Optimal Structure and Energy Prediction of Clusters based on GA-BP-TS Algorithm - Taking Au20 Gold Cluster as an Example

[1]Shiyu Zhao, [2]Jianing Bai and [3]Yanqun Li,
[1,2,3]Beijing Wuzi University, Beijing, China

Abstract: Aiming at the optimal structure and energy optimization of clusters, the genetic algorithm is used to optimize the initial weights and thresholds of BP neural network model. According to the existing data, the cluster energy is predicted, and the fitness function is constructed according to the prediction results. The GS-BP-TS algorithm is established to optimize the optimal structure of Au20 gold clusters. The experimental results show that the GA-BP model has higher prediction accuracy and better prediction performance, and the optimal structure stability of Au20 gold cluster obtained by GA-BP-TS algorithm is higher.

Keywords: *Cluster; Tabu Search Algorithm; GA-BP; Neural Network*

I. INTRODUCTION

Cluster is a relatively stable microscopic or submicroscopic aggregate composed of several or even thousands of atoms, molecules or ions through physical or chemical binding force; The spatial scale of clusters ranges from a few angstroms to several hundred angstroms, and their physical and chemical properties are determined by the number of atoms or molecules contained. Cluster is regarded as a new level of material structure between atoms, molecules and macroscopic solid matter. The research on the electronic structure, stability, aromaticity and bonding mode of clusters is a hot spot in the field of chemistry, and cluster science is a very important research direction in the field of condensed matter physics.

In recent years, metal nanoclusters have been widely used in various technical fields such as imaging, luminescent materials, detection equipment, biology and medicine. Metal nanoclusters have become a new type of materials in the field of nano-research, which has attracted more and more researchers' attention, and gold nanoclusters are one of the research hotspots in this field. The potential energy surfaces of clusters are too complex, and different potential functions will lead to different structures. The structural optimization problem of searching clusters is equivalent to the global optimization problem aiming at the lowest energy of clusters, and it is one of the important research topics in the field of computational chemistry. Determining the lowest energy structure of clusters is helpful for us to understand many physical and chemical properties of real clusters [1-3]. However, because its configuration space increases exponentially with the increase of cluster size, the solution of this kind of problem is usually NP-hard problem. Therefore, it is of great theoretical and practical significance to develop efficient global optimization algorithms for such problems.

In the current literature, a large number of algorithms have been used to solve this optimization problem. In 1997, Wales and Doye[4]found the lowest energy configuration of LJ69, LJ78 and LJ107 by basin-hopping algorithm. In 1999, Leary[5] used Monotonic Sequence Basis-Hopping (MSBH) algorithm to find the lowest energy configuration of LJ98. In 2006, Takeuchi[6]found the lowest energy configuration of LJ506, LJ521, LJ537, LJ538 and LJ541 with HA-SIO algorithm. At present, DLS method[7]and two-stage local search method[8]are two particularly effective optimization techniques for cluster structure optimization. DLS is one of the most effective optimization methods at present, which makes good use of the cluster structure information, and thus has high computing speed. On the other hand, the research shows that the two-stage local search can effectively improve the efficiency of the algorithm by using a transformed potential energy function[9], and has been successfully applied to many optimization problems.

Based on this, a GA-BP-TS model is established to predict the energy and optimal structure of three-dimensional clusters.

II. PREDICTION OF OPTIMAL STRUCTURE OF METAL CLUSTERS

A. Problem description

The optimal structure and energy calculation of large-size clusters mostly depend on potential functions. Common potential functions include LJ, Gupta, Sutton-Chen, EAM potential, etc. Different potential functions often lead to different optimal structures of clusters. Therefore, taking au cluster Au20 as an example, according to the known structure and energy of Au20, a GA-BP-TS model is constructed to predict the lowest energy structure. BP neural network is used to fit the functional relationship between 20 atomic coordinates and energy, so that energy can be predicted according to atomic structure, and tabu search algorithm is constructed according to the prediction results of the model, which is the adaptive value function in (TS). Considering the shortcomings of traditional BP neural network, such as slow convergence speed and easy to fall into local extremum, the initial weights and thresholds of BP neural network model are optimized by combining genetic algorithm The results show that the error between the prediction result of the lowest energy of Au20 and the actual energy based on the improved BP neural network of genetic algorithm is small, and the prediction accuracy is high. Therefore, the improved BP neural network based on genetic algorithm is combined with the tabu search algorithm (TS), and the fitness function of the tabu search algorithm is set according to the energy prediction value based on the improved BP neural network. The GA-BP-TS algorithm is designed to find the global optimal solution, that is, the optimal structure and energy of Au20.

