

RESEARCH ON GENERAL DISEASE AUXILIARY DIAGNOSIS TECHNOLOGY BASE ON CASE-BASED REASONING

ZHANG LIN¹, ZHANG JIE², YAO NANZHEN^{3,*}, WANG HONGHAI⁴, LIU YONGNING⁵

¹Institute of computer engineering, AnHui SanLian University, China - ²School of Information Engineering, Huainan United University, China - ³Library of University of Science and Technology of China, China - ⁴ChaoHu University, China - ⁵North China Electric Power University, China

ABSTRACT

Objective: In order to solve the problem of difficult and expensive medical treatment over the years, a series of social contradictions have been triggered.

Methods: This paper proposes a general disease diagnosis technique based on case-based reasoning. This technology designs the case recombination, retrieval, and retention technology through professional knowledge structured expression, scientific keyword weight measurement, and feature word distribution calculation with inter-class dispersion.

Results: The algorithm model can be applied to the auxiliary diagnosis based on the patient's symptoms and the auxiliary diagnosis based on the results obtained by the diagnosis and treatment instrument. It can not only be used for self-diagnosis and treatment of patients but also can be used as a reference to assist doctors in diagnosis and treatment.

Conclusion: It provides an effective method for the establishment of disease auxiliary diagnosis techniques and related models.

Keywords: Case-based reasoning, text frequency, retrieval model.

DOI: 10.19193/0393-6384_2023_4_143

Received January 15, 2023; Accepted April 20, 2023

Introduction

Over the years, the difficulty and high cost of medical treatment have caused a series of social conflicts. According to the survey, 41.66% of residents think it is difficult to see a doctor, 89.72% think the cost of seeing a doctor is high, 91% think the cost of medicine, 69.5% think the cost of examination, 54.4% think the cost of diagnosis and treatment, and 53.5% think the family income is low. Although the state has been striving to improve the social, medical security system, strengthen the government's role in public social undertakings, strengthen the supervision and management of medical institutions, standardize their medical

services, strengthen health education and health consultation services, the national basic medical insurance participation rate and the total income of the basic medical insurance fund have also increased steadily. However, the total amount of medical and health resources is insufficient, and the high-quality resources are few. People passively support the rapidly expanding, highly market-oriented, and internationalized medical and health services with limited income, making the high cost of medical treatment a cause of serious social instability. In addition, the distribution of medical resources is unbalanced. 80% of medical resources are concentrated in cities, and in cities, 80% of resources are concentrated in large hospitals, especially

excellent resources. In order to reduce the working pressure on medical institutions and medical staff and reduce the number of hospitalization examinations and diagnoses of patients, this paper proposes to use Case-Based Reasoning technology to assist disease diagnosis so that residents can make a preliminary diagnosis based on their own symptoms at home and obtain the corresponding auxiliary diagnosis and treatment plan, so as to reduce the frequency of going to the hospital for examination.

Case-Based Reasoning (CBR) technology originates from Roger Schank of Yale University, who described it in his monograph *Dynamic Memory* in 1982. It is an important knowledge-based problem-solving and learning method emerging relatively recently in the field of artificial intelligence⁽¹⁻¹¹⁾. CBR solves problems by constructing and assembling a rich case base and reusing solutions from existing cases or modifying solutions from existing similar cases. It turns out that this cognitive style, which mimics analogies in human decision-making process, can effectively solve problems in unstructured, knowledge-poor domains.

Related research

Through years of research on CBR, the author has designed and completed a diagnosis system for common diseases in children, a decision support system for urban road congestion relief, and a health assessment and diagnosis system for the elderly, and declared and authorized the corresponding software copyright. In addition, She also led the team to successively apply for and complete the key project of natural science research in colleges and universities of Anhui Province, "Study on the Diagnosis Method of Common Diseases in Children Based on Case-based Reasoning," the project of a Support plan for outstanding young Talents in colleges and universities of Anhui Province "Study on the health Evaluation Method of the Elderly based on case-based Reasoning" and the horizontal project "Research on Key technologies of personalized intervention for chronic Diseases based on big health Data."

In the previous research results, whether it is the diagnosis of common diseases in children, the health assessment of the elderly, or the personalized intervention technology for chronic diseases, the application field is relatively narrow, it is easy to amplify the role of rare attributes in the process of reasoning, and the description of the distribution of

different characteristic attributes in different cases is not ideal, and the diagnosis results are basically the use of the diagnosis results of known cases. In the absence of manual intervention, the intelligence of giving corresponding diagnosis results needs to be improved. In order to solve the above problems, the research group integrated the corresponding technologies of the previous research.

