TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/09 pp813–825
DOI: 10.26599/TST.2020.9010007
Volume 25, Number 6, December 2020

Traffic Clustering Algorithm of Urban Data Brain Based on a Hybrid-Augmented Architecture of Quantum Annealing and Brain-Inspired Cognitive Computing

Ning Wang, Gege Guo, Baonan Wang, and Chao Wang*

Abstract: In recent years, the urbanization process has brought modernity while also causing key issues, such as traffic congestion and parking conflicts. Therefore, cities need a more intelligent "brain" to form more intelligent and efficient transportation systems. At present, as a type of machine learning, the traditional clustering algorithm still has limitations. K-means algorithm is widely used to solve traffic clustering problems, but it has limitations, such as sensitivity to initial points and poor robustness. Therefore, based on the hybrid architecture of Quantum Annealing (QA) and brain-inspired cognitive computing, this study proposes QA and Brain-Inspired Clustering Algorithm (QABICA) to solve the problem of urban taxi-stand locations. Based on the traffic trajectory data of Xi'an and Chengdu provided by Didi Chuxing, the clustering results of our algorithm and K-means algorithm are compared. We find that the average taxi-stand location bias of the final result based on QABICA is smaller than that based on K-means, and the bias of our algorithm can effectively reduce the tradition K-means bias by approximately 42%, up to approximately 83%, with higher robustness. QA algorithm is able to jump out of the local suboptimal solutions and approach the global optimum, and brain-inspired cognitive computing provides search feedback and direction. Thus, we will further consider applying our algorithm to analyze urban traffic flow, and solve traffic congestion and other key problems in intelligent transportation.

Key words: cluster analysis; intelligent transportation; quantum annealing and brain-inspired clustering algorithm; K-means

1 Introduction

With the deepening of urbanization, certain problems are increasingly prominent, such as difficulty of maintaining resource dependence mode, worsening environmental problems, lack of sustainable economic development potential, and simplification of urban planning and construction mode. Therefore, urban revitalization necessitates a strong, intelligent brain to participate in the adjustment of the city's "metabolism".

IEEE Academician Xiansheng Hua has pointed out that the urban data brain is the only Artificial Intelligence (AI) system that can analyze the overall urban data in real time. The urban data brain integrates multi-source heterogeneous data, including the whole network open data, government and public institutions data, Internet Of Things (IOT) perception data, and others, and further gathers these massive data resources in real time, conducts overall real-time analysis and adjustment of the urban state, and optimizes social public resources in a global way. In the process of planning and construction

• Ning Wang, Gege Guo, Baonan Wang, and Chao Wang are with Key laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China and also with State Key Laboratory of Cryptology, Beijing 100878, China. E-mail: nana_shu@126.com; 2014422593@qq.com; 2693620328@qq.com; wangchao@shu.edu.cn.
• Chao Wang is also with Center for Quantum Computing, Peng Cheng Laboratory, Shenzhen 518000, China.
* To whom correspondence should be addressed.
Manuscript received: 2020-03-03; accepted: 2020-03-10

of smart cities, the construction of urban data brains will play an important role. If supply and demand are balanced, then roads will not be congested. Urban road traffic congestion may be due to many factors, and the root cause is that the traffic demand of travelers continues to grow, far exceeding the city's existing road traffic supply capacity. Therefore, the focus of modern urban transportation research is how to effectively mine and use data to improve the utilization efficiency of the road network and make the operation of the transportation system more rational and intelligent[1].

Data mining is an interdisciplinary subject. It reveals the potential laws and characteristics of the interaction between traffic data of various dimensions from random and large amounts of database data, and then helps decision-making[2]. The core technology of data mining — cluster analysis, is to divide the dataset into several clusters so that the similarity of objects in the same cluster is higher, and the similarity of different clusters is lower. Traditional clustering algorithm still has some problems, such as poor robustness, strong dependence on training parameters, lack of cognition, and so on. Existing research on clustering analysis can be divided into two categories: One is to optimize and improve the original traditional clustering algorithm to adapt to different problems; the other is to improve and integrate various clustering algorithms for different types of problems. Although the aforementioned improvements have minimal effect, they cannot fundamentally solve the limitations of the traditional clustering algorithm, and the application is extremely narrow in scope and does not have good universality.

To address the defects of traditional algorithms, we propose Quantum Annealing (QA) and Brain-Inspired Clustering Algorithm (QABICA) based on a hybrid data architecture of QA and brain-inspired cognitive computing, and verify the effectiveness and reliability of our algorithm by solving the urban taxi-stand location problem. This location problem is essentially a clustering problem concerning pick-up points of taxi passengers. To effectively evaluate the performance of our algorithm, we propose the average taxi-stand location bias to measure the bias of clustering results under different sample datasets in the same time period and same region. This study further conducts a comparison experiment with the classic K-means algorithm. The experimental results show that the average taxi-stand location bias of the final result based on QABICA is smaller than that based on K-means, and the bias of our algorithm can effectively reduce the traditional K-means bias by approximately 42%, up to approximately 83%, which proves that it has a better clustering effect and stronger robustness than traditional clustering algorithms.

Based on the hybrid architecture of quantum annealing and brain-inspired cognitive computing, the excellent performance of QA and brain-inspired clustering algorithm in solving the urban taxi-stand location problem proves that brain-inspired cognition and quantum computing are feasible, effective, and universal in solving current Machine Learning (ML) problems. Therefore, we can foresee that quantum clustering algorithm will be widely used in helping the urban data brain to control and analyze traffic flow direction and solve traffic congestion.

2 Cluster Analysis Based on Traffic Data Processing

With the rapid growth of data in various fields, clustering analysis has become an increasingly popular research direction. Cluster analysis mainly includes two aspects: similarity measurement and clustering algorithm. Similarity measurement aims to calculate the similarity between two samples. The most common method is distance calculation, such as the method of Euclidean distance.

2.1 Massive traffic data in cities

The modern urban transportation system is a dynamic and open system with a complex and large structure. The huge amount of traffic data are mainly reflected in the diversity of data types, including multi-sector and multi-industry heterogeneous data[3]. The realization of smart transportation requires the construction of massive data analysis and data management based on urban data brains, including processing of multi-source data, such as GPS data on vehicle driving and traffic sensor and special event data[4]. Based on the dynamic update frequency of data, existing traffic data can be divided into basic and real-time types. Basic data mainly include infrastructure data related to road traffic. Real-time data refer to the real-time dynamic data of people, cars, roads, and goods. The fusion of multi-source, heterogeneous, and multi-temporal road data need to integrate all types of data from various sources according to the needs so that irregular data fusion can be conducted[5].

