Cybersecurity Policy Classification, National Clustering, and Transnational Transfer Based on SVM and HCA

Yue Guo †, *, Yiling Yang †, Tianyu Liu †

Harbin Institute of Technology, Harbin, China

† These authors also contributed equally to this work

* Corresponding Author Email: yg_0312224@qq.com

Abstract. This paper proposes a cybersecurity policy analysis and transfer model integrating natural language processing (NLP), support vector machines (SVM), hierarchical cluster analysis (HCA), and network analysis. It focuses on exploring the patterns of utility flow in policy classification, country clustering, and policy transfer. First, a dataset is constructed by extracting policy text word vectors using TF-IDF. SVM classifies policies into four categories: prevention, accountability, cooperation, and emergency response. Policy effectiveness is calculated by comparing crime counts between control and experimental groups. Second, using per capita GDP, higher education enrollment rate, and internet penetration rate as indicators, countries undergo data standardization. Euclidean distance measures clustering distance, and HCA categorizes 141 countries into five groups. Finally, a multiple linear regression model linking countries and cybercrime was constructed. The policy transfer guidance coefficient (incorporating non-transfer scenarios and the weakest-link effect) was defined. Combined with cosine similarity calculations for country similarity, relative policy effect coefficients were derived. Network analysis visualized policy efficacy diffusion. This model enables precise policy classification, scientific country clustering, and efficient policy transfer. Its advantage lies in integrating multiple algorithms to enhance the scientific rigor of analysis and policy transfer.

Keywords: Support vector machine, hierarchical clustering analysis, multiple linear regression.

1. Introduction

This paper focuses on the classification modeling of global cybersecurity policies, the clustering of national characteristics, and the analysis of policy transfer effectiveness [1]. It aims to construct a technical framework using multiple algorithms to address the issues of imprecise policy classification and the difficulty in quantifying the impact of national differences on policy outcomes. First, TF-IDF is employed to extract policy text word vectors and construct the dataset. Support Vector Machines (SVM) with an RBF kernel classify policies into four categories: prevention, accountability, cooperation, and emergency response. Policy effectiveness is then calculated using a crime prediction function for control and experimental groups [2] [3]. Second, using per capita GDP, higher education enrollment rate, and internet penetration rate as indicators, we standardize data and measure clustering similarity via Euclidean distance. Hierarchical Cluster Analysis (HCA) divides 141 countries into five characteristic groups [4]. Finally, a linear regression model linking countries with cybercrime rates was established. A policy transfer guidance coefficient (accounting for non-transfer scenarios and the weakest-link effect) was defined. Combined with cosine similarity to calculate relative policy effectiveness coefficients, network analysis visualized utility flows [5] [6].

Experimental results demonstrate this framework enables precise policy classification and scientific country clustering, with controllable policy transfer utility losses, providing support for optimizing transnational cybersecurity policies.

2. Network Security-related Policy Analysis Based on TF-IDF and SVM

To classify the policy all over the world and model the corresponding utilities. So, in this section, first construct the policy dataset based on natural language processing (NLP). Use word vectors as features and employ the support vector machine (SVM) to determine the classification rules.

2.1. Construction of Cybersecurity Policy Dataset Based on TF-IDF

Construct the policy dataset through these main steps:

Step 1: Word Vector Features

Term - Frequency - Inverse Document - Frequency (TF - IDF) is a common method of text feature representation that can measure the importance of words in a document. It is calculated in two aspects:

(1) Term Frequency (TF), which is the frequency of the word appearing in the policy;

(2) Inverse Policy Frequency (IPF), which measures the rarity of words in the entire policy collection. If a word appears in many policies, it has a lower IDF value and vice - versa.

Define w as the word and d as the policy. Based on the definition of TF and IDF, also define the function $\mathrm{TF}(w,d)$ as the term frequency of word w in the policy d and $\mathrm{IPF}(w)$ as the inverse policy frequency of the word w. Furthermore, measure the importance of word by

$$\begin{aligned} &\mathrm{TF-IDF}(w,d) = \mathrm{TF}(w,d) \times \mathrm{IDF}(w), \\ &\mathrm{IDF}(w) = \log(N / \mathrm{DF}(w)) \end{aligned} \tag{1}$$

where N is the total number of policies, and $\mathrm{DF}(w)$ is the number of policies containing the word w.

