# Construction of soft power index model of urban network culture based on feature selection algorithm

Yifei Dong, Qixing Yin\*

[Handan](http://www.baidu.com/link?url=EGXhUUtga96-wwjnqQSV339xfvtWRAiSljh4rjSvlVIEcHT5XJl5FBm03eK8BD3j) University, Handan, 056001, China

Qixing Yin Email: [yinqixing@hdc.edu.cn](mailto:yinqixing@hdc.edu.cn)

Yifei Dong Email: [du2009204zij@163.com](mailto:du2009204zij@163.com)

**Abstract:** Cultural soft power (SP) represents a critical aspect of a nation’s overall strength, complementing hard power to collectively shape a country’s comprehensive capabilities. To enhance their cultural SP and gain a competitive edge internationally, researchers and academics worldwide have devoted significant effort to studying this concept. They have also developed various systems for evaluating cultural SP. After years of refinement, these systems have become increasingly robust and now play a pivotal role in assessing the development of cultural SP holistically. In the big data era, feature selection is a key step in data preprocessing. As a technique for dimensionality reduction, its primary goal is to identify the most relevant features from the original dataset. This helps reduce data dimensions, simplify learning tasks, and enhance model efficiency. While research on feature selection algorithms has made notable progress, substantial challenges persist, with the "curse of dimensionality" being a significant obstacle in feature selection and classification tasks. This paper first outlines the fundamental structure of feature selection algorithms, detailing four key processes: subset generation, subset evaluation, termination criteria, and result validation. It then reviews the methodologies and advancements in feature selection, categorizing and analyzing them based on evaluation strategies, search strategies, and supervisory information. Traditional approaches are compared, with their respective strengths and limitations highlighted. Lastly, the paper provides an overview of feature selection research and explores potential directions for future studies.

**Keywords:** Soft power, feature selection, machine learning, C4.5

## 1 Introduction

Although countries like Japan, South Korea, and Thailand do not possess the same level of comprehensive national strength as the United States or many European nations, they have garnered significant international attention in recent years. This has been achieved through cultural elements such as film and television, cuisine, animation, sports, and beauty trends, which have greatly enhanced their global image. In today’s era of increasing globalization, strengthening cultural soft power (SP) is a crucial strategy for China to secure a leading role within the international relations framework[1-4]. Across various societal sectors, there is broad support for advancing China's cultural influence and bolstering its ST. This collective enthusiasm has sparked active academic inquiry and discussion regarding cultural ST, its development, and related strategies. These studies aim to deepen understanding and offer insights into building a stronger cultural presence on the global stage.

ST in the cultural realm refers to the appeal, cohesion, absorption, innovation, and dissemination based on the culture in a country or region, as well as the resulting competitiveness and influence. In today's white-hot world of international competition, ST in the cultural realm will become a significant part of a country's ST. It has increasingly become an inexhaustible driving force and important source of a country's stability, prosperity, and sustainable development, and an important scale and backbone factor in measuring a country's comprehensive national strength and international competitiveness[5]. If any country wants to win the lasting initiative and obtain the preemptive advantage in the fierce international competition, it must vigorously improve its ST, especially its ST in the cultural realm, while continuously improving its development level of the economy, and science, technology, and military.

Compared with the economy, ST in the cultural realm can not be seen directly and is also an intangible "soft" power. In order to better promote the steady, rapid improvement and sustainable development of ST in the cultural realm, it is necessary to quantitatively evaluate the ST in the cultural realm of a country or region to provide an accurate basis for decision-making or provide "hard knowledge" for "ST." However, at present, a large number of studies on China's ST in the cultural realm are general theoretical analyses or logical deduction or just stay in the "soft" research. Therefore, it is particularly necessary to construct scientific evaluation indicators and conduct empirical research. At present, a few scholars have begun to carry out research work in this regard. For example, Yang Xinhong[6] believes that the ST in the cultural realm index system should be composed of value indicators, physical indicators, and relative indicators on the basis of relevant subject concepts and principles: Sun Liang focuses on the vision of nationality from the perspective of development level about the country and we can evaluate the ST in the cultural realm from six aspects: soft diplomatic power and soft communication power: Chen Yiyuan [7] proposed five principles for designing the ST in the cultural realm index system, made ten secondary indicators and twenty-four tertiary indicators to measure the ST in the cultural realm and launched the Ningbo ST in the cultural realm evaluation model. It should be said that these exploratory studies are valuable. However, these studies need to be further improved. Some index systems and models are too complicated, and the level system is not clear: some index systems are incomplete and fail to comprehensively and systematically reflect the situation of ST in the cultural realm. Based on the previous research, according to the scientific principles and methods, and based on the in-depth study of ST in the cultural realm, this paper will design the index system of ST in the cultural realm and construct the evaluation model so as to provide an empirical research basis for measuring the national ST in the cultural realm.

