Journal of Artificial Intelligence Practice (2020) Vol. 3: 31-37
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2020.030106
ISSN 2371-8412

A Co-word Analysis of the Applications of Machine Learning in China

Chen Fei

Business School, Shandong University of Technology, Zibo, Shandong, 255000

Keywords: Co-word analysis, Cluster analysis, Machine learning

Abstract: With the rapid development of technology, the influence of artificial intelligence on various industries is increasing. As one of the realization methods of artificial intelligence, machine learning has also received more and more attention and has been applied in many fields. This article was screened on CNKI, and obtained 4057 documents on machine learning applications from January 1, 2015 to December 31, 2019. Through co-word analysis of the keywords of the documents and further Based on statistical analysis, we have obtained 7 themes that are mainly used in machine learning at the moment, and explained these 7 themes.

1. Introduction

Machine learning is a branch and implementation of artificial intelligence. It allows computers to automatically "learn" algorithms to find laws in massive data and make certain predictions about the future based on the laws. Nowadays, machine learning is applied to more and more fields and attracts more and more scholars' attention. Therefore, it is necessary to dig out the hot spots of machine learning technology application, which can provide certain reference for later researchers. Pooja R. Makawana analyzed the literature on the application of machine learning in network security from 2015.01 to 2016.12 based on the bibliometric method, and showed the hot topics in the field during this period (Makawana & Jhaveri, 2018). EI-Sayed M. EI-Alfy used citation data and visualization technology to review the relevant literature on the application of machine learning in big data, and analyzed the main publishers and research topics (El-Alfy & Mohammed).

Although the above documents have quantitatively analyzed the frontiers of machine learning technology applications, they are mainly based on citation data. There is a certain time lag in citation analysis, and the citation data has not penetrated into the main content of the literature. Co-word analysis is a tool for statistical analysis of word pairs in the same document to study the characteristics of documents in related fields. It analyzes topic words based on the main content of the document, paying more attention to the current and more in-depth. The hypothesis was first put forward by Whittaker, mainly: the author takes the terms they use very seriously; when different terms appear in the same article, the author assumes that there is a certain relationship between the two; if there are enough Authors' recognition of this relationship means that this relationship has a certain meaning in the field (Whittaker, Courtial, & Law, 1989). Similarly, we generally believe that the more times a word pair appears in the same article, the closer the relationship between the two themes, and the fewer times it appears, the more distant the relationship. Under these premises,

co-word analysis has certain practical significance for observing the development trend of the research field and digging the frontier hotspots in the field. MJ Cobo showed the 3 stages of the evolution of the concept of intelligent transportation based on co-word analysis, and discovered 6 main research topics in this field, which provided a certain reference for subsequent research (Cobo, Chiclana, Collop, de Ona, & Herrera- Viedma, 2014). Marcus Matthias Keupp used co-word analysis, cluster analysis, and frequency analysis to quantitatively analyze the literature on strategic management from 1992 to 2010, providing a basis for future theoretical development in this field (Keupp, Palmie, & Gassmann, 2012).

This article is mainly divided into four parts. The first part gives a general introduction to the article; the second part carries on the keyword word frequency statistics and the construction of the high-frequency word co-word matrix; the third part carries out further statistics on the data Analysis; The fourth part draws corresponding conclusions based on the analysis results.

2. Analysis process

2.1 Data pre-processing

This article uses CNKI as the data source and uses "machine learning" as the subject term to search all the documents from January 1, 2015 to December 31, 2019, totaling 6000 articles. A total of 4057 articles were obtained after manual removal of no abstract, no keywords, and irrelevant documents.

2.2 Extract high-frequency keywords and count word frequencies

Imported keywords of 4057 articles into Excel, merged synonyms such as "random forest" and "random forest algorithm", "SVM" and "support vector machine", and deleted non-meaningful keywords such as "research" and "prospect", and then use related functions to count the frequency of keywords. Since this article is the focus of research in the field of machine learning, the "machine learning" keywords that completely overlap with the topic are eliminated, and then keywords with a word frequency of not less than 15 are selected, and a total of 61 high-frequency keywords and their word frequencies are obtained.