Step 2: Label Annotation

A large volume of policy documents on natural language processing (NLP) topics were studied and annotated with labels. Keywords for each policy category are shown in Table 1.

Table 1. Policy Type and Keywords

Policy type	Keywords	Label
Prevention	measure, protection, technology, infrastructure, implement, ensure	1
Prosecution	imprisonment, offence, law, illegal, punishment, prosecution	2
Partnership	share, cooperation, global, national, collaboration, prosecution	3
Emergency Response	incident, emergency, report, event, rapidly, affected	4

Step 3: PolicyClassification

After finishing the dataset construction, start to classify the policy into four types: Prevention, Prosecution, Partnership and Emergency Response. In this paper, the support vector machine (SVM) is employed to achieve the classification.

The Support Vector Machine (SVM) aims to find a hyperplane that best separates the data into different categories. The objective function for SVM is given by:

$$k(p) = w^T p + b_0 \tag{2}$$

where $\mathbf{w}$ is the weight vector, p is the input policy words vector, and b_0 is the bias term. The kernel function $k(\cdot)$ can be chosen depending on the nature of the data.

The SVM optimization problem is formulated as:

$$\begin{aligned} &\min_{w,b_0} \frac{1}{2} \| w \|^2 \\ &s.t.\ \phi_i (w^T p_i + b_0) \geq 1, \forall i \end{aligned} \tag{3}$$

where ϕ_i represents the class labels (either $+1$ or -1), and p_i is the policy word vector for the i-th policy. These constraints ensure that all policies are correctly classified, with a margin of at least 1 between the classes.

The RBF kernel function is used to classify policies into the classification set $\{P^{(1)}, ..., P^{(M)}\}$, where M denotes the total number of policy types, i.e., $M = 4$ in this model.

2.2. Modeling of Cybersecurity Policy Utility

To measure the efficiency of policy, policy utilities are required to model. In general, the difference between the expected effect before the implementation of a policy (control group) and the effect after the implementation (experimental group) is considered to be the policy's utility. In the context of cybercrime, the effect is often measured by the number of crimes. Therefore, predict the number of crimes at time t in the control group and the experimental group through the functions $f(t)$ and $g(t)$, respectively. Then the utilities $\{U^{(m)}, m=1,\ldots,M\}$ of the policy type m can be obtained:

$$U^{(m)} = \int_{t_0}^{\tau} \left(f^{(m)}(t) - g^{(m)}(t) \right) dt, m = 1,\ldots,M, \tag{4}$$

where t_0 is the policy start year.

3. National Classification Research Based on Hierarchical Clustering Analysis

Considering the existence of demographics differences (e.g., access to internet, wealth, education levels, etc.), and these indicators greatly affect the degree of national cybersecurity and the implementation effectiveness of national cybersecurity policies. Therefore, a Hierarchical Clustering Analysis (HCA) is used to classify the nations. HCA is a clustering method commonly used in the fields of data mining and machine learning. It aims to reveal the intrinsic associations between the data by building a hierarchical structure. In order to establish HCA, wealth level, education level and internet penetration level are chosen as indicators, measured by per capita GDP, higher education rate and Internet penetration rate respectively.

Let $\mathbf{X} \in ¡^{n \times p}$ represent the set of nations, where n is the total number of nations and p is the number of features. This paper defines a set of clusters $\mathbf{A} = \{A_1, A_2, \ldots, A_K\}$ partitioning the data into K groups, represented as $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}$, where each group corresponds to a cluster of nations. Each data point $x_{i,j}$ represents the j-th indicator of the i-th country.

To standardize the data, this paper apply the transformation $x_{i,j} = (x_{i,j} - \mu_j)/\sigma_j$, where μ_j and σ_j are the mean and standard deviation of the j-th feature across all nations, respectively.

