Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Research Hotspots and Path Evolution of Machine Learning in the Context of Big Data: Visual Analysis based on CiteSpace

Yuzan Dai[1, 2], Chengke Zhu[1, 2*], Wei Li[1, 2]
[1]College of Economics and Management, Hefei University, Hefei, Anhui, China
[2]Key Laboratory of Financial Big Data, Hefei University, Hefei 230601, China
*Corresponding Author.

Abstract:

With the further development of big data, there are more and more integration of of big data and machine learning. In this paper, CiteSpace is used to analyze the research hotspots and path evolution of machine learning in the context of big data. Taking big data and machine learning as keywords, the web of science database was searched, and a total of 4628 literature were obtained. Then, the research status of machine learning is analyzed through the number of articles published, author co-occurrence analysis, country co-occurrence analysis, institution co-occurrence analysis, journal co-citation analysis, etc., and the research hot spots and future research directions in this field are studied through keyword atlas and clustering.

Keywords*: Machine learning, Big data, CiteSpace, Research hotspot.*

I. INTRODUCTION

The connotation of big data is to describe data from the perspective of data volume, type and growth rate. Machine learning is an important method to obtain experience from data and improve system performance. The meaning of "learning" is to solve the experience of the closest truth, and the main theoretical basis is statistics. The research directions of traditional machine learning mainly include decision tree, random forest, artificial neural network, Bayesian learning and so on.

With the rapid development of the internet, big data technology has attracted much attention and successfully entered into various fields, and has brought better technical support for data conversion, data processing and data storage. In the era of big data, massive data processing is more complex, diverse and high-dimensional, and data mining is also very difficult. Therefore, traditional machine learning can not deal with big data problems well. How to deeply analyze complex and diverse data

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

based on machine learning and make more efficient use of information has become the main direction of machine learning research in the current big data environment. So, machine learning and big data are gradually integrated, and it has become an important source of intelligent data analysis technology. Under this background, the research of machine learning has become a research hotspot.

Machine learning is a technology and method to realize intelligence from data. Therefore, machine learning is at the heart of data science and the essence of modern artificial intelligence [1]. In order to have a more comprehensive understanding of the research status of machine learning in the context of big data, this paper combined CiteSpace with the method of bibliometrics to analyze the research, which can not only understand the development trend of the research, but also provide different perspectives for discovering new problems.

II. DATA SOURCES AND METHODS

2.1 Data Sources

In this paper, the core database in the Web of Science database is selected, and the subject contains "Big Data" and "Machine learning" as keywords for retrieval. The literature type is limited to "Article", the language is English, and the time is from 2013 to 2021. After data screening, a total of 4628 qualified literature were obtained.

2.2 Research Methods

In order to find out the research status and path evolution in this field under the background of big data, CiteSpace is used to visually analyze the literature in the field of machine learning. We analyze the knowledge map of the number of documents, author cooperation network, national cooperation network, keyword clustering and other issues to explore the research hotspot and path evolution of machine learning.

III. LITERATURE STATISTICAL ANALYSIS

3.1 Time Distribution

The number of papers published can reflect the research heat in a certain field to some extent. The number of papers on machine learning research from 2013 to 2021 in the Web of Science database is shown in Fig 1, the number of published papers has shown an exponential growth trend since 2013.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

