텍스트 분류의 성능 향상을 위한 나이브 베이즈 응용 기법 비교 연구

허재희*, 박은영**, 박영호*
*숙명여자대학교 멀티미디어과학과
**협성대학교 시각디자인학과
e-mail : jaehee@sm.ac.kr, parkey@uhs.ac.kr, yhpark@sm.ac.kr

A Comparison Study on the Application Method of Naive Bayes for Text Classification

Jae-Hee Heo*, Eun-Young Park**, Young-Ho Park*
*Dept. of Multimedia Science, Sookmyung Women's University
**Dept. of Visual Design, Hyupsung University

요 약

텍스트를 분류해내는 일이 점점 중요해지고 있는 현 시점에서 기계학습은 다른 기법들보다도 가장 효과적인 성능을 드러낸다. 그 중에서도 특히 나이브 베이즈 분류기는 간결하고 효율적으로 알려진 기계학습 모델 중에 하나이다. 본 논문은 보다 효과적인 텍스트 분류를 위해 나이브 베이즈의 기법들을 응용 및 개선하고자 한 기존의 연구들을 소개하고, 이를 분석하고자 한다.

1. 서론

기계학습이란 컴퓨터가 훈련 데이터를 바탕으로 스스로의 성능을 향상시켜 마치 인간처럼 학습하는 효과를 얻도록 하는 기술 분야를 의미하며[1, 2] 이러한 텍스트 분류이 중요성은 점점 높아지고 있다[3]. Almeida[4]는 이와 같은 분류 작업에는 다른 기법들보다도 기계 학습이 가장 효과적인 성능을 드러낸다고 밝혔다. [4]에 의하면 베이지안 기법은 간결함과 계산 복잡도, 그리고 정확성으로 인해 정교한 학습 알고리즘을 짜는 데에 효과적이며, 특히, 변수들이 모두 독립이라는 것을 가정한 베이지안 기법인 나이브 베이즈 분류기는 간결하고 효율적[3]이기 때문에 텍스트 분류에서 널리 사용되고 있다[5]. 이에, 본 논문은 나이브 베이즈 분류기를 응용하여 텍스트 분류의 성능 향상을 연구한 기존의 연구들을 소개하고 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 나이브 베이즈 분류기를 통해 텍스트 분류의 성능을 개선한 기존의 연구들을 소개한다. 3장에서는 개선된 나이브 베이즈 분류기 응용 기법들을 비교 분석한다. 4장에서는 향후 관련 연구에 대한 기대효과와 함께 결론을 내린다.

2. 관련 연구

본 장에서는 나이브 베이즈를 통해 텍스트 분류 성능을 효과적으로 개선시킨 관련 연구들을 조사한다. 텍스트 분류의 가장 큰 문제점으로 특징 공간의 고차원성이 있다[3, 4]. 이는 노이즈를 유발하여 텍스트 분류의 정확성을 떨어뜨리고, 분류기의 성능을 저하시킨다. 그러므로 텍스트 분류의 성능을 향상시키기 위해서는 차원 문제를 해결하는 것이 중요하다. 따라서, 본 논문에서는 나이브 베이즈 분류기를 통해 고차원성 문제 해결에 기여한 관련 연구들을 분석하고자 한다.

특징 공간의 고차원성을 해결하기 위한 대표적인 기법 중에 하나는 특징 선택 기법이다. Yang과 Pedersen[6]은 특징 선택 기법이 텍스트 분류에서의 특징 공간 고차원 감소 문제 해결의 효율성을 연구를 통해 검증했으며, 이후 Chen[3]이 특징 선택 기법의 효율성을 검증할 수 있는 새로운 메트릭스를 제안했다. Dhillon[7]은 특징 군집화 기법을 바탕으로 제안한 분리 알고리즘을 나이즈 베이즈와 접목시켜 텍스트 분류의 성능을 향상시켰다. 마지막으로 SB Kim[8]은 나이브 베이즈 분류기에 특징 가중치 기법을 적용시킴으로써 전통적인 나이브 베이즈 텍스트 분류의 문제점을 개선시켰다.

2.1 특징 선택 기법을 통한 텍스트 분류 성능 개선

Chen은 특징 공간의 차원을 줄여 효율성과 정확성을 향상시키기 위해 나이브 베이즈 분류기에 특징 선택 기법을 적용시킬 것을 제안했다. 또한 [4]를 통해 클래스 분할 지표(CDM: Class Discriminating Measure)와 다중 클래스 오즈비(MOR: Multi-class Odds Ratio)라는 두 개의 메트릭스를 제안하여 다중 클래스 데이터셋이 적용된 나이브 베이즈 분류기를 검증하는 방법을 제안하였다. 연구에 의하면 나이브 베이즈의 다항분포 모델에서 적은 데이터셋일 때는 SVM(Support Vector Machine)과 성능 면에서 큰 차이를 보이지 않았지만, 데이터셋이 커질 때는 시간이 단축되는 효과를 보였다.

2.2 특징 군집화 기법을 통한 텍스트 분류 성능 개선

Dhillon은 Baker과 McCallum을 인용하여 특징 군집화는 특징의 개수가 적을 경우에 특징 선택 기법보다 더 효율적이며, 텍스트 분류의 정확도를 높인다고 밝혔다. [7]의 연구에 의하면 Dhillon은 군집화에 의한 상호 정보량의 감소를 포착하는 분리 알고리즘을 제안했다.