B. Symbol description

X: Atomic coordinate informatio n set of Au20, $x_i \in X, i = 1,2,\dots,n$;

Y: Output collection of hidden layer, $y_i \in Y, j = 1,2,\dots,h$;

Z: Output collection of output layer, $z_k \in Z, k = 1,2,\dots,m$;

T: Output collection of the target, $t_k \in T, k = 1,2,3,\dots.,m$;

$\hat{X}$: Output collection of the target, $\hat{x}_i \in \hat{X}, i = 1,2,\dots,n$;

$\hat{T}$: Model prediction energy set of samples corresponding to the test set, $\hat{t}_k \in \hat{T}$

$\hat{Z}$: The actual energy set of the sample corresponding to the test set, $\hat{z}_k \in \hat{Z}$

D: Set of difference between actual energy and model predicted energy, $d_i \in D$

W: The weights of connecting neurons i and j, $w_{ij} \in W$

x_j: The state of neuron is 1if it is activated, otherwise it is 0 or -1

ε: Error between predicted energy and actual energy; α: Learning rate

III. ALGORITHM DESIGN

A. GA-BP neural network

Back-ProPagation Network (BP-NN), also known as back-propagation neural network, through the training of sample data, constantly modifies the weights and thresholds of the network, so that the error function decreases along the negative gradient direction and approaches the expected output. It is a widely used neural network model, which is mostly used for function approximation, model recognition and classification, data compression and time series prediction. Its structure is shown in Figure 1.

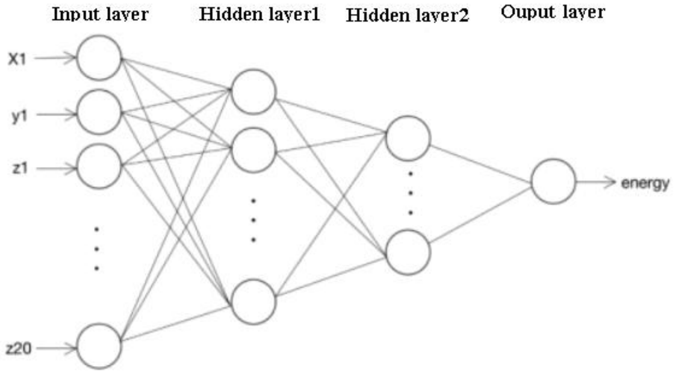


Figure 1: Neural network structure

Output unit of hidden layer y_j: $y_j = f(\sum_{i=1}^{n} w_{ij}x_i - \theta) = f(\sum_{i=0}^{n} w_{ij}x_i), w_{0j} = \theta, x_0 = -1$

Output unit of output layer z_k: $z_k = g(\sum_{j=0}^{h} w_{jk}y_j)$

Error between predicted energy and actual energy ε: $\varepsilon = \frac{1}{2}\sum_{k=1}^{m}(t_k - z_k)^2$

Iterative formula of weight: $w_{ij}(t+1) = w_{ij}(t) + \alpha(d_i - y_i)x_i(t)$

Optimizing BP neural network by genetic algorithm mainly includes three parts: determining the structure of BP neural network, optimizing the initial weights and thresholds of BP neural network by genetic algorithm, and forecasting by BP neural network. The specific steps of the algorithm are as follows:

1. Establish the structure of BP neural network, and randomly initialize the weights and thresholds in the model.
2. Individual coding and determination of population size. Individual adopts real coding method, and individual coding length s = nm+mh+h+m; There are many variables involved in the study, so the population size NP=45 is set.
3. Determination of fitness function. Individuals represent the initial weights and thresholds of BP neural network model. The purpose of the study is to make the prediction error of the output layer of GA-BP model as small as possible, so the fitness function is set as the reciprocal of the sum of squares of errors of the output layer of the model.
4. Individual selection, replication and variation.
5. circulating the above steps until the requirements are met or the maximum number of iterations is reached; Decoding the optimal chromosome, obtaining the optimal weights and thresholds, and giving them to BP neural network model.
6. Training the model and calculating the prediction error; Combined with the error, the weights and thresholds of the model are updated by using the training function.
7. When the error reaches the target requirement or the training frequency reaches the maximum number, the training is finished, and the model is used for prediction.

B. GA-BP-TS algorithm

In order to search and predict the global optimal structure of Au20 gold cluster better, the global optimization algorithm and machine learning methods are combined to train the relationship between cluster structure and energy, so as to predict the global optimal structure of Au20 cluster. In the research field of natural computing, tabu search algorithm avoids circuitous search with its flexible storage structure and corresponding tabu criteria, which is a global iterative optimization algorithm with strong local search ability. Therefore, this paper combines tabu search algorithm with improved BP neural network algorithm based on genetic algorithm, and designs GA- BP-TS heuristic algorithm to find the global optimal structure and energy of Au20 cluster. GA-BP-TS algorithm steps can be described as follows:

1. given the parameters of tabu search algorithm, randomly select one isomer with different structures from known Au20 cluster as the initial solution, and empty the tabu list; The taboo length is 8,(L= $\sqrt{n}$, n is the scale of the problem, and n = 60 in this topic); The number of neighborhood solutions Ca=1000 and the maximum iteration times G=200.
2. neighborhood function: x_near (I) = x_temp(i)+(2*rand-1)*w*(xu-xl); Adaptive value function: f=min(prediction result of ga-BP model); X_temp is the current solution, x_near is the neighborhood solution generated by the current solution, w is the adaptive weight coefficient, and xu and xl are the upper and lower limits of variables respectively.
3. Substituting the initial solution into the GA-BP prediction model, predicting the adaptation value of the initial solution, and recording it as the current optimal solution and the current solution; According to the neighborhood function, 500 neighborhood solutions of the current solution are generated, their adaptation values are calculated, and the optimal solution is set as the candidate solution.
4. judging whether the candidate solution is superior to the current solution: if the current solution is not improved, assigning the candidate solution to the current solution of the next iteration and updating the tabu list; If the candidate solution is better than the current solution, judge whether it is better than the current optimal solution; if it is better than the current optimal solution, assign the candidate solution to the current solution and the current optimal solution of the next iteration, and update the tabu

list; If the candidate solution is not superior to the current optimal solution, judge whether the candidate solution is in the tabu list, if so, regenerate the neighborhood solution with the current solution, and repeat step (4).

5. judging whether the termination condition is met: if yes, ending the search process and outputting an optimized value; If not, continue iterative optimization. The flow chart of GA-BP-TS algorithm is shown in the figure.