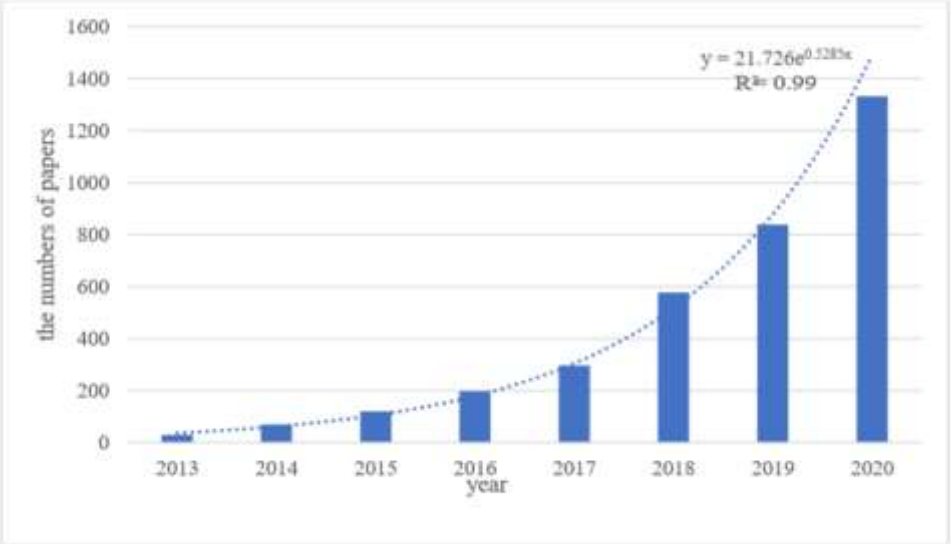

Fig 1: Time distribution

As can be seen from Fig 1, the equation of the fitting curve is:

$y = 21.726e^{0.5285x}, R^2 = 0.99$

Y represents the number of articles published each year, and x represents the year. The size of R^2 is 0.99, indicating a relatively high degree of fitting, which indicates that y and X present an exponential relationship. Especially after 2018, the number of articles on machine learning has increased rapidly, and it may grow even faster in the future, reflecting that machine learning is a hot research issue in recent years.

3.2 Author Co-authorship Analysis

Firstly, the authors' publications in the field of machine learning under the background of big data are analyzed. The authors' cooperative network is shown in Fig 2. There are 425 maps with a density of 0.0039, indicating that the authors in this field do not cooperate closely and mainly work in small teams.

In addition, according to Price's Law [2], $m = 0.749(n_{max})^{0.5}$, where m represents the total number of prolific authors and n represents the maximum amount of authors' publications. Among the obtained literature, the largest number of published papers was 32, and m was calculated to be 5. There are 25 authors with more than 5 papers. The information of productive authors is shown in Table I.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Fig 2: Author co-authorship analysis

TABLE I. Productive authors (Freq≥5)

FREQ	AUTHOR	YEAR	FREQ	AUTHOR	YEAR
32	Amir Mosavi	2019	6	Damminda Alahakoon	2018
10	Lei Wang	2019	6	Francisco Herrera	2017
9	Shahaboddin Shamshirband	2019	6	Hossein Moayedi	2021
8	Wei Wang	2018	6	Javier Del Ser	2020
7	Mazhar Javed Awwan	2021	6	Muhammad Imran	2018
7	Humbert Gonzalezdiaz	2018	5	Ivo D Dinvo	2019
7	JoelL J P C Rodrigues	2019	5	Azlan Mohd Zain	2021
7	Mehrbakhsh Nilashi	2021	5	Amirhosein Mosavi	2020
7	Haitham Nobanee	2021	5	Kyungyong Chung	2020
7	Sarminah Samad	2021	5	Qinghua Zheng	2016
7	Zhiqiang Ge	2019	5	Ala Alfuqaha	2018
7	Shahab S Band	2020	5	Khan Muhammad	2021
7	Awais Yasin	2021			

3.3 Country Co-authorship Analysis

Fig 3 reflects the co-authorship network of national cooperation. As can be seen from Fig 3, there are

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

126 nodes in total. The United States had the largest number of articles with 1378, followed by China with 1211. Machine learning has attracted wide attention in China and has made some achievements. UK and Germany are followed with more than 250.

