IMPACT: International Journal of Research in
Engineering & Technology
ISSN(P): 2347–4599; ISSN(E): 2321–8843
Vol. 8, Issue 2, Feb 2020, 1–10
© Impact Journals

ANALYSIS OF INNOVATION PERFORMANCE OF CHINESE HIGH-TECH ZONES BASED ON IMPROVED K-MEANS AND DOMINANCE ROUGH SET

Zhao Xiaoyu

Research Scholar, Department of Economics and Management, Nanjing University of Aeronautics and Astronautics, Jiangsu, China

Received: 12 Feb 2020 **Accepted:** 20 Feb 2020 **Published:** 29 Feb 2020

ABSTRACT

As an important platform for innovation, the Chinese High-tech Zone undertakes the important responsibility of using innovation to drive industry and economic progress. Starting from the multiple stages of innovation in the high-tech zone, we will examine the impact of innovation resources input, output of achievement and transformation of results on innovation performance. Relying on the cross-section data of 115 national key parks, we will select indicators for innovation performance analysis from the perspective of input, output, and transformation. The discrete method based on improved K-Means is used to discretize the data, and the dominance rough set is introduced to reduce the information system and explore the relationship between the various links of the innovation of the high-tech zone and the final innovation performance. The results show that the output and transformation stages are the main stages affecting the innovation performance of the high-tech zone. The R&D results and the level of conversion services invested in the transformation phase are important factors influencing the innovation performance of the high-tech zone. If the results and service levels of the high-tech zones are low, their innovation performance is generally at a relatively backward level in the country otherwise, the level of innovation performance is higher.

KEYWORDS: *Chinese High-Tech Zone, Innovation Performance, Dominance Rough Set, K-Means*

INTRODUCTION

As the first driving force for development, innovation is the source of promoting national development. As an important platform for carrying out innovation activities and promoting innovation development, high-tech zones have become the focus of attention from all walks of life. China's high-tech zones have many factors that influence innovation development, and have produced a series of considerable innovation results. Based on 2017 statistical data, in terms of the input of innovation resources, the high-tech zone has pooled more than 30% of China's R & D investment and gathered more than 50% of high-tech companies [1]. From the perspective of innovation achievements, the number of authorized invention patents obtained by enterprises in high-tech zones in 2017 accounted for 46.3% of the total number of authorized invention patents in China, and the number of domestic effective invention patents owned by employees in high-tech zones reached more than 8 times the national average [2]. In summary, the Chinese high-tech zones have become the main force for improving the level of innovation.

As a main force of national innovation, high-tech zones have great advantages in realizing industry and economic development through innovation. High-tech zone's innovation is a dynamic system process, which mainly includes two stages of technology research and development and achievement transformation, and it finally outputs products with economic value. The ultimate goal of innovation in high-tech zones is to obtain rich innovation results and to promote the continuous improvement of the level of industrial economy, which is an important dimension to measure the final innovation performance of high-tech zones. Therefore, identifying the factors that affect innovation performance from the two major stages of innovation development in high-tech zones, screening important indicators related to innovation performance in high-tech zones, and obtaining potential rules for innovation performance in high-tech zones can improve the innovation performance of high-tech zones.

There are various methods of performance analysis, including analytic hierarchy process, data envelopment analysis, factor analysis and so on. Dominance rough set has the advantages of knowledge induction, attribute reduction, rule extraction, and the results easy to understand. Compared with classic rough set theory, dominance rough sets can more effectively deal with decision analysis problems with preference attribute information, so the theory applied to the innovation performance analysis of high-tech zones is reasonable and applicable. In addition, the discretization results obtained by using the K-Means-based discretization method have high classification quality and approximate accuracy [3], so this paper uses it as the optimal data discretization method. Based on the above analysis, this article comprehensively considers the various stages of innovation development in high-tech zones, selects performance analysis indicators from multiple perspectives on innovation resource input, outcome output, and achievement transformation, and uses the advantage rough set to screen the indicators, and excavates important influence factors, Knowledge rules between these factors and innovation performance.

