Analysis of application of machine learning based on bibliometrics in pattern recognition

Jin Huang¹, Wenyan Tu¹ and Qing Yuan¹
¹Huazhong Univ Sci & Technol, Lib, Wuhan, Hubei, Peoples R China

Corresponding author and e-mail: Qing Yuan, yuanq@hust.edu.cn

Abstract. This paper analyzes the application of machine learning in pattern recognition by using bibliometrics methods, keyword cluster analysis, and burst terms detection. According to analysis, the number of published studies on the application of machine learning in pattern recognition has shown a rapid growth trend in recent years. The top five countries with published volumes are the United States, China, the United Kingdom, Germany, and India. According to the results of keyword co-occurrence cluster analysis, it is found that the hot research topics in this field mainly include feature extraction, support vector machines, and neural networks. Research topics include functional magnetic resonance imaging, deep learning and image processing, bioinformatics and fuzzy logic systems, semi-supervised learning and face recognition. According to the analysis of burst terms detection results, it is found that future research frontiers include big data, sensors, classifier integration, entropy, pattern classification, attribute reduction, image analysis, principal component analysis, fuzzy logic, Alzheimer’s disease, magnetic resonance imaging, etc.

1. Introduction

Pattern recognition refers to the process of processing and analyzing various information of characterizing things or phenomena to describe, identify, classify, and explain things or phenomena [1]. Machine learning is a science of artificial intelligence, which aims to improve the performance of specific algorithms in empirical learning, so as to extract the laws used to judge some unknown things [2].

Pattern recognition has emerged since the 1950s and is an important component of information science and artificial intelligence, which is mainly used in image analysis and processing [3-4], speech recognition [5], voice classification [6], communication [7], computer-aided diagnosis [8], data mining [9] and other aspects. The 21st century that is intelligent, informatization-based, and cyberization-based. With the rapid progress of large-scale parallel computing, big data, and machine learning algorithms, the bottleneck in the development of pattern recognition technology has gradually been broken through, that has set off a new wave of upsurge in the scientific research and industry of pattern recognition modernization and has profound influence on all aspects of social and economic development. Therefore, it is of great guiding significance to understand the latest research trends in the field, the core research countries and institutions, and the cooperation between them, to identify and analyze the hot research topics and frontiers of machine learning application in pattern recognition.
Keyword cluster analysis and burst terms detection are common methods used in bibliometrics to analyze the content of documents. So far, they have been well applied in the fields of scientometrics, information science, and information retrieval. It also produced a lot of results. Jiang Ying used the cluster analysis method to analyze the theme content of the papers in the field of global bibliometrics from 1995 to 2004, and found that the internal institutions in the field of bibliometrics are changing, the scope of research is further expanded, and the research objects are diversified. [10]Luo Mingying analyzed the front evolution of global walnut research using the burst terms detection method, and found that the international research front on walnuts presents a multi-level, multi-disciplinary, comprehensive development trend, and the research topics are becoming more and more abundant. [11]

This article combines Chen Chaomei’s definition of research frontiers, based on the Web of Science database, using bibliometrics methods, social network analysis methods, keyword clustering analysis methods, burst terms detection and other methods to carry out visual analysis on the application of machine learning in pattern recognition.

2. Data sources and methods

2.1. Data sources
Based on the Web of Science core collection data, this paper uses the topic retrieval method to analyzes the relevant literature of machine learning in pattern recognition. The search strategy is TS=“machine learning” and TS=“pattern recognition”, the publication time of the document is up to 2019, no language limitation, the article type is limited to article, review, and procedures paper, and 3537 related articles are finally retrieved.

2.2. Analysis methods and tools
This article uses bibliometrics methods to analyze and detect the 3537 retrieved papers from various angles such as the main body of the paper, the influence of the paper, and the research content of the paper. Using CiteSpace visualization software to visually analyze institutions and regional cooperation networks, keyword co-occurrence networks, etc., to reveal the important nodes and closeness of the cooperative relationship between the main bodies of paper publications, as well as research hotspots and research frontiers in this field.

