ATLANTIS PRESS
Advances in Computer Science Research, volume 89
2019 International Conference on Wireless Communication, Network and Multimedia Engineering (WCNME 2019)

Research Hotspots and Frontier Evolution in the Field of Machine Learning

Fujun Zhang, Wenbin Zhao, Quanhui Ye, Xue Gao and Hao Wan
Computer Science and Engineering College, Shandong University of Science and Technology, Qingdao, Shandong Province 266590, P. R. China

***Abstract*—Due to the emergence of Internet big data and hardware GPU, machine learning is out of the bottleneck period. Machine learning began to explode and began to become an independent popular subject and was applied to various fields. Various machine learning algorithms are emerging, and deep learning using deep neural networks has been further developed. At the same time, the flourishing of machine learning has also promoted the emergence of other branches, such as pattern recognition, data mining, bioinformatics and autopilot [1]. In this paper, papers and literatures in the field of machine learning in the core collection of Web of Science from 2008 to 2018 were selected as data sources, and CiteSpace bibliometric method was used for analysis. The results were as follows: The hotspots in the map co-occurrence and the information in the table for nearly ten years are analyzed according to the Algorithm and framework. The hotspots algorithms include classification, support vector machine, regression, neural network, random forest and so on. The popular frames are sorFlow, Caffe, PaddlePaddle and so on. Database, sequence, framework, deep learning, segmentation, mirroring, genetic algorithm, pattern recognition, timing arrangement, learning effect, decision tree, these mutant words together constitute the research frontier and emerging fields of machine learning in the past decade.**

Keywords—informatics; mapping knowledge Domain; machine learning; evolution of frontier

I. INTRODUCTION

With the continuous development of science and technology, the amount of data to be processed is also increasing geometrically. In the context of the complexity and rapid changes of massive data, people are searching for valuable information and bringing many new problems. The visual research method came into being [2]. It is to directly extract or statistically classify the related information involved in the complicated information, and display it with the visual image of the image, so that people can intuitively understand the hot spot of the target research field, the development status and the overall context. Currently, software for knowledge visualization in the world is being widely used [3],For example, Pajek, developed by Thomson Reuters developer HistCite, InCitesTM, Vladimir Batagelj [4], CiteSpace[5] developed by Professor Chen Chaomei and so on. In space, the atlas analysis software can analyze the structural relationship of the knowledge of the region, organization, author, etc. of the literature through co-occurrence and social network analysis; in time, through the dynamic time dimension, draw the "knowledge development process spectrum", and intuitively show the evolution of knowledge [6].With the increasing application of artificial intelligence and the continuous improvement of machine learning technology, only by constantly understanding the development of research in the discipline, and actively researching the subject hotspots can more accurately position itself and the direction of research. In addition, this paper combines the two disciplines of computer science and information science, and uses the bibliometrics method of information science to study hotspots and research frontiers in the field of excavator learning; In the hotspot frontier mining, this paper conducts co-occurrence analysis and mutation analysis on the keywords of machine learning literature to mine the research hotspot of machine learning; coupling analysis of machine learning literature and mining the frontier of the field. Therefore, it is very necessary to provide reference for the future research of researchers in related fields.

II. RESEARCH DESIGN

A. Data Source

The raw data studied in this paper is derived from the core collection database on the Web of Science platform (including: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, ESCI, CCR-EXPANDED, and IC).In order to ensure that the research data has a high degree of objectivity and comprehensiveness, choose machine learning as the search subject. The search strategy is as follows: First, the database selects the core collection of web of science, enters the subject word TS=“machine learning”, and selects the search time span from 2008 to 2018.Secondly, the literature type was refined into "ARTICLE", and 34,835 search results were obtained. Select the export data format to download in full-record format (each download item <=500), then CiteSpaces specifies the format download*.txt, save it in the same file data, as the basic data for drawing the map, and in the same file. Folder to build the project.