2.3 Create the high-frequency keyword co-occurrence matrix

Table 1 Part of the word co-occurrence matrix

	DL	AI	SVM	CNN	Big Data	Neural Network	Random Forest	Data mining
DL	444	65	9	87	17	42	4	4
AI	65	338	2	12	59	29	1	7
SVM	9	2	254	8	5	17	26	7
CNN	87	12	8	234	3	5	0	2
Big Data	17	59	5	3	218	2	7	28
Neural Network	42	29	17	5	2	167	3	0
Random Forest	4	1	26	0	7	3	139	7
Data mining	4	7	7	2	28	0	7	131

The co-word matrix is a matrix that reflects the number of common occurrences of keywords. And then uses the pivot table to obtain the co-word matrix of high-frequency keywords, as shown in Table 1. Show. The rows and columns in the table are the same keywords, the diagonal line is the frequency of the high-frequency word in the common word matrix, and the remaining cells indicate the number of times the row and column two keywords appear together. The more keywords

co-occur, the closer the relationship between keywords is. The construction of the co-word matrix provides a basis for further analysis.

2.4 Create the similar matrix

Since the co-occurrence frequency difference in the co-word matrix is very different, it is not conducive to multivariate statistical analysis. Therefore, this paper uses Ochiia coefficient to convert the co-word matrix into a similar matrix. The calculation formula is:

$$\frac{Num(k_1, k_2)}{\sqrt{N(k_1)} * \sqrt{N(k_2)}}$$

Among them, $Num(k_1, k_2)$represents the number of times that keyword k_1 and keyword k_2 appear together, $N(k_1)$represents the frequency of keyword k_1 in the co-word matrix, and $N(k_2)$ represents the frequency of keyword k_2 in the co-word matrix The frequency. The value in the similarity matrix indicates the degree of similarity between the two keywords. The closer to 1, the higher the degree of similarity between the keywords and the closer the distance; otherwise, the longer the distance between the keywords.

2.5 Create the dissimilarity matrix

The larger the value in the similarity matrix, the closer the distance between the two keywords. In order to facilitate understanding and reduce errors, all the values in the similarity matrix are subtracted from 1 to obtain the dissimilarity matrix. In the dissimilarity matrix, the larger the value and the closer to 1, the farther the two keywords are. The smaller the value, the closer the distance between the two keywords and the more similar they are. Some dissimilarity matrices are shown in Table 2.

Table 2 Part of the dissimilarity matrix

	DL	AI	SVM	CNN	Big Data	Neural Network	Random Forest	Data mining
DL	0.0000	0.8322	0.9732	0.7301	0.9454	0.8458	0.9839	0.9834
AI	0.8322	0.0000	0.9932	0.9573	0.7826	0.8779	0.9954	0.9667
SVM	0.9732	0.9932	0.0000	0.9672	0.9788	0.9175	0.8616	0.9616
CNN	0.7301	0.9573	0.9672	0.0000	0.9867	0.9747	1.0000	0.9886
Big Data	0.9454	0.7826	0.9788	0.9867	0.0000	0.9895	0.9598	0.8343
Neural Network	0.8458	0.8779	0.9175	0.9747	0.9895	0.0000	0.9803	1.0000
Random Forest	0.9839	0.9954	0.8616	1.0000	0.9598	0.9803	0.0000	0.9481
Data mining	0.9834	0.9667	0.9616	0.9886	0.8343	1.0000	0.9481	0.0000

3. Statistical data analysis

3.1 Factor analysis

Factor analysis is a multivariate statistical method that reduces the dimensionality of multiple variables with overlapping information, and uses a few unrelated common factors to represent most of the information. Generally speaking, it is to group multiple variables. The correlation between variables in the same group is strong, and the correlation between different groups is weak. Each group represents a common factor. Since there are many high-frequency keywords selected in this article, the method of factor analysis is used for dimensionality reduction, and common factors are

used to represent the research samples for further analysis.