In this paper, hierarchical clustering is used to group the nations based on their feature vectors. To evaluate the similarity between clusters, this paper defines two distance metrics:

1. Inter - cluster distance: This measures the similarity between different clusters. The distance between two nations $\mathbf{x}_i^{(l)}$ and $\mathbf{x}_i^{(m)}$ from clusters l and m, respectively, is given by the Euclidean distance:

$$d\left(\mathbf{x}_i^{(l)}, \mathbf{x}_i^{(m)}\right) = \sqrt{\sum_{j=1}^{p} \left(x_{i,j}^{(l)} - x_{i,j}^{(m)} \right)^2} \tag{5}$$

where $x_{i,j}^{(l)}$ and $x_{i,j}^{(m)}$ are the values of the j-th feature for the i-th country in clusters l and m.

2. Intra - cluster distance: This measures the similarity within the same cluster. The distance between two nations $\mathbf{x}_i^{(l)}$ and $\mathbf{x}_i^{(l)}$ within the same cluster l is calculated similarly:

$$d\left(\mathbf{x}_i^{(l)}, \mathbf{x}_i^{(l)}\right) = \sqrt{\sum_{j=1}^{p} \left(x_{i,j}^{(l)} - x_{i,j}^{(l)} \right)^2} \tag{6}$$

When performing the clustering, this paper aims to maximize the inter - cluster distance, ensuring that nations in different clusters are as dissimilar as possible, while minimizing the intra - cluster distance, ensuring that nations within the same cluster are as similar as possible. This approach helps in forming cohesive and well - separated clusters based on the nations' features.

The results of Hierarchical Clustering Analysis (HCA) are shown in Figure 1. This paper can see that the nations are grouped based on three axes: the X-axis (wealth), the Y-axis (internet penetration), and the Z-axis (education). This paper defines these three indicators as $e_1, e_2,$ and e_3, respectively. Based on their relative positions, this paper define a mapping H_q such that $\mathbf{x}^{(k)}$ represents the feature vector for the k-th country, with q denoting the q-th indicators in Nation, $q = 1, 2, 3$.

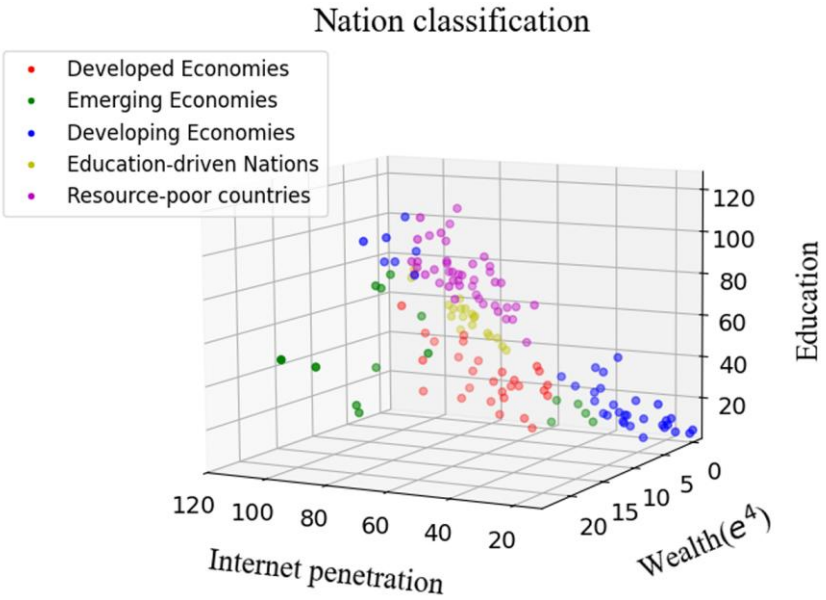


Figure 1. Country classification

Using the relative positions of the countries in this 3D space, this paper categorize the 141 nations into five subgroups, $\{A_k\}, k = 5$: Developed Economies (A_1), Emerging Economies (A_2), Developing Economies (A_3), Education - Poor (A_4) and Resource - poor Nations (A_5), which is named by the level of indictors e_q . The details can be seen in Table 2.

