Artificial Immune Network Clustering Based on a Cultural Algorithm

Liyuan Deng (dly198804140033@hotmail.com)
Xi' an Research Institute of High-Technology
Ping Yang
Xi'an Research Institute of High-Technology
Weidong Liu
Xi'an Research Institute of High-Technology

Research

Keywords: Clustering, Immune network, Culture algorithm, Immune defense, Shadow set theory

Posted Date: June 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-21168/v2

Version of Record: A version of this preprint was published on September 3rd, 2020. See the published version at https://doi.org/10.1186/s13638-020-01779-1.

Artificial Immune Network Clustering Based on a Cultural Algorithm

Liyuan Deng[1*], Ping Yang[1], Weidong Liu[1]

1 Xi'an Research Institute of High-Technology, China

{ dly198804140033@hotmail.com, 1689773440@qq.com, 2673826919@qq.com }

Corresponding author: Liyuan Deng

Abstract: Data mining technology has been applied in many fields. Prototype-based cluster analysis is an important data mining method, but its ability to discover knowledge is limited because of the need to know the number of target data categories and cluster prototypes in advance. Artificial immune evolutionary network clustering is a clustering method based on network structure. Compared with prototype-based cluster analysis, it has the advantage of realizing unsupervised learning and clustering without any prior knowledge of data. However, artificial immune evolutionary network clustering also has problems such as a lack of guidance in the clustering process, fuzzy boundary sensitivity, and difficulty in determining parameters. To solve these problems, an artificial immune network clustering algorithm based on a cultural algorithm is proposed. First, three kinds of knowledge are constructed: normative knowledge is used to regulate the spatial range of population initialization to avoid blindness; state knowledge is used to distinguish the type of antigen, and immune defense measures are taken to prevent the network structure caused by noise and boundaries from being unclear; topology knowledge is used to guide the antigen for optimal antibody search. Second, topology knowledge in the cultural algorithm is used to characterize the distribution of antigens and antibodies in space, and elite learning is used to improve the traditional clone mutation operator. Based on the shadow set theory, a method for adaptively determining the compression threshold is proposed. Finally, the results of simulation experiments show that the proposed algorithm can effectively overcome the above problems, and the clustering performances on a synthetic dataset and an actual dataset are satisfactory.

Keywords: Clustering; Immune network; Cultural algorithm; Immune defense; Shadow set theory

1. Introduction

Data mining is the process of discovering specific information hidden in massive databases through various algorithms [1]. With the accumulation of massive data brought by the development of information technology, the use of data mining technology to transform these data into useful information has been used in various fields [2-3]. In the field of cloud services, work [4] proposed a distributed cloud service method based on distributed sensitive hashing in multisource data. Work [5] proposed a big data-driven mashup building method that supports economic software developments. In the field of the Internet of Things, work [6] proposed a

multidimensional data processing and query method, work [7] studied IoT offloading utilities that support edge computing. In the field of business services, Bismita S. Jena [8] and others analyzed the business information mining technology used on transaction datasets. Zhang et al. [9-11] researched related data mining techniques such as business service recommendation and service quality query, Chen et al. [12-13] researched data for business intelligence and business service computing. Data mining technology is also widely used in e-commerce, media, energy services, automotive engineering and other fields [14-19].

Clustering analysis is an important method of knowledge discovery in data mining [20-22]. If we know the number of target data categories and clustering prototypes in advance, we can use a prototype-based cluster analysis method, but this will inevitably limit the ability of clustering analysis to discover knowledge. Therefore, a clustering method without any prior knowledge of data is of course preferable for clustering analysis [23-25]. Based on the artificial immune evolutionary network (aiNet) clustering algorithm, this paper proposes a new algorithm of artificial immune network clustering based on a cultural algorithm. It guides the aiNet clustering process by constructing normative knowledge and topological knowledge in the trust space and introduces the principle of immune defense in the human immune system into the algorithm. In this paper, the input taboo threshold is avoided, and the shadow set theory is used to realize the adaptive determination of the compression threshold.

The algorithm is called the cultural evolutionary artificial immune network (CaiNet). The algorithm uses topological knowledge in the cultural algorithm to characterize antigens and antibodies in space [26-27]. When an antigen searches for the antibody with the highest affinity, it searches through the antibodies in the topological unit where it is located. The algorithm uses immune defense to improve the flexibility of the application and improves the traditional clone mutation operator through elite learning.

Compared with the aiNet algorithm, the CaiNet algorithm has higher average accuracy and smaller variance. In the simulation experiment, the seed dataset is selected as the experimental object. The accuracy rate of the CaiNet algorithm is improved by 5.8%, and the variance is reduced by 3.71%. When the Wine dataset is selected as the experimental object, the accuracy rate of the CaiNet algorithm reaches 98.78%. The CaiNet algorithm has a certain improvement in balance, accuracy, recall rate, and hit rate and has better convergence.

The main innovations of this article are summarized as follows:

(1) We propose using the cultural algorithm to guide the aiNet clustering algorithm and use the topology knowledge in the cultural algorithm to characterize the distribution of antigens and antibodies in the space, which greatly reduces the complexity of the algorithm search.

(2) Based on the concept and theory of the shadow set, we propose a method for adaptive determination of the compression threshold based on the shadow set, which improves the ability of the algorithm to quickly solve the algorithm.

(3) Drawing on the immune defense suppression measures adopted in medicine to avoid excessive epidemic prevention, we propose a new algorithm immune defense mode, which improves the flexibility of algorithm application.

The organizational structure of the paper is as follows. First, we discussed related work in Section 2. Subsequently, we introduce the structure of the CaiNet algorithm in Section 3, define three kinds of knowledge in Section 3.1, design and improve the operation operator in Section 3.2, simplify the optimal antibody search for antigens through topology knowledge steps, formulate the mutation rules of the algorithm, propose a method of adaptive determination of the compression threshold based on the shadow set and formulate the immune defense criteria of the CaiNet algorithm. The algorithm steps of the CaiNet algorithm are summarized in Section 4. Finally, we select three types of datasets in Section 5 for simulation experiments and evaluate the stability and convergence performance of the CaiNet algorithm.

2. Related Work

Artificial immune evolutionary network clustering is a clustering method based on network structure [28]. Compared with the prototype clustering method, it can achieve real unsupervised learning and clustering. Based on the basic aiNet algorithm, Li Jie et al. introduced the concept of taboo cloning in immunology to the artificial immune network clustering algorithm, which solved the problem that aiNet cannot handle the fuzzy boundary of the sample subset [29]. Considering the problem of memory network dynamics and irregular changes caused by the lack of objective function guidance of the aiNet algorithm, Guo Jianhua et al. established the overall objectives and constraints of the memory network by defining quality evaluation standards, thus realizing the guidance of the algorithm, and discussed the value of the compression threshold [30]. To overcome the problem that the monoclonal algorithm easily falls into local optimum, Zhou Yang et al. proposed an evolutionary immune network clustering algorithm based on polyclonal algorithms [31]. Ma Li et al. applied a variety of artificial immune system operators to the clustering process. Based on the basic principles of biological immunity and cloning, they proposed an adaptive multiclone clustering algorithm that automatically adjusts clustering categories by setting affinity functions to increase the antibody population diversity of individuals to expand the search range of the solution and avoid precocity of the algorithm [32]. It has been found in experiments that for unbalanced datasets, clusters of small samples are easily undetectable when using a large taboo threshold. However, the improved aiNet algorithm does not have a unified understanding of the death threshold, compression threshold, and taboo threshold to be input. In many cases, it needs to be determined according to the characteristics of the data, which makes the algorithm more difficult to apply.