3. Empirical analysis

3.1. Posting trends
3537 papers related to the application of machine learning in pattern recognition were retrieved in Web of Science. The first paper was published in 1997. From 2003 to 2008, the number of papers grew slowly. After 2008, it entered a stage of rapid growth. In 2018, it showed explosive growth. Generally speaking, the number of papers published in this field and the frequency of citations have shown an exponential growth trend, which is still in a rising stage, indicating that its research heat is continuing to heat up.
3.2. Locations analysis

From 1997 to 2019, 71 countries or regions in the world published papers on the application of machine learning in pattern recognition. The top five countries in which articles are published are the United States, China, the United Kingdom, Germany, and India. Betweenness centrality is an indicator to measure the importance of nodes in the network. According to the global cooperation network shown in Figure 2, the centrality of the United States, the United Kingdom, France, Canada, and Poland is relatively high, indicating that these countries have an important connection and bridge role in the global cooperation network relationship network, which plays an important role in academic exchanges. Among them, the United States has cooperative relations with 31 countries, and the betweenness centrality is also the highest, indicating that the United States has the most active international cooperative network in this field.

| Country       | Number of papers | Betweenness centrality | cooperating countries Number |
|---------------|------------------|------------------------|-----------------------------|
| The United States | 769              | 0.2                    | 38                          |
| China         | 759              | 0.04                   | 25                          |
| The British   | 255              | 0.14                   | 31                          |
| Germany       | 189              | 0.07                   | 27                          |
| India         | 187              | 0.01                   | 13                          |

Further analysis of the cooperation networks of the top five countries in terms of the number of articles published shows that there are relatively few cooperation countries between China and India. In the rich network of cooperative relations in the United States, its network connection with Iran is thicker than other countries, indicating that they have closer cooperation in this field. Compared with the United States, China has relatively less cooperation with other countries, and the countries with closer cooperation are mainly the United Kingdom. In addition to maintaining a close cooperative relationship with China, the UK has also maintained good and close exchanges with Brazil, Germany, Spain and others. Germany's main cooperating countries in this field are the United Kingdom and Ireland. India's international cooperative relations network is relatively simple, and the differences in cooperation intensity with other countries are not obvious.
Further analysis of the cooperation networks of the top five countries in terms of the number of articles published shows that there are relatively few cooperation countries between China and India. In the rich network of cooperative relations in the United States, its network connection with Iran is thicker than other countries, indicating that they have closer cooperation in this field. Compared with the United States, China has relatively less cooperation with other countries, and the countries with closer cooperation are mainly the United Kingdom. In addition to maintaining a close cooperative relationship with China, the UK has also maintained good and close exchanges with Brazil, Germany, Spain and others. Germany’s main cooperating countries in this field are the United Kingdom and Ireland. India’s international cooperative relations network is relatively simple, and the differences in cooperation intensity with other countries are not obvious.

### 3.3. Organizations analysis

There are 453 institutions participating in the publication in this field, and the top five institutions in the volume of publications are: Chinese Academy of Sciences, University of Sao Paulo, Harbin Institute of Technology, Tsinghua University, Wroclaw Technical University.

Through a visual analysis of the cooperation relationship between the issuing organizations, it can be seen that there are 592 cooperation links between various agencies. The cooperation between agencies has the obvious characteristics of more intra-region cooperation and less cooperation with outside regions.

**Table 2.** The top five institutions in terms of the number of published papers.

| Organization name                  | Number of published papers | Betweenness centrality | Organization name |
|------------------------------------|---------------------------|------------------------|------------------|
| Chinese Academy of Sciences        | 50                        | 0.23                   | 42               |
| University of Sao Paulo            | 34                        | 0.1                    | 27               |
| Harbin Institute of Technology      | 23                        | 0.03                   | 9                |
| Tsinghua University                | 22                        | 0.01                   | 9                |
| Wroclaw Technical University       | 21                        | 0                      | 0                |
Further analysis of the cooperative network relationships among the three top five institutions in terms of the number of published papers shows that the Chinese Academy of Sciences has the highest intermediary centrality of 0.23. At the same time, the number of institutions it cooperates with is up to 42, including 24 domestic cooperative institutions and 18 foreign institutions. From this, it can be seen that in this research field, the academic research activities of the Chinese Academy of Sciences are relatively active and play an important role in cooperation and exchanges at home and abroad. Compared with the Chinese Academy of Sciences, Harbin Institute of Technology and Tsinghua University have fewer cooperative institutions, and most of them are domestic cooperation.