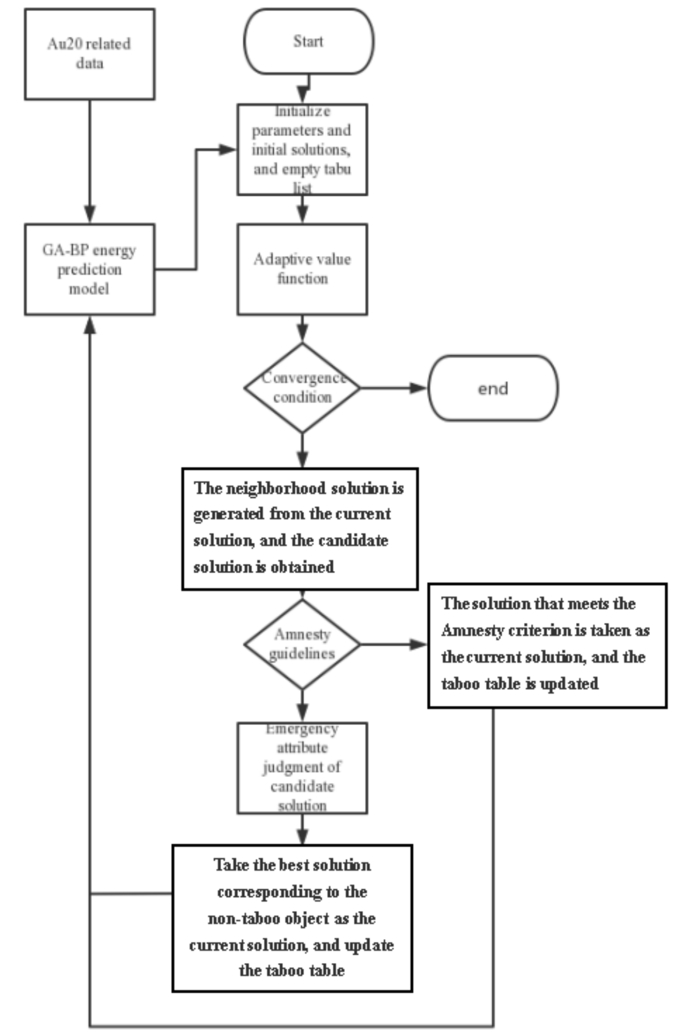

Au20 related data
Start
Initialize parameters and initial solutions, and empty tabu list
GA-BP energy prediction model
Adaptive value function
Convergence condition
end
The neighborhood solution is generated from the current solution, and the candidate solution is obtained
Amnesty guidelines
The solution that meets the Amnesty criterion is taken as the current solution, and the taboo table is updated
Emergency attribute judgment of candidate solution
Take the best solution corresponding to the non-taboo object as the current solution, and update the taboo table

Figure 2: Flow chart of GA-BP-TS algorithm

IV. DATA PROCESSING AND MODEL SIMULATION ANALYSIS

A. Data processing

The isomer related data of some Au20 is known, and the data includes 20 atomic coordinates and energy information. Since the isomer information of each Au20 is scattered in different files, it needs to be preliminarily processed, and the existing atomic coordinate information of 1000 Au20 gold clusters is sorted into a table, and then it is trained. The specific steps are as follows:

1. reading the data in xyz file; There are three parts of data in xyz file, the number of atoms in the first row, the energy of the second row and the remaining rows are all atomic position coordinate information, with one atom in each row; Use the read () function to read the file information, and use the split () slicing function to slice the data separated by spaces. Finally, classify them and save them in excel table for the next training.
2. carry out information statistics on the excel table initially obtained in the first step; The necessary information such as the minimum and maximum energy and the range of atomic coordinates are counted, and the data are standardized to prepare and reference for the subsequent establishment of the model.

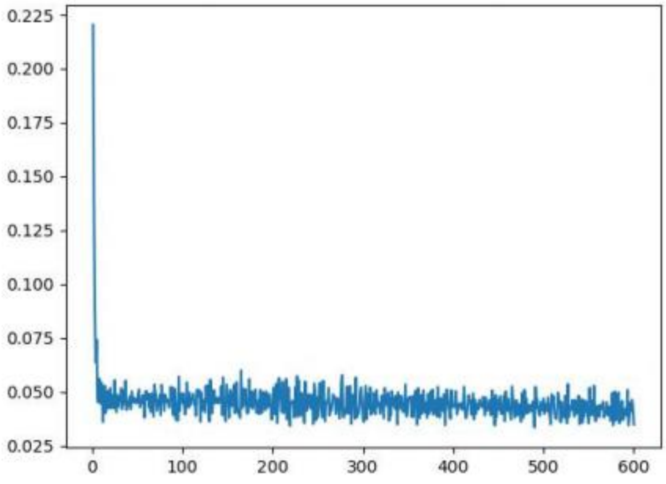
B. Analysis of simulation results

A total of 999 sample data were collected, of which 900 data were selected for model training, and the remaining 99 data were used as test sets to verify the prediction accuracy of GA-BP model. Average absolute percentage error (MAPE) and root mean square error (RMSE) were used to evaluate the prediction performance of the model. The calculation formulas are shown in formulas (1)-(2) respectively.

