section can basically show that the joint probability derived from quantum entanglement has certain effectiveness and superiority in classical classification tasks, and also show that the strong statistical correlation revealed by quantum entanglement has certain utility and contribution in specific tasks.

**Table 2.** Experiment results on Abalone, WQ (Red) and WQ (White) Dataset. The best-performed values for each dataset are in bold.

Dataset	Algorithm	F1-score	Accuracy	AUC
Abalone	LR	0.7690	0.7699	0.7708
	NBM	0.7392	0.7349	0.7345
	KNN	0.7747	0.7749	0.7842
	SVM	0.7670	0.7536	0.7542
	LDA	0.7769	0.7792	0.7601
	QDA	0.7429	0.7593	0.7558
	MLP	0.7617	0.7693	0.7584
ECA	<b>0.7971</b>	<b>0.7957</b>	<b>0.7958</b>	
WQ (Red)	LR	0.7568	0.7217	0.7404
	NBM	0.7361	0.7210	0.7292
	KNN	0.6718	0.6497	0.6428
	SVM	0.7281	0.7098	0.7095
	LDA	0.7506	0.7367	0.7224
	QDA	0.7430	0.7135	0.7146
	MLP	0.6930	0.7141	0.7438
ECA	<b>0.7898</b>	<b>0.7810</b>	<b>0.7818</b>	
WQ (White)	LR	0.8233	0.7380	0.6797
	NBM	0.7807	0.7066	0.6683
	KNN	0.7805	0.6939	0.6407
	SVM	0.8256	0.7472	0.6731
	LDA	0.8261	0.7527	0.6863
	QDA	0.8172	0.7452	0.6953
	MLP	0.7976	0.7057	0.6918
ECA	<b>0.8287</b>	<b>0.7672</b>	<b>0.7271</b>	

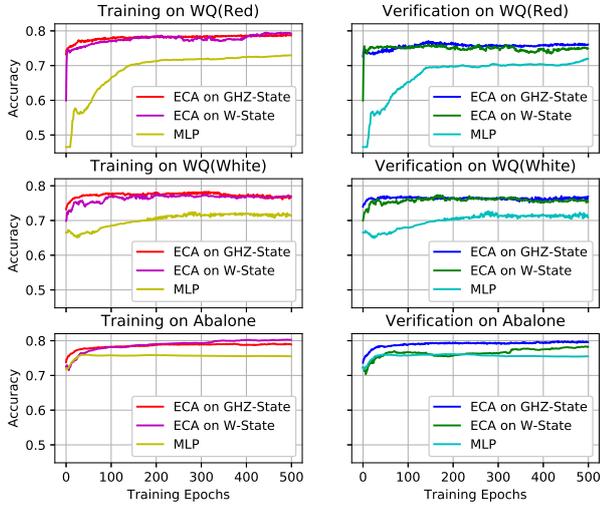
## 5.4 Comparison of Different Entangled Systems

To verify the impact of different entangled systems on the experimental results, we select the GHZ and W states respectively for the comparative experiment. The hyper-parameter setting of ECA in the experiment was exactly the same as that of Experiment 5.2.

The experimental results are exhibited in Fig. 3. It can be seen from the experimental results that the influence of different entanglement systems on the experimental results is not obvious. If the

<sup>9</sup> <https://scikit-learn.org/stable/index.html>

entangled system is a general superposition state, that is, there is no entanglement, ECA will not have the ability to learn. It also proves that ECA has the ability to learn entanglement features.



**Figure 3.** Under the three datasets Abalone, WQ (Red) and WQ (White), the accuracy curves of ECA on the GHZ and W states, respectively. The left column is the result of the training set, and the right column is the result of the validation set.

## 6 Discussions

In this section, we mainly discuss three questions (**Q**) and give the answers (**A**) from the author.