This paper defines three kinds of knowledge in the CaiNet algorithm: normative knowledge, topology knowledge and state knowledge. Normative knowledge provides a code of conduct for evolution, topological knowledge easily guides the expansion of the network in different spaces, and state knowledge is used to control the strength of

antigen activation networks in different states. In this paper, the topology unit is used to form the topology knowledge in the cultural algorithm, which simplifies the optimal antibody search step and formulates new mutation rules, which overcomes the limitations of the traditional algorithm. The determination of the compression threshold is the difficulty of most algorithms. Based on the concept and theory of the shadow set, this paper proposes an adaptive determination method for the compression threshold based on the shadow set, which is conducive to the rapid solution of the algorithm. To avoid the unclear structure of the immune network caused by the boundary data in the traditional algorithm, it may prevent the network from accurately expressing the distribution of antigens so that it does not activate the immune network. This article refers to the immune defense suppression measures taken to avoid excessive defense in medicine. For the noise, the boundary and the antigen inside the cluster, three different methods are used to treat them differently.

To test the effect of the new algorithm, we select three UCI datasets as the experimental objects and compare the average accuracy and variance in the algorithm. The experimental results show that the stability and convergence of the new algorithm and the performance significantly improve.

3. Method

Cultural algorithms use trust space and population space for double-layer evolution. The population space forms different types of knowledge through trial and error in the processing of trust space and then guides the evolution of the population space. The designed algorithm structure is shown in Fig. 1:

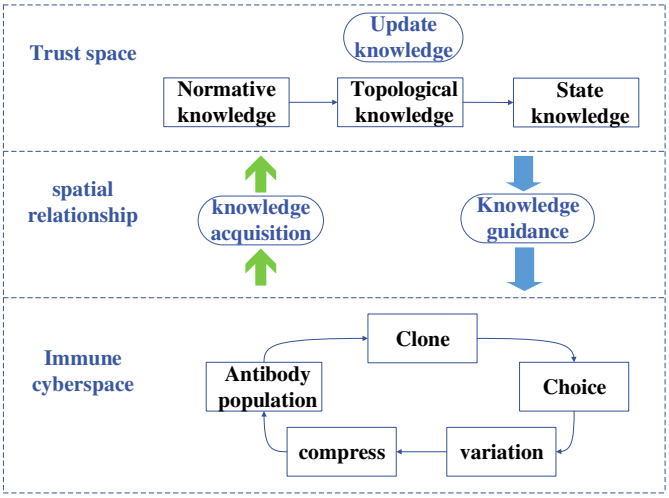

Fig. 1 AiNet clustering principle based on a cultural algorithm.

3.1 Background Knowledge

In this algorithm, three kinds of knowledge are defined. Normative knowledge defines the interval range of the antigen and each generation of antibodies and provides a behavioral rule for evolution. Topological knowledge expresses the distribution of antigens and antibodies in the search space and provides opinions and

recommendations for immune recognition suggestions that are helpful to guide the expansion of the network in different spaces. State knowledge records the different states that the antigens may be in and is used to control the strength of the antigen activation network in different states.

Definition 1. Antibody-antigen affinity. Antibody-antigen affinity is the measurement of affinity between the antibody and antigen and is described in detail in formula (1).

$$f(\mathbf{g},\mathbf{b}) = \frac{1}{1+\|\mathbf{g}-\mathbf{b}\|} \tag{1}$$

In the formula, $\|\cdot\|$ represents the Euclidean distance, G represents the antigen collection, and g_i represents a single data sample. B_k represents the immune network, that is, the antibody collection, and $b_{k,j}$ represents the k antibody in the j-th network.

Definition 2. Antibody-antibody affinity. Antibody-antibody affinity is expressed by the Euclidean distance $d_{i,j}$ between the antibodies, which can form the affinity matrix $D_k=(d_{i,j})_{N_k\times N_k}$ of the network, and N_k represents the number of antibody neurons in the *k-th* network.

Definition 3. Clone operation. The clone operation selects a part of the antibody with a high affinity to copy. For antibody b_i, the clone operation can be expressed as:

$$C(b_i) = [b_{i,0}, b_{i,1}, \cdots, b_{i,n-1}], \quad n = Int\left(N_c \times \frac{A_i}{\sum_{j=1}^{N} A_j} \right) \tag{2}$$

where N_c represents the total antibody size after cloning, A_i represents the affinity of the i antibody, and N represents the number of antibodies participating in the clone.

Definition 4. Normative knowledge. Normative knowledge records the spatial range of antibody production; one range is the value interval of each dimension of the antigen, and the other is the value interval of each dimension of the memory network antibody neuron, which is represented by N_0 and N_t, and its formal definition is:

$$N_0 = \{ l_1, u_1; l_2, u_2; \cdots; l_n, u_n \} \tag{3}$$

$$N_t = \{ l_1^t, u_1^t; l_2^t, u_2^t; \cdots; l_m^t, u_m^t \} \tag{4}$$

where l_i represents the lower bound, u_i represents the upper bound, and i represents the *i-th* dimension. N_0 is static knowledge and does not change throughout the clustering process; N_t is dynamic knowledge, which changes with each network change. The superscripts of l_i^t and μ_i^t in N_t represent the number of iterations.

In the internal image of antigens, antibody neurons are generally distributed in the space determined by all antigens; therefore, the antibody population should be within the space determined by N_0 during initialization. As the network evolves, it should gradually converge because such a network is more refined and clustering is more obvious. To achieve this goal, N_t is used to guide the initialization of the antibody population. When the population is initialized, most of the antibodies (80%) are generated in the specified space. To avoid a suboptimal algorithm solution, some of the antibodies (20%) are also generated in the residual set of N_t relative to N_0, forming a disturbance and preventing the network from falling into the local optimal solution.