At present, the common traffic network detection data

include fixed detector data and floating vehicle GPS data[6]. Among them, the latter has the advantages of wide collection area, low installation and maintenance cost, and relatively stable data acquisition, and has been more widely used[7]. However, much of the information are raw and rough, which cannot be effectively organized and utilized, resulting in the waste of traffic flow data resources[8]. To effectively deal with historical and real-time traffic data resources, we need to apply data mining technology to massive traffic data. Data mining can help to discover traffic information patterns and traffic laws and predict future traffic development trends as well as provide data and theoretical basis for relevant departments to make scientific decisions and develop intelligent transportation mechanisms.

2.2 Cluster analysis

Intelligent traffic data analysis can effectively learn the operating characteristics and laws of urban traffic conditions, and is an important step in establishing multi-level theories and methods for road traffic planning, design, control, and management[9]. With the availability of high-dimensional and massive traffic data, selecting a reasonable cluster analysis method to analyze the depth information of the traffic operation state in the space-time dimension is also a key means to adjust the traffic operation state in the existing transportation field.

Clustering refers to the process of dividing an object set into multiple categories: grouping samples based on the selected similarity measure, so that the data belonging to the same group are as similar as possible, and the data of different groups are as different as possible. Similarity measurement is a standard to compare the similarity between two objects in a dataset. According to the principle of similarity measurement, distance can be classified into Euclidean, trajectory, Manhattan distance, and others. Choosing a reasonable similarity measure is one of the key steps of clustering, in which trajectory is the most intuitive form of behavior of moving objects, which can be directly used as the behavior representation of moving objects. Clustering analysis based on the similarity measure of track distance[10] is the process of dividing tracks with similar motion law and close distance into the same category.

Gaffney and Smyth[11] proposed a trajectory clustering method that used a hybrid regression model to model trajectories. Lee et al.[12] analyzed the importance of trajectory subsegment to trajectory clustering analysis, constructed a trajectory clustering framework based on grouping, and proposed an effective trajectory clustering algorithm. The trajectory clustering analysis method proposed by Won et al.[13] proved that the method had good clustering accuracy through clustering experiments, but the algorithm implementation efficiency was not high. Hwang et al.[14] noted that in certain monitoring scenarios, the traditional European distance was not enough to be applied to the road network space. Therefore, an improved track similarity measurement method was proposed, which can be effectively applied to the data preprocessing stage of track clustering. Li et al.[15] applied the concept of micro-clustering to the analysis of moving target trajectory data, and proposed a clustering analysis method for mobile micro-clustering. In addition, many scholars have directly applied the classic clustering method to the clustering analysis of moving target trajectory data to achieve the division of such data.

Based on algorithm principles, cluster analysis methods can be roughly divided into partitioned, hierarchical, density-based, grid-based, and model-based clustering methods. Different clustering algorithms are suitable for various types of data.

Partitioned clustering algorithm is the most widely used in ML. Among them, K-means algorithm is one of the most classical partition algorithms. Owing to its advantages of simplicity, no prior knowledge, and strong applicability, it has become the main application algorithm of traffic state mining. The advantage of hierarchical clustering algorithm is that users can clearly express the hierarchical relationship between clusters without specifying the number of clusters in advance. The disadvantage is that the group formed in the previous level cannot be adjusted, and the calculation complexity of the method is extremely high. Density-based clustering algorithm is used to find clusters of arbitrary shape according to the distribution density of sample points when there is no requirement for the number and shape of clusters, so that noise can be eliminated. The grid-based clustering algorithm divides the clustering objects into grids and then clusters them. This type of clustering algorithm can only find horizontal or vertical clustering, and detecting the clustering of skew boundary is difficult. The model-based algorithm is based on the premise that the data satisfy the potential probability distribution assumption, so it is suitable for clustering the known data distribution. The algorithm applies various mathematical models to the known data, and then fits and optimizes them.

Although many studies and applications of clustering analysis have been conducted in the field of transportation, the following problems are observed in the comprehensive research:

(1) A large number of clustering methods have limitations and deficiencies, and the algorithm itself needs to be improved and perfected. For example, most of the studies adopt the methods of partition clustering and hierarchical clustering, but they do not deal with the initial data noise reduction, which leads to the lack of noise resistance of the algorithm and affects the clustering results.

(2) The research on the generality, stability, accuracy, and efficiency of clustering algorithms in the existing research needs to be strengthened and deepened.

(3) Most existing clustering algorithms only dig out the latent laws of urban road networks macroscopically in the time dimension. The lack of mining depth causes difficulty in obtaining the underlying information in the spatial dimension.

(4) Existing research lacks the ability to assist in decision-making on the congestion problem in intelligent transportation. Thus, a practical, effective, versatile, and robust clustering algorithm that can be applied to intelligent transportation is urgently needed.

In view of the limitations of the current clustering algorithm, such as sensitivity to initial points, vulnerability to training samples, and poor robustness, we propose a novel algorithm. This QA and brain-inspired clustering algorithm is based on the hybrid-augmented architecture of brain-inspired cognitive computing and QA. It is applied to the layout of urban taxi stops. Based on the taxi pick-up data provided by the Didi platform[16], the advantages of this framework in traffic clustering of urban data brain are verified.

The data in this study, which we use to analyze the urban taxi-stand location problem, are based on the local trajectory data of Xi'an and Chengdu in November 2016. The problem entails clustering the taxi pick-up points, so we extract the data of ride start time, pick-up latitude and longitude, and preprocess the time stamp, coordinate system conversion, and time and space division. The final format of the experimental data is shown in Table 1.

Table 1 Format of experimental data.

Field	Type	Sample	Comment
Number	Int	1	Sort by time
Ride start time	String	2016-11-01 7:02	Beijing time
Pick-up longitude	String	104.112 25	GCJ-02
Pick-up latitude	String	30.667 03	GCJ-02

To obtain the map of Xi'an and Chengdu, we load the online Gaode map to visualize taxi pick-up points based on QGIS software. The map of Xi'an city is divided into 5×5 grids and that of Chengdu city is divided into 20×20 grids according to the latitude and longitude coordinate range of pick-up points. Among them, each grid is approximately 4 km^2. By analyzing the data of taxi pick-up points in each grid, we can obtain the layout of taxi stops, which can greatly improve the computing efficiency and achieve the scalability of the architecture.

The similarity measure is based on the OSMnx software package[17] to calculate the driving distance and measure the distance similarity between the pick-up coordinates.

3 QA and Brain-Inspired Clustering Algorithm

In view of the defects of the current clustering algorithm, such as sensitivity to the initial cluster core, vulnerability to the training samples, and poor robustness, this study proposes a new hybrid-augmented architecture of QA and brain-inspired cognitive computing, which consists of QA, brain-inspired cognitive science, and classical computing.