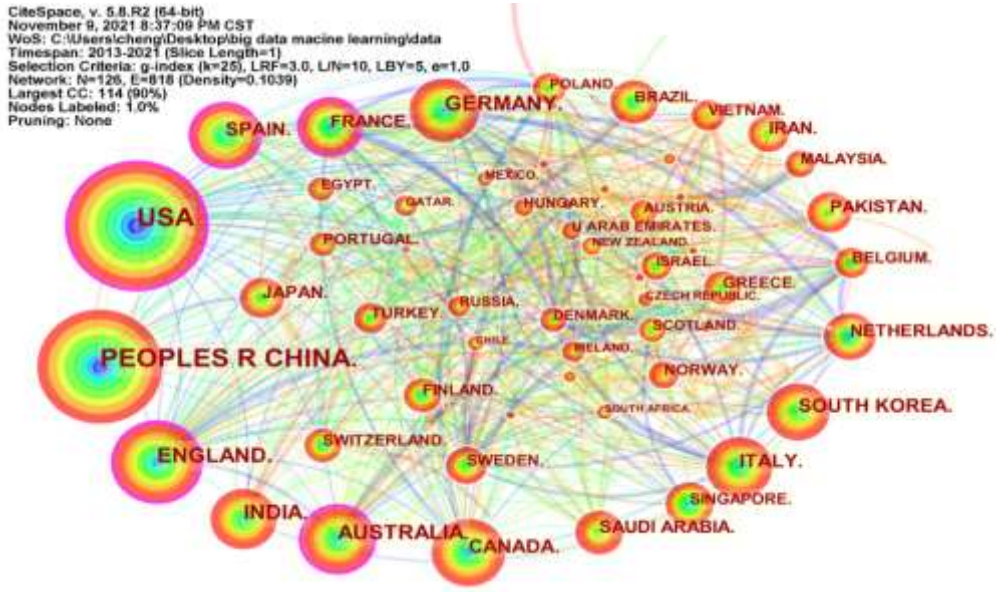

Fig 3: Country co-authorship analysis

3.4 Institution Co-authorship Analysis

In order to reflect the cooperation network of institutions, the node is set as “Institution”, and Fig 4 is obtained. In the WoS database, there are 112 organizations with more than 10 articles. Among them, the University of the Chinese Academy of Sciences has the largest number of articles, up to 100. Followed by Nanyang University of technology, Tsinghua University and King Saud University with 47, 45 and 44 times respectively. As can be seen from the density in the figure, the cooperation between universities is relatively close.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

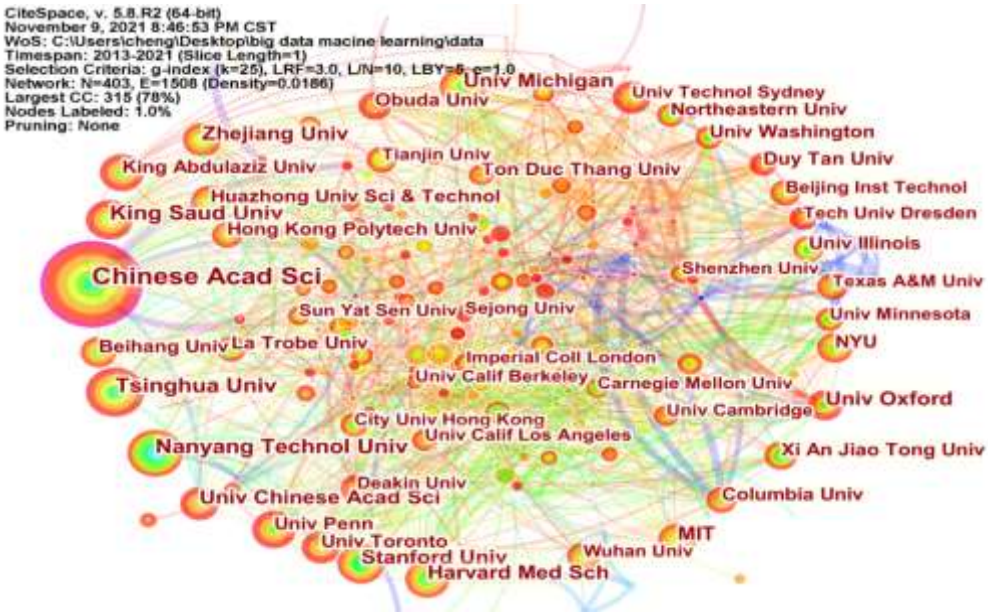

Fig 4: Institution co-authorship analysis

3.5 Co-Citation Analysis

3.5.1 Journal co-citation analysis

The journal co-citation map is shown in Fig 5. The co-cited journals of machine learning are mainly in computer science, and there are also some comprehensive journals, such as Journal of Machine Learning Research (J Mach Learn Res), Lecture Notes in Computer Science (Lect Notes Comput SC), Machine Learning (Mach Learn), Expert Systems With Applications (Expert Syst Appl) and other computer science journals, as well as Nature, Science and other international top comprehensive journals. Among them, the journal with the highest citation frequency is J Mach Learn Res, with a citation frequency of 1128 times, followed by Lect Notes Comput SC with a citation frequency of 1097 times, Nature, IEEE Access and Mach Learn are cited more than 900 times. These journals are important sources of literature in the field.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