## 2 Related Works

### 2.1 The ST of the country

Since the 20th century, the movement of social indicators has sprung up, and the ideological trend of new public management has affected the process of government reform. As a tool to measure effectiveness, indicators are more and more widely used in statistics and have become a means of assessing performance[8]. With the government's soft economic indicators and cultural indicators of the development of cultural works, such as the government's soft economic indicators and cultural indicators of cultural services have become more and more important.

At present, China is on the journey of building a socialist cultural power. Improving ST in the cultural realm is undoubtedly an important component of governing the country. Therefore, it is necessary to design a set of special performance evaluation index systems. At present, compared with developed countries, the research on cultural indicators in China lags behind and is still in its infancy. China needs to speed up the progress and design a set of systematic indicator systems with strong references and good implementation, which is mainly used to evaluate and assess the main work links and work performance of relevant departments of cultural construction in China, monitor and manage in an all-round way, and provide a statistical basis for measuring the level of ST in the cultural realm in China[9], help speed up the development speed of building a socialist cultural power. Although there is still a lack of mature theory on the construction of an index system in the field of culture in China, with the introduction of a large number of cultural policies in recent years, many policy documents have special requirements for cultural development, which provides a corresponding basis for the design of this index system. This design primarily draws upon the policy document that establishes a comprehensive set of cultural reform and development benchmarks. These indicators hold authoritative significance and provide essential guidance for constructing the proposed indicator system.

Comprehensive National Power. Ashley Tellisetal[10] defines Comprehensive National Power as the ability of a state to acquire resources over a period of time and over a geographical area and to achieve the intended purpose of its conduct. If this capability is used effectively by a country, it will contribute to the development of military capabilities, the creation of a political environment, and the cultivation of economic advantages, and then can make the country in the current and even in the future the international system to obtain a dominant position in the strategic conditions[11]. International Relations Kenneth's study of the Comprehensive National Power is done in a long-term, systematic way. In its research, it describes Comprehensive National Power as a combination of capacities: capacities to access and use strategic resources oriented toward national strategic objectives. He argues that what we mean by Comprehensive National Power is the same as a country's strategic resources and that this can be used interchangeably. As the most important measure of a country's resources and energy, the Comprehensive National Power should not include only some aspects of a country, especially military and economic aspects[12-14]. The scholar said it needs to include seemingly intangible factors such as science and technology, culture, art, and thought. Michelle Baud, an American scholar, put forward the Comprehensive National Power Resources. The connotation of each of these elements is very broad. The process of research needs different countries or regions in specific circumstances to carry out targeted thinking.

### 2.2 The feature selection algorithms

Feature selection (FS) serves as a technique for reducing data dimensionality. It involves exploring high-dimensional feature spaces and employing evaluation algorithms to assess candidate subsets, ultimately selecting the most optimal feature subset. The distinction between filter-based and wrapper-based methods lies in their differing evaluation approaches, whereas hybrid methods integrate the strengths of both. These types can be summarized as follows.

Uncertainty, symmetric, and symmetric, ie. Sosa et al. have already developed the principle of mutual information to measure and compute the symmetry of each feature, took into account the correlation between features, categories, and features, and screened features by setting thresholds. Mendoza et al. developed a distributed filtering method on the basis of ReliefF, which combines ReliefF and spark, evaluates the distinguishing ability of close samples according to the features, and is used to solve the problem of FS of huge data. There are many researchers who have also developed a filtering algorithm based on information gain ratio (GR): first, use the information gain ratio feature to sort; Then, FS is carried out according to the mutual information between features[15].