At present, QA and simulated annealing algorithms have made outstanding contributions in fields, such as cryptography[18–20]. Among these algorithms, QA, as the core algorithm of D-Wave machine, utilizes the quantum tunneling effects to present the tendency toward low-energy states based on the quantum adiabatic computing theorem[21] with the potential to approximate, or even achieve the global optimum[22], which can overcome the defect in which classical methods are easily trapped in the local optimal solution in large-scale cases. Based on the tendency of the QA algorithm toward low-energy states, the framework further considers the introduction of brain-inspired cognitive techniques, such as selective attention mechanism, which provides search direction and feedback for QA in the process of searching global optimum and enhances the exponential search advantage of quantum tunneling.

In this paper, we propose QABICA on the foundation of the proposed architecture. The algorithm adopts two-step quantum optimization for the urban taxi-stand location problem. The first step is to initialize a cluster core based on data density[23], and the second step is to iterate the clustering process based on cluster center and brain-inspired cognitive computing.

3.1 Initialization cluster core based on data density

NASA and D-Wave[23] jointly proposed the quantum spectrum clustering algorithm in 2012, but it was not used to solve practical problems at the time. Inspired by this algorithm, we use it to find the "center of gravity" in the case of different distributions of taxi pick-ups. The datasets can be divided into two clusters surrounded by each other through the quantum spectral clustering algorithm. We let the central coordinate point of the inner cluster be the "center of gravity" in this area. Then, we can obtain the initial cluster core by the "center of gravity" for the next step of the algorithm.

One measure of a good partitioning is how far apart the data points grouped together in one cluster are from each other. Based on MAXCUT partitioning, the quantum spectral clustering algorithm[23] aims to maximize the distance between clusters, that is, find the maximum of Eq. (1),

$$\sum_{c_0 \in C_0,\, c_1 \in C_1} d(c_0, c_1) \tag{1}$$

where the sum is the total distance between all pairs of points c_0 and c_1, which are in clusters C_0 and C_1, respectively.

Let $\{x_i\}$ be the set of all data points. Formula (1) can be written as a binary optimization problem, or say, Quadratic Unconstrained Binary Optimization (QUBO) problem, where it can be attacked by QA on a D-Wave quantum computer. The binary membership variable q_i indicates which cluster the point x_i is placed in. When x_i belongs to cluster C_0, q_i is 0; otherwise, q_i is 1, i.e.,

$$q_i = \begin{cases} 1, & \text{if } x_i \in C_1; \\ 0, & \text{if } x_i \in C_0 \end{cases} \tag{2}$$

In terms of the binary variable q_i, Formula (1) can be mapped to the QUBO problem:

$$E(q) = -\sum_{i,j} d_{ij} q_i (1 - q_j) = -\sum_i d_{ij} q_i + \sum_{i,j} d_{ij} q_i q_j \tag{3}$$

where each coefficient d_{ij} is the distance between the points x_i and x_j, $d_{ij} = d(x_i, x_j)$. Here, the data points are taxi pick-up points and d_{ij} is driving distance between each pair of pick-up points. When this cost function $E(q)$ reaches the minimum, it indicates that the distance between clusters reaches the maximum.

According to the results of the first step, we let the central coordinate point of the inner cluster be the "center of gravity" in this area. In the actual setting of taxi stands, the stands must be distributed around the center of gravity similar to the pick-up points. Thus, the initial cluster core can be set based on the center of gravity.

3.2 Iteration clustering process based on cluster center and brain-inspired cognitive computing

3.2.1 Algorithm framework

Step 1 Initialization cluster core. Through quantum spectral clustering[23], we can extract temporal and spatial distribution characteristics of taxi pick-up points to automatically obtain the center of gravity of the pick-up area from a global perspective. We can determine the initial cluster core by letting the longitude of the center of gravity remain unchanged, and the latitude of the center of gravity can add and subtract 0.002.

Step 2 Introduction of brain-inspired cognitive techniques. We focus on the pick-up points in the middle area of the cluster center by introducing selective attention mechanism[24]. In this study, we call these points key pick-up points to prepare for the implementation of the QUBO model.

Step 3 QUBO model. The experimental data are clustered by using iteration clustering model based on cluster center and brain-inspired cognitive computing. We can calculate the distance-based similarity between each pick-up point and every cluster core based on the driving distance, and use the angle measurement to calculate the angle-based similarity between the key pick-up points and other points. The QUBO model is established based on the distance-based similarity and angle-based similarity, and solved by quantum simulation. According to the results, each pick-up point is divided into the cluster where the cluster core is located.

Step 4 Updating cluster core. We update the cluster core based on the new clustering results. According to the experimental results of Step 3, a new cluster core can be obtained by calculating the average value of longitude and latitude of all pick-up points in each cluster.

Step 5 Repeating iteration until convergence. We need to repeat Steps 2 – 4 until the clustering result remains unchanged, and the final cluster core location is the taxi-stand location in this area. Through experiments, we find that the number of iterations is basically stable within 10 times.

3.2.2 Algorithm model

Let $\{x_i\}$ be the set of all taxi pick-up points, and the total number of pick-up points is N, $i \in \{1, 2, \ldots, N\}$. As in Section 3.1, the binary membership variable q_i indicates which cluster the taxi pick-up point x_i is placed in. When x_i belongs to cluster C_0, q_i is 0; otherwise, q_i is 1, i.e.,

$$q_i = \begin{cases} 1, & \text{if } x_i \in C_1; \\ 0, & \text{if } x_i \in C_0 \end{cases} \tag{4}$$

Figure 1 shows the variable diagram used in the study.

(1) Key pick-up points

Let the center of cluster C_0 be c_a, the center of cluster C_1 be c_b, and d_{ab} be the driving distance between the two centers. For each pick-up point x_i, d_{ia} is the driving distance between x_i and c_a, and d_{ib} is the driving distance between x_i and c_b. When x_i satisfies Eq. (5), we put x_i into set $\{x_k\}$; otherwise, we put x_i into set $\{x_n\}$, and the total number of elements of sets $\{x_k\}$ and $\{x_n\}$ is N. Through Eq. (5), we can judge whether the pick-up point x_i is in the middle area of cluster cores c_a and c_b; of course, the elements in set $\{x_k\}$ are key pick-up points.

$$\begin{cases} d_{ia} < d_{ab}, \\ d_{ib} < d_{ab}, \\ \dfrac{|d_{ia} - d_{ib}|}{d_{ab}} \leqslant \dfrac{1}{3} \end{cases} \tag{5}$$