![Figure 3. Global institutional cooperation network.](image)

### 3.4. Keyword clustering analysis

Keywords are the core summary of a thesis. Analyzing the keywords of the thesis can help you understand the relationship between research topics, and the generation and changes of research topics.

Cluster analysis refers to the analysis process of grouping a collection of physical or abstract objects into multiple classes composed of similar objects, based on the similarity of the analyzed objects. There are three clustering algorithms provided by CiteSpace, namely: LSI shallow semantic index, LLR log maximum likelihood ratio, and MI mutual information.

In this paper, the LLR algorithm is used to perform cluster analysis on keywords, and 9 clusters are obtained as shown in Table 3. The cluster ID order is from 0 to 8. The smaller the number, the more keywords are included in the cluster. Each cluster is composed of multiple closely related words. The larger the cluster size, the smaller the number. Size represents the number of members included in the cluster. Silhouette is an index to measure the homogeneity of the entire cluster members. The larger the value, the higher the similarity of the cluster members. Mean Year represents the average year in which keywords appear.

Cluster 0: functional magnetic resonance imaging. Since the end of 1990, researchers have begun to apply pattern recognition technology to brain science and cognitive science. In particular, the use of pattern recognition methods for medical image analysis has aroused the interest of researchers. With the deepening of research, the pattern recognition analysis method can be used to analyze the multiple types of discrimination problems presented in the magnetic resonance imaging data, thereby effectively assisting the research of human brain mechanism and clinical diagnosis of mental diseases [12].

Cluster 1: Deep learning and image processing. In 2006, the deep learning method was proposed, and it is widely used in image recognition systems. It belongs to a structure containing multiple layers of perceptrons, which can separate complex factors in an image through multiple layers of nonlinear mapping successfully. It has strong advantages in extracting global features and context information of the image [13].

Cluster 2: Bioinformatics and fuzzy logic system. Bioinformatics is an emerging discipline formed by the intersection of life science and computational science.
Cluster 3: Semi-supervised learning and face recognition. Semi-supervised learning (Semi-supervised Learning) is a new type of machine learning method developed in the past more than ten years. The idea is to avoid the problem of degradation of performance (or model) when traditional supervised learning is insufficient (inadequate learning) by introducing unlabeled samples in model training when the number of labeled samples is small [14]. With the continuous improvement and optimization of semi-supervised learning algorithms and success in many applications, it has also achieved good practical results in the application of face recognition.

Cluster 4: Feature extraction. When matching recognition or classifier classification recognition is performed in pattern recognition, the basis of judgment is the feature. The extracted features can be used to characterize the entire image content, and image targets can be classified according to feature matching. Common feature extraction algorithms are mainly divided into three categories: color, texture, and shape features. Commonly used feature extraction algorithms are LBP feature extraction algorithm, HOG feature extraction algorithm, Haar feature extraction algorithm and so on.

Cluster 5: Distributed sensors. Distributed sensor is a multi-sensor data processing system established by distributed computer as a reference.

Cluster 6: Machine translation and semantic analysis. Machine translation is an important branch of natural language processing. In recent years, with the increasing variety of data resources such as various vocabularies, semantic grammar dictionaries and corpora, word segmentation, part-of-speech tagging, syntactic analysis and other technologies have advanced rapidly. The continuous emergence of new theories, new methods, and new models has promoted natural In-depth study of language processing. At the same time, traditional machine learning has also begun to turn to deep learning. Deep learning algorithm models such as recurrent neural networks, recurrent neural networks, and convolutional neural networks have been used in natural language processing and machine translation [15].

Cluster 7: Multidimensional signal processing. With the development of modern computing technology, digital signal processing is a new research field. Multidimensional signal processing has only been developed in recent years due to the large amount of digital storage and computation required [16].