The first part of this article explains the important role of innovation in high-tech zones and the applicability of dominance rough set theory in innovation performance analysis. The second part gives the specific research methods. The third part explores the process of innovation in China's high-tech zones, uses relevant park data, combines K-Means and dominance rough set to mine knowledge rules for innovation performance in high-tech zones, and analyzes the results. Finally, conclusions are given.

RESEARCH METHOD

This article uses the dominance rough set method to evaluate and analyze the innovation performance of Chinese high-tech zones. The dominance rough set method can naturally and effectively maintain the partial order relationship of each attribute value, and it is an important method for analyzing information systems with preference order. Discretization of attribute values is a prerequisite for using rough set. Therefore, this paper first used data discretization method based on improved K-Means to discretize continuous data in the information system, and then used dominance rough set to perform knowledge mining on the information system. In the end, objective and reasonable performance analysis results are obtained.

Data Discretization Method based on Improved K-Means

Discretization of continuous attribute data is an important part of performance evaluation analysis using rough set method, and its effect will directly affect the subsequent evaluation analysis results [4]. K-Means algorithm (KCM) [5] is an unsupervised clustering algorithm that aggregates data objects with high similarity in the same class, thereby achieving the effect of discretizing continuous data.

The steps for discretizing continuous attribute data using the classic K-Means algorithm [6-7] are as follows:

Firstly, selecting a data object from the original data object set as the initial cluster center. Secondly, calculating the Euclidean distance between each data object and the cluster center, and assign each data object to the nearest category according to the calculation result. Then, updating the cluster center. Finally, repeating the above process until the evaluation function converges.

The classical K-Means clustering method has the problem that the cluster center and the number of clusters cannot be accurately determined. Therefore, cluster indicators are introduced. The cluster index is very sensitive to the number of clusters. When the cluster number is greater than or equal to the optimal number of clusters, the cluster index value will slowly decrease. When the value of the number of clusters is less than or equal to the optimal number of clusters, the value of the cluster index will drop sharply [8]. The average value of the average centroid distance of the clusters is selected as a cluster index for judging whether the number of clusters is appropriate, and the calculation formula is shown in formula (1)[9].

$$E = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum d\left(x_j, e_i\right) \tag{1}$$

Where, k is the number of clusters, n_i is the total number of data objects in the i-th classC_i, $d(\cdot)$is the Euclidean distance function, e_i is the cluster center of C_i.

The continuous attribute data discretization method based on improved K-Means first divides the continuous attribute data object into multiple clusters through the K-Means clustering method, and secondly introduces the cluster index to determine the optimal classification number. Finally, the class cluster label of each class is extracted and used to replace all continuous data in the cluster to achieve discretization.

Dominance Rough Set Theory

Rough set theory was pioneered by Polish scientist Z. Pawlak in 1982. As a mathematical analysis tool that can be effectively applied to uncertain knowledge expression systems, its main idea is to reduce knowledge and mine system classification rules on the premise of ensuring that the classification level of the knowledge expression system is unchanged. In the application of classic rough set theory, if there is a preference attribute in the knowledge expression system, it will lead to inconsistent decision-making in typical cases. Therefore, Greco and other scholars proposed the dominance rough set theory, which improved the deficiencies of the classic rough set by replacing the indistinguishable relationship with the dominating relationship [10].

The Basic Theoretical Method of Dominance Rough Set is as follows

Knowledge Expression System and Decision Table

If the knowledge expression system $S = (U, A, V, f)$, then U is a non-empty finite object set, also known as the universe object space;A is a non-empty finite attribute set, $A = C \cup D, C \cap D = \Phi$, where C is the conditional attribute set and D is the decision attribute set, and S is also called a decision table at this time; $V = \bigcup_{a \in A} V_a$, where $a \in A$, V_a is the range of attribute a's values; $f: U \times A \to V$ is an information function, for$\forall x \in U$, $\forall a \in A$, there is $f(x, a) \in V_a \subseteq V$ [11].