B. Research Methods and Tools

Bibliometric analysis mainly uses literature as the data source, including the three basic processes of collecting, sorting and analyzing. In this study, the 2008-2018 machine learning papers in the web of science database are used as data sources, and they are cleaned and cleaned with different perspectives and dimensions. Knowledge map analysis is a classification of visual analysis. The analysis results are presented in the form of nodes and connections. The nodes represent the objects being studied. The connection indicates the strength of the relationship between the two research objects, making the analysis results more beautiful and direct. The knowledge

mapping tool used in this paper is CiteSpace developed by Dr. Chen Chaomei from Drexel University in the United States based on the Java platform. The software can identify and visualize new trends and new trends in scientific literature, and is widely used in the field of scientific literature measurement. The combination of science and information science combines the telemetry of information science to the spatial and temporal distribution of the excavator learning field, the author's cooperation network, research hotspots, and research frontiers.

III. Machine Learning Research Hotspots and Frontier Trend Analysis

A. Research Hotspot Analysis

Only the subject areas familiar with machine learning can help to master the research of computer top technology, and the key words are the summary of the author's intention and subject matter, which is the core and essence of the literature. Research hotspots in a field often refer to the high-frequency, high-centrality, and high-intensity keywords that appear in the literature at a certain stage. The co-word analysis method is used to map the keywords of the machine learning field in the past ten years, and to collect the high-frequency keywords, to clarify the research hotspots in the field of machine learning in the past ten years, and then to analyze its evolution and development process. After importing the data into CiteSpace and adjusting the parameters, the result is shown in Figure 1.

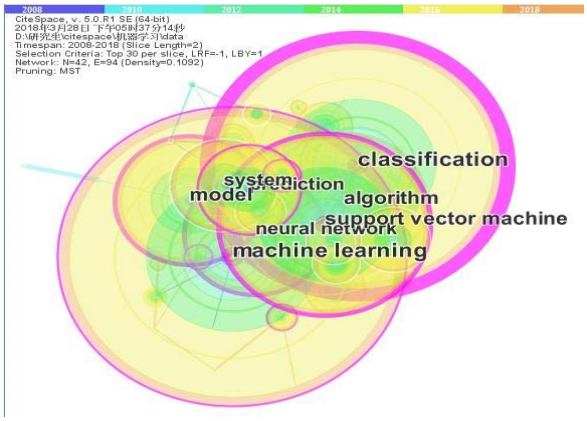

FIGURE I. KEY WORDS CO-OCCURRENCE KNOWLEDGE MAP IN THE FIELD OF MACHINE LEARNING FROM 2008 TO 2018

As shown in Figure 1, there are a total of N (42) nodes in the co-occurrence knowledge map of the machine learning domain, with E (94) and center (Density = 0.1092).First, the size of the circle represents the frequency of occurrence of the keyword in the 10 years from 2008 to 2018. The order from the largest to the smallest is: machine learning, classification, support vector machine, Algorithm, model, neural network, system, etc. Because the paper is mainly about machine learning, the first keyword has no analytical significance, that is to say, classification, support vector machine, algorithm, model, neural network, system, etc. constitute a research hotspot of machine learning for nearly ten years; Secondly, the heavier the purple at the outermost periphery of each circle proves that the centrality of the keyword is higher, which means that the keyword plays an indispensable role in the research of the past ten years. For example, the purples in the picture are classification, which means that the word classification is enough to be called a hot topic in the field, which has important contributions and represents a research hotspot. Then there is the connection between the point and the point. The color of the connection indicates the year in which the two words first co-occurred. The thick rules of the connection prove the closeness of the connection between the two keywords. The closer the connection is, the thicker the connection. For example, it can be seen from the figure that the connection between the two keywords categoryification and support vector machine is dark blue, which means that the first two keywords are co-occurring in 2008. Then, after clustering in Figure 1, Figure 2 will appear.

