Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2023.040106
ISSN 2790-0983 Vol. 4, Issue 1: 44-51

Automatic Text Classification Model Based on Machine Learning

Jianhua Li*

Philippine Christian University, Philippine

13513673366@163.com

**corresponding author*

Keywords: Machine Learning, Automatic Text, Text Classification, Automatic Model

Abstract: The artificial neural network is composed of RBF algorithm, multi-layer parallel model and other parts. Its working principle is to classify the input image through layer by layer training, and then develop and determine the attributes of each neuron with similar features. This paper studies them based on Bayesian classifier, a common method in machine learning. First of all, before text recognition, we need to select the sample library and data type, and design the parameters according to the requirements to form an efficient automatic text classification. Then, according to the functional modules required by the specific model, select the appropriate language and compile and generate program code to realize the whole process, and carry out simulation tests on the functions of the model. The test results show that this test has prepared different quantities of simple words, short sentences, long articles, professional terms and colloquial expressions for classification tests. The classification accuracy of this model is as high as 90%, the error rate is low, and the classification time is fast, This shows that the model meets the needs of users.

1. Introduction

In recent years, the research focus of artificial intelligence has gradually shifted from computer based and machine learning technology to other related fields [1-2]. As a new discipline, it has made many remarkable achievements in practice. However, because the traditional classification methods have some defects and deficiencies, the artificial neural network theory has not been fully mature, and the international exploration of big data fusion application of expert systems is still in a rapid development stage, its research progress is still of great significance.

With the increasingly mature research of machine learning, scholars at home and abroad are also constantly making a large number of in-depth discussions on the relevant technologies based on

artificial neural networks, natural language processing and other unstructured data mining technologies, and have made some achievements. Some scholars proposed a clustering analysis algorithm based on machine learning method to deal with the problem of Chinese automatic tagging corpus task, and combined traditional classification training with semantic recognition for in-depth similarities and differences research [3-4]. Some scholars have proposed automatic text classification algorithm based on Bayesian theory. The model is based on fuzzy mathematics, combined with expert experience for decision analysis and calculation, and finally realizes the feature extraction of target corpus content and the design of word segmentation. Other scholars discussed and designed a template matching algorithm to preprocess the dataset (keywords) based on the multi semantic model, so as to effectively identify important words and topic phrases [5-6]. Therefore, based on machine learning, this paper studies the automatic text classification model.

This paper mainly studies the use of machine learning as a data training set, and designs an interactive function for automatically labeling keyword features, using machine learning algorithms for semantic analysis, and puts forward some suggestions for improving the three models based on actual cases. The experimental results show that this method can significantly improve the recognition efficiency and result quality.

2. Research on Automatic Text Classification Model Based on Machine Learning

2.1. Automatic Text Classification

The key of automatic text classification is to gather a large number of content with the same or similar features from a field, and establish that the objects to be recognized in similar recognition are accurately and appropriately distinguished. At present, the research on machine learning at home and abroad mainly focuses on the following aspects. Based on knowledge management and database management systems, it can improve classification efficiency and use big data analysis capabilities to achieve effective recognition of artificial intelligence. In the preprocessing process, it mainly involves the word frequency extraction, feature selection and other issues. Among them, the most basic and most important is to extract the most effective information and facilitate the research of using machine learning algorithms to obtain key data sets for recognition, and then whether the text classification task can be realized and how to accurately distinguish text and field become a difficulty [7-8]. However, at present, some existing automated text categories that have achieved significant improvement are only used as reference indicators. In machine learning, the attribute set and the topic set are mapped vertically and equivalently, that is, the basis for automatic recognition. In text classification, we should first analyze the characteristic attributes of the target group and their internal relations, and apply the different methods identified to the text content. The classification is based on the size of the feature space. The category data contains multiple text paragraphs, keywords with certain correlation and high overlap between different types. These important information can be directly obtained from a database or extracted from other systems through identifiers to achieve the implicit parameters required in the training process of the artificial neural network, such as length, position and other indicators.

