ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

Analysis of the Research Status of SMOTE Algorithm in the Last Three Years—Statistical Analysis of Core Literature Based on CNKI 2022-2024

Guiyu Ou[1,*]

[1]*College of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong, China*
[*]*Corresponding author: ouguiyu@suse.edu.cn*

Abstract: *To clearly understand the research and application status of the SMOTE algorithm in China, In this paper, bibliometrics is used to study the core papers of the SMOTE algorithm published in CNKI in the last three years. This method takes the literature in the CNKI database from 2022 to 2024 as the retrieval object, and statistically analyzes the number of articles published by SMOTE algorithm in China, main authors, institutions, core journals, highly cited documents, research hotspots, and so on. The research status and hot spots in the last three years have been found. The research results provide a direction for the later research and application of the SMOTE algorithm.*

Keywords: *SMOTE algorithm; Bibliometric method; Algorithm improvement; Research on algorithm application*

1. Introduction

Unbalanced data exist in medicine, economics, the internet, and other fields. Because the number of unbalanced data is very small, it is very difficult to process data. The traditional method of dealing with unbalanced data sets is not ideal. To solve this problem, In 2002, Chawla proposed the SMOTE algorithm.[1] The principle of this method is to synthesize a new minority sample by linear interpolation between a minority sample and its neighboring samples. The SMOTE algorithm can deal with general unbalanced data, However, in the face of high-dimensional unbalanced data, the SMOTE algorithm can not complete the processing task well. Therefore, some scholars later made some improvements to the SMOTE algorithm. There are many papers on improving the SMOTE algorithm.

At present, there is no paper statistical analysis on SMOTE algorithm in the last three years. Therefore, this paper uses bibliometrics to analyze the application of SMOTE algorithm. The purpose of the main analysis is as follows: Firstly, the research status of SMOTE algorithm is analyzed in China. Secondly, find out the future research trend of SMOTE algorithm.

2. Method

Bibliometrics takes the literature system and literature-related media as the research objects. A method of quantitative research on the distribution, structure, quantitative relationship, and law of literature through mathematics and statistics. Through bibliometrics, this paper mainly analyzes the number of articles related to SMOTE algorithm, research topics, periodical distribution, author statistics and so on, show the inherent law of the research object.

2.1 Data source

In this paper, the related literature of the SMOTE algorithm during 2022-2024 is statistically analyzed, the data in this paper comes from CNKI. The main reason for choosing CNKI is that it is authoritative and complete in China. Advanced search in the column. Fill in the topic "SMOTE algorithm" in the search conditions, the retrieval time range is 2022-2024. The source categories of journals are Peking University Core and Nanjing University Core Journals. Under this condition, 103 papers related to the SMOTE algorithm can be retrieved. In this paper, the collected data resources are statistically analyzed by Excel.[2] Using the bibliometrics method, this paper analyzes the number and year distribution, author distribution, organization distribution, periodical source, keyword frequency,

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

highly cited papers, and so on. We will dig out the research status and industry representatives from it to provide convenience for further research in the future.

3. Statistical analysis of data

3.1 Distribution year of the number of papers

The number of papers is an important manifestation of the research object. From the number of papers each year, we can see the changes and characteristics of the research status of the SMOTE algorithm. In the last three years, there have still many core papers about the SMOTE algorithm, with a total of 103 articles. We will make a bibliometric analysis of the changes in the number of papers in each year in the past three years.