Cluster 8: Alzheimer's disease and resting state functional MAGNETIC resonance imaging. In recent years, the use of functional magnetic resonance imaging (fmri) combined with complex network theory based on graph theory found that alzheimer's patients brain function network abnormal changes of local and global topological properties, it is not only to understand the pathophysiology mechanism provides a new Angle of view, may also be new imaging marker for early diagnosis of [17].

### Table 3. Keyword clustering.

| Cluster ID | Size | Silhouette | Mean Year | Labels (LLR) |
|------------|------|------------|-----------|--------------|
| #0         | 67   | 0.677      | 2009      | fmri (9.94, 0.005); prediction (9.94, 0.005); magnetic resonance imaging (mri) (6.62, 0.05); chatter (6.62, 0.05); roc curves (6.62, 0.05); image processing (11.75, 0.001); deep learning (11.52, 0.001); biometrics (8.63, 0.005); signal processing (8.63, 0.005); rehabilitation robotics (8.63, 0.005); bioinformatics (11.97, 0.001); metabolic profiling (7.97, 0.005); fuzzy logic (7.97, 0.005); microarray (7.97, 0.005); remote sensing (7.77, 0.01); semi-supervised learning (17.21, 1.0E-4); face recognition (16.38, 1.0E-4); object recognition (12.88, 0.001); pca (8.63, 0.005); manifold learning (8.58, 0.005); |
| #1         | 59   | 0.558      | 2010      |              |
| #2         | 56   | 0.674      | 2006      |              |
| #3         | 53   | 0.683      | 2007      |              |
3.5. Research fronts

CiteSpace uses the Burst Detection algorithm to detect burst-terms with high frequency change rates from a large number of keywords. This algorithm is not only to analyze the level of word frequency, but to determine the Research fronts by the frequency and trend of word frequency. The basic principle is to count the frequency of words in the title and abstract of literature, and to determine the hot words in the research front according to the growth rate of these words.

In this paper, by using the function of CiteSpace burst terms detection, a visual analysis was conducted on 3,537 papers, and it was concluded that the year of mutation in the field of application research of machine learning in pattern recognition started in the last decade. After manual screening and judgment, a total of 19 research frontier phrases were obtained.

Table 4. Research frontiers of machine learning application in pattern recognition.

| Burst term           | Strength | Begin  | End    | 1997 - 2019                               |
|----------------------|----------|--------|--------|------------------------------------------|
| big data             | 7.2029   | 2014   | 2019   |                                          |
| sensor               | 6.9919   | 2017   | 2019   |                                          |
| classifier ensemble  | 4.5533   | 2010   | 2013   |                                          |
| entropy              | 4.5533   | 2010   | 2013   |                                          |
| pattern classification | 4.5449 | 2014   | 2015   |                                          |
| attribute reduction  | 4.2905   | 2012   | 2016   |                                          |
| image analysis       | 4.1723   | 2017   | 2019   |                                          |
| PCA                  | 4.048    | 2011   | 2012   |                                          |
| fuzzy logic          | 4.0467   | 2010   | 2013   |                                          |
| alzheimers disease   | 3.9194   | 2014   | 2019   |                                          |
| MRI                  | 3.8425   | 2015   | 2019   |                                          |
| signal processing    | 3.6813   | 2015   | 2017   |                                          |
| random forest        | 3.6454   | 2016   | 2017   |                                          |
| FMRI                 | 3.3965   | 2011   | 2014   |                                          |
supervised learning 3.3619 2015 2017
semi-supervised learning 3.2374 2009 2015
unsupervised learning 3.2166 2014 2016
breast cancer 3.2079 2011 2013
mutual information 3.1904 2012 2013

The Table 4 shows that machine learning applications in pattern recognition field mutation year for nearly a decade of research frontiers include: big data integration, sensor, classifier, entropy, pattern classification and attribute reduction, image analysis, principal component analysis, fuzzy logic, Alzheimer's disease, magnetic resonance imaging, signal processing, random forest, functional magnetic resonance imaging (fmri), supervision, a semi-supervised learning and unsupervised learning, breast cancer, and mutual information.