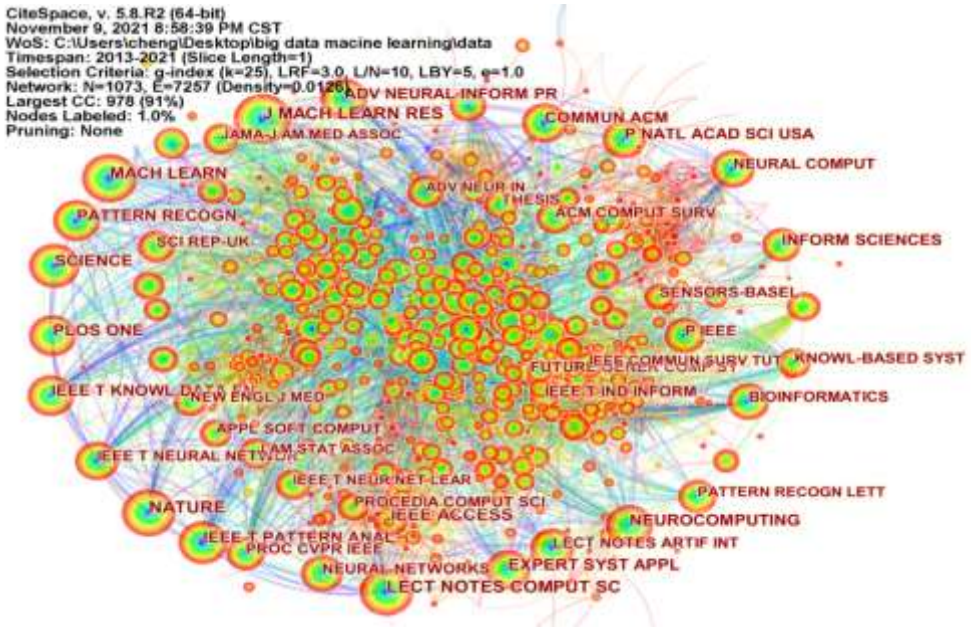

Fig 5: Journal co-citation analysis

3.5.2 Reference co-citation analysis

The co-citation analysis atlas of the references is shown in Fig 6. LeCun [3] has the highest citation frequency, up to 232 times, followed by Krizhevsky [4] and Goodfellow [5], reached 109 times and 91 times respectively.

Table II lists the top ten cited references. LeCun [3] believes that deep learning has improved the technical level of many fields, such as speech recognition, drug discovery and genomics. Krizhevsky [4] trained a large deep convolution neural network, which was significantly superior to the previous technical level. Obermeyer [6] believes that machine learning algorithms can screen a large number of variables to find combinations that can reliably predict results, which will improve prognosis, replace most of the work of radiologists and anatomical pathologists, and improve diagnostic accuracy. He [7] believes that deep-seated neural networks are more difficult to train. Silver [8] introduced a new computer Go method using deep neural network to enhance learning.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