Table 2. Three Scheme comparing

Nation type	Wealth	Education	Internet penetration
Developed Economies (A_1)	High	High	High
Emerging Economies (A_2)	Middle	High	High
Developing Economies (A_3)	Low	Middle	Middle
Education Power (A_4)	Low	High	Middle
Resource-poor nations (A_5)	Low	Low	Low

4. Cybersecurity Policy Transfer Research Integrating Multiple Linear Regression and Network Analysis

Policy Transfer refers to the process by which a policy formulated by country x_i is transferred and implemented in another country x_m, which shares similar characteristics or conditions. In this model, policy transfer occurs between nations within the same class.

In this section, first establish the relationship between Nation and Cybercrime. Then, leverage this relationship to guide the construction of a model for the association between Nation and Policy. Specifically, introduce the policy transfer and establish the corresponding criteria, including no-transfer and shortboard-effects, to determine when multiple policies are required. Finally, apply network analysis to construct the flow of policy utilities among nations within the same class.

4.1. Modeling the Relationship Between Nation and Cybercrime Based on Multiple Linear Regression

Perform classification based on $e_q, q = 1,2,3$. Use the previously defined nation indicators to predict the crime rate indicators. To enhance the interpretability of the fit, employ a simple multiple linear regression model for prediction.

Define the crime rate indicator set as $c = \{c_1, \ldots, c_D\}$, where D is the number of indicators, and $c_1, \ldots, c_D$ represent the various crime rate indicators described in Section 4. Then construct the prediction model as follows:

$$c_s = \alpha_{s1} e_1 + \alpha_{s2} e_2 + \alpha_{s3} e_3 + b, s = 1, \ldots, D, \tag{7}$$

where $\alpha_{s1}, \alpha_{s2}, \alpha_{s3}$ are the regression coefficients for the indicators e_1, e_2, e_3, respectively, and b is the bias term.

Take prosecution rate as an example and give the corresponding results in Figure 2.

