Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.6 No.2

Design of data preprocessing recommendation algorithm for deep learning model learning

Hyeonji Kim [1)], Yoosoo Oh [2,)]*

[1) 2)] Dept.of ICT Convergence, Dept. of AI, Daegu University, Gyeongsan-si, Republic of Korea

Abstract. In machine learning and deep learning models, accuracy is determined by preprocessing of data. So the data preprocessing process is critical to learning in machine learning and deep learning. Therefore, in the deep learning model learning data analysis process, the data preprocessing process must be carried out to use the data collected by users. To proceed with the data preprocessing procedure, the user must analyze the collected data directly. This paper proposes an algorithm that identifies the learning data type collected by users. In addition, to make meaningful data, users must repeat the task of finding the appropriate preprocessing of the files they want to use. Therefore, this paper proposes a data preprocessing method recommendation algorithm for deep learning model learning to minimize cumbersome tasks. Recommendation algorithms are recommended by knowledge-based filtering. This paper verifies the performance of the recommendation algorithm through the evaluation of the Titanic classification model.

Keywords; Preprocessing data, Knowledge base filtering, Recommendation, data, deep learning

*Corresponding: yoosoo.oh@daegu.ac.kr
Received: Aug 14, 2021; Accepted: Nov 30, 2021; Published: Jun 30, 2022

This research was supported by X-mind Corps program of National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2019H1D8A1109865)

1. Introduction

Garbage in, garbage out (GIGO)' When garbage is put in, garbage comes out.' In the information and communication technology field, computers accept meaningless input data and produce meaningless results. Depending on the data's suitability and worth before input, the results' reliability may change. Therefore, the data preprocessing process is required to make the data you want to use meaningful. To perform the data preprocessing process, users must analyze the data collected for model generation directly and complete the cumbersome process of finding the appropriate preprocessing method. Since analyzing data now includes subjective opinions, an inappropriate preprocessing procedure may be performed [1]. This paper proposes an algorithm that recommends the data preprocessing method for deep learning. The recommendation algorithm proposed in this paper uses knowledge-based filtering that can be recommended even with a small amount of data. Knowledge-based filtering algorithms recommend based on explicit knowledge data about items. This paper vectorizes data related to preprocessing method using word2Vec. It automatically analyzes the types of files users want to use and uses genism [2] to derive similarities between file types and vectorized knowledge data. The proposed recommendation algorithm recommends a preprocessing method with the highest similarity [6][7]. The purpose of this proposed algorithm is to eliminate the hassle of analyzing data themselves and recommending preprocessing methods to apply appropriate preprocessing methods to learning data.

2. Related Research

This chapter examines the existing research related to designing an algorithm that recommends the data preprocessing method for deep learning presented in this paper.

2.1 wor2vec

Boo-sik Kang used vectors derived from word2Vec without using the user's existing user ratings. Boo-sik Kang constructed a corpus using user purchase history and rating information. Boo-sik Kang used the collected corpus as input data and calculated user vectors using Word2Vec. To select neighboring users, we computed similarities between users and other users with user vectors. The evaluation scores of the recommended products were predicted and recommended to users in the order of the highly select products based on this. Boo-sik Kang verified the proposed method using Mean Absolute Error (MAE). Research has shown that user vector-based collaborative filtering methods using word2Vec have significantly increased recommendation accuracy and were effective [4].

Sang-hoon Yoon and Geun-hyung Kim propose an analysis method that applies word2vec to LDA algorithms, a topic modeling algorithm that analyzes textual data. They applied the word2vec technique to extract related words for each topic's discrimination word. Then, for each discrimination word, the associated words were extracted by the word2vec function of the gensim module. Finally, according to the study, the discriminatory words included in the topic and each related word were analyzed to effectively identify the topic's meaning and give it a name [10].

Go-eun Heo used word2vec to analyze the semantic relationship between uncertainty words. The algorithm proposed in this paper uses WordNet to analyze synonyms-based networks for upper, lower word relationships, and uncertain words. Goeun Heo confirmed the model-based similarity between uncertainty words using the word2vec CBOW algorithm. As a result, it was possible to grasp the characteristics of the uncertain terms and visually confirm the semantic similarity of the words. Furthermore, it identified the possibility of extending uncertainty words based on synonyms.

2.2 Recommended Systems

Rosa, Renata Lopes, proposed a knowledge-based recommendation (KBRS) that included mental health monitoring systems that recognized potential mental disorders such as depression and stress. The knowledge-based recommendation algorithm proposed by Rosa Renata Lopes uses sentences from the Online Social Network (OSN). Characterize and extract stress and depression sentences included in mental health monitoring. The collected sentences are analyzed using sentimental material, and the user's information, ontology, etc., are further analyzed. The analyzed results become knowledge data for knowledge-based recommendation algorithms. As a result, Rosa, Renata Lopes derived an accuracy of 0.89 to 0.90. In addition, the knowledge-based recommendation system proposed by Rosa, Renata Lopes led to 94% user satisfaction [8].