In the aforementioned filtering algorithms, feature relevance is evaluated using specific assessment methods. However, the recognition accuracy of the resulting feature subsets is often low, and threshold settings lack precision. To address this, some researchers suggest using classification accuracy from classification algorithms as the evaluation criterion for feature subsets, coupled with search algorithms. For instance, Zhang et al. introduced an encapsulated optimization technique leveraging particle swarm optimization (PSO) to identify and eliminate redundant features, with the C4.5 algorithm serving as the evaluation method. Similarly, other researchers have proposed wrapper-based approaches utilizing support vector machines (SVM) for FS. In these methods, genetic algorithms (GA) are employed to minimize the SVM classification error rate and identify optimal feature subsets. Enhancements to genetic algorithms have also been explored, including adding population extinction and transfer strategies during the evolutionary process to maintain diversity and mitigate issues such as premature convergence [16]. Encapsulated algorithms that integrate search strategies evaluate feature importance directly using classifier accuracy, effectively combining wrapper methods with search strategies. Although these approaches yield high-quality results, they often require considerable computational time. Furthermore, genetic algorithms can encounter "premature convergence" issues during feature subset searches. To overcome these limitations, hybrid algorithms combining filtering and encapsulation techniques have been proposed. For example, Uguz et al. developed a hybrid algorithm that begins by using the information gain method to rank document features. Features are then filtered based on a predefined threshold. The remaining feature subsets are refined through a combination of GA and PCA..

For an object composed of features, a total of feature subsets can be generated. The objective of FS is to identify the optimal subset that benefits a specific task. This subset should not only minimize the number of features but also maximize the classification performance of the trained classifier. The fundamental architecture of a FS algorithm is illustrated in Figure 1.



**Figure.1The basic structure diagram of the FS algorithm**

Figure 1 illustrates this point, and the FS algorithm is generally divided into the following four basic steps:

(1) Subset generation: this step is the process of searching feature subset space through a specific search strategy. The main task is to provide corresponding feature subsets for subsequent evaluation functions. The search algorithms used can be divided into three categories: global search, heuristic search, and random search.

(2) Subset evaluation: it involves the use of various evaluation functions to assess the quality of selected feature subsets. These criteria are typically categorized into independent criteria and association criteria based on their relationship with the learning algorithm. Independent criteria operate separately from the learning algorithm. They rely on the intrinsic properties of the features and are commonly employed in filter-based FS methods [18]. Examples of independent criteria include information metrics, distance metrics, relevance measures, and consistency measures. In contrast, association criteria are closely tied to the learning algorithm. They use the algorithm's performance as the evaluation metric during feature subset selection. These criteria are more frequently applied in wrapper-based FS methods. While the computational complexity of association criteria is relatively high and their generalizability is limited, their effectiveness in FS remains significant.

(3) Termination condition: closely related to the evaluation function, it refers to the specific requirements to be met in the search process of the termination algorithm. There are three stopping criteria: execution time, evaluation times, and setting threshold.

(4) Result verification: verify the effectiveness of the selected feature subset through a priori knowledge and the results of experiments on the verification set.

## 3 Hybrid FS algorithm

### 3.1 GRRGA: hybrid FS algorithm

Typical FS algorithms face several challenges, such as the arbitrary nature of threshold settings in filtering methods and the "premature convergence" issue often encountered in genetic algorithms used for encapsulation. Moreover, simply combining filtering and encapsulation techniques does not fully leverage their respective strengths [19]. So, we propose and implement a hybrid FS algorithm, GRRGA, which integrates gain ratio-based sorting with a grouping evolutionary genetic algorithm. The process flow of the GRRGA algorithm is illustrated in Figure 2.

As shown in Figure 2, the GRRGA algorithm consists of three main stages:

1. Sorting with the filter algorithm: The original features are ranked using the gain ratio as the criterion.

2. Feature grouping: Based on the principle of equal-density division, the sorted features are divided into 𝑘 feature groups of equal density.

3. Search and evaluation with the wrapper algorithm: The evolutionary GA is applied to search within the feature groups, with the C4.5 algorithm used as the evaluation metric within the encapsulation process.



**Figure 2. Flow chart of GRRGA algorithm**

Filtering effectively reduces feature redundancy by selecting important subsets through thresholding, but its thresholds are often subjective and may not suit later classification tasks. To improve this, we propose a sorting and grouping algorithm based on information gain ratio, which ranks features and groups them to ensure uniform information density for better alignment with subsequent processes.

Suppose sample set , non-category feature is The -th feature of sample set is and the -th feature is , where . All features need to be normalized.:

(1)

Where is the information gain ratio. Then, we have the density distance:

(2)

From formula (1) and formula (2), it can be seen that the greater the , the greater the density The closer the information density distance between features and . The greater the density between features, the better[24]. Finally, the feature groups are reordered according to the method of maximum feature weight in the group so that the representative feature groups rank first. The algorithm is described in Table 1.