2.2. Application of Text Classification

Text classification plays an increasingly important role in modern scientific research. The recognition analysis makes statistics and classification on the target information to be identified, and further determines the feature category, laying a good foundation for the next step of Web page

design. According to the acquired data, the most effective method is automatically selected to transform it into the optimal class or inefficient type, so as to meet the minimum time interval between the extraction of each attribute in the classified text. This is also one of the two main indicators to achieve high accuracy and accuracy [9-10]. There are a variety of tools for text classification, such as text, pictures and videos according to the content, text in real space according to the different recognition fields can be divided into paper documents, and a computer vision algorithm including the characteristics of a defined conceptual structure unit is designed to extract and classify each type of information with specific purposes in the target corpus, Thus, correct conclusions can be drawn when the input and output results are achieved [11-12].

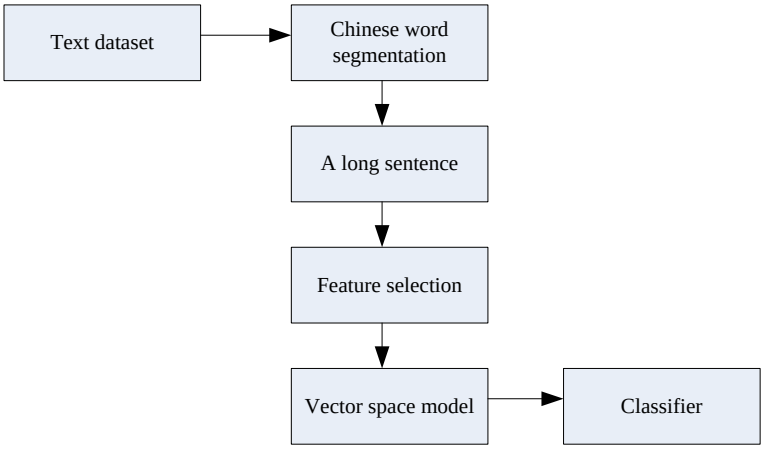

Figure 1. Text classifier basic operation

Text classification is a process of using machine thinking mode to complete the feature extraction, statistics and comparison of text content (as shown in Figure 1), and transform it into a specific form. In document retrieval, text sorting mainly adopts decision tree analysis and naive Bayes algorithm. The decision tree first constructs a sample set by constructing a branch or parent layer, and then calculates the number of sample subsets containing this characteristic component in the cluster center in the network according to the number of agglutinations (i.e. candidates) corresponding to the established ground class and this class, and judges which class it belongs to [13-14].

2.3. Machine Learning

Machine learning is a computer based ability to process large data sets by using existing program memory and methods. It can solve traditional uncertainty problems, such as neural networks. With the simplest text in human cognition as the main feature, it determines that there is a specific relationship between the members of the document by defining the data set identified by the corpus [15-16]. In this BIM attribute set, there are already one or more content candidates that can represent the nature of the entire text category as input objects. The process of machine learning includes two stages: perceptual training, organizational structure design and the ability to acquire knowledge from the outside world. The most popular research based on artificial intelligence is the artificial brain. Human beings create productivity by using mental labor. In this process, a large number of special processors are required to ensure whether the information output state (such as the working conditions of real-time workpieces) is normal or abnormal. Machine learning can overcome these problems and maximize their performance or application effects. At the same time,

it can improve decision-making ability and fault tolerance [17-18]. Figure 2 shows the machine learning process.

