텍스트 분류 기법의 발전

신광성[1] · 신성윤[2*]

[1]원광대학교 · [2]군산대학교

Enhancement of Text Classification Method

Kwang-Seong Shin[1] · Seong-Yoon Shin[2*]

[1]Wonkwang University · [2]Kunsan National University

E-mail : waver0920@wku.ac.kr/s3397220@kunsan.ac.kr

요 약

Classification and Regression Tree (CART), SVM (Support Vector Machine) 및 k-nearest neighbor classification (kNN)과 같은 기존 기계 학습 기반 감정 분석 방법은 정확성이 떨어졌습니다. 본 논문에서는 개선 된 kNN 분류 방법을 제안한다. 개선 된 방법 및 데이터 정규화를 통해 정확성 향상의 목적이 달성됩니다. 그 후, 3 가지 분류 알고리즘과 개선 된 알고리즘을 실험 데이터에 기초하여 비교 하였다.

ABSTRACT

Traditional machine learning based emotion analysis methods such as Classification and Regression Tree (CART), Support Vector Machine (SVM), and k-nearest neighbor classification (kNN) are less accurate. In this paper, we propose an improved kNN classification method. Improved methods and data normalization achieve the goal of improving accuracy. Then, three classification algorithms and an improved algorithm were compared based on experimental data.

키워드

Traditional Machine Learning, Support Vector Machine, k-Nearest Neighbor Classification

Ⅰ. 서 론

최근에는 기계 학습 기반의 정서 분류 방법이 아마존의 도서 추천 시스템, 북미 영화 박스 오피스 평가 시스템, 사용자 선호도 및 평가를 기반으로 한 대용량 데이터 분석 및 매서운 판매 사용자를 위한 추천 된 추천과 같은 특정 결과를 달성했다. 서적 및 인기 리뷰 영화는 서적 판매 및 영화 흥행에 크게 기여했다 [1, 2].

Ⅱ. SVM

SVM은 일반적인 차별 방법이다. 기계 학습 분야에서는 패턴 인식, 분류 및 회귀 분석에 일반적으로 사용되는 감독 학습 모델이다. Vapnik et al. 수년간의 통계 학습 이론에 기반 한 선형 분류기에 대한 또 다른 설계 최상의 기준을 제안했다. 이 원리는 또한 관점에서 선형이며, 선형 불가분성의 경우까지 확장된다. 심지어 비선형 함수를 사용하도록 확장 된 이 분류자를 SVM (Support Vector Machine)이라고 한다.

* corresponding author

III. 발전된 문서 분류

데이터에 대한 정규화 된 처리 방법은 특정 차원의 수치가 거리 계산에 영향을 주지 않도록 하는 것이다. 선형 함수 정규화 (최소 최대 스케일링)와 Z 점수 표준화의 두 가지 표준화 방법이 있다. kNN 알고리즘의 경우 사전 처리 된 데이터를 처리하기위한 Z 점수 표준화 방법이 사용되며 원시 데이터 평균 μ 및 표준 편차 σ가 데이터를 표준화하기 위해 제공된다.

IV. 실 험

이 논문에서는 THUCNews의 데이터 세트를 사용합니다. THUCNews 데이터 세트는 2005 년부터 2011 년까지 Sina News RSS 구독 채널의 과거 데이터 필터링 및 필터링에 따라 생성됩니다. kNN 방법 개선 전과 후의 분류 정확도 비교는 그림 1 에 나와 있습니다.

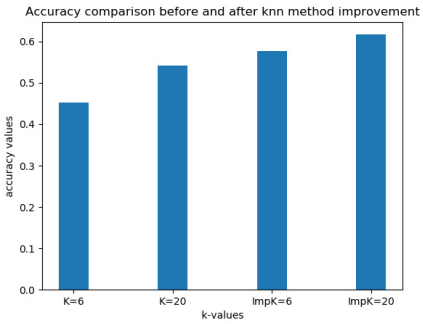

그림 1. kNN 방법 개선 전후의 분류 정확도 비교 K는 개선 전의 값을 나타내고, ImpK는 개선 된 KNN 방법의 K 값을 나타낸다.

V. 결 론

개선 된 kNN 분류의 정확성과 정확성이 향상되었고 실행 시간이 크게 줄어 들었다. 실험 결과에 따르면 kNN 모델의 정확도와 정확도는 감정 분류에서 11.5 %와 20.3 %로 증가했다. 특히 개선 된 kNN 방법 예측 시간이 36.5 % 단축되었다.

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.