Definition 5. Topology knowledge. The topological unit refers to the hypergeometric region with an antibody as the center and l_j as the edge length of the j-th dimension**.** The knowledge about antibody and antigen features contained in all topological units is called topology knowledge. The topological unit represented by antibody b_j can be expressed as:

$$T_i = \left\{ b_{i,1} - \frac{l_1}{2}, b_{i,1} + \frac{l_1}{2}; b_{i,2} - \frac{l_2}{2}, b_{i,2} + \frac{l_2}{2}; \cdots; b_{i,m} - \frac{l_m}{2}, b_{i,m} + \frac{l_m}{2} \right\} \tag{5}$$

In the formula, $b_{i,j}\text{-}\frac{l_j}{2}$ and $b_{i,j} + \frac{l_j}{2}$ represent the upper and lower bounds, respectively, of the topological unit represented by the antibody on the first dimension. By calculating the coordinates of the antibody and antigen in space, we can determine whether it belongs to a topological unit. If an antigen g_j belongs to a topological unit T_i**,** it is recorded as $g_j \in T_i$. There may be intersections between topological units. When an antigen belongs to more than one topological space, the distance between the antigen and the center of the topological unit is calculated, and the smallest distance is taken as the topological unit of the antigen. Due to the distribution of antibodies, some antigens may not be in any topological units. In this case, the distance between antigens and the center of all topological units is calculated, and the one with the smallest distance is taken as the topological unit.

When the topological unit is determined, the antigen can be mapped into the topological unit. Since antibody b_i is the center of topological unit T_i, and according to the principle of immune network clustering, the antibody is the inner image of the network. Therefore, we call antibody b_i the representative point of the antigen contained in topological unit T_i. In particular, when the topological unit does not contain any antigens, it is deleted.

Definition 6. State knowledge. Without losing generality, the data in the dataset are divided into noise, boundary and cluster internal points, and state knowledge is used to record the different antigen states.

Topological elements can be regarded as grids with knowledge characteristics. According to the existing grid-based clustering methods, noise and boundary points (including fuzzy boundaries) are significantly different from the data within the cluster. It has been found that the noise and boundary points of the data include but are not limited to the following features: the area where the noise and boundary points are located is generally sparse; the difference between the boundary points and the class interior is that the latter often has close neighbors in multiple directions, that is, the uniformity is relatively good; the density of the area where the boundary points are located generally has a jump. The difference between different point sets mainly lies in the density and uniformity, the noise density is small, the density at the boundary is small and uneven, and the data density inside the cluster is large and evenly distributed. The density is expressed by the number of antigens in the grid, i.e.,

$$\rho_j = \sum_{b_i \in T_j} 1 \tag{6}$$

The joint entropy method is used to measure the uniformity of the data distribution in the topological unit. For each antigen $b_{j,i}$ in the topological unit, the number of antigens is calculated in its ε neighborhood, and it is recorded as $\rho_{j,i}$, $\varepsilon = l_j / 4$. l_j is the length of the side of T_j, and

$$p_{j,i} = \frac{\rho_{j,i}}{\rho_j} \tag{7}$$

is recorded. The entropy of $b_{j,i}$ can be expressed as:

$$H_{j,i} = -p_{j,i} \log p_{j,i} \tag{8}$$

Furthermore, we can obtain the combined entropy of all antigens in T_j

$$H_j = \sum_{b_{j,i} \in T_j} H_{j,i} = -\sum_{b_{j,i} \in T_j} p_{j,i} \log p_{j,i} \tag{9}$$

The data can be divided into 3 categories according to prior knowledge, so this is a 2-dimensional clustering problem with a known number of categories, which can be solved well using methods such as fuzzy C-means. After the clustering is completed, the antigens in the corresponding topological units can be labeled as noise, boundary points, and cluster internal data.

When a data point is marked incorrectly, the algorithm may be guided in the wrong direction. The distribution of antibodies has randomness, and a clustering algorithm is not always effective. Therefore, misclassification always occurs. To avoid the impact of this situation, the idea of evidence accumulation is introduced. Evidence accumulation refers to adding 1 to the evidence value of an antigen if it is labeled in the same state in the adjacent time sequence, and 1 is subtracted from the evidence value if it is labeled in different states in the adjacent time sequence. Because of the

randomness of the antibody, this can greatly reduce the impact of misclassification. According to the above methods, state knowledge can be expressed as:

$$S = \{ S_1, D_1; S_2, D_2; \cdots; S_i, D_i; \cdots; S_n, D_n \} \tag{10}$$

where S_i represents the state of the i-th antigen, D_i represents the evidence of the state, and D_i is equal to 1 at the initialization phase.

3.2 AiNet Clustering Based on a Cultural Algorithm

3.2.1 Optimal Antibody Search

We use topological units (hypergeometry) to form topology knowledge in cultural algorithms. Topology knowledge includes two parts: antigen and antibody. Therefore, we hope to simplify the optimal antibody search by topology knowledge.

According to topological knowledge, antibodies can be regarded as representative points of antigens in antibody units, and the distance between antibodies with a high affinity and their representative points should be small. Therefore, we can first use a representative point antibody to find the $k'>k$ antibody with the smallest distance, then calculate the affinity between the k' antibody and antigen, and take the k antibody with the highest affinity as the optimal k antibody. Its pseudocode is:

Algorithm 1 Find the best antibody

Input: g_i, b_j, k', k

Output: The optimal antibody set

1: Find the b_j antibodies closest to k';

2: Calculate the affinity of g_i, and k' antibodies, and take the first k as the optimal antibody.

3: **Return** The optimal antibody set

For other antigens belonging to T_j, only Step 2 is needed to find the optimal K antibody, which can greatly reduce the complexity.

The value of k' should be greater than k because there is a certain distance deviation between b_j and g_i. In practice, the greater the difference between k' and k, the more accurate the results obtained, and the cost is the expansion of the search range. Considering the uniformity of antibody distribution in the network, k' is generally taken as $3k/2$.

3.2.2 Elite Learning Variation

In traditional aiNet clustering, antibody improvement is achieved by clone variation, expressed as

$$b_j = b_j - \alpha\left(b_j - g_i\right) \tag{11}$$

where α represents variability, and the value decreases with increasing b_j and g_i affinity. Formula (11) improves the antibody by reducing the distance between antibody b_j and antigen g_i, but this method still has some limitations, such as b_j being only close to the antigen and not focusing on learning from other antibodies. To make the target antibody obtain the advantage information of outstanding antibodies at the same time, the following variation rules are formulated:

$$b_j = b_j - \alpha\left[r_1\left(b_j - g_i\right) + r_2\left(b_j - b_0\right) \right] \tag{12}$$

where b_0 represents the antibody with the highest affinity with g_i. In the current network, r_1 and r_2 are weighting factors, meeting the requirement of $r_1 + r_2 = 1$; if $b_0 \equiv b_j$, $r_1 = 1$. In fact, when $r_1 = 1$, it degenerates into the mutation strategy of a traditional algorithm.

3.2.3 Compression Threshold Determination

There is no unified understanding of how to set the compression threshold. The general guidance is to take a very small compression threshold first, for example, 10^{-3}, and gradually increase it with the change in the network. There is little discussion on this in the existing literature. According to the concept and theory of shadow sets, we propose an adaptive method to determine the compression threshold.