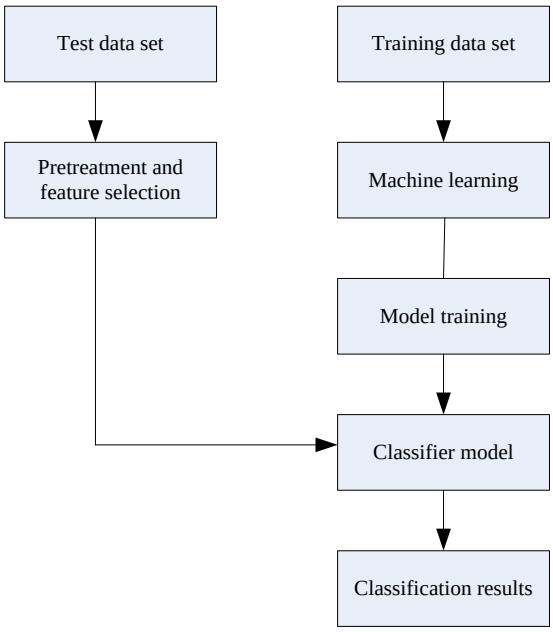

Figure 2. Machine learning process

It can solve the problems that are difficult to deal with under some specific conditions, and cannot be solved by using traditional methods when there is unknown or no clear representation in the application and it needs to be enumerated repeatedly. In an undetermined scheme, a variety of network image function sets with different variables or type characteristics are used, but they are connected with each other. There is an essential association between the training sample spaces (such as group distance relationship), and according to this connection, a backward layer by layer progressive learning process is formed. When an input occurs, reaching the output node state changes its storage unit, so that the entire organization is composed of countless small partners. In practical applications, the calculation of mutual information does not use the calculation method of probability in the formula, but uses the number of text occurrences in the training text set to replace the probability. The calculation formula is shown in (1):

$$MI(t_i, c_j) = \log \frac{XN}{(X+Y)(X+Z)} \tag{1}$$

Where, X represents the total number of texts with the characteristic term t in category c, Y represents the total number of texts without the characteristic term t in category c, Z represents the total number of texts with the characteristic term t in other categories except category c, and N represents the total number of all texts in the training corpus. It can be seen from Formula (2) that each category c can calculate one MI value for feature t. In actual use, for multiple categories, the MI value is generally the maximum. The calculation formula is as follows (2):

$$MI_{\max}(t_i) = \max\left(MI(t_i, c_j)\right) \tag{2}$$

The advantage of mutual information method is that it is easier to understand and implement, and the idea is simple and direct. However, the disadvantage of mutual information is that it does

not consider the frequency of feature entries, and it is easy to select some low-frequency features when selecting, which is greatly affected by the edge distribution of feature entries. When the P of two features are the same, the size of P (t) determines the size of mutual information. The larger the P (t) is, the greater the mutual information value will be, which will have a negative impact on the classification algorithm.

3. Experimental Process of Automatic Text Classification Model Based on Machine Learning

3.1. Model of Automatic Text Classification Model Based on Machine Learning

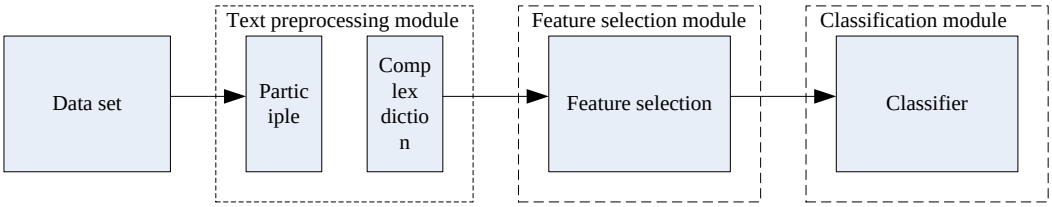

Figure 3. A model for automated text classification

In this paper, based on the text classification method of machine learning (as shown in Figure 3), combined with its inherent laws, automatic recognition and artificial neural network are combined to achieve the extraction, classification and statistics of all kinds of information involved in the process of natural language processing. Prognostic input into simple functions and then output into a specific library. The classification criteria are selected according to the input corpus of the model, and the classes are automatically generated through the machine learning training set, and then the text is annotated to the version table for statistical classification. Select corresponding type samples for corresponding objects in different categories based on classification rules. It includes: data characteristics, attribute variables and target fields; According to the clustering analysis method, all kinds of input languages are processed into high-frequency words in the distance dimension to express them. Based on the known information of one type of sample and the relevant data recorded in other databases of this type in another large type of database, the existing algorithm is used to obtain the final results and carry out a prediction comparison study. The algorithm constructs the recognition effect by introducing nonlinear neuron groups into the model to enhance the connection strength between feature parameters with high recognition rate, and combines text classification with other data points by using the mapping relationship between the training group and the tested object.