**Table 1. Althorithm flow**

|  |
| --- |
| Input: The original feature set , the number of feature groups |
| Output: The sorted feature groups |
| Step 01: Initialize sort feature set |
| Step 02: For i = 1,2,…,m-1 do |
| Step 03: compute the information gain ratio according to the equations below |
| Step 04: Add the feature after calculating the information gain ratio to the set |
| Step 05: Output sorted feature set |
| Step 06: End For |
| Step 07: For i = 1,2,…,m-1, do |
| Step 08: if then |
| Step 09: |
| Step 10: End If |
| Step 11: End For |
| Step 12: For i = 1,2,…,k do |
| Step 13: if then |
| Step 14: Use the following formulae to calculate the information density and density distance of features. |
| Step 15: Add feature into the -th feature group |
| Step 16: End If |
| Step 17: End For |

Feature coding in encapsulated FS involves determining whether a feature is selected or not, typically using binary coding. Specifically, binary coding is applied to each feature group in the GR, resulting in a binary string representing the selection status of the grouped features. : . is the combination of all individuals [25-26]. Then:

(3)

The length of the population is generally a multiple of the population size. In practical application, the population size is generally taken as a certain number between and . The initial population is generally selected by the random selection method. That is, the initial population composed of genes is generated by a random function[27].

In encapsulated FS, searches often rely on classification algorithms to model the sample set, using classification outcomes as the fitness function for evaluating feature sets. However, genetic algorithms frequently face the issue of local optima, where population diversity diminishes during evolution. To overcome this limitation, this paper introduces two population evaluation methods designed to assess initialized populations based on the sorted and grouped feature groups.

For individual within a population, the genetic algorithm is selected by the fitness of the individual, and individuals who are physically fit are more likely to be passed down to the following generation.The fitness function is:

(4)

Where, is the -th individual within the group , and represents the classification accuracy of each class in the class classification problem.

This paper adopts an improved fitness function:

(5)

Among them, is the individual subset screened within the group, and is a penalty factor that is used to keep the categorization accuracy in check and the size of the individual subset.

In order to make the individuals with the best fitness inherit, this paper selects a roulette strategy for selection operation. The specific process is as follows: assuming that the scale of the -th individual is , then means that the fitness of the individual is , then the probability of its selection is (such as formula (6)):

(6)

The selected individuals undergo evolutionary operations, with the expected number of parent individuals in the next generation determined by their fitness values .

Then single-point crossover and bit mutation strategies are applied. In single-point crossover, two parent individuals are randomly selected, and a crossover point is determined based on the crossover probability , resulting in two new offspring. For bit mutation, a gene location in each parent is chosen based on the mutation probability , and the gene is flipped from 0 to 1 or 1 to 0, creating a new individual (refer to the algorithm outlined in Table 2).

**Table 2. Grouping evolutionary genetic algorithm**

|  |
| --- |
| Input: Sorting grouped feature groups |
| Output: Optimal feature subset |
| Step 01: For i = 1,2,…,k do |
| Step 02: While or do |
| Step 03: according to equation (4) to compute for each individual in |
| Step 04: Generate a new generation of the population according to |
| Step 05: if |
| Step 06: |
| Step 07: Individuals are selected, crossed, and mutated |
| Step 08: Update: |
| Step 09: else |
| Step 10: =0.033 |
| Step 11: Individuals are selected, crossed, and mutated |
| Step 12: Update: |
| Step 13: End If |
| Step 14: End While |
| Step 15: Fitness evaluation of individuals outside the population: calculate for each group of individuals in according to formula (5) |
| Step 16: Save the best to array |
| Step 17: End For |
| Step 18: Returns the best subset of features from an array |

### 3.2 The C4.5 Algorithm

C4.5 is a decision tree algorithm in machine learning that recursively splits data based on attribute tests to maximize classification accuracy. The primary goal is to construct a decision tree with strong generalization ability using a "divide-and-conquer" approach. Selecting the optimal partition attribute is crucial, often relying on information gain, which measures sample set purity through information entropy calculated as per Equation (7):

(7)

If a discrete attribute with possible values , is used to divide a sample set , the information gain is calculated by weighting the information entropy of each branch node by its proportion of samples , as shown in Equation (8).:

(8)

While higher information gain indicates better partition attributes, the C4.5 algorithm mitigates the bias toward attributes with many values by using the gain ratio, as defined in Equations (9) and (10):

(9)

(10)

To address the gain rate's preference for attributes with fewer values, the C4.5 algorithm first identifies attributes with above-average information gain and then selects the one with the highest gain rate, while the CART decision tree uses the Gini index, defined in Equation (11), to measure dataset purity.