Shadow sets is a theory proposed by Pedrycz to address fuzzy problems, in which set levels 1,0 and [0,1] are used to describe and simplify fuzzy relationships. The sample points corresponding to level 1 belong to a set completely, [0,1] indicates whether the sample point belongs to a set or not. The 0 corresponding sample point does not belong to a collection at all. The above three levels correspond to the complements of the lower approximation, upper approximation and lower approximation relative to the upper approximation.

The purpose of network compression is to improve the affinity between antibody and antibody, that is, to increase the distance between antibodies and prevent network redundancy caused by a small distance. The smaller the distance, the more likely it is to be compressed, and the greater the distance, the more likely it is to not be compressed. Without losing generality, we use the normalization of distance to express the possibility membership degree of whether the antibody should be compressed. The possibility membership of whether the antibody should be compressed is defined as the mapping of the distance between the antibody and the antigen to the [0,1] closed interval, expressed by the formula:

$$u_{i,j} = \frac{d_{i,j} - d_{\min}}{d_{\max} - d_{\min}}, \quad i, j = 1, 2, \cdots, n \tag{13}$$

The objective function is defined as:

$$\arg\min_{\alpha} F(\alpha) = \left| \xi_1 + \xi_2 - \xi_3 \right|, \quad \alpha \in (0, 0.5] \tag{14}$$

where a is in the range of (0,0.5], $\xi_1 = \sum_{u_{i,j} \le \alpha} u_{i,j}$, $\xi_2 = \sum_{u_{i,j} \ge 1-\alpha} \left(1 - u_{i,j}\right)$, and $\xi_3 = card(I)$ represents modulo set A, and $I = \left\{ i \middle| \alpha < u_{ij} < 1-\alpha \right\}$. When the α value is determined, the part of the antibody satisfying $\mu_{i,j} \le \alpha$ needs to be compressed.

Obviously, according to this threshold determination method, a certain number of antibodies are compressed each time, which is consistent with the actual situation of network compression in the algorithm. In addition, $F(\alpha)$ is a simple step-like unimodal function that can be quickly solved by methods such as dichotomy.

3.2.4 Immune Defense

According to the traditional aiNet clustering method, regardless of the nature of the antigen, the antibody can generate an immune response and then activate the antibody network. This is the main reason for the unclear structure of the immune network due to "abnormal" data such as noise and fuzzy boundaries.

The immune defense mechanism means that the immune system can attack, destroy and clear “alien components” such as bacteria, viruses and foreign bodies, which is a very important protection mechanism for the human body. We simulate this process in the algorithm.

To defend against “alien elements”, we must first identify the “alien elements” according to the state knowledge constructed in the cultural algorithm. It is convenient to determine the “alien component”, that is, the parts marked as the noise and boundary in the state knowledge.

In the clustering problem, because the boundary data easily cause the immune network structure to be unclear, it does not activate the immune network, which creates the problem that it may make the network unable to accurately express the distribution of the antigens.

To avoid this problem, three different methods are adopted to treat the antigens in noise, boundary and cluster according to the immune defense inhibition measures taken to avoid overdefense in medicine, namely,

$$\begin{cases} g_i \in S^0, & \text{Clonal dominant antibody selection and variation} \\ g_i \in S^1, & \text{dominat antibody selection and mutation} \\ g_i \in S^2, & \text{do not operate} \end{cases} \tag{15}$$

where S^0, S^1 and S^2 represent the interior, boundary and noise antigen set of the cluster, respectively, and the noise and boundary points are defended differently by the immune defense mechanism guided by state knowledge. If noise is no longer involved

in the immune process, it is eliminated directly. Boundary points do not participate in the process of cloning to avoid the generation of a large number of cloned antibodies at the boundary and prevent the blurring of network structure at the boundary. The reason why boundary points participate in the selection and variation is to avoid the excessive movement of antibodies to the clustering center, resulting in a lack of affinity between the boundary and antibody network, thus leading to the problem of boundary point misclassification.

3.3 Specific Steps

For the final immune network, the minimum spanning tree is generated according to its connected graph. There is a larger weight between the representative antibodies of two different clusters. According to the set pruning threshold, the m connections with larger weights are removed so that m+1 clusters can be obtained. The steps of the CaiNet algorithm are shown as:

Algorithm 2 Cultural evolution immune network clustering

Input: G is the antigen set; N_t is dynamic knowledge; N_0 is static knowledge; $l_{\max}$ is the number of iterations; N_C is the cloning scale; σ_d is the death threshold

Output: the memory network M

Initialize: $N_t \leftarrow T_i \leftarrow S_i \leftarrow \emptyset$, $l \leftarrow 1$

Update topology knowledge and state knowledge
$G_1 \leftarrow$(Recognize the alien for immune defense and get "self")
$B_1 \leftarrow$(Recognize the alien for immune defense and get effective antibody set)
For $g_i \in G_i$ **do**
    $k \leftarrow$(Get the optimal antibody of g_i according to algorithm 1)
    Sort k antibodies in descending order and perform cloning operation according to formula (2);
    Perform elite learning mutation-improving antibodies according to formula (12);
    Calculate the improved affinity, and form the temporary memory Network m_p with the k antibodies with the highest affinity;
    Natural death;
    Calculate compression threshold and perform network compression;
    $M \leftarrow [M, M_P]$
end for
    Recalculate compression threshold and perform network compression
    Update normative knowledge N_t;
    while the termination condition is not reached **do**
        $l=l+1$ and return to Step1;
    end while

After the data points in these units are eliminated, the dataset is recorded as $X = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$, and the clustering is recorded as $C_1, C_2, \cdots, C_j, \cdots, C_m$. Next, determine the type of data based on the distance between the data point and the antibody, that is,

$$d^2(x_i,b_l)=\min\left\{d^2(x_i,b_k),k=1,2,\cdots,m\right\},b_l\in C_j\Rightarrow x_i\in C_j \tag{16}$$

4. Experiments

The running configurations include hardware settings (2.70 GHz CPU, 8.0 GB RAM) and software settings (Windows 10 and Python 3.6). Each test is executed 50 times to record their average performances.

4.1 Experimental Results on a Synthetic Dataset

High-dimensional data are not easy to display intuitively, so we use a 2-dimensional synthetic dataset to verify the proposed clustering algorithm. There are three clusters in the dataset, two of which have more samples, and the other contains fewer samples. There are fuzzy boundaries between the three clusters, and they contain many instances of sample noise.