**Q:** Can quantum mechanics theory be applied to the field of machine learning (or pattern recognition)?

**A:** Although quantum mechanics theory is generally regarded as microphysical theory, its connotation is about information rather than physics. Since Hardy [42, 26], the information nature of quantum mechanics has been increasingly clarified: it can be proved that quantum mechanics can be formally derived from 4-5 general information processing axioms which are conceptually natural and technically concise. Moreover, if a specific limitation is imposed on the set of information processing axioms derived from quantum mechanics, a special case of quantum mechanics, namely classical probability theory, can be derived. Therefore, the law of quantum mechanics should not only be regarded as the law of the microphysical world, but should be regarded as the general law of information processing.

**Q:** Can classical computers simulate (or reproduce) QE?

**A:** Although it has experienced rapid development in recent years, quantum computer technology is still in its infancy compared with mature classic computer technology. Recently, Google’s paper published in Nature magazine announced the realization of quantum hegemony, but it is only a prototype, and it will take some time to achieve the goal of serving general scientific research. Recurring QE or performing quantum computation on a classical computer requires more computational power than quantum computers to construct associations and constraints between computational units, but it does not indicate that classical computers cannot simulate QE.

**Q:** How to prove the existence of QE in the classification algorithm?

**A:** The study of QE is still in its infancy, many of the theories are still not perfect, and many phenomena are still unclear. The verification of QE at this stage can only be proved by the violation of the inequality proposed by the classical theory, that is, Bell inequality. However, it is a sufficient but unnecessary condition to prove the existence of QE by violating Bell inequality. For this difficulty, this paper can only be used to construct the classification algorithm from the formalization of QE that violates Bell inequality, and prove that the algorithm’s infrastructure can violate Bell inequality.

## 7 Summary and Outlook

Quantum entanglement is an important quantum resource and has no classical counterpart and therefore has been receiving continuously theoretical attention ever since the birth of quantum mechanics. In this paper, in order to clarify the manifestation and action mechanism of the strong statistical correlation revealed by quantum entanglement, we quantify it as a quantum interference term for classical probability based on the separability of the density matrix of entangled system. In fact, some other quantum phenomena in quantum mechanics have taken the same approach. Moreover, we use the formal framework of quantum entanglement to construct a classification algorithm (called ECA) to verify the role of strong statistical correlation in classical classification tasks. Due to the limitations of simulating quantum algorithms on classical computers, ECA can only be tested on some lightweight machine learning datasets. The experimental results show that ECA is superior to most representative classification algorithms in multiple evaluation metrics.

The research in this paper has achieved some meaningful results, but there are also some shortcomings. We list the shortcomings of the algorithm to facilitate colleagues to study together. 1) Need to propose better optimization algorithms and optimizers in the complex domain. 2) Further clarify the physical meaning of the quantum interference term or further theoretical analysis. 3) Excavate the applicable scene of the entanglement phenomenon and test the utility of the quantum interference term.

## ACKNOWLEDGEMENTS

This work is funded in part of the National Key R&D Program of China (2017YEF0111900), the National Natural Science Foundation of China (61876129), the National Natural Science Foundation of China (Key Program, U1636203), the Alibaba Innovation Research Foundation 2017 and the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321. Part of the work was performed when Junwei Zhang visited the Alibaba Inc. in 2018.