(2) Distance-based similarity

For any ordinary pick-up point x_n, d_{na} is the driving distance between x_n and c_a, and d_{nb} is the driving distance between x_n and c_b. We let the distance-based similarity D_{na} between x_n and c_a equal d_{na}, and the distance-based similarity D_{nb} between x_n and c_b equal d_{nb}. For the key pick-up point x_k, d_{ka} is the driving distance between x_k and c_a, and d_{kb} is the driving distance between x_k and c_b. We let the distance-based similarity D_{ka} between x_k and c_a equal $3d_{ka}/d_{ab}$, and the distance-based similarity D_{kb} between x_k and c_b equal $3d_{kb}/d_{ab}$. The absolute value $|3d_{ka} - 3d_{kb}|/d_{ab}$ of the difference between distance-based similarity D_{ka} and D_{kb} has been normalized to [0, 1]. The reason is that for the key pick-up points, the proportion of distance-based similarity decreases, and angle-based similarity plays a more important role in the model. The smaller the value of distance-based similarity, the higher the similarity between the pick-up point and cluster core, and the greater the probability that the pick-up point belongs to this cluster.

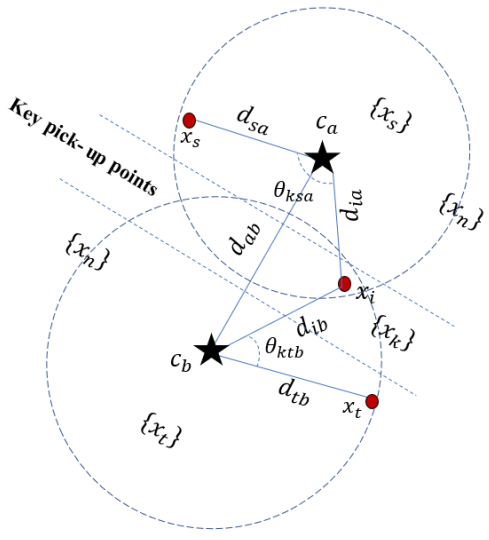


Fig. 1 Variable diagram.

The key pick-up points are in the middle of the two cluster cores. The distance-based similarity between these points and the cluster core is not obvious; thus, the distance-based similarity is not the necessary condition to judge which cluster these points are placed in. Therefore, based on human intuitive experience, the angle-based similarity is introduced to distinguish these points from the direction to achieve the goal of traffic diversion effect.

(3) Angle-based similarity

Case 1: Taking cluster core c_a as reference point.

In set $\{x_k\}$, the key pick-up point x_k is selected in turn, and d_{ka} is the driving distance between x_k and c_a. For each x_k, compared with each x_i in set $\{x_i\}$, d_{ia} is the driving distance between x_i and c_a, and d_{kia} is the absolute value of the difference between d_{ka} and d_{ia}. When d_{kia} is less than the threshold ε (Eq. (6)), we put x_i in $\{x_s\}$. Thus, we can obtain a set $\{x_s\}$ through each x_k.

$$d_{kia} = |d_{ka} - d_{ia}| \leqslant \varepsilon, \quad i \in \{1, 2, \ldots, N\} \tag{6}$$

The threshold ε aims to screen out x_i with a small gap between d_{ia} and d_{ka}. In this study, the threshold ε is related to the total number of pick-up points N, which aims to reduce the calculation amount properly without affecting the overall situation, in meter (m). The threshold ε is given as

$$\varepsilon = \begin{cases} 150, & 0 < N < 500; \\ 120, & 500 \leqslant N < 1000; \\ 90, & 1000 \leqslant N < 1500; \\ 60, & N \geqslant 1000 \end{cases} \tag{7}$$

d_{ksa} is the absolute value of the difference between d_{ka} and d_{sa}. Each x_k corresponds to a set $\{x_s\}$ and also to a set $\{d_{ksa}\}$. In each set $\{d_{ksa}\}$, the maximum value is $\max(d_{ksa})$, and the minimum value is $\min(d_{ksa})$. $\text{norm}(d_{ksa})$ (Eq. (8)) is normalized by d_{ksa}. We can determine the weight of length W_d^{ksa} (Eq. (9)) in angle-based similarity by $\text{norm}(d_{ksa})$, which is normalized to

[0, 1]. The smaller the absolute value of the difference between d_{ka} and d_{sa} is, the greater the weight of length W_d^{ksa} is.

$$\text{norm}(d_{ksa}) = \frac{d_{ksa} - \min(d_{ksa})}{\max(d_{ksa}) - \min(d_{ksa})} \tag{8}$$

$$W_d^{ksa} = \begin{cases} 1, & \text{norm}(d_{ksa}) \leqslant 0.3; \\ -1.429 \times \text{norm}(d_{ksa}) + 1.429, & \\ & 0.3 < \text{norm}(d_{ksa}) \leqslant 1 \end{cases} \tag{9}$$

With the cluster core c_a as the reference point, θ_{ksa} is the angle between x_k and x_s . According to the angle θ_{ksa}, we can obtain the weight of angle W_θ^{ksa} (Eq. (10)) in the angle-based similarity, which is normalized to $[-1, 1]$.

$$W_\theta^{ksa} = \begin{cases} 1, & 0° \leqslant \theta_{ksa} \leqslant 30°; \\ -\dfrac{1}{\tan(60°)} \tan(\theta_1 - 90°), & 30° < \theta_{ksa} \leqslant 150°; \\ -1, & 150° < \theta_{ksa} \leqslant 180° \end{cases} \tag{10}$$

Finally, when x_k and x_s take c_a as the reference point, the angle-based similarity W_{ksa} (Eq. (11)) is decided by W_d^{ksa} and W_θ^{ksa}, and the values of W_d^{ksa} and W_θ^{ksa} are set as shown in Figs. 2 and 3, respectively.

$$W_{ksa} = W_d^{ksa} \times W_\theta^{ksa} \tag{11}$$

Case 2: Taking cluster core c_b as reference point.

Using cluster core c_b as reference point is the same as using cluster core c_a as the reference point. Similarly,

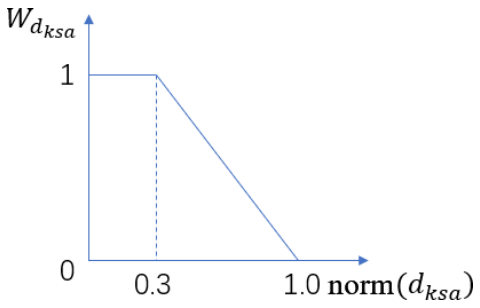


Fig. 2 Weight of length W_d^{ksa}.