4. Conclusions
In this paper, bibliometric analysis, social network analysis, co-word analysis, burst terms detection and other methods are used to visually analyze the application of machine learning in pattern recognition in Web of Science. The analysis draws the following conclusions: research on the application of machine learning in pattern recognition has achieved fruitful results. In recent years, the number of published papers has shown a trend of rapid growth, and the research interest is getting higher and higher. The United States, China, The United Kingdom, Germany and India were the top five countries in terms of publication volume. From the perspective of publication volume, the United States is at the top level in pattern recognition research, followed by China with a small gap. From the perspective of keyword co-occurrence and cluster analysis, the research hotspots in this field mainly include feature extraction, support vector machine, neural network, etc. Research topics include functional magnetic resonance imaging, deep learning and image processing, bioinformatics and fuzzy logic systems, semi-supervised learning and face recognition. Future research frontiers include big data, sensors, classifier integration, entropy, pattern classification, attribute reduction, image analysis, principal component analysis, fuzzy logic, Alzheimer's disease, and magnetic resonance imaging. With the development of technology, pattern recognition research will involve more fields.

References
[1] Zhang Xuegong. Pattern Recognition (the third edition). Beijing: Tsinghua University Press, 2010.
[2] Chen Haihong, HUANG Biao, Liu Feng, et al. Principle and application of machine learning. Chengdu: University of Electronic Science and Technology Press, 2017.
[3] Jiaping Li, Wendi Wang, Na Deng, et al. A novel digital method for weave pattern recognition based on photometric differential analysis [J]. Measurement, 2020, 152.
[4] Sun Wei, Wang Xiaowei, You Shijun, et al. Application of pattern recognition in feature extraction of medical ultrasonic digital image [J]. Chinese Medical Equipment, 2020, 17(02):1-5.
[5] Turker Tuncer, Sengul Dogan. Novel dynamic center based binary and ternary pattern network using M4 pooling for real world voice recognition [J]. Applied Acoustics,2019,156.
[6] Maria Barbarosou, Ioannis Paraskevas, Amr Ahmed. Military aircrafts' classification based on their sound signature [J]. Aircraft Engineering and Aerospace Technology, 2016, 88(1).
[7] He Jiai, Zhang Wengi. Modulation Identification technology and its Development of Communication Signals [J]. High Technology Communications, 2016, 26(02): 157-165.
[8] Zhang Long, WU Rongzheng, LEI Bing, et al. Extension Intelligent Identification of Rolling Bearing Fault based on multi-scale entropy [J]. Noise and Vibration Control, 2019,
9

39(06):200-205.

[9] Li Weifeng, Zhou Wenjin. Research on Vibration Data Mining and Fault Location of Ship Mechanical Equipment based on spectral Characteristics [J]. Automation and Instrumentation, 2019(07):10-13.

[10] Jiang Ying. Co-word analysis of bibliometrics research from 1995 to 2004 [J]. Journal of Information, 2006, 20(4):504-512.

[11] Luo Mingying. Hot spot analysis of world walnut research front based on burst terms detection [J] 2016, 28(2):60-63.

[12] Yu Yang. Multi-class model analysis of FMRI data [D]. National University of Defense Technology, 2012.

[13] Xia Changlin. Application of deep learning in image recognition [J]. Computer knowledge and technology, 2019, 15(33):185-186.

[14] Tu Enmei, Yang Jie. Overview of Semi-supervised Learning theory and its Research Progress [J]. Journal of Shanghai Jiao Tong University, 2016, 52(10):1280-1291.

[15] Wang Yu. Natural Language Processing and its Application in Machine Translation [J]. Modern Chinese, 2019(05):187-191.

[16] Han Yanbin. Application research of pattern recognition Theory in Multi-dimensional Signal Feature Extraction [D]. Shandong: Jinan University, 2005.

[17] Liu Xue-na, Zhang Xinqing. Brain network graph analysis based on functional magnetic resonance imaging and its application in alzheimer's disease [J]. Chinese journal of modern neurological diseases, 2012, 12(04):483-487.