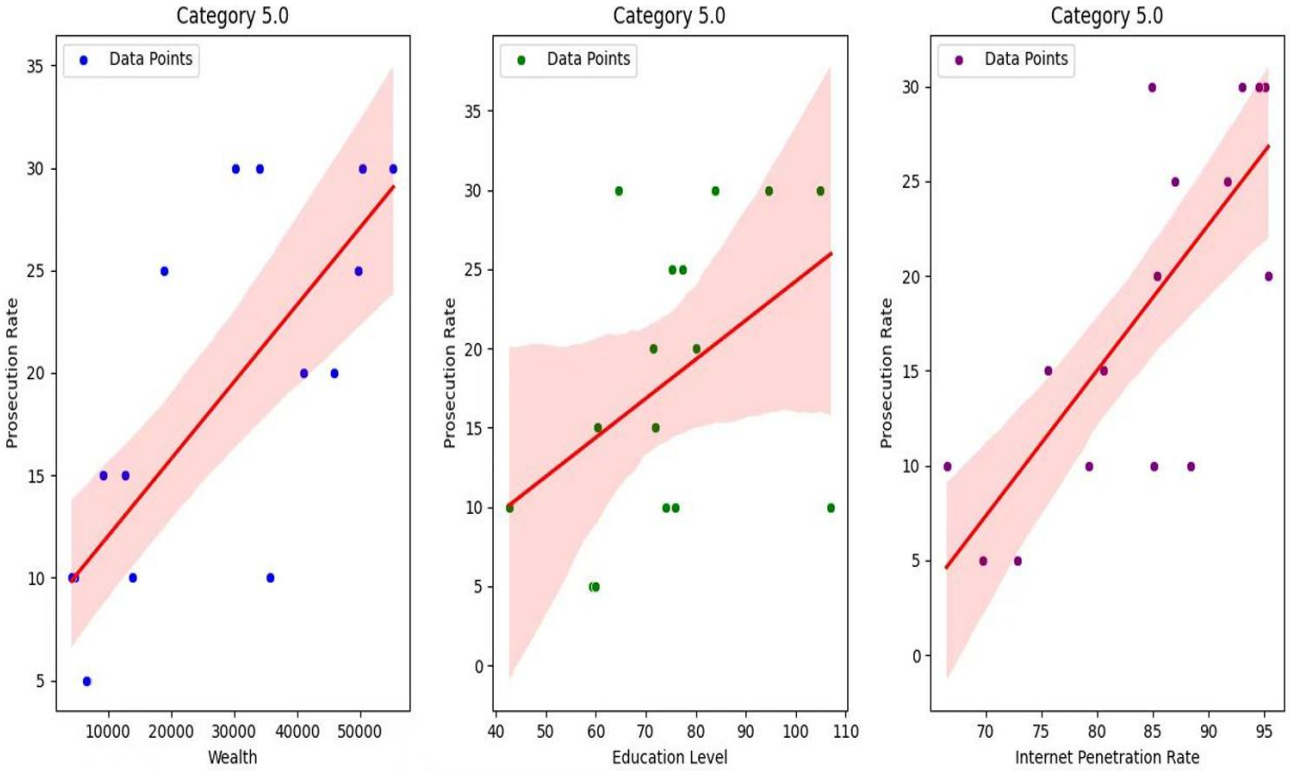


Figure 2. Using wealth, internet penetration, and education in Nation to predict the prosecution rate in Cybercrime

4.2. Construction of Guiding Coefficient for Cybersecurity Policy Transfer

This paper aims to utilize the relationship between crime rate indicators in Cybercrime and indicators in Nation to guide the construction of the policy transfer model between nations. This relationship helps establish the effectiveness or utility of policy transfers across nations.

Define the guidance coefficient δ_i as the parameter that governs the effectiveness of policy transfer from country x_i to a similar country x_m, based on the relationship between the crime rate indicator c_i and Nation. δ_i plays a key role in adjusting the policy transfer between nations.

To determine the guiding coefficients, construct the criteria for Policy Transfer.

Definition 1. Ideal no-transfer scenario: A country x_k is considered to be in a no-transfer situation if and only if the condition $s_{i,j} = 0$ holds for all pairs of indicators i and j. Specifically, $s_{i,j}$ is defined as:

$$s_{i,j} = \begin{cases} 0, & \text{if } \alpha_{i,j} > 0 \text{ and } H_k(e_j) = \text{Low}, \\ 0, & \text{if } \alpha_{i,j} < 0 \text{ and } H_k(e_j) = \text{High}, \\ 1, & \text{otherwise}, \end{cases} \tag{8}$$

where $H_k(e_j) \in \{\text{High}, \text{Middle}, \text{Low}\}$ represents the level of the indicator e_j in country k, and $\alpha_{i,j}$ is the coefficient representing the relationship between policy p_i and the crime rate indicator c_j for country k. If $s_{i,j} = 0$ for all i and j, no policy intervention is required, meaning that the country does not need policy p_i to be transferred.

If there exists i, j such that $s_{i,j} = 0$, no policy intervention is required for the corresponding policy. This paper only focusses on the cases where $s_{i,j} = 1$, and specifically on the coefficients of $e_{i,j}$ that fail to meet the conditions. When multiple indicators do not satisfy the ideal conditions, consider the most critical indicator, which is the "shortboard effect" and focus on transferring the policy to address this indicator.

Definition 2. Shortboard Effect: When $s_{i,j} = 1$, indicating that the policy transfer is needed, focus on the indicator with the largest absolute value of $\alpha_{i,j}$. That is, define the guidance coefficient δ_i as: $\delta_i = \max |\alpha_{i,j}|$

4.3. Analysis of Policy Utility Flow Based on Network Analysis and Cosine Similarity

This paper aims to analyze the flow of policy utility through network analysis. Simply put, the higher the similarity between two countries, the less the policy's utility is reduced when implemented in the second country compared to the original policy - making country. Conversely, if the countries are less similar, the policy's utility is reduced more. However, analyzing the flow directly from countries to policies might be too simplistic, so this paper modifies this based on the guidance coefficient previously defined.