Dominating Relationship and Dominating Set

Let $\mathrm{P} \subseteq \mathrm{C}$, $\mathrm{x}, \mathrm{y} \in \mathrm{U}$, if $\forall \mathrm{q} \in \mathrm{P}$, $\mathrm{f}(\mathrm{y}, \mathrm{q}) \geq \mathrm{f}(\mathrm{x}, \mathrm{q})$ are all true, then y is better than x on the attribute P, denoted as $\mathrm{yD_px}$, $\mathrm{yD_px}$ is the dominant relationship.

Given $\mathrm{P} \subseteq \mathrm{C}$ and $\mathrm{x}, \mathrm{y} \in \mathrm{U}$, define the P-dominating sets of x as[12]:

$$\mathrm{D} +_{\mathrm{P}} (\mathrm{x}) = \{\mathrm{yD_P x}\} \tag{2}$$

Dominant Rough Approximation

According to the value of the decision attribute D, the universe object space can be divided: $\mathrm{U/D} = \mathrm{Cl} = \{\mathrm{Cl_t}, \mathrm{t} = 1,2, \cdots \mathrm{n}\}$. Where $\mathrm{Cl_t}$ is the t-th equivalent class, and $\mathrm{Cl_n} > \cdots > Cl_{\mathrm{t}} > \cdots > Cl_1$, if $\mathrm{Cl_t}$ is combined up or down, then:

$$\mathrm{Cl} \geq_{\mathrm{t}} = \cup_{\mathrm{s} \geq \mathrm{t}} \mathrm{Cl_s}, \quad \mathrm{Cl} \leq_{\mathrm{t}} = \cup_{\mathrm{s} \leq \mathrm{t}} \mathrm{Cl_s}, \quad \mathrm{t}, \mathrm{s} \in \{1,2, \cdots, \mathrm{n}\} \tag{3}$$

The lower and upper approximations of $\mathrm{Cl}_{\mathrm{t}}^{\geqq}$ are recorded as:

$$\underline{\mathrm{apr}}_{\mathrm{P}}(\mathrm{Cl}_{\mathrm{t}}^{\geqq}) = \cup \{\mathrm{x} \in \mathrm{U}: \mathrm{D}_{\mathrm{P}}^{+}(\mathrm{x}) \subseteq \mathrm{Cl}_{\mathrm{t}}^{\geqq}\} \tag{4}$$

$$\overline{\mathrm{apr}}_{\mathrm{P}}(\mathrm{Cl}_{\mathrm{t}}^{\geqq}) = \cup \{\mathrm{x} \in \mathrm{U}: \mathrm{D}_{\mathrm{P}}^{-}(\mathrm{x}) \cap \mathrm{Cl}_{\mathrm{t}}^{\geqq} \neq \phi \} \tag{5}$$

Similarly, the lower and upper approximations of $\mathrm{Cl}_{\mathrm{t}}^{\leqq}$ are:

$$\underline{\mathrm{apr}}_{\mathrm{P}}(\mathrm{Cl}_{\mathrm{t}}^{\leqq}) = \cup \{\mathrm{x} \in \mathrm{U}: \mathrm{D}_{\mathrm{P}}^{-}(\mathrm{x}) \subseteq \mathrm{Cl}_{\mathrm{t}}^{\leqq}\} \tag{6}$$

$$\overline{\mathrm{apr}}_{\mathrm{P}}(\mathrm{Cl}_{\mathrm{t}}^{\leqq}) = \cup \{\mathrm{x} \in \mathrm{U}: \mathrm{D}_{\mathrm{P}}^{+}(\mathrm{x}) \cap \mathrm{Cl}_{\mathrm{t}}^{\leqq} \neq \phi \} \tag{7}$$