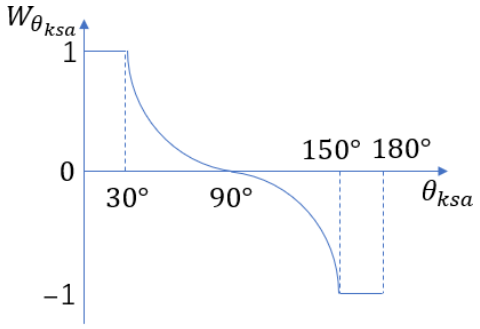


Fig. 3 Weight of angle W_θ^{ksa}.

d_{kb} is the driving distance between x_k and c_b. For each x_k, compared with each x_i in set $\{x_i\}$, d_{ib} is the driving distance between x_i and c_b, and d_{kib} is the absolute value of difference between d_{kb} and d_{ib}. When d_{kib} is less than the threshold ε, we put x_i in $\{x_t\}$. Thus, we can obtain a set $\{x_t\}$ by each x_k. We can also determine the weight of length W_d^{ktb} by d_{ktb} and the weight of angle W_θ^{ktb} by θ_{ktb}. Finally, we obtain the angle-based similarity W_{ktb} between x_k and x_t.

(4) QUBO model

Based on distance-based and angle-based similarity, the QUBO model is constructed. Firstly, the QUBO cost function Obj_1 obtained by the driving distance-based similarity is shown in Eq. (12). When the coefficient of q_n or q_k is greater than 0, the pick-up point x_n or x_k belongs to the cluster C_0; otherwise, when the coefficient is less than 0, the pick-up point x_n or x_k belongs to the cluster C_1.

$$\text{Obj}_1 = \sum_n (D_{na} - D_{nb}) q_n + \sum_k (D_{ka} - D_{kb})\, q_k \tag{12}$$

Then, the QUBO cost function Obj_2 obtained by the angle-based similarity is shown in Eq. (13). The closer the angle-based similarity is to 1, the greater the probability that the two pick-up points are in different clusters is; the closer the angle-based similarity is to -1, the greater the probability that the two pick-up points are in the same cluster is. If the angle-based similarity is close to 1, then the angle θ between the two pick-up points is small, and the probability of meeting when the two pick-up points go to the same taxi stand is large. Thus, to achieve the purpose of traffic diversion, two points should belong to different clusters and go to different stands.

$$\begin{aligned} \text{Obj}_2 = & \sum_{k,s} W_{ksa}(-q_k - q_s + 2q_k q_s) + \\ & \sum_{k,t} W_{ktb}(-q_k - q_t + 2q_k q_t) \end{aligned} \tag{13}$$

Finally, the global cost function Obj for the QUBO model with iteration clustering process based on cluster center and brain-inspired cognitive computing is given as Eq. (14), which is solved by quantum simulation.

$$\text{Obj} = \text{Obj}_1 + \text{Obj}_2 \tag{14}$$

4 Experiment

We propose QA and brain-inspired clustering algorithm to solve the urban taxi-stand location problem, which can improve the utilization rate of the road network through traffic diversion and promote the development

of intelligent transportation of the city brain. The urban taxi-stand location problem is essentially a clustering problem of taxi pick-up points. Therefore, we innovatively propose the clustering algorithm, and prove its advantages in robustness and generality compared with the traditional clustering algorithm through comparative experiments.

Inspired by distributed computing, we divide the Chengdu map into 20×20 grids and Xi'an into 5×5 grids (each grid is approximately 4 km^2) according to the range of taxi pick-up coordinates to achieve efficient computing and scalable architecture. That is, we can determine the taxi-stand location in each grid.

4.1 Instance

In this section, we take a dataset in Xi'an as an example to analyze the feasibility of QABICA based on the hybrid-augmented architecture of QA and brain-inspired cognitive computing from the specific experimental results.

(1) Dataset

A total of 223 taxi pick-up points are found in the designated area (longitude 108.927°–108.946° and latitude 34.2280°–34.2452°) at 07:00:00–9:59:59 on November 1, 2016. These points are marked as blue in QGIS as shown in Fig. 4.

(2) Comparative experimental results based on QABICA and K-means

We compare the method used in this study with the K-means algorithm, one of the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining[25]. The K-means algorithm is widely used in clustering problems, particularly taxi-stand location problem, which is the best-known partition clustering algorithm with simplicity and efficiency. However, K-means as a traditional ML algorithm has certain defects, such as sensitivity to initial state, low robustness to noise, and strong dependence on training data.

Based on a group of datasets of Xi'an, the feasibility and effectiveness of our algorithm are analyzed by comparing the experimental results. The final clustering result based on QABICA is shown in Fig. 5a. A total of 223 pick-up points are in the selected area, among which the yellow pentagrams are the taxi stands according to the final clustering results. The red cluster includes 111 data points and the green cluster includes 112 data points. These data points in the cluster are evenly distributed. The final clustering result based on K-means is shown in Fig. 5b. Similarly, 223 pick-up points are in the selected area, among which the yellow pentagrams are the taxi stands in this area according to the final

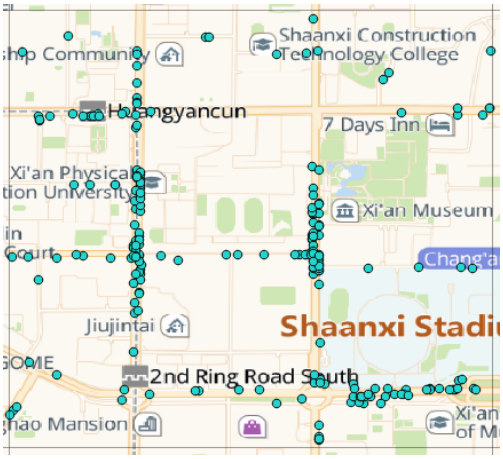
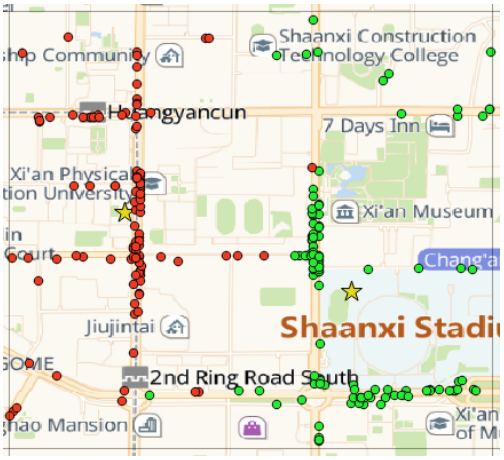
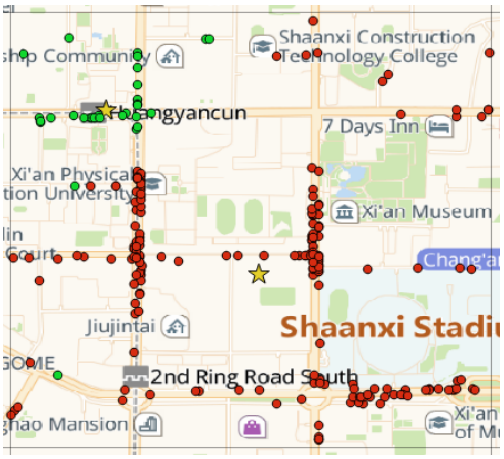


Fig. 4 Visualization results of taxi pick-up points.