First, define the similarity function using the cosine similarity function:

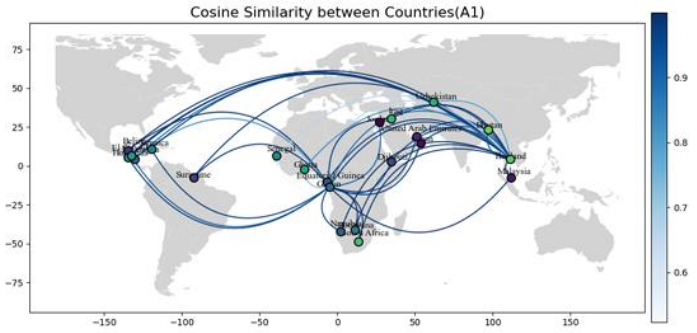
$$\sin(e_{m,q}, e_{l,q}) = \frac{\sum_{q=1}^{3} e_{m,q} \cdot e_{l,q}}{\| e_m \| \ \| e_l \|} \tag{9}$$

where $e_{m,q}$ and $e_{l,q}$ represent the q-th indicator (e.g., wealth, internet penetration, education) for the countries m and l, respectively, and q runs from 1 to 3 for the three indicators.

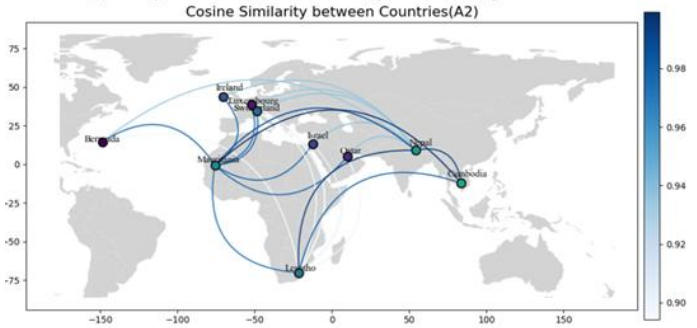
Next, define the relative effect coefficient for policy p_i in country i as:

$\text{Effect}_i(p_i) = \sin(e_{m,q}, e_{l,q}) \cdot \delta_i$ where δ_i is the guidance coefficient, and the similarity term adjusts the policy's relative effect in country i based on its similarity to the country where the policy was originally implemented.

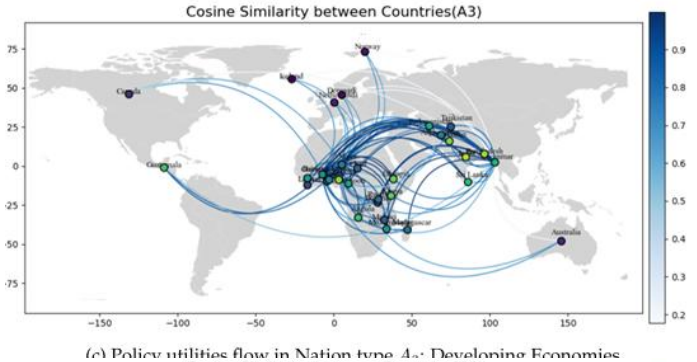
Finally, represent the intensity of the policy's utility through the thickness of the network lines. The stronger the policy's utility, the darker the color and the thicker the line, as shown in Figure 3.



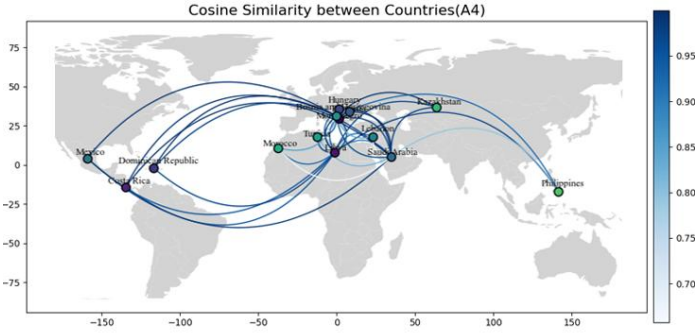
(a) Policy utilities flow in Nation type A_1: Developed Economies