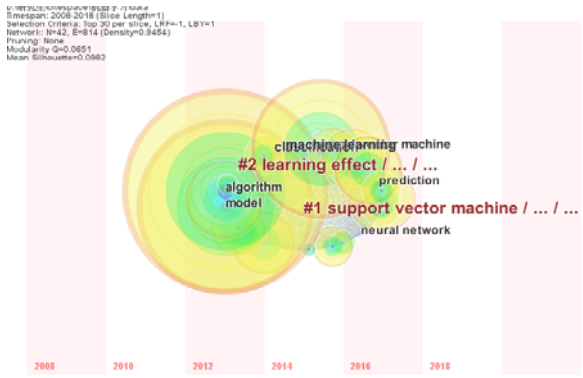

FIGURE II. KEYWORDS CO-OCCURRENCE KNOWLEDGE MAP IN THE FIELD OF MACHINE LEARNING FROM 2008 TO 2018

After clustering the keywords, there are six categories, the largest of which (#0) has 9 members with a contour value of 0.488. It is marked as user skills level by LLR, classifying by the TFIDF algorithm, and the most active cluster reference is paper 0.44Wang, JB (2010) "Scheduling jobs with an exponential sum-of-actual-processing-time- Based learning effect [12].

The second largest cluster (#1) has 9 members with a contour value of 0.671. It is called classifying human physical activity by the LLID algorithm and is marked as machine by the TFIDF algorithm. One of the most active articles is 0.33 Mannini, A (2010) "Machine learning methods for classifying human physical activity from on-body accelerometers" [13].

TABLE I. HIGH-FREQUENCY KEYWORD STATISTICS IN THE FIELD OF MACHINE LEARNING, 2008-2018

ranking	frequency	Key word	Earliest citation time	ranking	frequency	Key word	Earliest citation time
1	7022	machine learning	2008	13	1192	identification	2008
2	6049	classification	2008	14	1146	optimization	2008
3	4368	support vector machine	2008	15	1009	classifier	2008
4	3584	algorithm	2008	16	894	random forest	2012
5	3078	Model	2008	17	746	data mining	2008
6	2962	neural network	2008	18	496	genetic algorithm	2008
7	2245	prediction	2008	19	414	pattern recognition	2008
8	1673	regression	2008	20	393	deep learning	2016
9	1554	recognition	2008	21	291	image	2016
10	1423	feature selection	2008	22	270	sequence	2008
11	1290	extreme learning machine	2012	23	258	segmentation	2016
12	1210	artificial neural network	2008	24	186	framework	2014

The data provided in the above CiteSpaces is organized into Table 1. Since the subject of this research is the machine learning field, the machine learning ranked first can be ignored. Through the above chart, the hot words can be divided into two categories: ①Algorithm②framework. These hot words are only the most basic vocabulary, and will be analyzed one by one in order to find research hotspots in the field of machine learning in the past decade.

The following is a popular machine learning algorithm in the past ten years. The first is the second hot topic classification (classification algorithm) in the table. There is a very simple and currently popular algorithm in the classification algorithm for Naive Bayes classification. The basic basis of Naïve Bayes is this: for the given item to be classified, the probability of occurrence of each category under the condition of this occurrence, and which is the largest, is considered to belong to which category. Its practical use examples are: classification processing of papers, public opinion analysis, and so on.

Followed by the third-ranked hot topic support vector machine in the table, which is a binary classification algorithm. Given a set of two types of N-dimensional local points, the SVM produces an (N-1) dimensional hyper plane to divide these points into two groups. Suppose you have two types of points, and they are linearly separable. The SVM will find a line that divides the points into 2 types, and this line will be as far away as possible from all points. The main problems currently dealt with using the support vector machine are display of commercial advertisements, processing of face recognition splice sites, image processing with large data difference, and the like.

Next is the eighth hotspot regression (regression algorithm) in the table. Logistic regression in the regression algorithm is a powerful statistical method. By estimating the probability of using a logical operation, measuring the relationship between a categorical dependent variable and one (or more) independent variables is a cumulative logical distribution. At present, logistic regression is mainly used for traffic analysis, usage scoring, measuring the success rate of marketing activities, and so on.