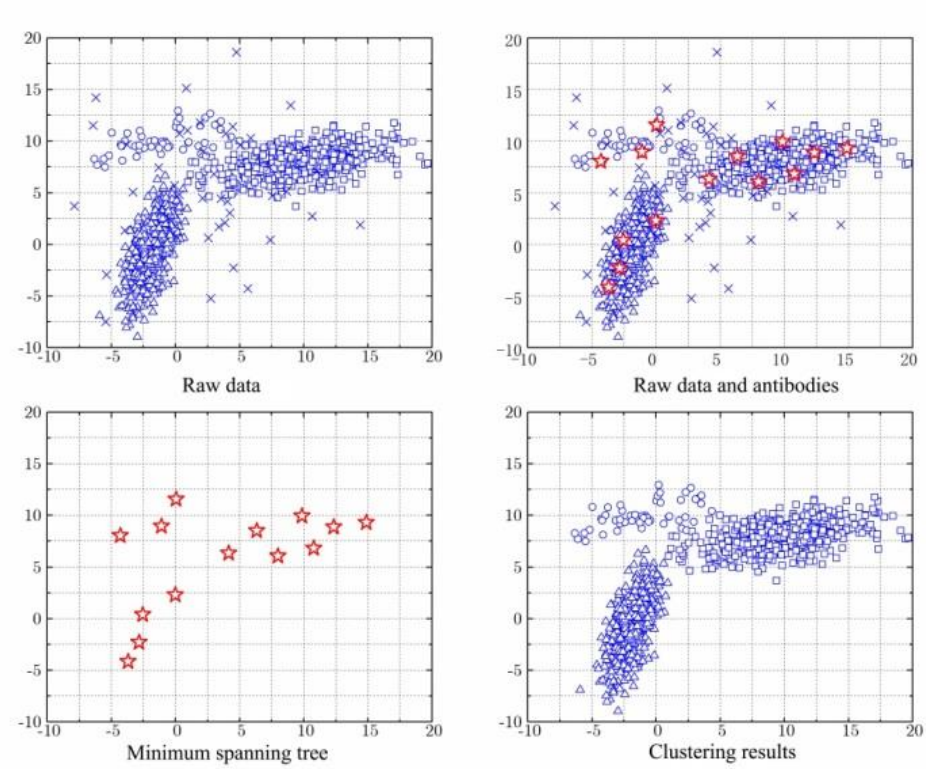

Fig. 2 Clustering effect on a synthetic dataset.

As shown in Fig. 2, the minimum spanning tree obtained by CaiNet can be divided into three distinct categories. The nodes of the tree can better reflect the data distribution of the dataset, and the nodes are relatively uniform. According to the algorithm, the last operation before obtaining the final minimum spanning tree is network compression. Therefore, the uniform distribution of the nodes is related to the selection of the compression threshold, which also shows that select the threshold using the shadow sets method is effective and can avoid the blindness of choosing a fixed compression threshold. The taboo clone method is not used in the algorithm, but the algorithm is also effective for datasets with fuzzy boundaries, indicating that the

immune defense principle can achieve the same effect as the taboo clone.

The algorithm clusters the noise, boundary and normal data and explicitly eliminate the noise. From the results, we can see that most of the noise in the data can be identified by the algorithm. Since taboo cloning is not used, the new algorithm does not need to set the taboo threshold in advance, which is very convenient and effective in practice.

4.2 Experimental Results on a Real-world Dataset

To test the clustering effect of the algorithm on actual high-dimensional data, we choose the iris, wine and seeds UCI datasets as experimental objects.

Table 1. Comparison of clustering performance between the two algorithms for three datasets.

Dataset	Sample size	Clustering number	Dimension	Algorithm	Average accuracy	Variance
IRIS	150	3	4	CaiNet	92.16	2.39
				aiNet	89.41	5.87
Wine	178	3	13	CaiNet	98.78	0.66
				aiNet	94.15	2.72
Seeds	210	3	7	CaiNet	88.24	2.51
				aiNet	82.44	6.22

The average correct rate represents the proportion of the data that the algorithm classifies in the cluster correctly. To test the stability of the algorithm, we test the specified algorithm 50 times on the datasets to obtain a variance in the accuracy after 50 times. Obviously, the smaller the variance is, the higher the stability of the algorithm. For these three datasets, the comparison between the CaiNet algorithm and the aiNet clustering algorithm is shown in Table 1. As seen from the table, in the comparative experiment results, the variance in the CaiNet algorithm is smaller than that of the aiNet clustering algorithm, and the average correct rate of the CaiNet algorithm is higher than the average correct rate of the aiNet clustering algorithm, which shows that the CaiNet algorithm is more stable. When the seeds dataset is selected as the experimental object, the variance in the CaiNet algorithm decreases the most, which is 3.71% less than the variance in the aiNet clustering algorithm. The average accuracy of the CaiNet algorithm is the highest, which is 5.8% higher than the average accuracy of the aiNet clustering algorithm. When the wine dataset is selected as the experimental object, the variance in the CaiNet algorithm is the smallest at only 0.66%; at the same time, the average correct rate of the CaiNet algorithm reaches 98.78%. It can be seen that the variance and average correct rate of the CaiNet algorithm are affected by the selected dataset, and the degree of improvement of its algorithm stability is also related to the selected dataset.

In addition, we test the algorithm convergence performances. In the running time of the simulation experiment, we choose to perform 100 simulation operations. The results are shown in Fig. 3 to Fig. 6. As seen in the figures, the CaiNet algorithm has the best balance and the highest accuracy. The CaiNet algorithm also has a higher

recall and hit rate than the other methods. Therefore, the CaiNet algorithm has better convergence performance.

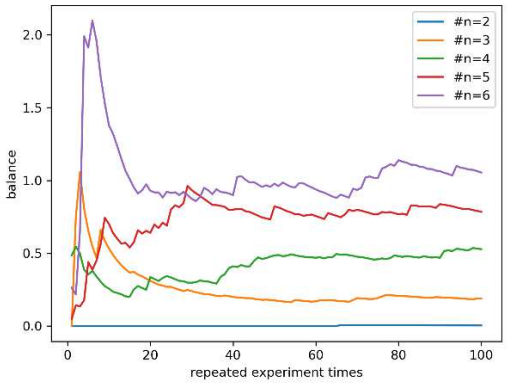

Fig. 3. The comparison of algorithm balance

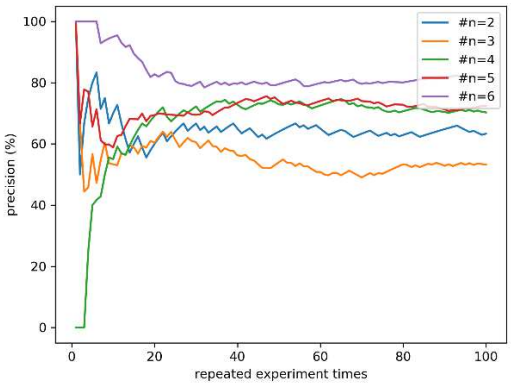

Fig. 4. The comparison of algorithm precision

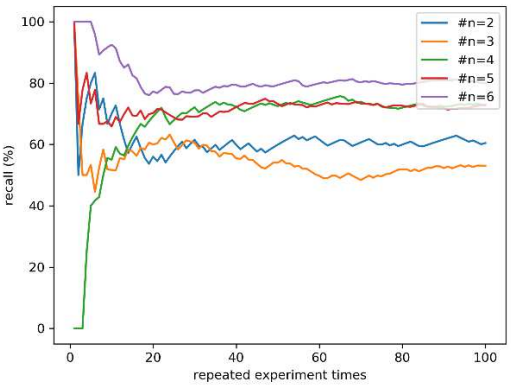

Fig. 5 The comparison of algorithm recall

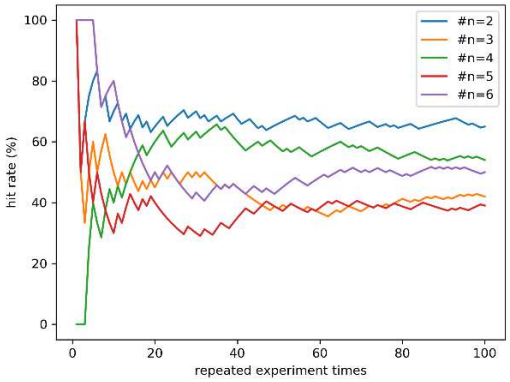

Fig. 6 The comparison of algorithm hit rate

4.3 Discussion

We tested and evaluated our proposed CaiNet method with a baseline method named aiNet to prove the advantages of our method. However, several additional points should be noted and further analyzed in detail, which are specified below.