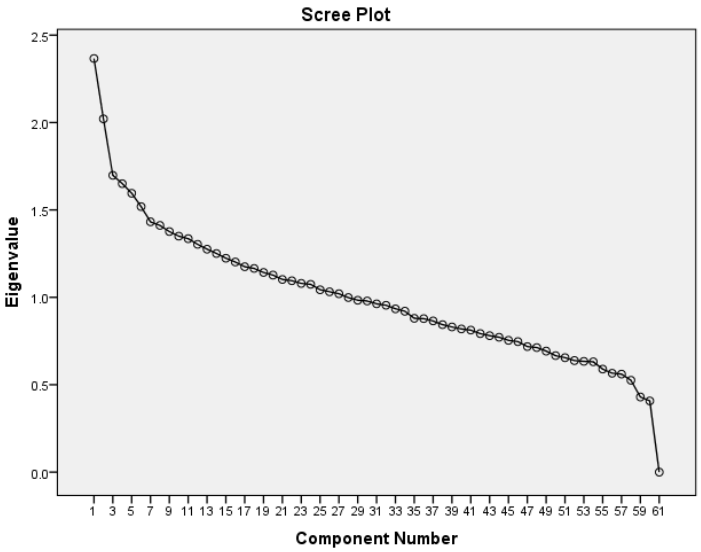

Figure 1. Scree Plot

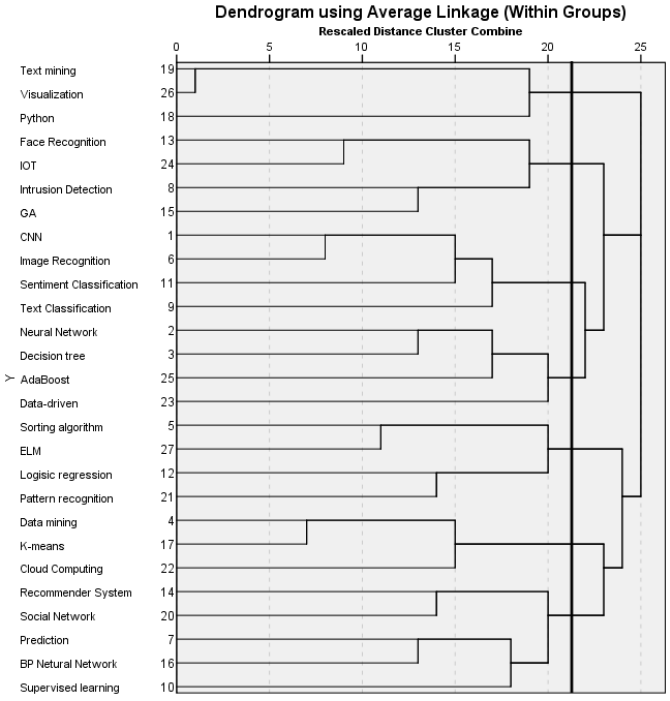

Figure 2. The result of clustering of factors

Import the dissimilarity matrix into SPSS for factor analysis. From the total variance explanation table in the output result, it can be observed that the eigenvalues of the first 27 factors are all greater than 1, and the cumulative sum of squares of the rotating load is 59.126%, as shown in Table 5. Observing the lithotripsy chart in Figure 1 again, it can be found that the lithotripsy chart gradually becomes smooth from the 28th factor, so the first 27 factors are selected as the main factors.

This paper selects the keyword with the largest load in each common factor to represent the

common factor, and

3.2 Cluster analysis

Clustering is a statistical data analysis method that groups a group of samples into different clusters according to certain characteristics. In theory, samples with similar attributes will enter the same group, and samples with larger attributes will enter different groups. In this paper, the main factor keywords are clustered in order to obtain the hot areas of machine learning applications.

The dissimilarity matrix of the main factor keywords obtained after screening is imported into SPSS to perform systematic clustering of the inter-group linkage method, and the resulting pedigree diagram is shown in Figure 2.

4. Conclusion

In this paper, through factor analysis and cluster analysis of the keywords of 4057 machine learning-related documents, 7 machine learning technology application hot spots are finally obtained. In this part of this article, these 7 hot topics will be explained.

4.1 Visual text mining

Text mining is a method used to discover potential and important information in a large amount of text data. It can be applied in many fields, such as identifying spam, collecting and analyzing competitive intelligence, managing customer needs, and text news classification (Kaushik & Naithani, 2016). With the development of machine learning in recent years, there have been many innovations in text mining. For example, in the mining process, a word matrix is constructed, and the information is also expressed in the matrix, and then machine learning techniques, such as clustering, association rules, etc., are used for analysis to identify potentially valuable information. Finally, these results will be visualized. Because visualization presents the results in simple ways such as pictures, and is easy to understand, visualization is becoming more and more common, and text mining is no exception.