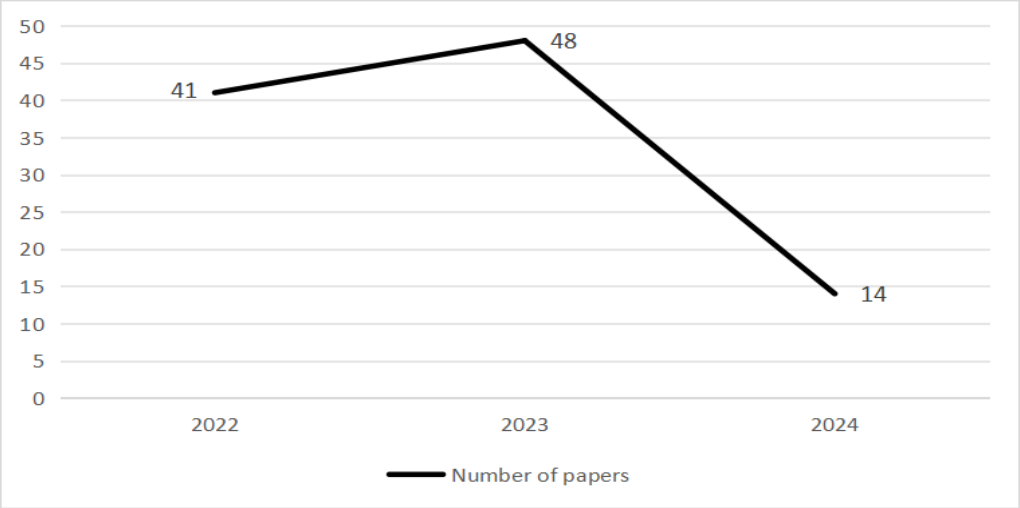

Figure 1: 2022-2024 Number of Papers Published by SMOTE Algorithm in China

In Figure 1, there are 7 more papers published in 2023 than in 2022. The year 2024 is not over yet, so the number of articles published cannot be used as the final reference. We can find that the number of papers published is still increasing. Judging from the specific number of papers, in 2022, 41 papers related to the SMOTE algorithm were published, and in 2023, 48 papers related to the SMOTE algorithm were published. In 2024, 14 papers related to SMOTE algorithm were published, But 2024 is less than half. Therefore, it is impossible to determine the number of papers related to SMOTE algorithm published in 2024.

3.2 Author's analysis

Through the statistical analysis of the author, we can find out the scholars who have studied the SMOTE algorithm deeply in the last three years. The number of papers published by the author shows that the author is authoritative in the field of the SMOTE algorithm, so the leader in this field can be found quickly. In this paper, the top 10 scholars with a large number of papers are selected for in-depth investigation and analysis.

Table 1: Rank of the number of papers published by authors related to SMOTE algorithm

author	Number of papers published	percentage
Haixiang Lin	3	2.9%
Yuming Bai	1	0.9%
Linguo Cai	1	0.9%
Hailong Chen	1	0.9%
Yang Chen	1	0.9%
Yulong Chen	1	0.9%
Xin Cui	1	0.9%
Shan Du	1	0.9%
Zaipeng Duan	1	0.9%

In Table 1, we find that there are few core papers on the SMOTE algorithm published by scholars

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

on CNKI. Except for Haixiang Lin, who published three papers in the last three years, every other scholar published only one core paper on CNKI in the last three years. The main reason for this result is that CNKI mainly includes the papers of Chinese scholars. CNKI mainly includes the core journals of Peking University, SCI journals are higher than the core journals of Peking University, so China scholars mainly publish SCI journals in foreign languages. Haixiang Lin is from Lanzhou Jiaotong University, Her three articles mainly focus on the application of the SMOTE algorithm in high-speed railways.[3][4][5] From the author's analysis and statistics, it can be seen that the focus of papers published by China scholars is on SCI journals, not CNKI's core journals of Peking University.

3.3 Analysis of Research Institutions

By analyzing the research institutions of the SMOTE algorithm, we can quickly know which institutions have conducted in-depth research on the SMOTE algorithm. The specific data are shown in Table 2.