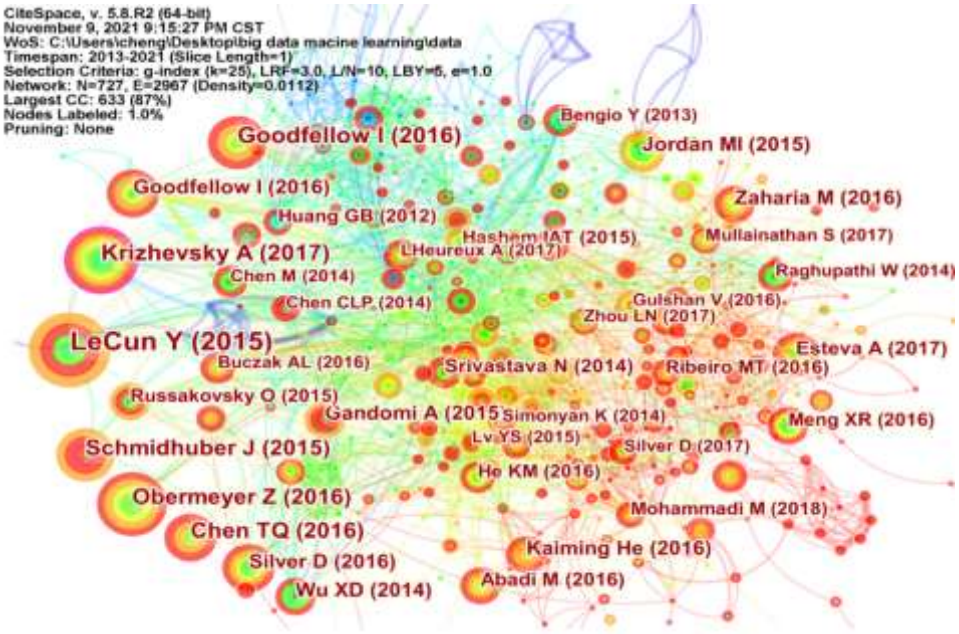

Fig 6: Reference co-citation analysis

TABLE II. The top ten cited references

RANK	FREQ	AUTHOR	YEAR	SOURCE
1	232	LeCun Y	2015	Nature
2	106	Krizhevsky A	2017	Commun ACM
3	91	Goodfellow I	2016	Adapt Comput Mach Le
4	81	Obermeyer Z	2016	New Engl J Med
5	80	Chen TQ	2016	Proceedings of The 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining
6	79	Schmidhuber J	2015	Neural Networks
7	57	Kaiming He	2016	2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
8	56	Goodfellow I	2016	Deep Learning
9	55	Silver D	2016	Nature
10	53	Wu XD	2014	Ieee T Knowl Data En

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

3.5.3 Author co-citation analysis

Firstly, we set the node to "Cited Author", so as to determine the authors with high citation frequency, and we can obtain the co-citation map of the authors, as shown in Fig 7.

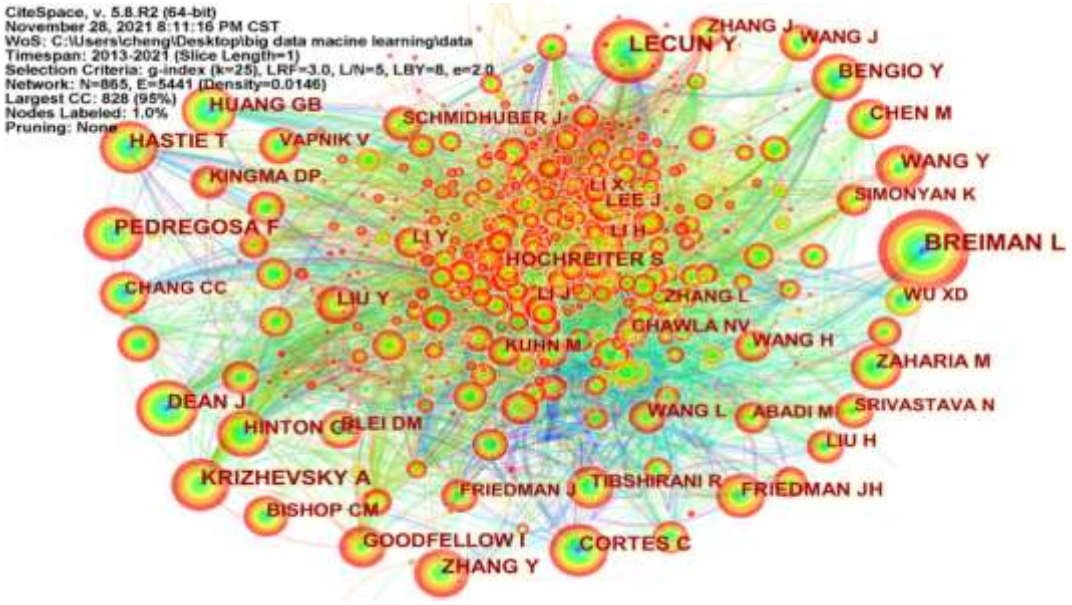

Fig 7: Author's co-cited Atlas

TABLE III. The top ten cited authers

RANK	FREQ	CENTRALITY	AUTHOR
1	568	0.01	Breiman L
2	433	0.04	LeCun Y
3	265	0.01	Pedregosa F
4	257	0.03	Krizhevsky A
5	226	0.06	Dean J
6	224	0.01	Zhang Y
7	220	0.05	Hastie T
8	215	0.04	Bengio Y
9	199	0	Wang Y
10	194	0.03	Cortes C

Table III lists the top ten cited authors. Leo Breiman is a distinguished statistician at the University