구체적으로, 특징 및 단어 군집화를 통해 텍스트 분류 성능을 개선시켰는데, 훈련 데이터가 희박하거나 특징의 개수가 적을 때 특징 선택 기법의 IG보다 더 높은 정확도를 얻었다. 따라서, Dhillon의 알고리즘은 나이브 베이즈 분류기와 결합했을 때 계층적 복잡성을 줄였다.

2.3 특징 가중치 기법을 통한 텍스트 분류 성능 개선

특징 선택 기법은 선별된 단어들의 중요도를 모두 같다고 설정한다. 이는 기계학습의 복잡성을 줄여준다는 장점을 가지고 있다. 그러나 텍스트 분류기로 새로운 훈련 데이터들이 지속적으로 유입될 경우, 그 때마다 수정된 단어 통계에 의해 특징들을 재정의해야 한다는 번거로움을 가지고 있다[8, 9].

이러한 문제를 해결하기 위해서, SB Kim은 대신에 특징 가중치 기법을 적용시킬 것을 제안했다[9]. 특징 가중치 기법을 통해 특징 데이터셋은 동적으로 활용되었으며, 비가중치 나이브 베이즈 분류기와 비교했을 때 더 좋은 성능을 보였다.

3. 비교 및 분석

본 장에서는 2장에서 설명한 기존 연구의 특징들을 <표 1> 에서와 같이 간략히 비교하고 텍스트 분류의 성능 향상을 모색한다. Chen은 새롭게 제안한 두 개의 메트릭스를 통해 효율적인 특징 선택 기법을 나이브 베이즈 분류기에 적용하여 성능을 검증했다. Dhillon은 군집화에 의한 상호 정보량의 감소를 포착하는 알고리즘을 제안하여, 이를 나이브 베이즈와 결합시킴으로써 텍스트 분류 성능을 개선시켰다. SB Kim은 특징 가중치 기법을 나이브 베이즈 분류기에 적용하여 훈련 데이터가 적은 경우에도 데이터셋을 동적으로 활용하여 정확도를 향상시켰다.

<표 1> 텍스트 분류 성능 개선을 위한 기법 비교

2.1 Chen	특징 선택 기법을 검증하는 두 개의 메트릭스를 제안하여 다중 클래스 데이터셋이 적용된 나이브 베이즈 분류기의 성능을 개선시켰다.
2.2 Dhillon	군집화에 의한 상호 정보량의 감소를 포착하는 분리 알고리즘을 제안했다. 계층적 복잡성을 줄이는 데에 기여했다.
2.3 SB Kim	특징 가중치 기법을 통해 데이터셋을 동적으로 활용했다. 비가중치 나이브 베이즈 분류기와 비교했을 때 성능이 개선되었다.

4. 결론 및 향후 기대효과

본 논문에서는 고차원성의 문제를 해결하기 위해 각각 나이브 베이즈에 특징 선택 기법, 특징 군집화 기법 그리고 특징 가중치 기법을 적용시켜 텍스트 분류 성능 향상에 기여한 기존 연구들을 소개하고 분석하였다. 강한 독립을 가정하는 나이브 베이즈 분류기는 효율적으로 텍스트를 분류하는 기계학습 기법 중에 하나이다. 그러나 특징 공간의 고차원성은 텍스트 분류의 정확도를 위해 해결해야할 문제이며, 이를 줄이기 위해 향후에는 고차원의 훈련 데이터를 단순화하는 방법을 강구하는 연구가 더욱 필요할 것이다.

향후 연구로는 고차원 문제를 해결하는 방법들을 다양한 조건에서 비교 실험하는 연구를 구체적인 방향으로 진행하고자 한다.

사사문구

본 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임. (2012003797)

참고문헌

[1] AL. Samuel, “Some studies in machine learning using the game of checkers,” Published in IBM Journal of research and development, 2000.

[2] 장병탁, “차세대 기계학습 기술,” 정보과학회지 제25권 제3호 통권 제214호 (2007년 3월) pp.96-107, 2007.

[3] J. Chen, H. Huang, S. Tian, Y. Qu “Feature selection for text classification with Naive Bayes,” Published in Journal Expert Systems with Applications: An International Journal archive, Volume 36 Issue 3, pp.5432-5435, 2009.

[4] Tiago A. Almeida, Jurandy Almeida, Akebo Yamakami “Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers,” Published in Journal of Internet Services and Applications Volume 1, Issue 3, pp.183-200, 2011.

[5] KM Schneider, “Techniques for Improving the Performance of Naive Bayes for Text Classification,” Proceedings of CICLing'05 Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing, pp.682-693, 2005.

[6] Y. Yang, Jan O. Pedersen “A Comparative Study on Feature Selection in Text Categorization,” Proceedings of ICML, 1997.

[7] I S. Dhillon, “Divisive information-theoretic feature clustering algorithm for Text Classification,” Proceedings of The Journal of Machine Learning Research archive, Volume 3, pp.1265-1287, 2003.

[8] SB. Kim, HC. Rim, DS Yook, HS Lim "Effective

Methods for Improving Naive Bayes Text Classifiers," Proceedings of PRICAI 2002: Trends in Artificial Intelligence, Lecture Notes in Computer Science Volume 2417, pp.414-423, 2002.
[9] SB. Kim, KS. Han, HC. Rim, SH. Myaeng "Some effective techniques for naive bayes text classification," Proceedings of the Knowledge and Data Engineering, IEEE Transactions on Volume:18, Issue: 11, Pages 1457-1466, 2006.