Table 2: Institutional Rankings of Papers Related to Smote Algorithm

Fuzhou University	4
Lanzhou Jiaotong University	4
Jiangnan University	3
University of Shanghai for Science and Technology	3
Southwest petroleum university	3
Zhengzhou University	3
China metrology university	3
Hubei University Of Technology	2
North China Electric Power University	2
Liaoning Project Technology University	2
Sichuan University of Light Chemical Technology	2
Taiyuan University of Technology	2
Xidian University	2
Yunnan University	2

In Table 2, Fuzhou University and Lanzhou Jiaotong University are the hottest places to study SMOTE algorithms in China. The core papers published by these two universities in the last three years are all four on CNKI. The research directions of the four papers of Fuzhou University are scattered, such as biology, algorithm improvement, machine learning, safety in production, etc. The research on the SMOTE algorithm at Lanzhou Jiaotong University is relatively concentrated, three papers are about the SMOTE algorithm for high-speed railways, and the other paper is about improving the algorithm.

3.4 Periodical source analysis

Listing high-yield core journals can help scholars to better learn the related literature of the SMOTE algorithm. At the same time, it can also let scholars find published journals suitable for their the SMOTE algorithm papers. Table 3 is the ranking of the number of articles published by periodicals, the top 10 journals are listed in the table.

Table 3: Ranking of the number of published papers in SMOTE algorithm journals

serial number	journal	Number of published papers	Percentage
1	Computer engineering and application	5	4.9%
2	science technology and engineering	5	4.9%
3	application research of computers	4	3.9%
4	Journal of Chongqing University of Technology (Natural Science)	4	3.9%
5	Computer engineering and design	3	2.9%
6	computer application	3	2.9%
7	Chinese journal of health statistics	3	2.9%
8	journal of safety and environment	2	1.9%
9	journal of electronic measurement and instrument	2	1.9%
10	Journal of Fuzhou University (Natural Science Edition)	2	1.9%

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

In Table 3, we can find that the number of papers published by “Computer Engineering and Application” and “Science Technology and Engineering” are tied for the first place. In the last three years, these two journals have published five papers on the SMOTE algorithm. Among these 10 periodicals, 5 of them are periodicals on computer application, This shows that the main research focus of the SMOTE algorithm is in the field of computer application, the remaining five journals are in different fields. The analysis results show that the research focus of SMOTE algorithm is in the field of computer application.

3.5 Keyword analysis

Keyword analysis can reflect the research trend and related applications in a certain field. By analyzing the keywords of core papers in the CNKI database, we can find out the research focus in China. Table 4 is the ranking list of keyword frequency.

Table 4: Keyword Frequency Ranking

ranking	keywords	frequency
1	Oversampling	21
2	Random forest	20
3	imbalance	15
4	sample	15
5	Sampling algorithm	15
6	machine learning	14
7	neural network	12
8	support vector machine	11
9	model	11
10	ensemble learning	10

In Table 4, the keywords with high frequency are "oversampling", "random forest", "imbalance" and "sampling algorithm". The keywords with less frequency are "machine learning" and "neural network". Therefore, we can find that Chinese scholars pay more attention to the improvement of the SMOTE algorithm itself.

3.6 Analysis of papers with high citation rate

The citation rate of papers refers to the number of citations of papers to documents. Literature with a high citation rate allows readers to quickly find groundbreaking papers or related review papers with great influence in this field. This paper selects the top ten kinds of literatures in the CNKI database, and makes an in-depth analysis of the citation frequency, author, institution, periodical source of the paper, and year. Table 5 is as follows:

Table 5: CNKI documents with high citation rate

serial number	title	author	institution	Periodical source	frequency
1	Research on Improved Random Forest Algorithm Based on Mixed Sampling and Feature Selection	Wang Lichun; Liu shuisheng	Nanjing Institute of Technology	Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)	28
2	Intrusion detection model based on CNN and BiGRU fusion neural network	Zhang Anlin; Zhang Qikun; Huang Daoying; Liu Jianghao; Li Jianchun; Chen xiaowen	Zhengzhou University of Light Industry	Journal of Zhengzhou University (Engineering Edition)	21
3	Summary of classification methods of multi-class unbalanced data	Li Ang; Han Meng; Mu Dongliang; High wisdom; Liu Shujuan	North Minzu University	application research of computers	19
4	Transient security state perception of smart grid based on improved deep	Li Haiying; Shen Yitao; Luo yuhang	University of Shanghai for Science and	Power system protection and control	15