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

of California, Berkeley, and a member of the National Academy of Sciences. Breiman's research includes probability theory, mathematical statistics, and cutting-edge statistical computing. He believes that the great adventure of statistics lies in collecting and using data to solve interesting and important real-world problems. LeCun is the director of artificial intelligence research at Facebook and Professor of Computer Science, Neuroscience, and Electrical Engineering at New York University, affiliated with NYU's Center for Data Science, Courant Institute for Mathematical Sciences, and Center for Neuroscience. Krizhevsky A is best known for his work on artificial neural networks and deep learning, particularly a deep convolutional neural network called AlexNet. Krizhevsky achieved a milestone in image recognition in ImageNet Challenge 2012 by using AlexNet, which revolutionized the field of computer vision and led to the current AI boom. Zhang was selected as one of "the World's Highly Cited Scientists" in 2019. In the past five years, he has published more than 90 papers, which have been cited for nearly 2000 times in SCI database and more than 3500 times in Google Scholar, and 14 papers have been selected as ESI highly cited papers.

IV. RESEARCH HOTSPOTS AND PATH EVOLUTION

4.1 Research Hotspots

The keyword map is shown in Fig 8. As can be seen from Fig 8, keywords with high frequency include "Big Data", "Model", "Classification", "Prediction" and "Algorithm", etc. Keyword clustering atlas is shown in Fig 9. This paper mainly analyzes the main clusters, which are "Deep Learning", "Internet of Things", "Cancer", "Fault Detection" and "Sentiment Analysis". These clusters reflect research hotspots in different stages of machine learning.

"Deep Learning" is the largest cluster, containing 84 literatures, which mainly reflects the research situation of deep learning. The "Internet of Things" cluster contains 74 articles, mainly reflecting the connection between machine learning and the Internet of Things (IoT). Machine learning requires huge amounts of data, often from the myriad sensors in the IoT, which makes for better machine learning. Machine learning and the IoT stimulate each other's potential and improve machine learning. The application of the IoT will benefit from it. The "Cancer" cluster contains 61 articles reflecting the role of machine learning in cancer research. From the existing literature, machine learning theories such as artificial neural network (ANN), Bayesian network (BN), support vector machine (SVM) and decision tree (DT) have been widely used in cancer research to develop predictive models, thus deriving effective and accurate decisions. The "Fault Detection" cluster contains 49 references. The "Sentiment Analysis" cluster contains 46 articles.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