(a) QABICA clustering result



(b) K-means clustering result

Fig. 5 Clustering result based on QABICA and K-means.

clustering results. However, the red cluster includes 192 data points, and the green cluster includes 31 data points. The distribution of data points in the cluster is extremely uneven.

Based on the above comparative experimental results, the following conclusions can be drawn: Compared with the K-means algorithm, the QA and brain-inspired clustering algorithm can divide the pick-up points into two clusters with more uniform distribution, thereby effectively avoiding an excessive number of pick-up points around one taxi stop to achieve the effect of traffic diversion and promoting research on intelligent transportation of urban data brains.

4.2 Evaluation

Based on the QA and brain-inspired clustering algorithm, the final taxi stops can be obtained by calculating the datasets obtained from multiple sampling of data in the same time period and in the designated area.

To improve the performance evaluation of our algorithm, we propose the average taxi-stand location bias to measure the bias of clustering results under different sample datasets in the same time and region. To ensure the same basic conditions, we use the classical K-means algorithm, which applies driving distance as the distance-based measurement. Here, we propose the average taxi-stand location bias, which can be calculated by clustering results of multiple datasets. The lower the bias based on different datasets is, the higher the robustness of the algorithm is. Otherwise, the algorithm has a large dependence on the initial core and datasets, and therefore cannot effectively deal with the urban taxi-stand location problem under the distribution of complex datasets.

$$\text{excursion\_var} = \sum_{i=1}^{2} \frac{1}{m_i} \sum_{k_j \in c_i, j=1}^{m_i} \text{dist}^2(c_i, k_j) \quad (15)$$

where i ($i = \{1, 2\}$) is the group of clustering results, and two groups exist. c_i is the cluster i core, k_j is the data point in the corresponding cluster, $j = \{1, \ldots, m_i\}$, m_i is the number of data points in c_i, and $\text{dist}(c_i, k_j)$ is the distance between k_j and c_i in cluster i. excursion_var represents the average distance variance between data points in each cluster and cluster core, in square meter (m^2). The arithmetic square root of excursion_var can be quantified as the average taxi-stand location bias of the final result, in meter (m). In the actual experiment process, we conduct comparative experiments based on the datasets of the designated grid area and period in Xi'an and Chengdu, respectively. The following is a detailed analysis of the experimental results.

4.2.1 Xi'an datasets

A total of 4808 taxi pick-up points are in the designated area (longitude 108.927°–108.946° and latitude 34.2280°–34.2452°) at 07:00:00–9:59:59 during all weekdays in November 2016. Through random sampling method, we obtain 10 datasets, and each dataset consists of 500 random sampling points, that is, 5000 data points are used in this experiment.

A total of 1233 taxi pick-up points are in the designated area (longitude 108.927°–108.946° and latitude 34.2280°–34.2452°) at 07:00:00–9:59:59 during all weekends in November 2016. Through the random sampling method, we obtain 10 datasets, and each dataset consists of 500 random sampling points, that is, 5000 data points are used in this experiment.

(1) Comparative experimental results based on weekday datasets of Xi'an

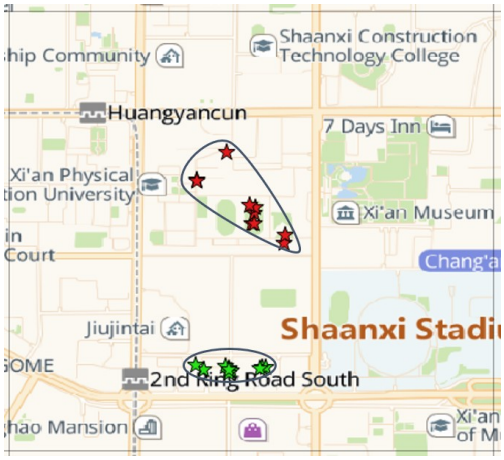
This experiment is based on the weekday datasets of Xi'an given above. By calculating the average taxi-stand location bias based on 10 datasets, we obtain the experimental results (see Table 2).

The comparative experimental results based on the weekday datasets of Xi'an show that the average taxi-stand location bias of the final result based on QABICA is smaller than that based on K-means, and the average taxi-stand location bias of our algorithm can effectively reduce that of the traditional K-means by approximately 42%, with higher robustness.

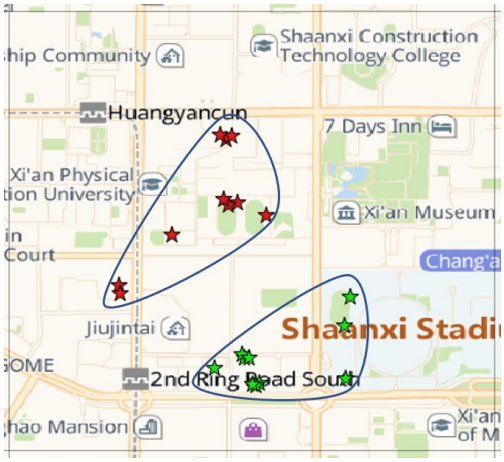
To show the advantage of our algorithm in the average taxi-stand location bias more intuitively, we visualize the experimental results. In Fig. 6, the pentagrams represent clustering results based on 10 datasets, and the red and green represent different clusters. The 10 groups of datasets can obtain 10 groups of clustering results. Through the comparison in Fig. 6, the clustering results based on QABICA are closer and the algorithm performance is stronger. Conversely, the clustering results obtained by K-means are scattered and greatly affected by the training data with poor robustness.

Table 2 Average taxi-stand location bias based on weekday datasets of Xi'an.

Algorithm	excursion_var (m^2)	Average taxi-stand location bias (m)
QABICA	268 575.1	518.24
K-Means	785 640.4	886.36



(a) QABICA clustering result



(b) K-means clustering result

Fig. 6 QABICA clustering result and K-means clustering result based on weekday datasets of Xi'an.

(2) Comparative experimental results based on weekend datasets of Xi'an

This experiment is based on weekend datasets of Xi'an given above. We calculate the average taxi-stand location bias based on 10 datasets to obtain Table 3.