This paper proposes the research of general disease auxiliary diagnosis technology based on Case-Based Reasoning. On the basis of the original application of TF-IDF and information entropy, the calculation of inter-class dispersion is added. The aim is to give a general model of the assisted diagnosis of disease and make it applicable to most fields where Case-Based Reasoning is possible.

A framework for the general disease auxiliary diagnostic technique based on case-based reasoning

Examination of social networks was one approach through the case data introduced in different fields and corresponding diagnosis and treatment plans to establish the initial case base, and then through the segmentation of the attributes in the initial case base to establish a general case base.

When unknown cases occur, similarity retrieval is carried out in the case base first according to the attributes of unknown case selection, to find out similar cases whose similarity meets the threshold requirement or ranking, and then the auxiliary diagnosis and treatment plan of unknown cases is obtained according to the analysis of diagnosis and treatment plan of similar cases. The framework of the general disease auxiliary diagnosis technology model based on Case-Based Reasoning is shown in Figure 1 below. In the general disease auxiliary diagnosis technology model based on Case-Based Reasoning, the collected cases should be first described in a standardized way, including characteristic attributes and diagnosis and treatment plan, and the description method or standard should be unified.

For example, some medical institutions use age when recording patient information, while others use date of birth, so unified description standards are required when standardized description of cases is carried out. Secondly, it is necessary to calculate the weight of each feature attribute and describe each case in the form of the feature vector. When new cases are added to the case base, the weights of feature attributes are recalculated. The weight of characteristic attributes determines the importance of corresponding attributes in case similarity retrieval.

Finally, the similarity retrieval algorithm is used to find all cases in the case base whose similarity to the new case is within the threshold range or meets the requirement of similarity ranking.

If there are cases in similar cases that are identical to the new cases or whose similarity meets the threshold requirements, the case is reused; that is, the solution of similar cases is used directly to solve the problem described by the new case; otherwise, the solution of the problem described by the new case is fitted by the algorithm with the solution of the most similar cases. If the new case solution has been validated and is typical, the new case is added to the case base.

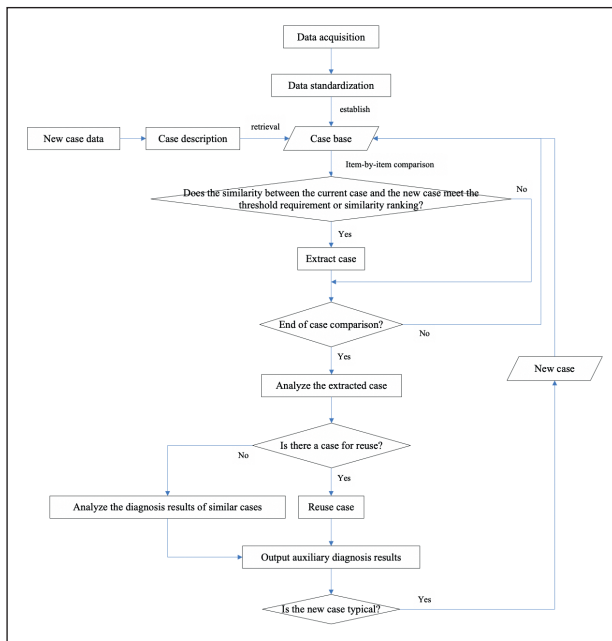


Figure 1: A framework for the general disease auxiliary diagnostic technique based on Case-Based Reasoning.

Key technologies of general case retrieval model based on case-based reasoning

As mentioned above, the process of general disease auxiliary diagnosis technology based on Case-Based Reasoning mainly includes four key technologies: standardized description of case knowledge, calculation of weight of case characteristic attributes, case similarity retrieval, and fitting of case diagnosis results and diagnosis and treatment plan.

Standardized description of case knowledge

Before case-based reasoning, cases should be selected first, and those with typical diagnosis results and treatment plans that have been unanimously recognized should be selected. In addition, it is necessary to clean and standardize the data,

eliminate the obviously unreasonable data, fill in the blank data, and describe each feature attribute with a unified standard to form a basic data case base, as shown in Table 1 below. In the basic data case base, use a variety of standardized data or text for case description.