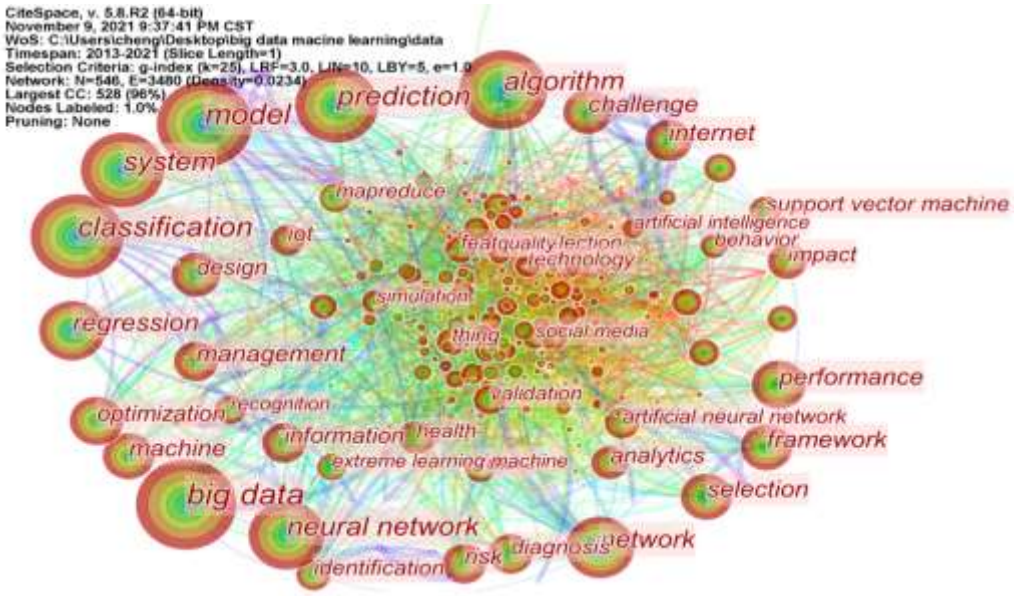

Fig 8: Keyword co-occurrence analysis

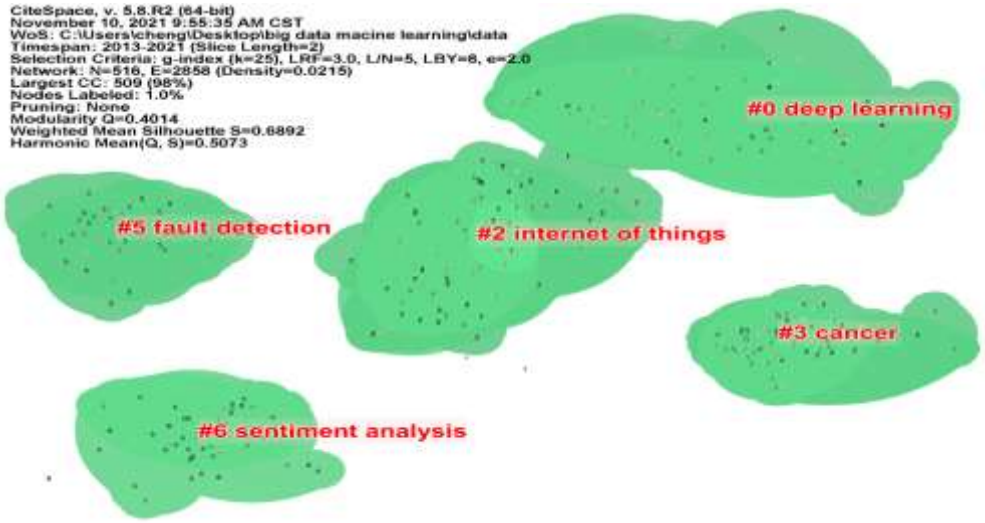

Fig 9: Keyword clusters

4.2 Path Evolution

In order to study the research path evolution of machine learning, timezone view is used for analysis,

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

as shown in Fig 10. As can be seen from Fig 10, the first stage is mainly about the basic concepts and theories of machine learning, including neural network, support vector machine, etc. Neural network is a method to achieve machine learning tasks. In the field of machine learning, neural network is generally referred to as "neural network learning". With the increasing amount of data and the optimization of algorithms, the layers of neural network become more and more, and the learning effect becomes better and better. This is deep learning, which is essentially a deep neural network. Support vector machine (SVM) is a supervised learning model, which is mainly used to solve the problem of data classification and has many applications in text classification, image classification, biological sequence analysis and biological data mining, handwritten character recognition and other fields [9].