## REFERENCES

- [1] Jeremy Adcock, Euan Allen, Matthew Day, Stefan Frick, Janna Hinchliff, Mack Johnson, Sam Morley-Short, Sam Pallister, Alasdair Price, and Stasja Stanisic, 'Advances in quantum machine learning', *arXiv preprint arXiv:1512.02900*, (2015).
- [2] Esma Aïmeur, Gilles Brassard, and Sébastien Gambs, 'Quantum speedup for unsupervised learning', *Machine Learning*, **90**(2), 261–287, (2013).
- [3] John S Bell, 'On the einstein podolsky rosen paradox', *Physica Physique Fizika*, **1**(3), 195, (1964).
- [4] Ingemar Bengtsson and Karol Życzkowski, *Geometry of quantum states. An introduction to quantum entanglement*, 2008.
- [5] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters, 'Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels', *Phys. Rev. Lett.*, **70**, 1895–1899, (Mar 1993).
- [6] Charles H Bennett, Gilles Brassard, and Artur K Ekert, 'Quantum cryptography', *Scientific American*, **267**(4), 50–57, (1992).
- [7] Charles H Bennett and David P DiVincenzo, 'Quantum information and computation', *Nature*, **404**(6775), 247, (2000).
- [8] Charles H Bennett, David P DiVincenzo, John A Smolin, and William K Wootters, 'Mixed-state entanglement and quantum error correction', *Physical Review A*, **54**(5), 3824, (1996).
- [9] Arit Kumar Bishwas, Ashish Mani, and Vasile Palade, 'Big data quantum support vector clustering', *arXiv preprint arXiv:1804.10905*, (2018).
- [10] Niels Bohr et al., *The quantum postulate and the recent development of atomic theory*, volume 3, Printed in Great Britain by R. & R. Clarke, Limited, 1928.
- [11] Iris Cong, Soonwon Choi, and Mikhail D Lukin, 'Quantum convolutional neural networks', *Nature Physics*, **15**(12), 1273–1278, (2019).
- [12] David G Cory, MD Price, W Maas, E Knill, Raymond Laflamme, Wojciech H Zurek, Timothy F Havel, and SS Somaroo, 'Experimental quantum error correction', *Physical Review Letters*, **81**(10), 2152, (1998).
- [13] Thomas Cover and Peter Hart, 'Nearest neighbor pattern classification', *IEEE transactions on information theory*, **13**(1), 21–27, (1967).
- [14] D. R. Cox, 'The regression analysis of binary sequences', *Journal of the Royal Statistical Society Series B (Methodological)*, **21**(1), 238–238.
- [15] Andreas De Vries, *Quantum computation: An introduction for engineers and computer scientists*, BoD–Books on Demand, 2012.
- [16] Vasil S Denchev, Nan Ding, SVN Vishwanathan, and Hartmut Neven, 'Robust classification with adiabatic quantum optimization', *arXiv preprint arXiv:1205.1148*, (2012).
- [17] David P DiVincenzo, 'Quantum computation', *Science*, **270**(5234), 255–261, (1995).
- [18] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [19] Vedran Dunjko and Hans J Briegel, 'Machine learning & artificial intelligence in the quantum domain', *arXiv preprint arXiv:1709.02779*, (2017).
- [20] Vedran Dunjko and Hans J Briegel, 'Machine learning & artificial intelligence in the quantum domain: a review of recent progress', *Reports on Progress in Physics*, **81**(7), 074001, (2018).
- [21] Vedran Dunjko, Jacob M Taylor, and Hans J Briegel, 'Quantum-enhanced machine learning', *Physical review letters*, **117**(13), 130501, (2016).
- [22] W. Dür, G. Vidal, and J. I. Cirac, 'Three qubits can be entangled in two inequivalent ways', *Phys. Rev. A*, **62**, 062314, (Nov 2000).
- [23] Ronald A. Fisher, 'The use of multiple measurements in taxonomic problems', *Annals of Human Genetics*, **7**(7), 179–188, (1936).
- [24] Nicolas Gisin, Grégoire Ribordy, Wolfgang Tittel, and Hugo Zbinden, 'Quantum cryptography', *Reviews of modern physics*, **74**(1), 145, (2002).
- [25] Daniel M. Greenberger, Michael A. Horne, and Anton Zeilinger, *Going Beyond Bell's Theorem*, 69–72, Springer Netherlands, Dordrecht, 1989.
- [26] Lucien Hardy, 'Quantum theory from five reasonable axioms', *Physica*, (2001).