The comparative experimental results based on weekend datasets of Xi'an show that the average taxi-stand location bias of the final result based on QABICA is smaller than that based on K-means. The average taxi-stand location bias of our algorithm can effectively reduce that of the traditional K-means by approximately 81%, with higher robustness.

Table 3 Average taxi-stand location bias based on weekend datasets of Xi'an.

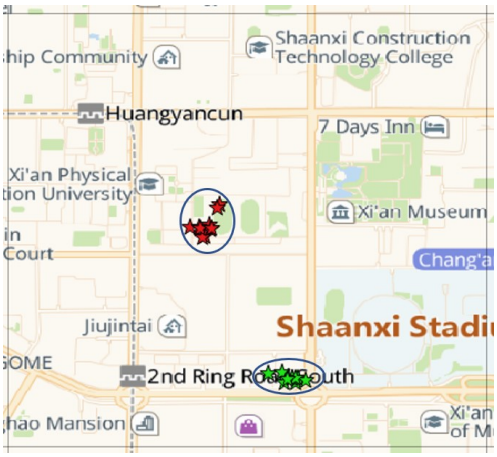
Algorithm	excursion_var (m^2)	Average taxi-stand location bias (m)
QABICA	16 399.84	128.06
K-Means	463 891.58	681.10

Similarly, we visualize the clustering results to obtain Fig. 7. The preceding experimental results based on data in the designated area of Xi'an show that compared with the K-means algorithm, our algorithm is more effective and superior in solving the urban taxi-stand location problem. To further reflect the universality, reliability, and applicability of this algorithm, we continue to conduct the same comparative experiment based on the experimental data in the designated area of Chengdu.

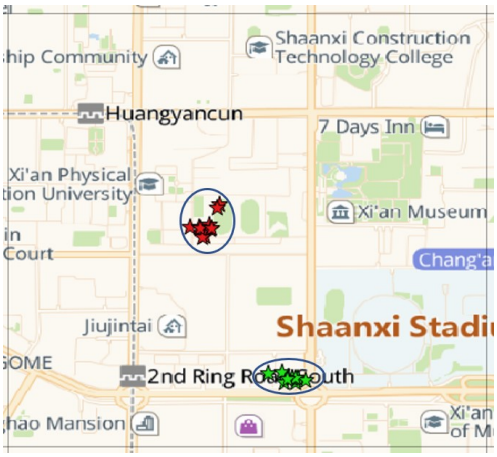
4.2.2 Chengdu datasets

A total of 30 413 taxi pick-up points are in the designated area (longitude 104.063°–104.082° and latitude 30.6504°–30.6676°) at 07:00:00–9:59:59 during all weekdays in November 2016. Through the random sampling method, we obtain 10 datasets, and each dataset consists of 500 random sampling points, that is, 5000 data points are used in this experiment.

A total of 8877 taxi pick-up points are in the



(a) QABICA clustering result



(b) K-means clustering result

Fig. 7 QABICA clustering result and K-means clustering result based on weekend datasets of Xi'an.

designated area (longitude 104.063°–104.082° and latitude 30.6504°–30.6676°) at 07:00:00–9:59:59 during all weekends in November 2016. Through the random sampling method, we obtain 10 datasets, and each dataset consists of 500 random sampling points, that is, 5000 data points are used in this experiment.

(1) Comparative experimental results based on weekday datasets of Chengdu

This experiment is based on the weekday datasets of Chengdu given above. We calculate the average taxi-stand location bias based on 10 datasets to obtain Table 4.

The comparative experimental results based on the weekday datasets of Chengdu show that the average taxi-stand location bias of the final result based on QABICA is smaller than that based on K-means, and the average taxi-stand location bias of our algorithm can effectively reduce that of the traditional K-means by approximately 83%, with higher robustness.

(2) Comparative experimental results based on weekend datasets of Chengdu

This experiment is based on the weekend datasets of Chengdu given above. We calculate the average taxi-stand location bias based on 10 datasets to obtain Table 5.

Similarly, the average taxi-stand location bias of our algorithm can effectively reduce that of the traditional K-means by approximately 57%, with higher robustness. By combining all the experimental results for Chengdu and Xi'an, we find that the QABICA bias can effectively reduce the traditional K-means bias by approximately 42%–83%. Through the average taxi-stand location bias, we can fully show that our algorithm is more robust and universal than the traditional ML clustering method.

Table 4 Average taxi-stand location bias based on weekday datasets of Chengdu.

Algorithm	excursion_var (m^2)	Average taxi-stand location bias (m)
QABICA	14 807.05	121.68
K-Means	509 762.75	713.98

Table 5 Average taxi-stand location bias based on weekend datasets of Chengdu.

Algorithm	excursion_var (m^2)	Average taxi-stand location bias (m)
QABICA	98 928.88	314.53
K-Means	537 102.16	732.87

5 Conclusion

As an important branch of ML, cluster analysis is widely used in various fields of AI. At present, the algorithm research on clustering analysis is mainly divided into two parts: the improvement based on the classical algorithm itself and the combination and fusion of different algorithms, which cannot fundamentally solve the current difficulties of ML. The K-means algorithm is a classical and widely used clustering algorithm in ML. However, it has certain limitations, such as sensitivity to initial point selection, being easily affected by training samples, poor robustness against noise, and inability to recognize spherical clusters.

Therefore, we propose a hybrid-augmented architecture of QA and brain-inspired cognitive computing. This architecture consists of QA, brain-inspired cognition, and classical computation. The introduction of the QA algorithm can help us avoid the local optimal solution and obtain the global optimal solution. Among them, the quantum tunneling effect in the QA process provides an exponential search advantage for the algorithm, which overcomes the limitations of the traditional ML algorithm, such as high computational complexity and long-time consumption. The introduction of brain-inspired cognition can provide search direction and feedback in the process of searching for a global optimal solution through the QA algorithm. This approach can also assist algorithm optimization, which is difficult to achieve in ML and current quantum computing.

To prove the efficiency and generality of the algorithm proposed in this study, we compare it with the best-known partition clustering algorithm (K-means), based on 45 331 pick-up points in Xi'an and Chengdu. The average taxi-stand location bias is proposed to evaluate the robustness of the algorithm. The experimental results show that the clustering deviation based on our algorithm is much lower than that calculated by the K-means algorithm, which can be reduced by at least 42% and at most 83%, and has strong robustness.

Based on the hybrid architecture of brain-inspired cognition and QA, the QA and brain-inspired clustering algorithm performs effectively in solving the problem of taxi-stand layout. This condition proves that brain-inspired cognition and quantum computing are feasible, effective, and universal in solving the current issues in ML. Therefore, these methods can be widely used in transportation, such as in the analysis of urban traffic

flow direction, road network traffic planning, traffic flow prediction, and road traffic congestion, which can provide guidance for optimizing the layout of urban buildings. In the future, we intend to apply QA and brain-inspired clustering algorithm to additional fields and promote the optimization of urban data brain construction.