Classification Quality and Attribute Reduction

The ratio of the number of correctly classified objects to the total number of objects in the knowledge expression system is called classification quality. The classification quality calculation formula of Clis:

$$\gamma_{\mathrm{P}}(\mathrm{Cl}) = \frac{|\mathrm{U} - ((\cup \mathrm{bnd}(\mathrm{Cl}_{\mathrm{t}}^{\geqq})) \cup (\cup \mathrm{bnd}(\mathrm{Cl}_{\mathrm{t}}^{\leqq})))|}{|\mathrm{U}|} \tag{8}$$

The smallest subset $\mathrm{P} \subseteq \mathrm{C}$ that satisfies $\gamma_{\mathrm{P}}(\mathrm{Cl}) = \gamma_{\mathrm{C}}(\mathrm{Cl})$ is called a reduction of C with respect to Cl, and is denoted as $\mathrm{RED_{Cl}(P)}$.

Preference Decision Rules

After obtaining the dominant rough approximation, the following preference decision rules can be derived:

If $\mathrm{f}(\mathrm{x}, \mathrm{q}_1) \geq \mathrm{r}_{\mathrm{q}_1} \wedge \mathrm{f}(\mathrm{x}, \mathrm{q}_2) \geq \mathrm{r}_{\mathrm{q}_2} \cdots \wedge \mathrm{f}(\mathrm{x}, \mathrm{q}_\mathrm{p}) \geq \mathrm{r}_{\mathrm{q}_\mathrm{p}}$thenx $\in \mathrm{Cl}_{\mathrm{t}}^{\geqq}$;

If $\mathrm{f}(\mathrm{x}, \mathrm{q}_1) \leq \mathrm{r}_{\mathrm{q}_1} \wedge \mathrm{f}(\mathrm{x}, \mathrm{q}_2) \leq \mathrm{r}_{\mathrm{q}_2} \cdots \wedge \mathrm{f}(\mathrm{x}, \mathrm{q}_\mathrm{p}) \leq \mathrm{r}_{\mathrm{q}_\mathrm{p}}$thenx $\in \mathrm{Cl}_{\mathrm{t}}^{\leqq}$.

Where, $(\mathrm{q}_1, \mathrm{q}_2, \cdots, \mathrm{q}_\mathrm{p}) \subseteq \mathrm{C}$, $(\mathrm{r}_{\mathrm{q}_1}, \mathrm{r}_{\mathrm{q}_2}, \cdots, \mathrm{r}_{\mathrm{q}_\mathrm{p}}) \in \mathrm{V}_{\mathrm{q}_1} \times \mathrm{V}_{\mathrm{q}_2} \times \cdots \times \mathrm{V}_{\mathrm{q}_\mathrm{p}}$, $\mathrm{t} \in (1,2, \cdots, \mathrm{n})$.

ANALYSIS OF INNOVATION PERFORMANCE OF CHINESE HIGH-TECH ZONES

Indicator Selection and Data Source

As a typical institution of cross-organizational cooperation integrating enterprises, scientific research institutions and intermediary service platforms, the high-tech zone includes a number of activity links in its innovation process. The specific process is: resource input → technology research and development → achievement output → achievement transformation → production application → the company's operating income or industrial economic level has improved. In the research of innovation performance of high-tech zones, if we analyze from the traditional input and output perspectives, it is easy to ignore the multi-stage nature of high-tech zones' innovation activities and the link relationship between various stages. This article comprehensively analyzes the innovation performance of high-tech zones from multiple dimensions of resource input, output and transformation of results, and further explores the relationship between innovation factors and performance at each stage.

Innovation in the high-tech zone mainly includes the technological research and development stage and the achievement transformation stage. During the technological research and development stage, the high-tech zone invests in scientific research personnel and funds to produce scientific and technological innovation achievements. Then invest in the scientific and technological innovation services related to the transformation of results, and finally obtain the innovation performance represented by the operating income of high-tech industries. Starting from the relationship between the two major stages of innovation in high-tech zones, it is not difficult to find that the results of scientific and technological innovation can only be used as intermediate output, and its final performance needs to be measured by the relevant industrial economic level. Therefore, this article uses the high-tech industry's operating income as the decision attribute index of rough set analysis, and takes the resource input, outcome output and transformation factors involved in the innovation process as conditional attribute indicators. The specific attribute indicator settings are shown in Table 1.