(1) For the three compared datasets in the experiments in subsection 4.2, i.e., iris, wine and seeds, their data volumes are all not large enough (the three sample sizes are 150, 178 and 210, respectively). Therefore, in future work, we need to investigate more appropriate and larger datasets to validate the feasibility of our model and method, especially in the big data environment.

(2) Although our CaiNet method performs better than the compared baseline aiNet method, the accuracy of the CaiNet method is still not very high (92.16%, 98.78%, 88.24%). Therefore, we need to seek more efficient improvements to refine our work in this paper.

(3) Clustering is often a time-consuming task that requires a high time complexity, which is often not very suitable for the big data environment. Therefore, lightweight clustering methods are often required. We will further optimize our method to accommodate big data volume.

5. Conclusion

In this paper, cultural knowledge is used to guide the clustering of aiNet, and topology knowledge of the cultural algorithm is used to represent the distribution of antigens and antibodies in the space. Antigens only need to search using the antibodies in the topological unit when finding the highest affinity antibody, which greatly reduces the complexity. Through immune defense, the flexibility of algorithm application is improved. According to the theory of shadow sets, an adaptive method to determine the compression threshold is proposed; the traditional clonal mutation operator is improved by elite learning, which speeds up the convergence of the network. From the simulation experiment, we can see that the accuracy and stability of the improved algorithm have been improved, which proves its effectiveness.

In the future, we will continue to refine our work by considering more complex scenarios, such as multidimensional clustering problems. In addition, how to adapt our method to big data application requirements is another open question that requires intensive study.

List of Abbreviations

(1) aiNet: Artificial immune evolutionary network

(2) CaiNet: Cultural evolutionary artificial immune network

Availability of data and material

The recruited experiment dataset is the Clustering analysis of a synthetic dataset

Competing interests

We declare that there is no conflict of interest regarding this submission.

Funding

This work was supported by Xi'an Research Institute of High-Technology.

Authors' contributions

Liyuan Deng finished the algorithm and English writing of the paper. Ping Yang and Weidong Liu finished the experiments.

Acknowledgements

Not applicable.

References

[1] C. Zhang, M. Yang, J. Lv, W. Yang. An Improved Hybrid Collaborative Filtering Algorithm Based on Tags and Time Factor. Big Data Mining and Analytics, 1(2): 128-136, 2018.

[2] H. Liu, H. Kou, C. Yan, L. Qi. Link prediction in Paper Citation Network to Construct Paper Correlated Graph. EURASIP Journal on Wireless Communications and Networking, Article number: 233, 2019.

[3] F. Marcantoni, M. Diamantaris, S. Ioannidis, J. Polakis. A Large-Scale Study on the Risks of the Html5 WebAPI for Mobile Sensor-based Attacks. In Proc. of World Wide Web Conference (WWW'19), ACM Press, New York, USA, pp. 3063-3071.

[4] X. Chi, C. Yan, H. Wang, W. Rafique, L. Qi. Amplified LSH-based Recommender Systems with Privacy Protection. Concurrency and Computation: Practice and Experience, 2020. DOI: 10.1002/CPE.5681.

[5] N. Almarimi, A. Ouni, S. Bouktif, M. W. Mkaouer, R. G. Kula, M. A. Saied. Web Service API Recommendation for Automated Mashup Creation Using Multi-Objective Evolutionary Search. Applied Soft Computing, 85, 105830, 2019.

[6] W. Zhong, X. Yin, X. Zhang, S. Li, W. Dou, R. Wang, L. Qi. Multi-Dimensional Quality-Driven Service Recommendation with Privacy-Preservation in Mobile Edge Environment. Computer Communications, 2020. DOI: 10.1016/j.comcom.2020.04.018.

[7] X. Xu, R. Mo, F. Dai, W. Lin, S. Wan, W. Dou. Dynamic Resource Provisioning with Fault Tolerance for Data-Intensive Meteorological Workflows in Cloud. IEEE Transactions on Industrial Informatics, 2019, DOI: 10.1109/TII.2019.2959258.

[8] B. S. Jena, C. Khan, R. Sunderraman. High Performance Frequent Subgraph Mining on Transaction Datasets: A Survey and Performance Comparison. Big Data Mining and Analytics, 2(3): 159-180, 2019.

[9] Y. Zhang, K. Wang, Q. He. Covering-based Web Service Quality Prediction via Neighborhood-aware Matrix Factorization. IEEE Transactions on Services Computing, DOI:10.1109/TSC.2019.2891517, 2019.

[10] Y. Zhang, G. Cui, S. Deng. Efficient Query of Quality Correlation for Service Composition. IEEE Transactions on Services Computing, DOI:10.1109/TSC.2018.2830773,2018.

[11] Y. Zhang, C. Yin, Q. Wu, et al. Location-aware Deep Collaborative Filtering for Service Recommendation, IEEE Transactions on Systems, Man, and Cybernetics: Systems. DOI: 10.1109/TSMC.2019.2931723, 2019.

[12] Y. Chen, N. Zhang, Y. Zhang, X. Chen, W. Wu, X. S. Shen. Energy Efficient Dynamic Offloading in Mobile Edge Computing for Internet of Things. IEEE Transactions on Cloud Computing, 2019. DOI 10.1109/TCC.2019.2898657.

[13] J. Li, T. Cai, K. Deng, X. Wang, T. Sellis, F. Xia. Community-diversified Influence Maximization in Social Networks. Information Systems, Vol. 92, pp. 1-12, 2020.

[14] T. Cai, J. Li, A. S. Mian, R. Li, T. Sellis and J. X. Yu. Target-aware Holistic Influence Maximization in Spatial Social Networks. IEEE Transactions on Knowledge and Data Engineering, 2020. DOI: 10.1109/TKDE.2020.3003047.

[15] J. He, M. Han, S. Ji, T. Du, Z. Li. Spreading Social Influence with both Positive and Negative Opinions in Online Networks. Big Data Mining and Analytics, 2(2): 100–117, 2019.