Followed by the sixth hot topic in the table, neural network, there are many algorithms belonging to the neural network. In the past ten years, the recursive neural network is more prominent. In fact, the neural network is the general term of two artificial neural networks. Recurrent Neural Network, the other is the Recursive Neural Network...And with the continuous improvement of computer hardware, the number of neural network layers that can be processed is also deepening, which also lays a foundation for the subsequent study of deep learning.

The last one is the 16th hot topic random forest in the table. The random forest algorithm combines multiple trees and uses a randomly selected subset of data to improve the analysis accuracy of the decision tree. The advantage of the random deep forest algorithm is its ability to handle large-scale data sets, as well as a large number of seemingly unrelated data that can be used for risk assessment and customer information analysis.

After analyzing the hotspot algorithm for nearly ten years, the next analysis is the deep learning of the current machine learning field. Next, we will introduce the four major frameworks commonly used in the field of deep learning:① TensorFlow, originally developed by researchers and engineers at Google Brain Team in Google's Machine Intelligence research organization. The main role of the framework is: first, to help researchers in the machine learning field of algorithm research faster and easier; the second simplifies the transformation process from research model to actual production.②Neon, Neon is a Python-based deep learning library developed by Nervana. It's easy to use and its performance is at its highest level.③Caffe, Caffe is a deep learning framework focused on expressiveness, speed and modularity, developed by Berkeley Vision and Learning Center and community contributors. ④ DeepLearning4J, DeepLearning4J, like ND4J, DataVec, Arbiter and RL4J, is part of the Skymind Intelligence Layer.

B. Analysis of the Evolution of Research Frontiers

Using the Burst Detection function in CiteSpace, the terminology of the mutated words detected in the keywords of the literature in the past ten years, using the time distribution of the word frequency, the trend of change, and the combination of word frequency, finds out the evolution of research frontiers in this field [14].

TABLE II. LIST OF MUTATION WORDS IN THE FIELD OF MACHINE LEARNING, 2008-2018

Key word	Burst Detection	Beginning year	End year	2008 - 2018
database	104.9884	2008	2013	
deep learning	99.4643	2016	2018	
sequence	98.3826	2008	2013	
segmentation	77.7649	2016	2018	
image	73.469	2016	2018	
genetic algorithm	67.9653	2008	2015	
framework	63.5458	2014	2015	
pattern recognition	56.6286	2008	2015	
scheduling	44.1588	2010	2011	
learning effect	38.8508	2010	2011	
pattern	28.9851	2008	2009	
decision tree	27.1329	2008	2009	

From 2008 to 2018, there were 12 mutation words: database, sequence, framework, deep learning, segmentation, image, genetic algorithm, pattern recognition, scheduling, learning effect, pattern, decision tree, These mutations together constitute the research frontier and research emerging field in the field of machine learning for nearly a decade.

Dividing these 12 keywords into two time periods to show the evolution of machine learning in the past decade. First of all, from 2008 to 2009, the study of machine learning was only in the study of decision tree and pattern. The main reason was that the hardware equipment at that time did not reach a particularly high data operation. However, with the continuous updating of research technology and the continuous development of computer hardware equipment, the research focus of machine learning has changed between 2010 and 2011. The terms scheduling and learning effect occupy the frontier of research at that time. Machine learning has ushered in a whole new field of research. Over time, the framework from 2014 to 2015 has become a research hotspot at that time. The main research frameworks at that time were: 1TensorFlow, 2Keras, 3 Caffe, etc. These frameworks are mainly for researchers to develop better. The neural network lays the foundation for future deep learning.

The next step is from 2016 to 2018. At this time, researchers are more inclined to further explore machine learning, which is, to continue to explore the multi-layer neural network, and the machine learning field has come to the era of deep learning. Many scholars have proposed new algorithm models, such as convolutional neural networks, deep neural networks, deep belief networks, etc. At the same time, the majority of scholars continue to apply deep learning to different fields such as image object classification, image, segmentation, and pattern recognition. It can be seen that with the development of the field of machine learning, the current research focuses on the field of deep learning, but deep learning is still evolving and gradually entering the application stage. It can be foreseen that the field of deep learning will be faster and more convenient. The right algorithm, and will continue to be used in different fields to benefit the majority of researchers.