- [27] Tin Kam Ho, 'Random decision forests', in *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, (1995).
- [28] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki, 'General teleportation channel, singlet fraction, and quasidistillation', *Phys. Rev. A*, **60**, 1888–1898, (Sep 1999).
- [29] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki, 'Quantum entanglement', *Reviews of modern physics*, **81**(2), 865, (2009).
- [30] Andrei Khrennikov, 'Chsh inequality: Quantum probabilities as classical conditional probabilities.', *Foundations of Physics*, **45**(7), 1–15, (2014).
- [31] Yoav Levine, Or Sharir, Nadav Cohen, and Amnon Shashua, 'Quantum entanglement in deep learning architectures', *Physical review letters*, **122**(6), 065301, (2019).
- [32] Xiaoying Li, Qing Pan, Jietai Jing, Jing Zhang, Changde Xie, and Kunchi Peng, 'Quantum dense coding exploiting a bright einstein-podolsky-rosen beam', *Physical review letters*, **88**(4), 047904, (2002).
- [33] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, 'Quantum algorithms for supervised and unsupervised machine learning', *arXiv preprint arXiv:1307.0411*, (2013).
- [34] Klaus Mattle, Harald Weinfurter, Paul G Kwiat, and Anton Zeilinger, 'Dense coding in experimental quantum communication', *Physical Review Letters*, **76**(25), 4656, (1996).
- [35] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.
- [36] Giuseppe Davide Paparo, Vedran Dunjko, Adi Makmal, Miguel Angel Martin-Delgado, and Hans J Briegel, 'Quantum speedup for active learning agents', *Physical Review X*, **4**(3), 031002, (2014).
- [37] Sandu Popescu and Daniel Rohrlich, 'Quantum nonlocality as an axiom', *Foundations of Physics*, **24**(3), 379–385, (1994).
- [38] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd, 'Quantum support vector machine for big data classification', *Physical review letters*, **113**(13), 130503, (2014).
- [39] Frank Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*, Cornell Aeronautical Laboratory, 1957.
- [40] Bruce Rosenblum and Fred Kuttner, 'The observer in the quantum experiment', *Foundations of Physics*, **32**(8), 1273–1293, (2002).
- [41] Stuart J Russell and Peter Norvig, *Artificial intelligence: a modern approach*, Malaysia; Pearson Education Limited., 2016.
- [42] Rüdiger Schack, 'Quantum theory from four of hardy's axioms', *Foundations of Physics*, **33**(10), 1461–1468, (Oct 2003).
- [43] Maria Schuld and Nathan Killoran, 'Quantum machine learning in feature hilbert spaces', *Physical review letters*, **122**(4), 040504, (2019).
- [44] Zhigang Song, J-Q Liang, and L-F Wei, 'Quantum non-localities and correlation-measurement-induced berry phases for spin-singlet states', *arXiv preprint arXiv:1207.3379*, (2012).
- [45] Johan AK Suykens and Joos Vandewalle, 'Least squares support vector machine classifiers', *Neural processing letters*, **9**(3), 293–300, (1999).
- [46] Francesco Tacchino, Chiara Macchiavello, Dario Gerace, and Daniele Bajoni, 'An artificial neuron implemented on an actual quantum processor', *npj Quantum Information*, **5**(1), 26, (2019).
- [47] V.N. Vapnik and A.Ya. Chervonenkis, 'Algorithms with complete memory and recurrent algorithms in pattern recognition learning', *Automation and Remote Control*, **1968**(4), (1968).
- [48] Peter Wittek, *Quantum machine learning: what quantum computing means to data mining*, Academic Press, 2014.
- [49] Ting Yu and JH Eberly, 'Qubit disentanglement and decoherence via dephasing', *Physical Review B*, **68**(16), 165322, (2003).
- [50] Haifeng Zhang, Jianhua Wang, Zhigang Song, J-Q Liang, and L-F Wei, 'Spin-parity effect in violation of bell's inequalities for entangled states of parallel polarization', *Modern Physics Letters B*, **31**(04), 1750032, (2017).