Sung-seop Kim, Sung-woo Han, Ha-eun Mok, and Hye-bong Choi proposed collaborative filtering and hybrid methods. At this time, the core contents of the book were extracted and used as data using text mining techniques such as Latent Dirichlet Allocation (LDA). As a result, the proposed recommendation algorithm reflects the movie rating data and book rating data, and it confirmed that the recommendation accuracy improved between 0.3 and 0.5 [13].

Recommendation algorithms are recommended through data analysis of sentences and words in the data. Word2vec is suitable for determining the correlation between a sentence and a word. However, collaborative filtering or hybrids, the recommended algorithms used in the study above, require a large amount of data. Moreover, the above recommendation algorithms predict the central word through the context. It is not

appropriate for a method of predicting context with a single main word. In this paper, we apply the skip-gram of word2vec[9]. Using vectors derived from this, we figure out the correlation between the words in the data and use them as data for knowledge-based filtering. This paper proposes an algorithm that recommends the data preprocessing methods for deep learning using knowledge-based filtering that can be recommended even with a small amount of data.

3. Recommendation algorithm for pre-processing deep learning data

This paper proposes an algorithm identifying the learning data type and recommends the data preprocessing method for deep learning using knowledge-based filtering. The procedure for this is shown in Figure 1 below.

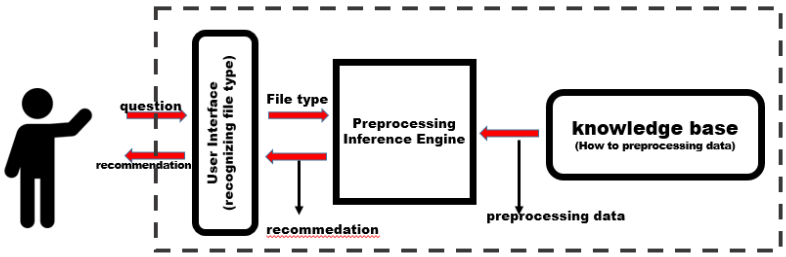

Fig. 1. Conceptual diagram of Recommend Preprocessing Algorithm

The preprocessing recommendation model is learning from preprocessing method data, built on explicit knowledge. The trained preprocessing recommendation model receives a file from the user. It analyzes the input files and recommends an appropriate preprocessing process.

3.1 Learning Data Type Identification Algorithm

An algorithm identifying the learning data type classifies collected data as numerical files, string files, and image files using file extensions to determine the file type. In the case of numeration and string files, the algorithm proposed in this paper checks the data type to identify whether it is an integer or a string. In the case of an integer, it is recognized as a numeral file, and in the case of a string, it is recognized as a string file in the case of an integer. The numeration file has to navigate for outliers, process preprocessing that removes outliers, or replace them with other values. Therefore, the standardization score (Z-score), Interquartile Range, chi-squared are used to detect outliers [2]. In this paper, to detect outliers using interquartile ranges. The process of calculating the quartile range is as follows. The quartile is the data sample divided by

four equal parts: Q1 (25% of the data is less than or equal to Q1), Q2 (50% of the data is less than or equal to Q2), and Q3 (75% of the data is less than or equal to Q3). Thus, the interquartile range can be obtained by IQR=Q3-Q1[2]. Outliers detect using a quartile range defines observations that exceed 1.5 times the quartile range as outliers. Using a quartile range, outlier detection uses a boxplot visualization graph. A boxplot visualizes the outliers of data collected using quartiles [2]. Figure 2 explores the outliers of fine dust (PM10) data and presents them as boxplots. If there is an outlier, as shown in Figure 2, an algorithm that identifies the data type proposed in this paper recognizes that preprocessing should be performed.

A string file should determine whether a missing value or a 'NaN' value exists. If there is a missing value or a 'NaN' value, the proposed algorithm recognizes that preprocessing should be performed.

Image files perform preprocessing according to image characteristics demanded by deep learning models, such as the size and color of images. The proposed algorithm that identifies data type, in this paper, recognizes the deep learning model entered by the user and the characteristics of the image file to be used and outputs whether preprocessing is required or not.