[16] G. Li, S. Peng, C. Wang, J. Niu, Y. Yuan. An Energy-Efficient Data Collection Scheme Using Denoising Autoencoder in Wireless Sensor Networks. Tsinghua Science and Technology, 24(1):86-96, 2019.

[17] L. Liu, X. Chen, Z. Lu, L. Wang, X. Wen. Mobile-Edge Computing Framework with Data Compression for Wireless Network in Energy Internet. Tsinghua Science and Technology, 24(3): 271-280, 2019.

[18] X. Xu, Y. Chen, X. Zhang, Q. Liu, X. Liu, L. Qi. A Blockchain-based Computation Offloading Method for Edge Computing in 5G Networks. Software: Practice and Experience, 2019. DOI:10.1002/spe.2749.

[19] Y. Huang, Y. Chai, Y. Liu, J Shen. Architecture of Next-Generation E-Commerce Platform. Tsinghua Science and Technology, 24(1): 18-29, 2019.

[20] K. Yang, K. Maginu, H. Nomura. Cultural Algorithm and Their Application. International Journal of Computer Mathematics, 87(10): 2143-2157, 2010.

[21] H. Liu, H. Kou, C. Yan and L. Qi. Keywords-Driven and Popularity-Aware Paper Recommendation Based on Undirected Paper Citation Graph. Complexity, Volume 2020, Article ID 2085638, 15 pages, 2020.

[22] J. Qian, M. Ji. A Quantum-inspired Evolutionary Algorithm Based on Culture and Knowledge. System Engineering-Theory & Practice, 35(1):228-238, 2015.

[23] L. Qi, Q. He, F. Chen, X. Zhang, W. Dou, Q. Ni. Data-Driven Web APIs Recommendation for Building Web Applications. IEEE Transactions on Big Data, 2020. DOI: 10.1109/TBDATA.2020.2975587.

[24] M. Daneshyari, G. G. Yen. Culture-based Multiobjective Particle Swarm Optimization. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 41(2): 553-567, 2011.

[25] B. Z. Qiu, Y. Yang, X. W. Du. BRINK: An algorithm of boundary points of clusters detection based on local qualitative factor. Journal of Zhengzhou University: Engineering Sciences, 2012, 33(3): 117-121.

[26] X. Li, P. Geng, B. Qiu. Clustering boundary points detection technology for attribute data set. Control and Decision, 30(1): 171-175, 2015.

[27] X. Xu, X. Zhang, X. Liu, J. Jiang, L. Qi, M. Z. A. Bhuiyan. Adaptive Computation Offloading with Edge for 5G-Envisioned Internet of Connected Vehicles. IEEE Transactions on Intelligent Transportation Systems, 2020. DOI: 10.1109/TITS.2020.2982186.

[28] B. Z. Qiu, J. Y. Shen. Grid-based and Extend-based Clustering Algorithm for Multi-Density. Control and Decision, 2006, 21(9): 1011-1014.

[29] B. Z. Qiu, T. Yu. Boundary Points Detecting based Gradient of Grid. Microelectronics and Computer, 2008, 25(3): 77-80.

[30] B. Z. Qiu, F. Yue, J. Y. Shen. BRIM: An Efficient Boundary Points Detecting Algorithm. Advances in Knowledge Discovery and Data Mining. Berlin: Springer, pp. 761-768, 2007.

[31] B. Z. Qiu, S. Wang. A Boundary Detection Algorithm of Clusters Based on Dual Threshold Segmentation. The 7th International Conference on Computational Intelligence and Security. Sanya: IEEE: 1246-1250, 2011.

[32] X. Li, P. Geng, B. Z. Qiu. Clustering Boundary Points Detection Technology for Attribute Data Set. Control and Decision, 30(1):171-175, 2015.

Figure Title and Legend

Figure 1

short title of figure: AiNet clustering principle based on a cultural algorithm.

detailed legend: Trust space, spatial relationship, Immune cyber space.

Figure 2

short title of figure: Clustering effect on a synthetic dataset.

detailed legend: Raw data, Raw data and antibodies, Minimum spanning tree, Clustering results.

Figure 3

short title of figure: The comparison of algorithm balance.

detailed legend: balance, n=2, n=3, n=4, n=5, n=6.

Figure 4

short title of figure: The comparison of algorithm precision.

detailed legend: precision, n=2, n=3, n=4, n=5, n=6.

Figure 5

short title of figure: The comparison of algorithm recall.

detailed legend: recall, n=2, n=3, n=4, n=5, n=6.

Figure 6

short title of figure: The comparison of algorithm hit rate.

detailed legend: hit rate, n=2, n=3, n=4, n=5, n=6.

Table Title and Legend

Table 1

short title of table: Comparison of clustering performance between the two algorithms for three datasets.

detailed legend: IRIS, Wine, Seeds.