IV. CONCLUSIONS AND PROSPECTS

This paper builds the machine learning knowledge map by downloading the literature data from 2008 to 2018 in the core collection of Web of Science, combining machine learning, knowledge map, and research related theories and techniques. The main conclusions are as follows: 1 According to the knowledge map co-occurrence and the information in the table, the hotspots of nearly ten years is analyzed according to two categories: Algorithm and framework. The hotspot algorithms are: classification, support vector machine, regression, neural network, random forest, etc.; Hotspot frameworks are: sorFlow, Caffe, PaddlePaddle, and more. Databases, sequences, frameworks, deep learning, segmentation, image, genetic algorithms, pattern recognition, scheduling, learning effects, and decision trees all form the forefront of research and emerging fields in the field of machine learning for nearly a decade. In the early preparation work, since the authors of the papers provided by the database are all shorthand, this will lead to cumbersome and error-prone data checking. I hope that in the Web of Science database, the full name of the author can be used when the article is included. The accuracy of the query.

REFERENCES

[1] Research on the status quo and application of machine learning development Reference URL: http://www.fx361.com/page/2018/0601/3608221.shtml

[2] Shi Jiyuan. Knowledge map analysis based on CiteSpaceIII transfusion medicine research field[D] Fourth Military Medical University 2015.

[3] Du Wenlong. Comparative Analysis of the Application of Citation Analysis Software[D]. Northwest University.2013

[4] Li Jie Chen Chaomei. CiteSpace technology text mining and visualization [M]. Capital University of Economics and Business Press.2016:115.

[5] Liu Beiyuan, Chen Chaomei, Hou Haiyan, Wang Xianwen, Towards the era of scientific change.[J]. Science and science and technology management. 2009, 30(7):5–12.

[6] Zhao Yupeng. Frontier Analysis of Machine Learning Research Based on Knowledge Mapping[J]. Intelligence magazine. 2012, 31(4):28-31.

[7] Jiao Licheng, Yang Shuyuan, Liu Fang, Wang Shigang,. Neural Network Seventy Years: Retrospect and Prospect[J]. Journal of Computer, 2016,39(08):1697-1716.

[8] Zhang Fujun, Ye Quanhui, Yu Luyun. Analysis of technical opportunities in the field of marine science based on knowledge maps [J]. Science and technology management research. 2017, 37(24):165-170.

[9] Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8[J]. Proteins Structure Function & Bioinformatics, 2009, 77(Supplement S9):100–113.

[10] Liu Q, Wang J. A one-layer recurrent neural network with a discontinuous hard-limiting activation function for quadratic programming.[J]. IEEE Transactions on Neural Networks, 2008, 19(4):558-70.

[11] Mallapragada P K, Jin R, Jain A K, et al. Semi Boost: boosting for semi-supervised learning.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(11):2000-2014.

[12] [12]Wang J B, Sun L H, Sun L Y. Scheduling jobs with an exponential sum-of-actual-processing-time-based learning effect[J]. Computers & Mathematics with Applications, 2010, 60(9):2673-2678.

[13] nnini A, Sabatini A M. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers[J]. Sensors, 2010, 10(2):1154-1175.

[14] Chang C C, Lin C J. LIBSVM: A library for support vector machines[M]. ACM, 2011.

[15] Depristo M A, Banks E, Poplin R E, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data[J]. Nature Genetics, 2011, 43(5):491-8.

[16] Huang G B, Zhou H, Ding X, et al. Extreme Learning Machine for Regression and Multiclass Classification[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2012, 42(2):513-529.

[17] Zhang Fujun, Liu Guiren, Liu Qian, Li Ruiqing. Shandong Province domestic patent bibliometric analysis [J]. Science and technology management research.2013, 33(01):60-63.

[18] Zhang Rui, Wang Yongbin. Machine learning and its algorithm and development research [J]. Journal of Communication University of China: Natural Science Edition, 2016, 23(2):10-18.