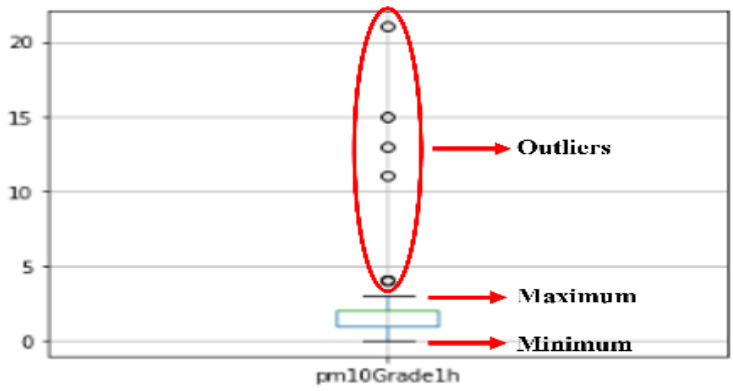

Fig.2.Outliers and boxplot visualization

3.2 Deep Learning Model Pre-processing Recommendation Algorithm

Recommended algorithms include collaborative filtering, content-based filtering, knowledge-based filtering, etc. [5]. This paper proposes an algorithm that recommends appropriate preprocessing methods based on numeric and string files and image files using a knowledge-based filtering model [5][12]. Explicit knowledge data for knowledge-based filtering is constructed using word2vec. Word2vec uses the gensim included in the Python library. Gensim's word2vec package is one of how words with similar meanings are vectorized to have similar orientation and scalar vectors [6]. This paper collected processing method data from Scikit-learn and GitHub to perform word2vec. The collected data are vectorized using word2vec. Figure 3 is a graph that visualizes vectorization derived from word2vec. Data derived from Word2vec are applied to preprocessing recommendation models. The preprocessing recommended learning model is learned by using Skip-gram, which predicts peripheral words as a central word. Skip-gram performs better because it can predict multiple words in one word. The recommended algorithm proposed in this paper was learned by applying Skip-gram from the central word to the left and right 10 words. In addition, words that appeared at least twice in the entire document were learned [9]. This paper calculates similarities between words with a learned model. The proposed algorithm designs algorithms to recommend preprocessing methods that have the highest similarity. The similarity was used as the cosine similarity [5].

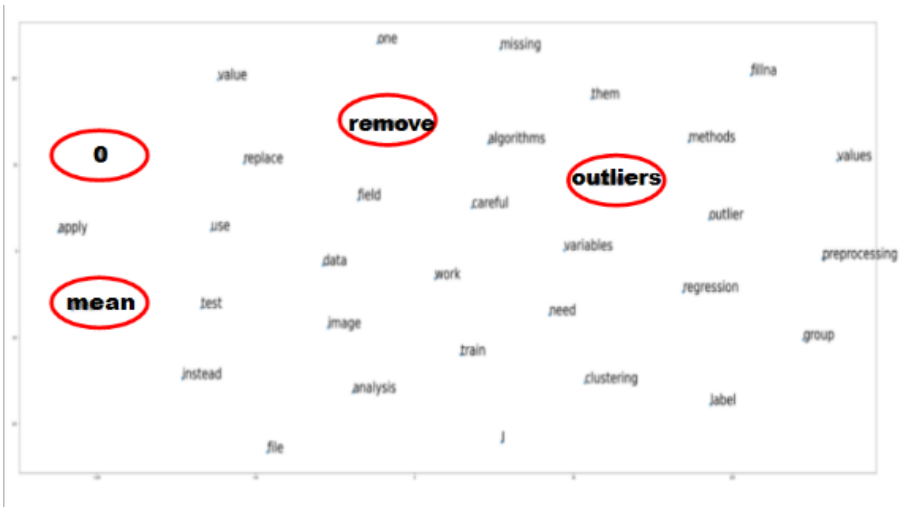

Fig.3.Word2vec visualization

4. Experiments and Results

This paper compares and analyzes pre-processing methods by the user and pre-processing by the recommended algorithm for evaluating the proposed algorithm. Titanic data was used as a dataset for evaluation. This paper split Titanic data into training data sets and test datasets for evaluation. Preprocessing was performed for classification model evaluation. This paper applied the preprocessing method recommended using the recommendation algorithm and other preprocessing methods that do not use the recommendation algorithm. The experimental steps are as follows. First, the path and extension name of the file are inputted by the user. This paper identifies that it is numerate data by applying the proposed algorithm. After identifying the data type, it detects for outlier values and outputs whether preprocessing is necessary. As shown in Figure 4, the learned recommendation model recommends being replaced with outliers to mean value. Finally, evaluate the classification model after performing alternative preprocessing with the mean of the outliers.

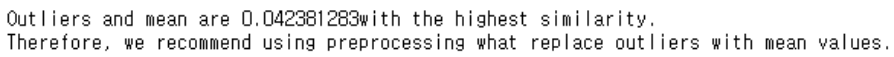

Fig.4.Sentences that recommend a preprocessing method

Table 1 below shows the Accuracy, Precision, Recall, F1 evaluation index for the test data set when the classification model is evaluated after preprocessing with the preprocessing method (Rp) recommended by the recommendation algorithm.