4.2 GA in Feature Selection

Feature selection is to select the optimal subset from the candidate feature set to eliminate irrelevant and useless features to achieve the purpose of dimensionality reduction and facilitate subsequent analysis. Feature selection is very important for machine learning, and accurate features can improve the performance of the model in machine learning. For example, in the process of face recognition, the selection of face feature extraction will affect the accuracy of the face recognition result. Genetic algorithm is one of the important methods in machine learning. It is a search algorithm that simulates Darwinian evolution in biology to find the optimal solution to the problem. In feature selection, compared with traditional search algorithms, genetic algorithm can find the global optimal subset instead of the local optimal solution. Therefore, feature selection based on genetic algorithm is also very common.

4.3 CNN for NLP

There are many unstructured data in the network, such as text, images, videos, etc., but these data also contain a lot of information. Natural language processing is a method of processing and analyzing the text. It is the link between the contact and the computer, allowing the computer to

learn to understand human language and analyze human emotions. Therefore, the main applications of natural language processing include emotion classification, text classification, and text recognition. In the past, natural language processing used machine learning methods such as SVM and logistic regression, but now this trend has turned to CNN. CNN is very efficient in mining contextual semantic clues. CNN may become the best method to deal with NLP problems (Li, Li, Wang, & IEEE, 2018).

4.4 Ensemble Learning

Ensemble learning refers to a method of combining multiple models to solve specific machine learning problems. Algorithms such as decision trees and AdaBoost all belong to ensemble learning. Since ensemble learning is a combination of multiple models for training, it will average different hypotheses, thereby reducing the risk of choosing incorrect hypotheses individually, so that the final result is for the whole, rather than a single hypothesis. Secondly, the main advantage of ensemble learning is to avoid overfitting. Class imbalance, Concept drift and the problem of dimensionality will appear in machine learning algorithms, and the application of integrated learning will reduce these problems (Sagi & Rokach, 2018). Therefore, integrated learning is now receiving more and more attention.

4.5 Classification

Classification is one of the important applications of machine learning. It is a kind of supervised learning, which will be labeled before classification. In the end, which categories are known. Logistic regression is a kind of classification algorithm. It is often used for two classifications. It is used by many people due to its simplicity and strong interpretability. For example, it can be judged whether a tumor is benign or malignant in medicine. Also, for example, SVM and ELM are all classified algorithms.

4.6 Unsupervised algorithms

There are three types of machine learning training methods: supervised learning, semi-supervised learning and unsupervised learning. Unsupervised learning is a method of classifying data. Unlike supervised learning, unsupervised learning is not labeled before classification, and the entire classification process is not very purposeful, but people often do not pay attention to what this category is when using unsupervised learning, but just hope to group similar ones into one. Class. With the rapid development of technology nowadays, human beings and society are becoming more and more complex. Therefore, it is extremely difficult and difficult to label massive data before classification. Therefore, the application of unsupervised learning is becoming more and more common. For example, according to users like to classify users, quickly find abnormal users, etc. Clustering and dimensionality reduction are commonly used algorithms for unsupervised learning.

4.7 Recommender System

The recommendation system is a way to find out which products users are really interested in. Nowadays, with the rapid development of technology, the amount of data is also increasing. People will feel at a loss when facing massive amounts of data. At this time, the recommendation system can use the user's behavior to predict the products that the user is interested in, which saves the user's time and helps Increase the exposure of the product. BP neural network is a kind of

supervised machine learning algorithm, it is a kind of multi-layer feedforward neural network trained according to error back propagation algorithm. BP network constantly corrects the weight relationship between input and output according to the error of back propagation, and finally makes the prediction of the model more accurate. The recommendation system based on BP network will also be more accurate due to this feature.

References

[1] Makawana, P. R., & Jhaveri, R. H. (2018). A Bibliometric Analysis of Recent Research on Machine Learning for Cyber Security. In Y. C. Hu, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Eds.), Intelligent Communication and Computational Technologies (Vol. 19, pp. 213-226).

[2] El-Alfy, E. M., & Mohammed, S. A. A review of machine learning for big data analytics: bibliometric approach. Technology Analysis & Strategic Management.

[3] Whittaker, J., Courtial, J. P., & Law, J. (1989). CREATIVITY AND CONFORMITY IN SCIENCE - TITLES, KEYWORDS AND CO-WORD ANALYSIS. Social Studies of Science, 19 (3), 473-496.

[4] Cobo, M. J., Chiclana, F., Collop, A., de Ona, J., & Herrera-Viedma, E. (2014). A Bibliometric Analysis of the Intelligent Transportation Systems Research Based on Science Mapping. IEEE Transactions on Intelligent Transportation Systems, 15 (2), 901-908.