3.2. Model Function Test of Automatic Text Classification Model Based on Machine Learning

In the automatic text classification model based on machine learning, the content to be implemented needs to be tested first, because it directly determines whether the classification results are accurate. Secondly, the input vector position, parameter value and threshold ratio comparison are constantly adjusted manually. Through a large number of training sets to identify feature points, the optimal answer is obtained. In order to ensure the smooth completion of the testing process and achieve better results to achieve the desired goal, this paper divides the classification object into several subsets of the item product, and then takes the ratio of N samples of candidate categories contained in each text as whether the set within the current region belongs to a certain category. If these candidate categories have extensive coverage values in the field they belong to, which

category they have been classified into is the basis for selecting the optimal class rule.

4. Experimental Analysis of Automatic Text Classification Model Based on Machine Learning

4.1. Model Function Test Analysis of Automatic Text Classification Model Based on Machine Learning

Table 1 shows the model function test data of automatic text classification.

Table 1. Model function for automatic text classification

Classification	The test number	Classification accuracy(%)	Classification time(s)	Classification error rate(%)
Simple words	43	100	4	0
Short sentence	26	99	5	1
Long article	4	90	3	10
Technical term	54	96	4	4
Folk adage	47	98	6	2

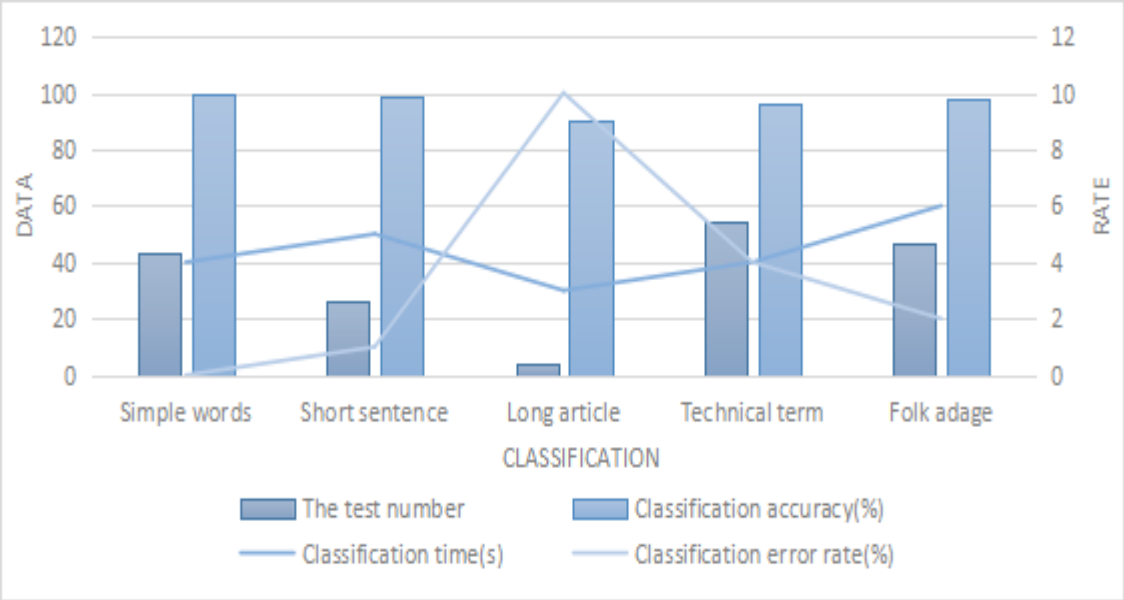

Figure 4. Model function of automatic text classification based on machine learning

Automatic text classification model testing based on machine learning mainly includes functional modules, flow charts and data analysis. In this part, after feature extraction, input a semantic code as the basic corpus, and design the semantic information expressed by different levels of languages according to the word bag structure to conduct hierarchical classification. Then, identify the topic category, attribute type, key code and other contents by using manual annotation forms. Finally, use machine learning to statistically classify the text classification results so as to intuitively show their correctness and effectiveness. It can be seen from Figure 4 that different numbers of simple words, short sentences, long articles, professional terms and colloquial expressions have been prepared for classification tests in this test. The classification accuracy of this model is up to more than 90%, the error rate is low and the classification time is fast, which

shows that this model meets the user's needs.