ID	Gender	Age	Height (m)	Weight (kg)	BIM	Consciousness	...	Diagnosis	Treatment plan
1	Male	70	1.75	60	19.59	vaguer	...	Diagnosis	Treatment plan
2	Female	28	1.65	50	18.37	clear	...	Diagnosis	Treatment plan
3	Male	46	1.77	63	20.11	Relatively clear	...	Diagnosis	Treatment plan
...

Table 1: Basic data case base.

However, because continuous numerical data or textual data are not easy to compare in case similarity retrieval, each characteristic attribute is classified according to medical classification on the basis of a basic data case base and represented by a Boolean data case base.

Of course, each case has one and only one class in the same attribute class that values True and all others that value False. Diagnosis results and treatment plans are also divided into several single-symptom diagnosis results and a single drug use plan or treatment method according to the actual situation. Of course, multiple diagnosis results in the same case can be True, indicating that the case suffers from multiple diseases at the same time. Similarly, the same case can be True for multiple choices in the diagnosis and treatment plan. Accordingly, the Boolean data case base can be obtained in Table 1, as shown in Table 2 below.

ID	Gender	Emaciation	Overweight	Fat	Normal	...	Treatment plan 1	...	Treatment plan n
1	True	False	False	False	True	...	False	...	True
2	False	True	False	False	False	...	True	...	False
3	True	False	False	False	True	...	False	...	True
...

Table 2: Boolean data case base.

Table 1 and Table 2 are associated by ID. Table 2 is used for case similarity retrieval and fitting of diagnosis and treatment plan.

Similar retrieval results are associated with relevant case descriptions in Table 1 by ID, and the case descriptions in Table 1 are displayed to users. As described in Table 2, each case can be represented by an n-dimensional feature vector FV, $FV=(FV_1, FV_2, \dots, FV_n)$, where n is the number of Boolean attributes after characteristic attribute classification

in Table 2, excluding the number of Boolean attributes after diagnosis results and diagnosis and treatment scheme decomposition, and the value of each FV_i is only 1 or 0, that is, when the value of corresponding Boolean classification attribute is True, the value of FV_i is 1, otherwise it is 0.

Calculation of weight of case characteristic attributes

In the process of disease diagnosis, different characteristic attributes have different influence on the diagnosis results. For example, many patients have cough symptoms. It is difficult to determine the specific disease of patients only by coughing. Other symptoms are also needed for reference; that is to say, cough is not a strong symptom. Of course, there are some symptoms that are very typical. Therefore, different weights should be assigned to each characteristic attribute to reflect the typicality of the symptom in the course of disease diagnosis.

The determination of symptom typicality is similar to the concept of Term frequency-Inverse Document Frequency (TF-IDF) in information theory; that is, the more times a symptom appears in the whole case, the less typical it is, and conversely, the more typical it is. Combined with the calculation method of information entropy, that is, the calculation method of information bits needed to determine the uncertainty of information, we can calculate the weight of each feature classification attribute in the case: $W_i = \log_2(N/N_i)$. W_i represents the weight of the i th feature classification attribute, N represents the total number of cases, and N_i represents the number of the i th feature classification attribute whose value is True. For example, there are 10,000 cases in the case base, of which 5,254 cases have gender attribute value True, that is, 4,982 males in 10,000 cases. So the attribute weight of gender $W_{gender} = \log_2(10,000/5254) \approx \log_2(1.903) \approx 0.928$. Using the same method, the number of cases whose attribute classification value is True is counted, and their weights are calculated, then the weights of the Boolean data case base can be obtained as shown in Table 3.

Attribute type	Gender	Emaciation	Overweight	Body fat	Orthotypic	...	Treatment plan 1	...	Treatment plan n
weight	0.928	2.869	2.198	3.987	0.784
2	False	True	False	False	False	...	True	...	False

Table 3: Boolean data case base weight.

From Table 3, we can get the weight vector $W=(W_1, W_2, \dots, W_n) = (2.869, 2.198, \dots, 1.202)$, and thus calculate the weighted vector

$WV_i = FV \times WT = (FV_1, FV_2, \dots, FV_n) \times (W_1, W_2, \dots, W_n) T = (1, 0, \dots, 0) \times (2.869, 2.198, \dots, 1.202) T = (2.869, 0, \dots, 0)$. Although the introduction of the concept of IDF and the calculation method of information entropy makes the weighted vector of the case show a good application effect in the distribution of the weight of the eigenvalue of the case, the original intention of the introduction of IDF is to suppress the negative impact of meaningless high-frequency attributes in the case, highlighting the role of low-frequency attributes.