Figures

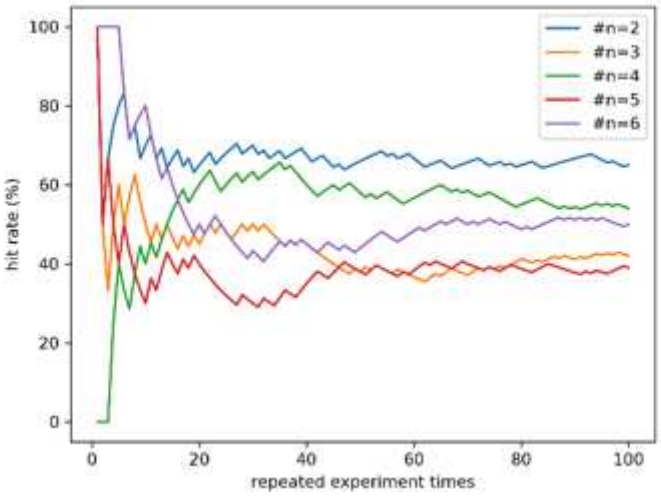

Figure 1

The comparison of algorithm hit rate

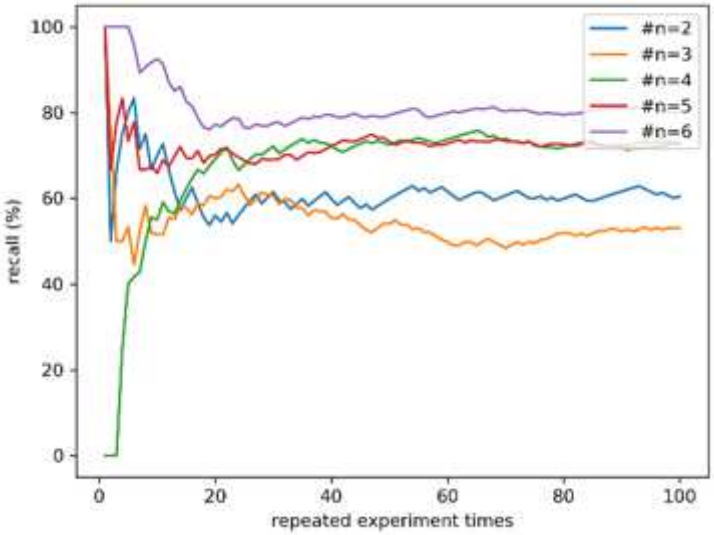

Figure 2

The comparison of algorithm recall

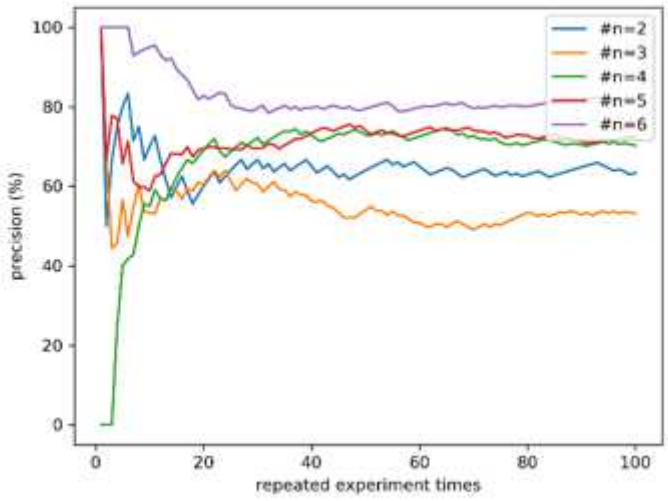

Figure 3

The comparison of algorithm precision

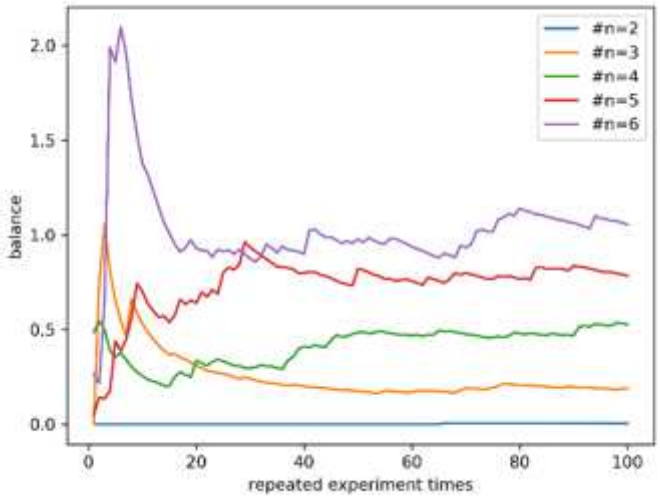

Figure 4

The comparison of algorithm balance

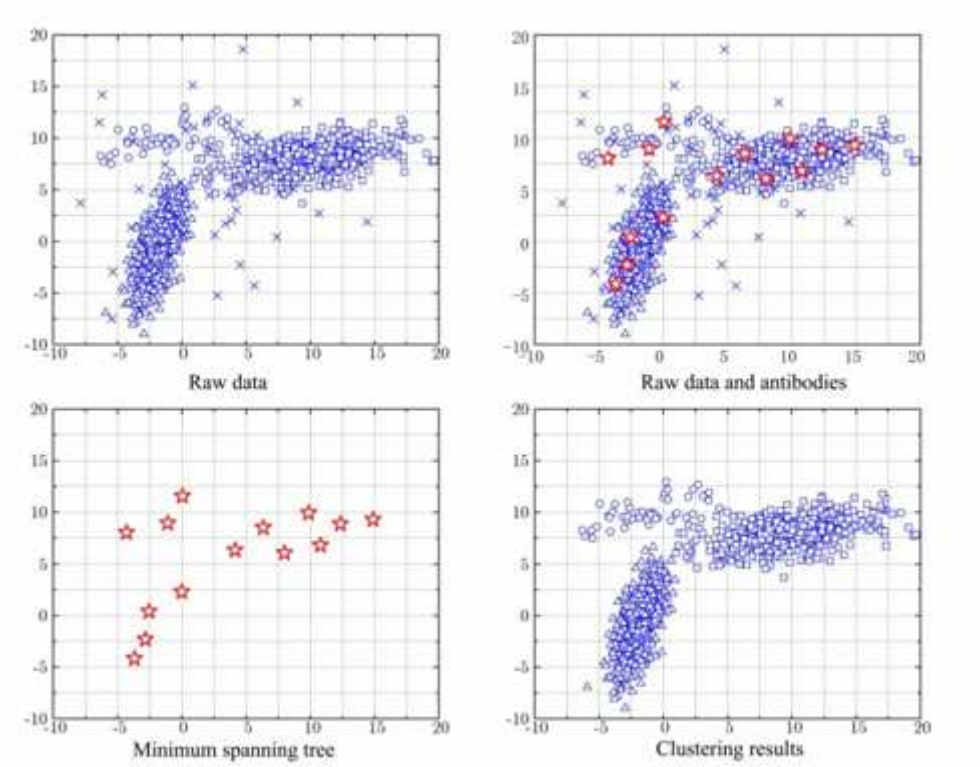

Figure 5

Clustering effect on a synthetic dataset.

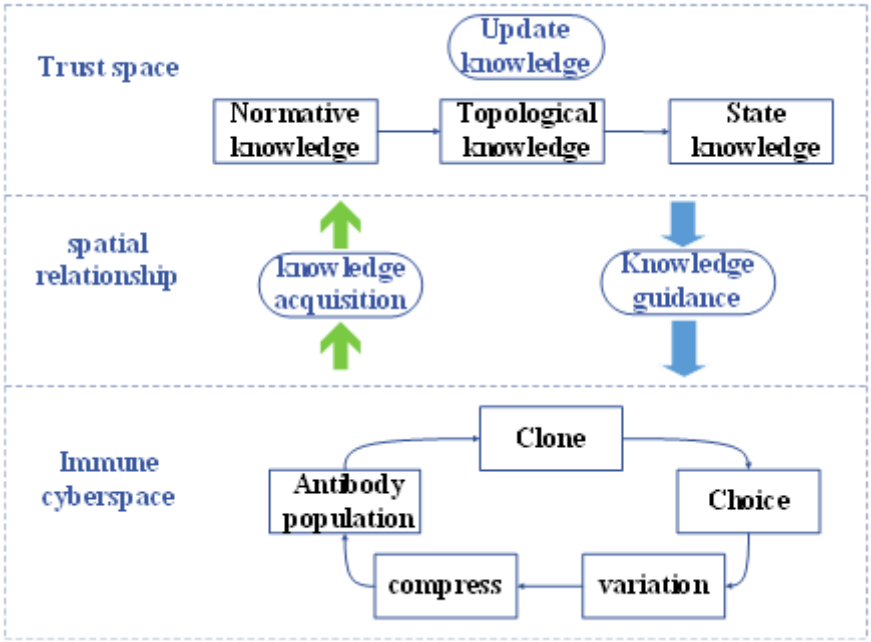

Figure 6

AiNet clustering principle based on a cultural algorithm.