Table 1.Recommendation algorithm performance evaluation index

	Accuracy	Precision	Recall	F1
Rp	0.80	0.756	0.673	0.712

Table 2 below shows the model was evaluated after processing, replacing outliers with 0 (OP_0) or removing outliers (OP_r) by applying a preprocessing method other than the one recommended preprocessing through the recommended algorithm.

Table 2. Evaluated by applying different preprocessing methods

	Accuracy	Precision	Recall	F1
Op_r	0.79	0.723	0.663	0.708
Op_0	0.79	0.726	0.665	0.709

Table 3 shows a comparison of evaluation indicators and Area Under the Curve (AUC) when pretreated with a recommended preprocessing method (Rp) and other preprocessing methods (op). In this experiment, the data applying the recommended

preprocessing method through the proposed algorithm is higher than the evaluation score of the classification model of the data using other preprocessing methods.

As a result, the recommended preprocessing methods through recommendation algorithms were suitable for improving learning model accuracy.

Table 3.Overall evaluation indicators including AUC

	Accuracy	Precision	Recall	F1	AUC
Rp:m*	0.80	0.756	0.673	0.712	0.852
Op:r	0.79	0.723	0.663	0.708	0.802
Op:0	0.79	0.726	0.665	0.709	0.806

Conclusion

While it is essential to collect many data, making the collected data meaningful is also important. A preprocessing process is critical to produce meaningful data. In this paper, we propose an algorithm that recommended the data preprocessing method utilizing knowledge-based filtering. In addition, an algorithm for identifying learning data types was proposed to solve the hassle of users having to analyze data themselves. We can confirm that preprocessing through the proposed recommendation algorithm is more effective than user preprocessing through evaluating the Titanic classification model.

However, the results may not be accurate because the proposed recommendation algorithm has been trained with limited data. Future research is expected to improve accuracy if more data on preprocessing methods are collected and learned in the proposed recommendation algorithm. In addition, in this paper, it is expected that although word2vec of the gensim function is used, the accuracy of recommendation algorithms learned with word2vec and recommendation algorithms learned with Doc2vec algorithm can be compared.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Lee Se-hoon, Kim Ki-tae, Baek Min-joo and Yoo Chae-won (2020). Data preprocessing blocks for the Deep aI Yourself hands-on platform based on educational programming languages. Journal of the Korean Computer Information Society, 28(2), 297-298.

[9] Choi Ho-sik, and Park Chang. "Collective analysis of fine dust data using reduced boxplots." Journal of the Korean Data Analysis Society 18.5 (2016): 2435-2443.

[10] Kang Hyung-seok, Yang Jang-hoon. (2019). Semantic relationship analysis of word vectors learned by Word2vec models. Journal of Information Science, 46(10), 1088-1093.

[11] Kang, BooSik (2018). Improving Predictive Accuracy of User-based Collaborative Filtering Using Word2Vec. Journal of Knowledge Information Technology and Systems,13(1), 169-176.

[12] Minsung Kim, Moon Soo Cha, Jaeyeon Lee, and Son Kyung-ah. "Comparison of recommended system algorithms based on data scarcity." Journal of the Korea Communications Association's Academic Conference 2016.1 (2016): 70-71.

[13] Gensim,"What is Gensim? (2021) , https://radimrehurek.com/gensim/ intro.html# what-is-gensim

[14] Kim Sung-chul (Sungchul Kim), Kim Jung-hwan (Jeong-hwan Kim), Kim Na-young (Na-yeong Kim), Kim Tae-hoon (Taehoon Kim), and Yoo Hwan-jo (Hwanjo Yu). "Investigation of top-k related pair search methods using cosine similarity techniques." Journal of the Korea Information Processing Association, 24.1 (2017): 808-809.

[15] Rosa Renata Lopes. "A Knowledge-Based Recommendation System That Includes Sentiment Analysis and Deep Learning" IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS : 2124-2135.

[16] (Dongjae Kim), (Doangjoo Synn),and (Jong-kook Kim). "SG-Drop: Fast Skip-Gram by Dropping Context Words." Journal of the Korean Information Processing Society 27.2 (2020): 1014-1017.

[17] Yoon Sang Hoon, Kim Geun Hyung. (2021). An example of the expansion and analysis of topic modeling using Word2Vec. Information Systems Research, 30(1), 45-64. Yoon Sang Hun, Kim Keun Hyun. (2021). Expansion of Topic Modeling

with Word2Vec and Case Analysis. The Journal of Information Systems, 30(1), 45-64.

[18] Heo, G. E. (2019). A study on the network analysis between Word2Vec and WordNet-based uncertainty words. Journal of Literature and Information, 53(3), 247–271. https://doi.org/10.4275/KSLIS.2019.53.3.247

[19] Security Engineering Research Support Center (IJSIA). "Spam Filtering based on Knowledge Transfer Learning." International Journal of Security and Its Applications 9.10 (2015): 341-352.