However, in the process of disease-assisted diagnosis, the classification attributes of common cases are not necessarily meaningless. On the contrary, some inherent living habits of some chronic patients often lead to some inherent changes in the health indicators of the body. By the same token, low-frequency attributes are not necessarily important attributes and should not all have high weights. In view of these problems, we consider adding a measure to represent the distribution of attributes among different classes between original cases-inter-class dispersion, that is, to calculate the frequency of each attribute in different classes, so as to determine whether the attribute can become a characteristic attribute of cases.

Firstly, cases are classified according to diagnosis results. Cases diagnosed with a certain disease are classified into one category. If a case is diagnosed with multiple diseases at the same time, the case can appear in multiple categories at the same time. Through case classification, we can see that the feature attributes distributed in a certain type of cases tend to have strong category differentiation ability; that is, the feature attributes will have a strong inter-class dispersion degree. For example, patients with COVID-19 often test positive for nucleic acid or antigen, so we can use nucleic acid or antigen tests to determine if someone has COVID-19. Assuming that all cases can be divided into n categories, $F(i)$ represents the frequency of feature attribute i in a certain category of cases and represents the average frequency of feature i in all types of cases. That is:

$$\overline{F(i)} = \frac{1}{n} \sum_{k=1}^n F_k(i) \quad (1)$$

The formula for calculating the overall inter-class dispersion $Id(i)$ is:

$$Id(i) = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (F_k(i) - \overline{F(i)})^2} \quad (2)$$

Substitute equation (1) into equation (2) to get:

$$Id(i) = \frac{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (F_k(i) - \frac{1}{n} \sum_{k=1}^n F_k(i))^2}}{\frac{1}{n} \sum_{k=1}^n F_k(i)} \quad (3)$$

Combined with the previous idea of using IDF and information entropy to calculate weight, if the characteristic attribute in formula (3) only appears in a certain type of case, then the attribute has the strongest classification ability, and the value of Id(i) is 1. Similarly, if a feature attribute does not have classification ability, that is, the frequency of occurrence of the attribute in each type of case is equal, then the feature is a useless feature.

In this case, the value of Id(i) is 0 and can be directly discarded. Thus, the value of Id(i) is between [0,1]. After considering the dispersion between classes, the weight calculation becomes:

$$W_i = (\log_2 \frac{N}{N_i}) \times \frac{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (F_k(i) - \frac{1}{n} \sum_{k=1}^n F_k(i))^2}}{\frac{1}{n} \sum_{k=1}^n F_k(i)} \quad (4)$$

The weight with inter-class dispersion measures the distribution of each attribute among different classes. However, if there are two characteristic attributes in the same class of cases, the distribution of these two fault characteristics still cannot be accurately judged.

Therefore, we define the information entropy in the same type of cases; that is, the more evenly distributed a certain feature attribute is in a certain type of case, the greater the information entropy in this type of case, and the more the feature attribute can reflect the feature information of this type of case. The calculation formula of information entropy in the same type of case is:

$$I_{cc}(C_j) = - \sum_k \frac{FC_{jk}}{FC_j} \lg \frac{FC_{jk}}{FC_j} \quad (5)$$

Where FC_{jk} represents the frequency of occurrence of the kth value (0 or 1) of feature attribute I in class C_j cases, and FC_j represents the total frequency of occurrence of feature attribute I in class C_j cases. Finally, a relatively accurate calculation method to determine the weight of feature attributes is obtained for the calculation of case classification based on the inter-class dispersion and intra-class information entropy:

$$W_i = (\log_2 \frac{N}{N_i}) \times \frac{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (F_k(i) - \frac{1}{n} \sum_{k=1}^n F_k(i))^2}}{\frac{1}{n} \sum_{k=1}^n F_k(i)} \times (- \quad (6)$$

According to the above formula, before each Case-Based Reasoning, the weight of each feature attribute is calculated first, and then several feature attributes with the largest weight are selected to form the feature vector of each case.

Then, the accuracy and efficiency of case retrieval can be improved by using these feature vectors selected by weight.

Case similarity retrieval

The so-called Case-Based Reasoning is to find similar historical cases and use specific knowledge in existing experience or results to solve new problems. Therefore, the core of Case-Based Reasoning is case similarity retrieval, which aims to retrieve as few similar cases as possible from massive historical cases and has reference significance to solve the current problem so as to serve as the basis for solving the current problem.