(11)

Intuitively, represents the likelihood that two randomly chosen samples from dataset 𝐷 belong to different categories, with smaller values indicating higher purity. The Gini index for attribute 𝑎 is defined in Equation (12).

(12)

The attribute that minimizes the Gini index after partitioning is selected as the optimal partition attribute from the candidate set , that is, .

## 4 Experiment and Results

To test the effectiveness of the GRRGA algorithm suggested in this paper, we selected six groups of UCI general data sets and two groups of UCI medical data sets for experiments. Among them, six groups of UCI common data sets come from different fields. For example, the spam base is mainly used to identify and classify spam. The comma is mainly used to monitor logistics and transportation, SCADA is mainly used to test the self-care of disabled children, etc. The two groups of UCI medical data sets were arrhythmia and cancer. Among them, arrhythmia is an arrhythmia data set with 452 samples, including 279 characteristics such as age, gender, height, and heart rate, which is mainly used to distinguish whether patients have arrhythmia; Cancer is a lung cancer data set with 32 samples and 56 features, which is mainly used to distinguish whether patients have cancer or not. See Table 3 for details of all data:

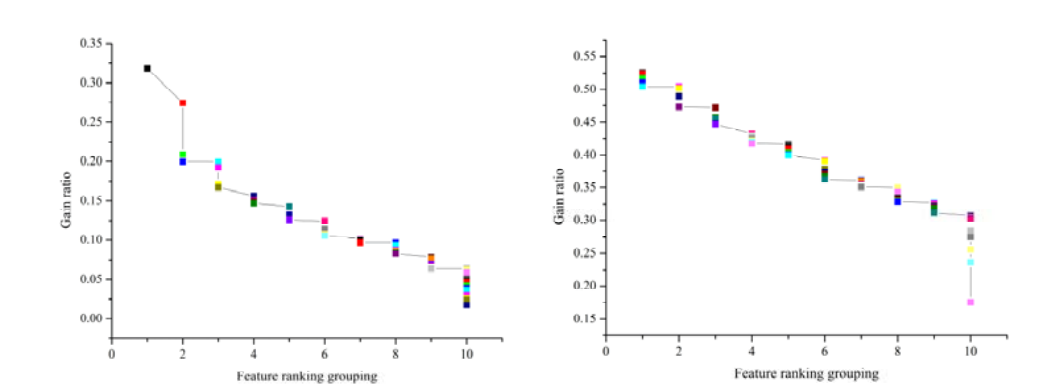
**Table 3. Experimental data set description**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset Attribute | Dataset Name | Count | Features | class |
| Validation Data | eighthr | 4736 | 72 | 2 |
| Spambase | 4601 | 57 | 2 |
| Comma | 3942 | 96 | 5 |
| Meu | 2856 | 71 | 63 |
| Urban | 675 | 147 | 8 |
| Scadi | 70 | 205 | 11 |
| Medical data | Arrhythmia | 452 | 279 | 3 |
| Cancer | 32 | 56 | 8 |

The FS algorithm evaluates filtering subsets based on classifier accuracy. In the experiments, the classifier's accuracy is assessed on both the original dataset and the dataset after FS. The study includes validation and application experiments.

