향상된 텍스트 분류

왕광싱[*], 신성윤[O], 신광성[**], 이현창[**]
[O*]군산대학교 컴퓨터정보통신공학부
[**]원광대학교 디지털콘텐츠공학과
e-mail: s3397220@kunsan.ac.kr[O], {waver, hclglory}@wku.ac.kr[**]

An Improved Text Classification

Guangxing Wang[*], Seong-Yoon Shin[O], Kwang-Weong-Shin[**], Hyun-Chang Lee[**]
[O*]School of Com. Inf. & Comm. Eng., Kunsan National University
[**]Dept. of Digital Contents Eng., Wonkwang University

● 요 약 ●

In this paper, we propose an improved kNN classification method. Through improved the mothed and normalizing the data, the purpose of improving the accuracy is achieved. Then we compared the three classification algorithms and the improved algorithm by experimental data.

키워드: Improved kNN classification method, normalizing, algorithm

I. Introduction

In recent years, the machine learning-based sentiment classification method has achieved certain results, such as Amazon's book recommendation system, North American movie box office evaluation system, analysis of big data based on user preferences and evaluation, and targeted recommendation to users for hot sales. Books and hot reviews movies have greatly increased book sales and movie box office attendance [1, 2].

II. Classification Method

1. CART model

The CART algorithm is an implementation form of the decision tree. Usually there are three main implementations of decision trees, namely ID3 algorithm, CART algorithm and C4.5 algorithm [3,4]. The CART algorithm is a binary recursive segmentation technique.

2. SVM model

SVM is a common method of discrimination. In the field of machine learning, it is a supervised learning model that is commonly used for pattern recognition, classification, and regression analysis. Vapnik et al. proposed another design best criterion for linear classifiers based on years of research on statistical learning theory [5]. The principle is also linear from the point of view, and then extended to the case of linear indivisibility. Even extended to use nonlinear functions, this classifier is called Support Vector Machine (SVM).

III. Improved Method

Research and improvement methods for the kNN method have continued, such as the cluster-based CLKNN improvement algorithm proposed by Lijuan et al. [12], and the weight-based kNN improvement algorithm proposed by Halil Yigit et al. [13]. Due to the use of kNN algorithm for classification, it is necessary to calculate the similarity between the test text and each training text, which undoubtedly greatly increases the calculation amount of the classification, and the classification speed cannot be improved. Therefore, in the case of more training texts, how to reduce the amount of calculation and improve the classification accuracy is a key issue. Therefore, we process the data again during the process of using the kNN algorithm.

IV. Experiments

Our experiment is divided into two steps. First, prepare the data set and preprocess the data set. The pre-processed data sets are imported into the CART, SVM, and kNN models for prediction. Then, compare the accuracy and precision rate of the three models in the classification of sentiment analysis texts. In the second step, the improved method is used for prediction and comparison.

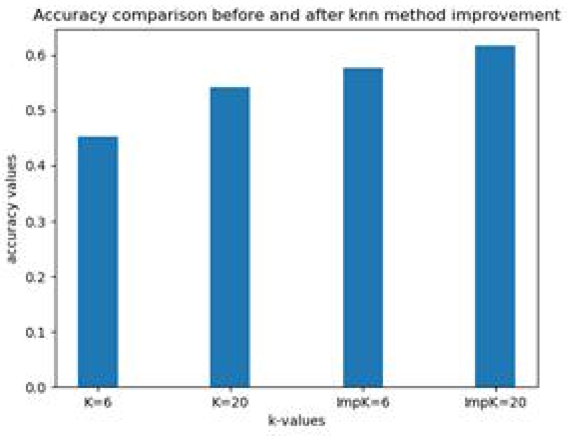

Fig. 1. Comparison of classification accuracy before and after kNN method improvement. K represents the value before the improvement, and ImpK represents the value of K of the improved KNN method.

V. Conclusions

This paper presented an improved sentiment classification method based on kNN, and elaborated on the improved methods, algorithms and implementation process. Based on the extracted small datasets and compared with the CART and SVM sentiment classification methods, the improved kNN method performed well in the experiment.

REFERENCES

[1] Brent Smith, Greg Linden, “Two Decades of Recommender Systems at Amazon.com,” IEEE Internet Computing, Vol.21, no.3, pp.12-18, 2017. DOI:10.1109/MIC.2017.72.

[2] Sajal Halder, Md. Samiullah, A. M. Jehad Sarkar, Young-Koo Lee, “Movie swarm: Information mining technique for movie recommendation system,” 2012 7th International Conference on Electrical and Computer Engineering, pp.462-465, 2013. DOI: 10.1109/ICECE.2012.6471587.

[3] Cai Yu, “Adaptive Japanese Teaching Optimization Based on Classification and Regression Tree,” 2017 International Conference on Robots & Intelligent System (ICRIS), pp.15-18, 2017. DOI: 10.1109/ICRIS.2017.12.

[4] Ruimin Li, Xiaoqiang Zhao, Xinxin Yu, Junwei Li, Nan Cheng, Jie Zhang, “Incident Duration Model on Urban Freeways Using Three Different Algorithms of Decision Tree,” 2010 International Conference on Intelligent Computation Technology and Automation, pp.526-528, 2010. DOI: 10.1109/ICICTA.2010.602.

[5] Rauf Izmailov, Vladimir Vapnik, Akshay Vashist, “Multidimensional splines with infinite number of knots as SVM kernels,” The 2013 International Joint Conference on Neural Networks (IJCNN), pp.1-7, 2013. DOI: 10.1109/IJCNN.2013.6706860.