After the standardized description of the case knowledge and the calculation of the weight of the feature attributes of the case, each case can be represented by the weighted feature vector constructed by the selected feature attributes with the greatest weight. Since the weights of the feature vectors are all positive, we can determine the similarity between the historical case and the current case by calculating the Angle included by the vector through the cosine theorem, and the cosine value of the Angle is between 0 and 1. When the cosine value of the Angle between two vectors is larger, the similarity between two vectors is higher. When the cosine value of the Angle between two vectors is 1, the two vectors coincide; that is, the two cases are identical. On the contrary, the less similar the two vectors are, when the cosine of the included Angle is 0, the two vectors are completely different.

Suppose that the weighted feature vector of the historical case is (x₁, x₂... , x_n), the weighted feature vector of the current case is (y₁, y₂... y_n), then the cosine of the Angle between the two vectors is:

$$\cos \theta = \frac{x_1 y_1 + x_2 y_2 + \dots + x_n y_n}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \cdot \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}} \quad (7)$$

Whenever a new case appears, we only need to calculate the cosine value of the Angle between the

new case and each historical case in the case base. If there is a historical case with the Angle cosine value of 1, we can directly reuse the diagnosis results and treatment plans of the case. If there is no historical case with an Angle cosine value of 1 between the new case, several historical cases whose threshold meets the requirements can be selected by setting the threshold value, or several historical cases most similar to the new case can be selected by similarity ranking, and then auxiliary disease diagnosis results and treatment plans can be given by fitting the diagnosis results and treatment plans.

Fitting of case diagnosis results and diagnosis and treatment plan

When there is no historical case with 100% similarity to the new case in the historical case database, we select the most similar cases and then fit the diagnosis result and treatment plan of the new case according to the diagnosis result and treatment plan of these cases.

For example, the diagnosis results and treatment plans of the top 10 cases selected by us are shown in Table 4 below.

ID	Similarity	Diagnostic result 1	Diagnostic result 2	...	Diagnostic result n	Treatment plan 1	Treatment plan 2	...	Treatment plan m
1023	98.62%	True	False	...	True	True	True	...	True
789	97.98%	False	True	...	True	True	False	...	False
254	97.33%	True	True	...	True	True	True	...	True
3	96.75%	True	True	...	True	True	True	...	True
699	96.02%	True	True	...	True	False	True	...	True
382	95.88%	True	True	...	True	True	True	...	True
543	95.54%	True	True	...	True	True	False	...	True
189	94.99%	True	True	...	True	True	True	...	True
42	94.63%	True	True	...	False	True	True	...	True
907	94.12%	False	True	...	True	False	True	...	True

Table 4: Diagnosis and treatment of similar cases.

The following formula can be used to calculate the diagnostic results and application degree of diagnosis and treatment plan:

$$New(A_i) = \frac{\sum_{i=1}^n S_i A_i}{\sum_{i=1}^n S_i} \times 100\% \quad (8)$$

Where S_i is the case similarity, A_i is the value of the corresponding diagnosis result or diagnosis and treatment scheme. If the value is True, A_i is 1; otherwise, it is 0; n is the number of similar cases.

Take the above table as an example, $n=10$, and:

$$New(\text{Diagnostic result}_1) = \frac{98.26\% \times 1 + 97.98\% \times 1 + \dots + 94.12\% \times 0}{98.26\% + 97.98\% + \dots + 94.12\%} \times 100\% \quad (9)$$

$$= 90.21\%$$

$$New(\text{Treatment Plan}_m) = \frac{98.26\% \times 1 + 97.98\% \times 0 + \dots + 94.12\% \times 1}{98.26\% + 97.98\% + \dots + 94.12\%} \times 100\% \quad (10)$$

$$= 89.81\%$$

After calculating the application degree of each diagnosis and treatment scheme of the new case, although the final calculated value may not be the accurate prevalence rate, the size of the application value can reflect the high and low prevalence rate and the relative applicability of the treatment scheme.

At this point, through manual intervention or threshold setting, the diagnosis results of the new case and auxiliary treatment plan can be given.