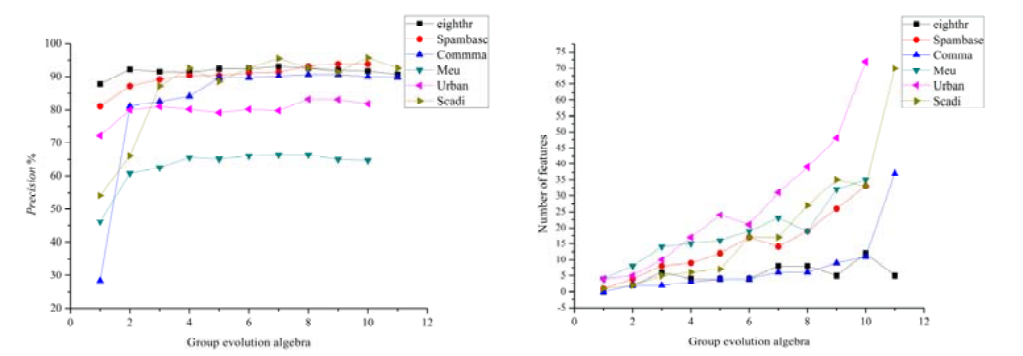
In the validation experiment, the GRRGA algorithm is tested, observing its filter and encapsulation stages by introducing interrupts. Comparisons are made between GRRGA and traditional FS methods. Additionally, the C4.5 algorithm used in the encapsulation stage is compared with four other classification algorithms, and ensemble methods are applied to reduce decision tree instability. The same algorithms and parameters are used for the application experiment.

The GRRGA algorithm experiment involves three steps: (1) feature sorting by calculating and ranking information gain ratios, (2) grouping features using the density bisection method, and (3) applying a grouping evolutionary genetic algorithm with C4.5 as the evaluation algorithm. Figure 3 illustrates the grouping results for the spam base and MEU datasets.



**Figure 3. Group diagram after sorting the characteristics of the data set**

The ordered features are as shown in Figure 3, and each group of the y-axis represents the feature subset in the group. For example, when feature ranking grouping = ten on the spam base dataset, it means that the 10th feature group has 14 features. Looking at Fig. 3, the smaller the information gain ratio of features, it means that more features are divided into a group. The larger the information gain ratio, it means that only one feature may be divided into a group (for example, feature ranking grouping = 1 on the spam base dataset). In addition, the features with an information gain ratio of 0 are not grouped, but they are still used in the process of population initialization. Then, the out of population evolution process of grouping evolutionary genetic algorithms can be given in Figure 4:



**Figure 4. Out of group evolution process in group evolutionary genetic algorithm**

Figure 4 illustrates this. With the increase of evolutionary algebra, the Precision index of the C4.5 algorithm on six groups of UCI data sets shows a trend of rising at first and then decreasing. When the evolutionary algebra begins, the Precision index of the C4.5 algorithm reaches the lowest value. This is mainly due to the fact that there are too few individuals in the population, which leads to the phenomenon of precocious; in the evolution of genetic algorithms, which leads to falling into the optimal local solution. Looking at figure 4, it is found that with the increase of evolutionary algebra, the number of feature subsets shows an upward trend and decreases on individual data sets but still increases as a whole. Therefore, this paper comprehensively considers the Precision of the C4.5 algorithm and the number of feature subsets. For example, the eighth generation of the Urban data set is selected as the optimal population size.

To compare the GRRGA algorithm's advantages, this section selects ten traditional FS algorithms for comparative experiments. Among them, the filtering algorithms are GR, ReliefF, and SU, respectively, the encapsulated algorithms are GA, PSO, and EA, respectively, and the hybrid algorithms are GRGA, GREA, and GRPSO, respectively. In order to improve the GRRGA algorithm's screening ability, in the process of FS, the same threshold parameters are used in the filtering stage of the traditional hybrid algorithm. The thresholds of three traditional filtering algorithms are set according to experience. Use the C4.5 algorithm tests the data set after FS. Tables 4 and 5 illustrate the results of the experiments.

**Table 4. Classification accuracy of GRRGA algorithm and traditional FS algorithm on six groups of UCI data sets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Original** | **Filter** | | |
| **GR** | **ReliefF** | **SY** |
| eighthr | 92.1 | 91.4 | 91.1 | 91.4 |
| Spambase | 93.0 | 90.9 | 91.0 | 92.7 |
| Comma | 89.5 | 88.8 | 88.4 | 88.7 |
| Meu | 64.2 | 65.2 | 42.8 | 64.7 |
| Urban | 80.2 | 82.1 | 77.6 | 77.9 |
| Scadi | 88.8 | 90.0 | 90.0 | 90.0 |
| Average | 84.63 | 84.73 | 80.15 | 84.23 |

**Table 5. Candidate feature subset size of GRRGA algorithm and traditional FS algorithm on six groups of UCI data sets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Original** | **Filter** | | |
| **GR** | **ReliefF** | **SY** |
| eighthr | 72 | 37 | 39 | 40 |
| Spambase | 57 | 24 | 25 | 18 |
| Comma | 96 | 41 | 64 | 41 |
| Meu | 71 | 28 | 42 | 30 |
| Urban | 147 | 64 | 35 | 35 |
| Scadi | 208 | 141 | 124 | 141 |
| Average | 108.00 | 55.83 | 54.83 | 50.83 |