5. Conclusion

With the development of the times and the continuous improvement and updating of automatic text classification system by artificial intelligence technology in all walks of life at home and abroad, machine learning methods have been developed based on content recognition and retrieval models to help high-level talents extract effective information. It mainly includes three aspects: artificial neural network, BP algorithm and interactive database. This paper will focus on the research ideas and key technologies used in each module of the model of using artificial neural network to design and implement an efficient and scalable automatic text annotation template.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Kadhim A I. Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 2019,52(1):273-292.

[2] Janani R, Vijayarani S. Automatic text classification using machine learning and optimization algorithms. Soft computing: A fusion of foundations, methodologies and applications, 2021,25(2):1129-1145.

[3] Asogwa D C, Anigbogu S O, Onyenwe I E, et al. Text Classification Using Hybrid Machine Learning Algorithms on Big Data.2021,6(5):127-131.

[4] Wiedemann G. Proportional Classification Revisited: Automatic Content Analysis of Political Manifestos Using Active Learning. Social science computer review, 2019, 37(2):135-159.

[5] Mumivand H, Piri R S, Kheiraei F. A New Model for Automatic Text Classification. Electrical Science and Engineering, 2021, 3(1):10-15.

[6] Stasak B, Epps J, Goecke R. Automatic Depression Classification Based on Affective Read Sentences: Opportunities for Text-Dependent Analysis. Speech Communication, 2019, 115(10):1-14.

[7] Kragelj M, Borstnar M K. Automatic classification of older electronic texts into the Universal Decimal Classification-UDC. Journal of Documentation, 2020,77(3):755-776.

[8] Janani R, Vijayarani S. Automatic text classification using machine learning and optimization algorithms. Soft computing: A fusion of foundations, methodologies and applications, 2021,25(2):1129-1145.

[9] Satj N U, Ordin B. Application of the Polyhedral Conic Functions Method in the Text Classification and Comparative Analysis. Scientific Programming, 2018, 2018(2),1-11.

[10] Mou S, Du P, Cheng Z. A Brain-inspired Information Processing Algorithm and Its Application in Text Classification. Expert Systems with Applications, 2021, 177(9):1-7.

[11] Wei Z, Gui Z, Zhang M, et al. Text GCN-SW-KNN:a Novel Collaborative Training Multi-Label Classification Method for Wms Application Themes by Considering Geographic Semantics. Earth Big Data (English), 2021, 5(1):66-89.

[12] Wang X, Tong Y. Application of an emotional classification model in e-commerce text based on an improved transformer model. PLoS ONE, 2021, 16(3):1-16.

[13] Sui Z. Application of machine learning method in text classification. Basic & clinical pharmacology & toxicology.2019,124(S1):119-120.

[14] Ngamsuriyaroj S, Taninpong P. Tree-based text stream clustering with application to spam mail classification. International Journal of Data Mining, Modelling and Management, 2018, 10(4):353-370.

[15] Camacho D M, Collins K M, Powers R K, et al. Next-Generation Machine Learning for Biological Networks. Cell, 2018,173(7):1581-1592.

[16] Benjamin S L, AG Al án. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 2018, 361(27):360-365.

[17] Butler K T, Davies D W, Hugh C, et al. Machine learning for molecular and materials science. Nature, 2018, 559(25):547-555.

[18] JF Hern ández, Z D áz, Segovia M J, et al. Machine Learning and Statistical Techniques. An Application to the Prediction of Insolvency in Spanish Non-life Insurance Companies. The International Journal of Digital Accounting Research, 2020, 9(5):1-45.