Conclusion

In this paper, a general disease auxiliary diagnosis technique based on Case-Based Reasoning is proposed, and the standardized and Boolean expression of cases, the similarity retrieval technique of cases, and the auxiliary diagnosis fitting technique are given. This algorithm model is applicable to the auxiliary diagnosis based on the patient's symptoms, such as whether the patient has cough, fever, eye secretions, skin rashes, etc., as well as the auxiliary diagnosis based on the results obtained by diagnostic instruments, such as blood test and CT scan data. It can not only be used for self-diagnosis and treatment of patients but also can be used as a reference to assist doctors in diagnosis and treatment.

In real life, we tend to trust experience. For example, if a newly graduated doctor of medicine and a doctor with low education but rich experience in seeing patients have different diagnosis results, we tend to follow the latter's diagnosis results because the latter has rich experience in seeing patients. Although experience is not always right, it is often an important reference for people to judge an outcome. In this paper, we propose a common case-based diagnosis technique for disease assistance, which makes use of such experience, not just one or two people's experience, but the experience of most professionals. The algorithm proposed in this paper has high applicability and feasibility, but there are still some problems: first of all, regarding the construction of Boolean attributes, there is no problem if the attributes are classified uniformly

for different diseases. However, if different diseases have different criteria for the same attribute, the classification of the attribute will produce ambiguity. How to solve this ambiguity is one of the future research directions.

Secondly, because the classification of attributes in the Boolean case database is based on the classification of attributes in the basic case database according to characteristics, and in different classifications of a case with the same attribute, only one classification is True, while other classifications are False. Therefore, the matrix formed by the Boolean representation of cases is a sparse matrix. How to improve algorithm efficiency through sparse matrix characteristics is also one of the future research directions.

References

- 1) Cai Ping. Discussion on the Intervention Effect of Individualized Community Chronic Disease Management [J]. *Medicine and health care*, 2018, 12(02): 284.
- 2) Li Ji-Qiang, Li Xing-guo, Gu Dong-xiao, Feng Shuai. "Case-Based Reasoning ISP Knowledge Reuse Method", *Computer Engineering*, vol. 36, January 2010: 36-39.
- 3) Liu Li, Fan Xiaozhong, Qi Quan, et al. Ontology-based question expansion for question similarity calculation[J]. *Journal of Beijing Institute of Technology*, 2011, 20(2): 244-248.
- 4) Han Min, Shen Li-hua, "Case-based Reasoning Based on FCM and Neural Network," *Control and Decision*, vol. 27, September 2012: 1421-1424.
- 5) Machine Learning; Findings from Helwan University Broaden Understanding of Machine Learning (Solving the motion planning problem using learning experience through case-based reasoning and machine learning algorithms)[J]. *Journal of Engineering*, 2020, 36(4): 251-256.
- 6) Designing participatory decision support systems: towards meta-decision making analytics in the generation of ecological economics [M]. Edward Elgar Publishing: 2020-05-07.
- 7) Mohamed Benamina, Baghdad Atmani, Sofia Benbelkacem, Abdelhak Mansoul. Fuzzy Adaptation of Surveillance Plans of Patients with Diabetes[M]. Springer International Publishing, 2019-12-05.
- 8) Hao Zhang. Research on Case-based Reasoning for Urban Road Traffic Congestion Safety Decision Support Technology [D]. Anhui University of Science and Technology, 2018.
- 9) Lin Zhang, Deqing Zhang. A Novel Diagnosis Method for Paediatric Common Disease Using Case-Based Reasoning[J]. *International Journal of Simulation Systems, Science & Technology*, 2016, 12(17): 37.1-37.5.
- 10) Science - Automation Science; Researchers at North China Electric Power University Detail Findings in Automation Science (Case-based reasoning based on grey-relational theory for the optimization of boiler combustion systems)[J]. *Energy Weekly News*, 2020, 352(5): 374-378.
- 11) Jun Wu. The beauty of mathematics[M]. People's Posts and Telecommunications Press, 2012.

Acknowledgments:

This work is supported by key research and development program of Key project of the Provincial Natural Science Research in Anhui Province, under Grant NO.2022AH051985; Anhui Sanlian University Collaborative Innovation Center key project, under Grant NO. zjqr23001, NO. zjqr23002; The Provincial Natural Science Research Program of Higher Education Institutions of Anhui province, under Grant NO. KJ2021A1030, NO. KJ2021A1189; The Quality Improvement Project of Chaohu University on Discipline Construction, under Grant NO. kj21gczx03; Special Support Plan for Innovation and Entrepreneurship Leaders in Anhui Province.

Corresponding Author:

YAO NANZHEN
School of Science and Technology University, 230026, China
Email: 864792734@qq.com
(China)