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

	confidence network		Technology		
5	Classification Strategy of Unbalanced Data in Manufacturing Process Based on Improved SMOTE	Li Xu; Chen jiadui; Wu Yongming; Zong Wenze	Guizhou University	Computer engineering and application	15
6	LDBSMOTE oversampling method for unbalanced data set classification	Wang Yongxin; Zhang Dabin; Che Daqing; Lv jianqiu	South China Agricultural University	statistics and decision	14
7	Research on the Prediction of Public Opinion Inversion Based on the Evolution Analysis of Public Opinion Events and the Improved KE-SMOTE Algorithm	Wang Nan; Haron Lee; Tan Shuru	Jilin university of finance and economics	Data analysis and knowledge discovery	12
8	Transformer fault diagnosis method based on improved extreme learning machine integrating minority oversampling balanced multi-classification data	Wang Yan; Li Wei; Zhao Hongshan; Shen Zongwang; Wang yinchu	North China Electric Power University	Power grid technology	11
9	Research on the Default Risk Prediction of Corporate Bonds Based on the Oversampling Technology of Generative Confrontation Network under Unbalanced Samples	Yao Xiao; Li ke; Yu lean	The Central University Of Finance and Economics	Theory and Practice of Systems Engineering Theory and Practice of Systems Engineering	10
10	Research on machine learning strategy under small sample unbalanced equipment data	Chen Yang; Liu Qinming; Liang Yaoxu	University of Shanghai for Science and Technology	Journal of University of Shanghai for Science and Technology	9

The literature with a high citation rates in Table 5 excavates the in-depth research fields of the SMOTE algorithm in China at present. The main research directions are improvement of the SMOTE algorithm, application of neural network, comprehensive research, application of smart grid, application in economic field, machine learning, and so on. The research of improved random forest algorithm has the highest citation rate.[6] This algorithm is an improved research on the SMOTE algorithm for processing unbalanced data.

4. Hot spot analysis of SMOTE algorithm in China

In this paper, the bibliometrics method is used to analyze the core documents of CNKI. According to the statistical results of keywords and literature citation rate, we dig out the research hotspots of the SMOTE algorithm in China, there are two main aspects: one is the improvement of the SMOTE algorithm; And the second is the application research of SMOTE algorithm.

4.1 Improvement of SMOTE algorithm

From the statistical papers, we found that some articles studied the improvement of the SMOTE algorithm. For example, the improvement of the oversampling algorithm by changing boundary classification, improved neural network detection model, Research on Deep Convolution Neural Networks, Improved stochastic forest integration model, Improved Algorithm of Rotating Balanced Forest, Improvement of ensemble learning algorithm, and so on. They all made some improvements to the SMOTE algorithm.

Through the collation of papers, it is found that most scholars who study algorithm improvement come from key universities in China. For example, Professor Luo Chaoyueling of Huazhong University of Science and Technology published research on the improved algorithm of Borderline-SMOTE-IHT-mixed sampling.[7] Research on resampling algorithms for unbalanced data

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

published by Professor Zhu Shen of Jiangnan University.[8] Research on neural network classification model based on chaotic Tianniu algorithm optimization published by Professor Wang Li of the Guilin University of Technology and so on.[9] They all conducted in-depth research on the improvement of the SMOTE algorithm.

4.2 Application research of SMOTE algorithm

There is much researches on the application of the SMOTE algorithm in 103 papers. The applied research mainly focuses on the following aspects: Vehicle signaling equipment, corporate debt, and stock issues, Liquor grade research, medical field, rockburst prediction research, Crop safety research, transformer fault research, groundwater research, Credit card fraud risk prediction research, railway signal equipment failure research, etc.[10][11][12] From the statistical analysis of applied research fields, it is found that the SMOTE algorithm has applications in many fields, which can be said to be very widely used. There are many applications in the medical field, with 8 papers in total. There are also many applications in the field of geology, with 10 papers in total. It can be seen that the SMOTE algorithm is widely used, and it has great research significance.