Table 1: Attribute Indicators of High-Tech Zones' Innovation Performance Analysis

Indicator Properties	Indicator Dimension	Indicator Name	Unit	Code
Conditional attribute indicator	Investment in scientific and technological innovation	R & D personnel full-time equivalent	Man-year	C_1
		R & D internal expenditure	Ten thousand yuan	C_2
	Scientific and technological innovation achievements	Formation of national or industry standards and participation in the development of international standards	Pieces	C_3
		Number of significant intellectual property rights	Pieces	C_4
	Scientific and technological innovation service	Number of innovation service agencies	Each	C_5
		Number of companies in technology incubators and accelerators	Each	C_6
		High-tech service industry employees	people	C_7
Decision attribute indicator	Scientific and technological innovation performance	High-tech industry operating income	Ten thousand yuan	D

The research data comes from the "National Key Park Innovation Test Report" [13] recently released by the Ministry of Science and Technology of the People's Republic of China. From this report, relevant statistical data of 115 national key high-tech zones are selected to form an information system for the innovation performance analysis of national high-tech zones.

Data Discrete and Construction of Multi-Criteria Decision Table based on Improved K-Means

Because the numerical value of each attribute index in the information system is large (that is, the difference between the maximum and minimum values of the attribute index is large), it will affect the effect of K-Means classification. Standardized processing to eliminate the influence of the size and value of the attribute index itself. The raw data is processed using the standardization method, and the calculation formula is as follows:

$$x' = \frac{\max - x}{\max - \min} \tag{9}$$

Where, max is the maximum value of the attribute index, min is the minimum value of the attribute index, and x is the original data.

Then, according to the data discretization method based on the improved K-Means proposed in the previous chapter, the standardized data is discretized to obtain the multi-standard decision table shown in Table 2.

Table 2: Multi-Criteria Decision Table for High-Tech Zones' Innovation Performance Analysis

High-tech Zone	C_1	C_2	C_3	C_4	C_5	C_6	C_7	D
Beijing Zhongguancun	High	High	High	High	High	High	High	High
Tianjin	Low	Medium	Second High	Second High	Second High	Second High	Low	Second High
Xinjiang Production and Construction Corps	Low	Low	Low	Low	Low	Low	Low	Low

Attribute Reduction and Rule Generation Based on Dominance Rough Set

As can be seen from Table 2, both the condition attribute and the decision attribute in the decision tables have preference information. According to the decision attribute, four preference order classes can be obtained: $\mathrm{Cl}_1 = \{\mathrm{Low}\}, \mathrm{Cl}_2 = \{\mathrm{Second\ Low}\}, \mathrm{Cl}_3 = \{\mathrm{Second\ High}\}, \mathrm{Cl}_4 = \{\mathrm{High}\}$. And then the following decision classes are obtained:

- $\mathrm{Cl}_1^{\leqq} = \mathrm{Cl}_1$, indicating that the innovation performance of high-tech zones is low;
- $\mathrm{Cl}_2^{\leqq} = \mathrm{Cl}_1 \cup \mathrm{Cl}_2$, indicating that the innovation performance of high-tech zones is at most second low;
- $\mathrm{Cl}_2^{\geqq} = \mathrm{Cl}_2 \cup \mathrm{Cl}_3 \cup \mathrm{Cl}_4$, indicating that the innovation performance of high-tech zones is at least second low;
- $\mathrm{Cl}_3^{\leqq} = \mathrm{Cl}_1 \cup \mathrm{Cl}_2 \cup \mathrm{Cl}_3$, indicating that the innovation performance of high-tech zones is at most second high;
- $\mathrm{Cl}_3^{\geqq} = \mathrm{Cl}_3 \cup \mathrm{Cl}_4$, indicating that the innovation performance of high-tech zones is at least second high;
- $\mathrm{Cl}_4^{\geqq} = \mathrm{Cl}_4$, indicating that the innovation performance of high-tech zones is high.