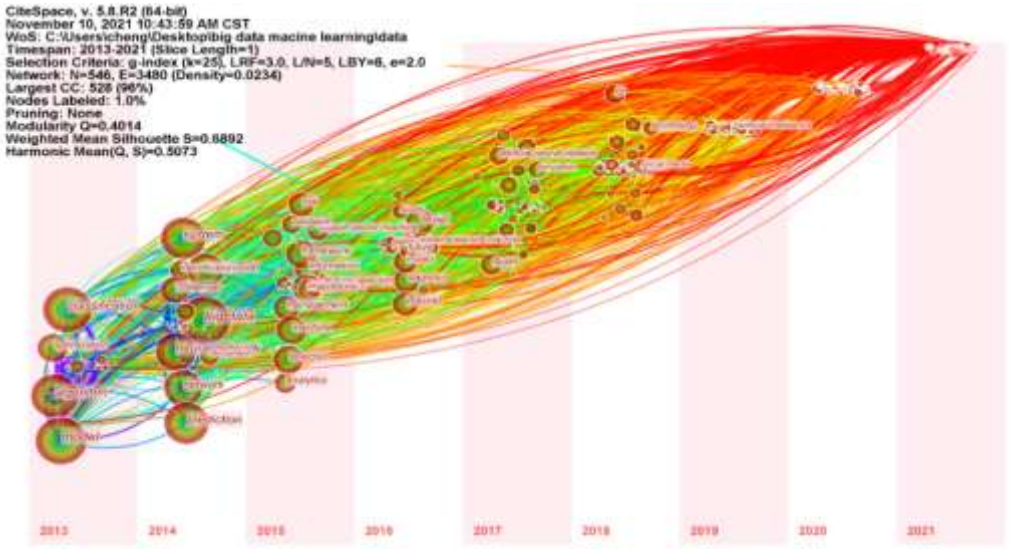

Fig 10: Timezone map

The second stage mainly studies the application fields of machine learning. Machine learning has been widely used in various fields, including prediction, model diagnosis, risk management, etc. Predicting by analyzing historical data is one of the core functions of big data analysis. Machine learning may seem like a constant process of repeatedly collecting, storing and analyzing data, but with improved algorithms and computing power compared to traditional production methods, it can filter out unnecessary data, identify inconsistent data, and find new data to support it. The prediction method based on machine learning is easy to use, simple to operate, and has high prediction accuracy. It has been widely used in the industry [10], such as disease prediction, price prediction, etc.

The third phase is mainly about the possible future application areas of machine learning, such as the combination of technology and artificial intelligence. With the deep integration of machine learning and technology, many revolutionary technologies have been produced. Take natural language generation for

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

example, which is a popular technique for transforming structured data into a local language. Programming using machine learning algorithms transforms the data into the format which the user wants. In this case, manual intervention will be greatly reduced because the data will be converted to the desired format. Data can be visualized in charts, graphs, and so on. From the relationship between machine learning and artificial intelligence, machine learning is one of the most important ways to realize artificial intelligence [11]. Machine learning is the use of algorithms that allow computers to "learn" automatically, analyzing patterns in data and then using those patterns to make predictions about new samples.

V. CONCLUSIONS

The development of big data requires scientific and reasonable machine learning algorithms to meet social needs and improve data processing efficiency. In order to make a breakthrough in the field of big data, the traditional machine algorithms should also be optimized and upgraded to comprehensively improve the data processing capacity. Therefore, machine learning in the context of big data has become a hot research issue. CiteSpace makes a literature visual analysis on the field of machine learning under the background of big data, which provides a new research direction for the research of machine learning. This paper selects 4628 documents from 2013 to 2021 as samples, and uses CiteSpace to analyze the research status of machine learning through the number of documents, author co-authorship analysis, national co-authorship analysis, institutional co-authorship analysis and journal co-citation analysis. Keyword atlas and cluster analysis are used to study the research hotspots and future research directions in this field.

From the above research, we can get the following conclusions: First, the number of articles published in this field shows exponential growth. In particular, the growth rate has increased significantly since 2018. We can predict that this field will remain a research hotspot in the future. In the existing research, the cooperation among authors in the field is not close, mainly small team's cooperation; American academia pays the highest attention to this field, and Chinese scholars also pay extensive attention to machine learning; In addition, the research cooperation between universities is relatively close. Second, according to the results of literature clustering, the research in this field is mainly in deep learning, IoT, cancer research and so on, among which deep learning is the largest cluster. Thirdly, it can be seen from the time zone diagram that machine learning research can be roughly divided into three stages, including basic theory, basic application fields and possible application fields in the future.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 452-465
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

ACKNOWLEDGMENTS

This research is supported by Talent Project of Hefei University (Grant No. 20RC63).

REFERENCES

[1] Pillow J, Sahani M (2019) Editorial overview: Machine learning, big data, and neuroscience. Current Opinion in Neurobiology (55): iii-iv
[2] Solla D, Price DJ (1963) Little Science Big Science. Columbia: Columbia University Press
[3] Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553): 436
[4] Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Communications of the ACM 60(6): 84-90
[5] Goodfellow I, Bengio Y, Courville (2016) A deep learning. MIT Press
[6] Obermeyer Z, Emanuel EJ (2016) Predicting the Future-Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 375(13): 1216-1219
[7] He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: 770-778
[8] Silver D, Huang A, Maddison CJ, et al. (2016) Mastering the game of go with deep neural networks and tree search. Nature (529): 484-489
[9] Saigal P, Chandra S, Rastogi R (2019) Multi-category ternion support vector machine. Engineering Applications of Artificial Intelligence 85(10): 229-242
[10] Garzón MB, Blazek R, Neteler M, et al. (2006) Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula. Ecological Modelling 197(3-4): 383-393
[11] Jones LD, Golan D, Hanna SA, et al. (2018) Artificial intelligence, machine learning and the evolution of healthcare. Bone & Joint Research 7(3): 223-225