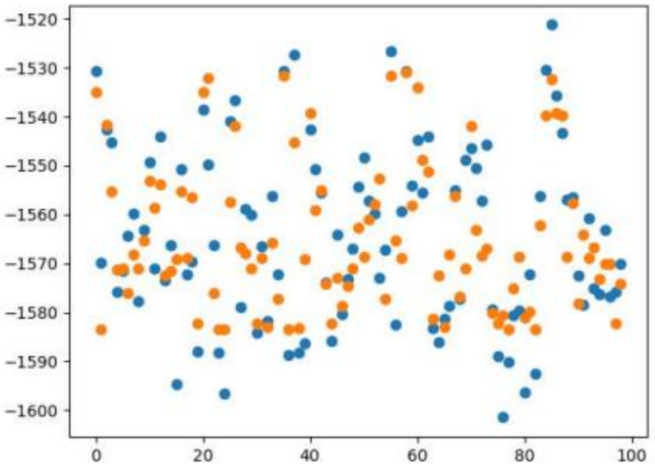
$$MAPE = \sum_{k=1}^{n} \frac{1}{n} \frac{\left|z^k - t^k\right|}{z^k} \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n} (t^k - z^k)^2}{n}} \qquad (2)$$

The change process of training loss of BP neural network is shown in Figure 3 (left), in which the horizontal axis is the number of training rounds and the vertical axis is MSE loss. The number of training sets is 900, the batch-size is 300, and a total of 200 epoch are conducted. The MSE error after data normalization is maintained at about 0.04. Fig. 3 (right) shows the comparison between the predicted results and the original data. The number of test sets is 99. The energy values predicted by the model are shown as blue dots, and the energy values of the original data are shown as orange dots. It can be seen from the images that the predicted values are basically maintained in the same energy range as the actual values. The experimental results show that MAPE = 0.012357； ; MASE = 3.4512, the prediction performance of GA-BP model is better, but there are still some errors.



(a)



(b)

Figure 3: Training loss process diagram of ga-BP neural network (a) and comparison diagram between predicted results and actual results (b)

Based on this, the atomic coordinate information of the global optimal structure of Au20 cluster obtained by GA-BP-TS algorithm is shown in Table 1, and its energy is -1532.7. Figure 4 is a visual graph.

Table 1: Prediction results of optimal structure of Au 20 gold cluster

团簇	x	y	z	能量
Au	0.95069415	-0.995015	0.97293995	-1532.7
Au	-0.95069414	0.995015	0.97293995	
Au	-0.99501488	-0.95069403	-0.97294011	
Au	0.99501488	0.95069403	-0.97294011	
Au	-0.9549562	-0.91280836	2.83755584	
Au	0.9549562	0.91280836	2.83755584	
Au	-0.91280835	0.9549562	-2.83755569	
Au	0.91280835	-0.9549562	-2.83755569	
Au	0.99826484	2.81536038	0.93395014	
Au	-0.99826484	-2.81536039	0.93395014	
Au	2.81536027	-0.99826484	-0.93395014	
Au	-2.81536027	0.99826484	-0.93395014	
Au	-0.86915009	2.85788332	-0.93460113	
Au	0.86915009	-2.85788332	-0.93460113	
Au	2.85788344	0.86915008	0.93460114	
Au	-2.85788344	-0.86915008	0.93460114	
Au	2.71205222	-2.83756085	-2.77547148	
Au	-2.71205222	2.83756085	-2.77547148	
Au	-2.83756085	-2.71205222	2.77547148	
Au	2.83756085	2.71205222	2.77547148	