Acknowledgment

Data source: Didi Chuxing GAIA Initiative. This work was supported by the Special Zone Project of National Defense Innovation, the National Natural Science Foundation of China (Nos. 61572304 and 61272096), the Key Program of the National Natural Science Foundation of China (No. 61332019), and Open Research Fund of State Key Laboratory of Cryptology.

References

[1] F. Neukart, G. Compostella, C. Seidel, D. Von Dollen, S. Yarkoni, and B. Parney, Traffic flow optimization using a quantum annealer, *Frontiers in ICT*, vol. 4, p. 29, 2017.

[2] E. F. Freitas, F. F. Martins, A. Oliveira, I. R. Segundo, and H. Torres, Traffic noise and pavement distresses: Modelling and assessment of input parameters influence through data mining techniques, *Applied Acoustics*, vol. 138, pp. 147–155, 2018.

[3] Y. Djenouri and A. Zimek, Outlier detection in urban traffic data, in *Proc. 8th Int. Conf. Web Intelligence*, *Mining and Semantics*, Novi Sad, Serbia, 2018, pp. 1–12.

[4] B. N. Silva, M. Khan, C. Jung, J. Seo, D. Muhammad, J. H. Han, Y. Yoon, and K. J. Han, Urban planning and smart city decision management empowered by real-time data processing using big data analytics, *Sensors*, vol. 18, no. 9, p. 2994, 2018.

[5] J. Brainard, What's coming up in 2018, *Science*, vol. 359, no. 6371, pp. 10–12, 2018.

[6] E. V. Sekar, J. Anuradha, A. Arya, B. Balusamy, and V. Chang, A framework for smart traffic management using hybrid clustering techniques, *Cluster Computing*, vol. 21, no. 1, pp. 347–362, 2018.

[7] M. Saeedmanesh and N. Geroliminis, Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks, *Transportation Research Part B*: *Methodological*, vol. 105, pp. 193–211, 2017.

[8] A. Gregoriades and A. Chrystodoulides, Extracting traffic safety knowledge from historical accident data, in *Adjunct Proc. 14th Int. Conf. Location Based Services*, Zurich, Switzerland, 2018, pp. 109–114.

[9] K. K. Santhosh, D. P. Dogra, and P. P. Roy, Temporal unknown incremental clustering model for analysis of traffic surveillance videos, *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1762–1773, 2019.

[10] P. X. Zhao, X. T. Liu, J. W. Shen, and M. Chen, A network distance and graph-partitioning-based clustering method for improving the accuracy of urban hotspot detection, *Geocarto International*, vol. 34, no. 3, pp. 293–315, 2019.

[11] S. Gaffney and P. Smyth, Trajectory clustering with mixtures of regression models, in *Proc. 5th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Diego, CA, USA, 1999, pp. 63–72.

[12] J. G. Lee, J. W. Han, and K. Y. Whang, Trajectory clustering: A partition-and-group framework, in *Proc. 2007 ACM SIGMOD Int. Conf. Management of Data*, Beijing, China, 2007, pp. 593–604.

[13] J. I. Won, S. W. Kim, J. H. Baek, and J. Lee, Trajectory clustering in road network environment, presented at 2009 IEEE Symp. Computational Intelligence and Data Mining, Nashville, TN, USA, 2009, pp. 299–305.

[14] J. R. Hwang, H. Y. Kang, and K. J. Li, Spatio-temporal similarity analysis between trajectories on road networks, presented at International Conference on Conceptual Modeling, Berlin, Germany: Springer, 2005, pp. 280–289.

[15] Y. F. Li, J. W. Han, and J. Yang, Clustering moving objects, in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004, pp. 617–622.

[16] Didi Chuxing GAIA Initiative, https://gaia.didichuxing.com, 2019.

[17] G. Boeing, OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, *Computers*, *Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[18] B. N. Wang, F. Hu, and C. Wang, Optimization of quantum computing models inspired by D-wave quantum annealing, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 508–515, 2020.

[19] C. Wang, F. Hu, H. G. Zhang, and J. Wu, Evolutionary cryptography theory-based generating method for secure ECs, *Tsinghua Science and Technology*, vol. 22, no. 5, pp. 499–510, 2017.

[20] F. Hu, C. Wang, H. G. Zhang, and J. Wu, Simple method for realizingweil theorem in secure ECC generation, *Tsinghua Science and Technology*, vol. 22, no. 5, pp. 511–519, 2017.

[21] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem, *Science*, vol. 292, no. 5516, pp. 472–475, 2001.

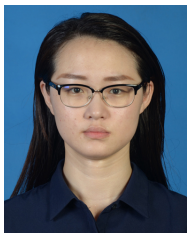
[22] A. B. Finnila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, Quantum annealing: A new method for minimizing multidimensional functions, *Chemical Physics Letters*, vol. 219, nos. 5&6, pp. 343–348, 1994.

[23] V. N. Smelyanskiy, E. G. Rieffel, S. I. Knysh, C. P. Williams, M. W. Johnson, M. C. Thom, W. G. Macready, and K. L. Pudenz, A near-term quantum computing approach for hard computational problems in space exploration, arXiv preprint arXiv: 1204.2821v2, 2012.

[24] N. N. Zheng, Z. Y. Liu, P. J. Ren, Y. Q. Ma, S. T. Chen, S. Y. Yu, J. R. Xue, B. D. Chen, and F. Y. Wang, Hybrid-augmented intelligence: Collaboration and cognition, *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 2, pp. 153–179, 2017.

[25] X. D. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, et al., Top 10 algorithms in data mining, *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

Ning Wang is an MS student at Signal and Information Processing Department, Shanghai University. Her main research interests include intelligent transportation and quantum computing.



Gege Guo is an MS student at Communication and Information Engineering Department, Shanghai University. Her main research interests include intelligent transportation and quantum computing.

Baonan Wang is a PhD student at Electronic and Information Engineering Department, Shanghai University. Her main research interests include information security and quantum computing cryptography.

Chao Wang received the PhD degree from Tongji University in 1999. He is a professor and senior member of CCF. He is an IEEE senior member, vice chair of IEEE China Council, council member of China Institute of Electronic and China Association of AI, deputy director of Information Security Experts Committee (China Institute of Electronic), vice chair of IEEE Shanghai Computer Chapter, and committeeman of the Sixth Shanghai Expert Committee for Informatization. His research interests include AI, smart city, and quantum computing.