115 analysis objects were included in the training set and testing set. Based on the statistical data of a large sample, the amount of data must not be less than 30, so the first 85 analysis objects are included in the training set, and the testing set contains the last 30 analysis objects.

The most popular algorithm for dominance rough set is DOMLEM, which can be implemented using software 4e Mka2. First enter the training set data in 4e Mka2 to determine the dominant relationship; then search the reduction and the kernel of the reduction for the training set; and finally derive the preference decision rules based on the reduction.

A total of three reductions in the training set were searched, which are $\{C_4, C_5, C_7\}, \{C_3, C_4, C_6, C_7\}, \{C_2, C_3, C_4, C_7\}$.The core of the reduction is $\{C_4, C_7\}$. According to the check of the reduction, it can be seen that the number of important intellectual property rights and employees of high-tech service industries are important factors affecting the innovation performance of high-tech zones, that is, the investment in the transformation stage is the key to determine the innovation performance of high-tech zones, and the impact of investment in the technology research and development stage on innovation performance in high-tech zones is relatively small.

The preference decision rule sets derived from the reduction $\{C_4, C_5, C_7\}$ are shown in Table 3 and Table 4.

Table 3: $D_{\leq}$Preference Decision Rule Set

No.	Preference Decision Rule	Support
1	If number of significant intellectual property rights is low and number of innovation service agencies is at most second low and number of employees in high-tech service industry is at most second low, then the innovation performance of high-tech zones is low.	65
2	If number of significant intellectual property rights is at most second low and number of innovation service agencies is at most second high and number of employees in high-tech service industry is at most second low, then the innovation performance of high-tech zones is at most second low.	12
3	If number of significant intellectual property rights is at most second high and number of innovation service agencies is at most second high and number of employees in high-tech service industry is at most second high, then the innovation performance of high-tech zones is at most second high.	2
4	If number of significant intellectual property rights is high and number of innovation service agencies is high and number of employees in high-tech service industry is high, then the innovation performance of high-tech zones is high.	1

Table 4: $D_{\geq}$Preference Decision Rule Set

No.	Preference decision rule	Support
1	If number of significant intellectual property rights is low and number of innovation service agencies is at most second low and number of employees in high-tech service industry is at most second low, then the innovation performance of high-tech zones is low.	65
2	If number of significant intellectual property rights is at least second low, then the innovation performance of high-tech zones is at least second low.	9
3	If number of significant intellectual property rights is at least second high and number of employees in high-tech service industry is at least second low, then the innovation performance of high-tech zones is at least second high.	2
4	If number of significant intellectual property rights is high and number of innovation service agencies is high and number of employees in high-tech service industry is high, then the innovation performance of high-tech zones is high.	1

It can be known from Table 3 or Table 4 that the derived preference decision rules have correctly classified most of the high-tech zones in the training set, the classification quality is higher than 90%, and the classification accuracy is extremely high. Reading the rule set, we can find:

- The level of innovation performance of national high-tech zones is in a state where very few key parks (such as Beijing Zhongguancun) are far ahead of most parks.
- The output level of important intellectual property rights is a key factor that determines whether innovation performance can break second lower level.
- In the stage of transformation of innovation results, if the investment results and service of high-tech zones are low, the innovation performance is generally at a relatively backward level; otherwise, the innovation performance level is higher.

Applying 30 analysis objects in the testing set for rule matching, the results show that 29 objects completely match the decision rule set, that is, the information contained in the testing set basically matches the rule sets shown in Table 3 and Table 4. Therefore, the rule set mines and displays most of the knowledge in the information system, and can reasonably classify and evaluate the overall innovation performance of high-tech zones. In addition, the consistency of the analysis results between the training set and the testing set indicates that the dominance rough set theory is universal and effective in the performance analysis of high-tech zones.