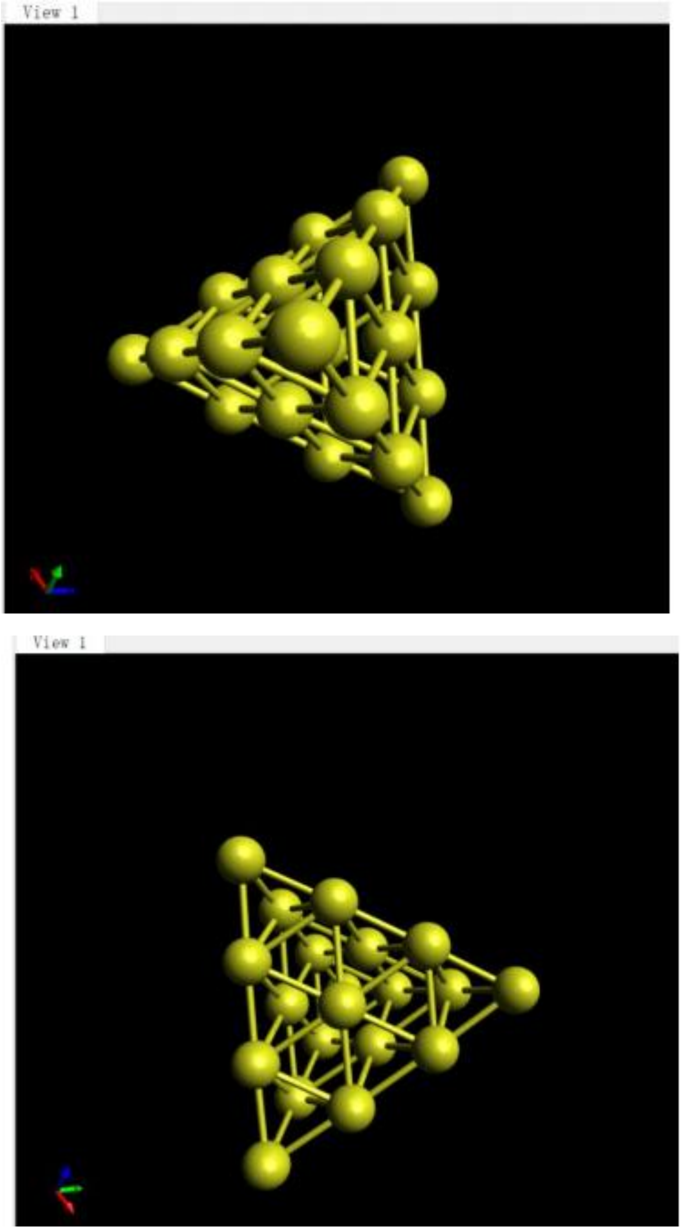


Figure 4: Optimal structure diagram of au20 gold cluster

References

[1] Doye JPK. Physical perspectives on the global optimization of atomic clusters. In: Pinter JD, eds. Global Optimization: Scientific and Engineering Case Studies. Berlin: Springer-Verlag, 2006, 103–139.

[2] Cheng LJ, Feng Y, Yang J, Yang JL. Funnel hopping: Searching the cluster potential energy surface over the funnels. J Chem Phys, 2009, 130: 214112.

[3] Fa W, Luo CF, Dong JM. Bulk fragment and tubelike structures of AuN (N = 2–26). Phys Rev B, 2005, 72: 205428.

[4] Wales DJ, Doye JPK. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A, 1997, 101: 5111–5116

[5] Leary RH. Tetrahedral global minimum for the 98-atom Lennard-Jones cluster. Phys Rev E, 1999, 60: R6320–R6322.

[6] Takeuchi H. Clever and efficient method for searching optimal geometries of Lennard-Jones clusters. J Chem Inf Model, 2006, 46: 2066–2070.

[7] Shao XG, Cheng LJ, Cai WS. A dynamic lattice searching method for fast optimization of Lennard-Jones clusters. J Comput Chem, 2004, 25: 1693–1698.

[8] Locatelli M, Schoen F. Efficient algorithms for large scale global optimization: Lennard-Jones clusters. Comput Optim Appl, 2003, 26: 173–190.

[9] Doye JPK, Leary RH, Locatelli M, Schoen F. The global optimization of Morse clusters by potential energy transformations. INFORMS J Comput, 2004, 16: 371–379.