5. Conclusion

In this paper, the core papers related to the SMOTE algorithm in the CNKI database are statistically analyzed. This paper only counts the relevant core papers in the last three years, which is a major deficiency of this paper. Through this analysis and statistics, we found that the research focus of the SMOTE algorithm in the last three years is the research of algorithm improvement and algorithm application, Major research institutions, scholars, major periodical sources, highly cited documents, etc. From the results of statistical analysis, we can see that the SMOTE algorithm has penetrated various fields and played an important role.

Acknowledgments

This work was supported by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (Grant Number 2021QYY04).

References

[1] Ma He, Song Mei, Zhu Yi. borderline-SMOTE oversampling method with improved boundary classification[J]. Journal of Nanjing University (Natural Science), 2023, 59(06): 1004-1006.
[2] Li Ang, Han Meng, Mu Dongliang, Gao Zhihui, Liu Shujuan. summary of classification methods of multi-class unbalanced data[J]. application research of computers, 2022, 39(12): 3535-3536.
[3] Haixiang Lin, Zhao Zhengxiang, Lu Renjie, Lu Ran, Bai Wansheng, Hu Nana. multi-level fault diagnosis combined model of high-speed rail turnout based on word fusion[J]. journal of electronic measurement and instrument, 2022, 36(10): 217-226.
[4] Haixiang Lin, Lu Ran, Lu Renjie, Li Xinqin, Zhao Zhengxiang, Bai Wansheng. Fault diagnosis of high-speed rail vehicle-mounted equipment based on BiLSTM-CBA combined model[J]. china safety science journal, 2022, 32(06): 79-86.
[5] Haixiang Lin, Lu Renjie, Lu Ran, Xu Li. Automatic fault classification method of railway signal equipment based on text mining[J]. Journal of Yunnan University (Natural Science Edition), 2022, 44(02): 281-289.
[6] Wang Lichun, Liu Shuisheng. Research on Improved Random Forest Algorithm Based on Mixed Sampling and Feature Selection[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2022, 42(01): 81-89.
[7] Luo Chaoyueling, Zheng Yunxin, Xu Jiyu, Xie Yulong, Dai Mingcheng, Li Li. Improved GWO-SVM transformer fault diagnosis method based on Borderline-SMOTE-IHT mixed sampling[J]. Wisdom electric power, 2023, 51(07): 108-114.
[8] Zhu Shen, Xu Hua, Cheng Jinhai. Resampling Algorithm for Unbalanced Data[J]. Journal of chinese computer systems, 2024, 45(03): 542-548.
[9] Wang Li, Chen Jili, Xie Xiaolan, Xu Rongan. Neural Network Classification Model Based on Chaotic Tianniu Algorithm Optimization[J]. Science technology and engineering, 2022, 22(12): 4854-4863.

ISSN 2616-5805 Vol. 5, Issue 3: 1-7, DOI: 10.25236/AJMS.2024.050301

[10] Li Ruiping, Zhu Junjie. prediction of coronary heart disease based on improved Borderline-Smote-GBDT[J]. Chinese journal of medical physics, 2023, 40(10): 1278-1284.
[11] Sheng Jianlong, Qiao Yu, Wang Ping, Yu Donghua, Zhang Yanwen. Study on risk prediction of mine karst collapse under the influence of groundwater based on LOF-SMOTE algorithm[J]. nonferrous metal science and engineering, 2023, 14(03): 372-380.
[12] Zhou Zhihao, Chen Lei, Wu Xiang, Qiu Dongliang, Liang Guangsheng, Zeng Fanqiao. intrusion detection algorithm of vehicle-mounted CAN bus based on SMOTE-SDSAE-SVM[J]. computer science, 2022, 49(s1): 562-570.