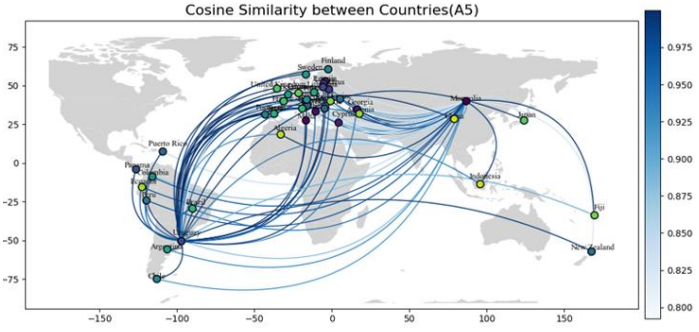
(b) Policy utilities flow in Nation type A_2: Emerging Economies



(c) Policy utilities flow in Nation type A_3: Developing Economies



(a) Policy utilities flow in Nation type A_4: Eduction Power



(b) Policy utilities flow in Nation type A_5: Resource-poor nations

Figure 3. Network analysis in policy utilities flow

5. Conclusion

This paper proposes a cybersecurity policy analysis and transnational transfer model integrating NLP, SVM, HCA, and network analysis. Its advantages lie in accurately processing policy texts, scientifically categorizing countries, and dynamically evaluating policy effectiveness. This approach addresses the issues of "ambiguous classification" and "poor country adaptation" in global policies, providing support for transnational policy coordination. First, TF-IDF is used to extract policy term vectors. Combined with annotations of four categories of policy keywords, SVM (RBF kernel) achieves classification. Policy effectiveness is then assessed objectively by predicting outcomes using a control group and experimental group prediction function, with cybercrime statistics as the metric. Second, three socioeconomic indicators including GDP per capita are selected. After standardizing data from 141 countries, Euclidean distance measures similarity, and HCA clusters nations into five categories, clarifying how country differences impact policy outcomes. A linear regression model linking countries and crime rates defines the policy transfer guidance coefficient. Combined with cosine similarity to adjust policy effect coefficients, this visualizes policy efficacy flows among countries in the same category, reducing efficacy loss. Finally, the model assists countries in selecting tailored policies and supports international organizations in implementing differentiated policies, enhancing policy implementation precision. Future research may further expand policy evaluation indicators or increase the sample size of countries.

References

[1] Zhang Chenfang, Xia Zhijie, Wang Yiming. Quantitative Analysis of China's Cybersecurity Policy Content from a Policy Instrument Perspective: Based on National Policy Texts from 2015 to 2020 [J]. Journal of Information Resources Management, 2021, 11 (03): 99-109+120. DOI:10.13365/j.jirm. 2021.03.099.

[2] Jiang Haoda, Zhao Chunlei, Chen Han, et al. A Method for Constructing Domain-Specific Sentiment Dictionaries Based on Improved TF-IDF and BERT [J]. Computer Science, 2024, 51 (S1): 162-170. DOI: CNKI:SUN:JSJA.0.2024-S1-020.

[3] Guo Li, Sun Hua. A Traffic Classification Method for Power Data Communication Networks Based on K-means and Support Vector Machines (SVM). Network Security Technology and Application, 2024, (04): 64-66. DOI: CNKI:SUN:WLAQ.0.2024-04-022.

[4] Guo Fang, Liu Xinyong, Zhang Jun, et al. Analysis of Water Quality Variation Characteristics in Typical Years of the Central Route Main Canal of the South-to-North Water Diversion Project Based on Hierarchical Clustering and Water Quality Index Method [J]. Journal of Environmental Engineering, 2024, 18 (03): 644-652.

[5] Xiong Jianfang, Feng Wen, Gao Ji, et al. Design and Implementation of a Linear Regression-Based Meteorological Data Forecasting and Visualization System [J]. Modern Information Technology, 2024, 8 (23): 133-137+144. DOI:10.19850/j.cnki.2096-4706.2024.23.026.

[6] Ji Shan, Wei Jiang, and Xin Jing. "A Power Load Forecasting Method Based on Cosine Similarity and Graph Convolutional Networks." Zhejiang Electric Power 44.01 (2025): 68-75. doi:10.19585/j.zjdl.202501007.