Tables 4 and 5 show that the C4.5 algorithm achieves an average Precision of only 84.63% on six original datasets, mainly due to the negative impact of high-dimensional feature spaces and redundancy on its classification performance. Among traditional filtering algorithms, performance varies across datasets, but the G algorithm generally yields better results. For encapsulation algorithms, using PSO as the search strategy achieves the best results, with an average Precision of 85.72% across six datasets, offering a slight improvement. However, in both filtering and encapsulation methods, the recognition accuracy of feature subsets tends to decline after FS.

## 5 Conclusion

Based on previous studies, this paper divides ST in the cultural realm into two parts: internal ST in the cultural realm and external ST in the cultural realm. In the process of selecting evaluation indicators, we strictly follow the principles of the combination of qualitative and quantitative, comprehensiveness and systematicness, data accessibility and operability, guidance and development, and calculate the weight of the selected evaluation indicators by using AHP analytic hierarchy process through consultation and investigation of experts, so as to build a set of scientific, reasonable, practical and operable evaluation index system of ST in the cultural realm, The system includes six primary indicators, and these primary indicators are further refined into 22 secondary indicators. Then the fuzzy comprehensive evaluation method is used to comprehensively evaluate the index system of China's ST in the cultural realm, judge the level of China's ST in the cultural realm.

FS is a classical problem in the field of machine learning. With the continuous development of the era of big data, new applications emerge in endlessly, and the data generated increases exponentially. How select the features of huge data sets has always been one of the hot topics concerned by scientific researchers. Firstly, this paper introduces the basic framework of the FS algorithm and focuses on the four basic steps of FS, namely subset generation, subset evaluation, termination conditions, and result verification; Then, it expounds on the research methods and research results in the field of FS, classifies the FS algorithms from three aspects: evaluation criteria, search strategy, and supervision information compares the sub methods of various methods, and points out their advantages and disadvantages respectively.

There are many kinds of existing FS algorithms, but the defects are still inevitable. The increasing complexity of the research object makes the performance of the existing algorithms poor. How to design supporting schemes for practical problems needs to be solved urgently. Through the summary and analysis of FS algorithms, some challenges and technical means faced by FS are analyzed and prospected below.

One of the most difficult difficulties in the field of FS is how to make high-quality FS for high-dimensional data. At present, the main research idea is to use a multi-stage FS algorithm, which can not get satisfactory FS results. The embedded method based on evaluation strategy and the sequential search method based on search strategy, supplemented by resampling, incremental learning, and other technical means, maybe one of the better research ideas.

Designing a FS algorithm suitable for a big data environment has a lot of research space and great application value. At present, some work uses deep neural network technology to improve the robustness and stability of the algorithm. In the high-dimensional big data environment, how to combine deep learning technology with swarm intelligence algorithm and apply it to the field of FS is worthy of further exploration.

**Availability of data and material**

The data used to support the findings of this study are available from the corresponding author upon request.

**Competing interests**

Declares that he has no conflict of interest.