CONCLUSIONS

Because the innovation performance analysis of high-tech zones is relatively complicated, there are various factors affecting innovation performance, and there may be mutual constraints or dependencies between these factors. It is necessary to select relevant performance analysis indicators as comprehensively and reasonably as possible. Therefore, on the basis of considering the multi-stage characteristics of innovation in high-tech zones, this paper selects relevant indicators from the dimensions of innovation resource input, output and transformation for the analysis of innovation performance in high-tech zones. In addition, this article applies the dominance rough set theory to the analysis of the innovation performance of national high-tech zones, making full use of the information of the data itself, and obtaining a more objective set of decision rules. For the evaluation analysis problem with preference information, such as innovation performance analysis, the dominance rough set theory considers the impact of preference information on the knowledge system, which is not only close to the objective reality, but also simplifies the complexity of the rules.

Because the K-Means discretization method cannot obtain the numerical intervals of various clusters, this makes the rules of dominance rough set generation not specific to the interval values. In addition, this article treats the high-tech industry operating income as the ultimate innovation performance of high-tech zones, and there are still some innovation performances to be studied, such as the number and output value of new products.

REFERENCES

1. *National Bureau of Statistics. China statistical yearbook [J]. Beijing: China Statistics Press, 2018: 669–671.*
2. *CaiHong, L. (2014). Operation Research and Decision Making The development of inference machine model for vocation psychology based on rough set theory. COMPUTER MODELLING & NEW TECHNOLOGIES, 18(12C), 16–22.*
3. *Torch High-tech Industrial Development Center of the Ministry of Science and Technology, China High-tech Zone Research Center of the Academy of Science and Technology Strategic Consulting, Chinese Academy of Sciences. National high-tech zone innovation capability evaluation report [R]. Beijing: Science and Technology Literature Press, 2018: 9.*
4. *Xie Ming, Ji Weizhuo. Comparative study of several Discretization Methods in Rough Set Theory [J]. Fuzzy Systems and Mathematics, 2016,30 (04): 135–143.*
5. *Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features[C]. Proceedings of the 12th International Conference on Machine Learning. San Francisco: Morgan Kaufmann,1995: 194–202.*

6. *Lansey K. Optimization of water distribution network design using the shuffled frog leaping Algorithm[J]. Journal of Water Resources Planning and Management, 2003,129(3):210–225.*

7. *Shun Long, Ku Tao, Zhou Hao. Accelerated K-Means clustering algorithm for large data sets with multiple cluster centers [J]. Application Research of Computers, 2016,33 (2): 413–416.*

8. *ShahrivarS , Jalili S. Single-pass and linear-time k-means clustering based on Map Reduce[J]. Information Systems,2016,60(C):1–12.*

9. *Rajaraman, Ullman. Big data: large-scale data mining and distributed processing on the Internet [M]. Wang Bin, Translation. Beijing: People's Posts and Telecommunications Press, 2012,187–189.*

10. *Zhou Weiben, Shi Yuexiang. Optimization algorithm of K-Means clustering center selection based on density [J]. Application Research of Computers, 2012, 29 (5): 1726–1728.*

11. *Gao Jianshan, Lu Shiwen. Research on quality evaluation of electronic resources based on dominance rough set and grey clustering [J], Computer Applications and Software, 2010, 27 (9), 149–159.*

12. *Jian Lirong, Liu Sifeng, Xie Naiming. Probabilistic decision-making method of heterogeneous gray clustering and extended advantage rough set [J]. Journal of Systems Engineering, 2010, 25 (04): 554–560.*

13. *Ziarko, W. Variable precision rough set model[J]. Journal of Computer and System Sciences,1993,46(1):39–59.*

14. *Ministry of Science and Technology of the People's Republic of China. Innovation monitoring report of national key parks [R].Beijing: Science and Technology Literature Publishing House, 2016: 6–11.*

AUTHOR PROFILE

Zhao Xiaoyu, a postgraduate student at the School of Economics and Management, Nanjing University of Aeronautics and Astronautics from 2017 to 2020 .The research direction during the school was mainly the management of scientific and technological innovation.