## Reference

1. Joseph S.Nye, Jr., Bound to Lead: The Changing Nature of American Power.New York: Basic Books,1990.
2. Yasushi, Watanad and David L.McConnell eds, Soft Power Superpowers: Cultural and National Assets of Japan and the United States.M.E.Sharpe Inc, 2008.
3. Greif, A.Genoa and Maghribi Traders: Historical Comparative Institutional Analysis.Cambridge: Cambridge University Press,1999.
4. Smith. A.The Theory of Moral Sentiments.Oxford: Oxford University Press,1976.
5. Joseph S.Nye, Jr., Bound to Lead: The Changing Nature of American Power,1990:29-35.
6. Consuelo Cruz.Identity and Persuasion: How Nations Remember Their Pasts and Make Their Futures.World Politics[J].2000,52(3).
7. Akiralriye, "Culture and International History," in Michael J.Hogan and Thomas G.Paterson(eds), Explaining the History of American Foreign Relations, Cambridge University Press,1991.
8. KOLLER D, SAHAMI M.Toward optimal feature selection[C]//Proceedings of the Thirteenth International Conference on International Conference on Machine Learning.Bari Italy,1996.
9. ALMUALLIM H,DIETTERICH T G.Learning Boolean concepts in the presence of many irrelevant features[J].Artificial Intelligence,1994,69 (1-2): 279-305.D0I:10.1016/0004-3702(94)90084-1.
10. DASH M,LIU H.Consistency-based search in fea- ture selection[J.Artificial Intelligence,2003,151(1- 2):155-176.D0I:10.1016/S0004-3702(03)00079-1.
11. KWAK N,CHOI C H.Input feature selection by mu- tual information based on Parzen window [J].IEEE Transactions on Pattern Analysis and Machine Intelli- gence,2002,24(12):1667-1671.D0I:10.1109/ TPAMI.2002.1114861.
12. PENG H C,LONG F H,DING C.Feature selection based on mutual information criteria of max-dependen- cy,max-relevance,and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226-1238.D0I:10.1109/ TPAMI.2005.159.
13. ESTEVEZ P A,TESMER M,PEREZ C A,et al. Normalized mutual information feature selection[J]. IEEE Transactions on Neural Networks,2009,20 (2):189201.D0I:10.1109/TNN.2008.2005601.
14. BRUNO M, BATTISTI R.X-MIFS: exact mutual information for feature selection [C]//Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN).Vancouver: IEEE,2016.DOI:10. 1109/JCNN.2016.7727644.
15. KOHAVI R,JOHN G H.Wrappers for feature sub- set selection[J].Artificial Intelligence,1997,97 (1- 2):273-324. D0I:10.1016/S0004-3702(97) 00043-X.
16. SINDHWANI V,RAKSHIT S,DEODHARE D,et al.Feature selection in MLPs and SVMs based on maximum output information[J].IEEE Transactions on Neural Networks,2004,15 (4):937-948.DOI: 10.1109/TNN.2004.828772.
17. YANG J B,SHEN K Q,ONG C J,et al.Feature se- lection for MLP neural network:the use of random permutation of probabilistic outputs[J].IEEE Trans- actions on Neural Networks,2009,20 (12):1911- 1922.D0I:10.1109/TNN.2009.2032543.
18. Cao TT, ZHANG M J, ANDREAE P, et al.A wrapper feature selection approach to classification with missing data[C]//Proceedings of the 19th European Conference on the Applications of Evolutionary Computation.Springer International Publishing.Por- to, Portugal: Springer,2016.DOI:10.1007/978-3- 31931204-044.
19. QUINLAN J R.Introduction of decision trees [J]. Machine Learning,1986,1(1):81-106.
20. BREIMAN L,FRIEDMAN J H,OLSHEN R A,et al.Classification and regression trees[J].Biometrics, 1984,40(3):874.D0I:10.2307/2530946.
21. Z. Sadeghian, E. Akbari, H. Nematzadeh, H. Motameni.A review of feature selection methods based on meta-heuristic algorithms J. Experiment. Theor. Artif. Intell., 35 (1) (2023), pp. 1-51
22. H. Cui Enhanced harris hawks optimization integrated with coot bird optimization for solving continuous numerical optimization problems. Comput. Model. Eng. Sci., 137 (2) (2023), pp. 1-41
23. L. Peng, Z. Cai, A. A. Heidari, L. Zhang, H. Chen. Hierarchical Harris hawks optimizer for feature selection J. Adv. Res., 53 (1) (2023), pp. 261-278
24. S. X. Y. X. Tang. Sensitive feature selection for industrial flotation process soft sensor based on multiswarm PSO with collaborative search. IEEE Sens. J., 24 (10) (2024), pp. 17159-17168
25. D. Yilmaz Eroglu, U. Akcan. An adapted ant colony optimization for feature selection Appl. Artif. Intell., 38 (1) (2024, Art. no. 2335098)
26. S. Zhang, K. Liu, T. Xu, X. Yang, A. Zhang. A meta-heuristic feature selection algorithm combining random sampling accelerator and ensemble using data perturbation. Appl. Intell., 53 (24) (2023), pp. 29781-29798
27. R. Jiao, B. Xue, M. Zhang. Learning to preselection: A filter-based performance predictor for multiobjective feature selection in classification. IEEE Trans. Evol. Comput., 28 (1) (2024), pp. 1-15
28. S. Zhuo, J. Qiu, C. Wang, S. Huang. Online feature selection with varying feature spaces. IEEE Trans. Knowl. Data Eng., 36 (9) (2024), pp. 1-14