단어의 불순도를 고려한 특징 선택 방법 연구

강진범[0] 양재영 최중민
지능시스템 연구실, 한양대학교
{jbkang[0], jyyang , jmchoi}@cse.hanyang.ac.kr

An Enhanced Feature Selection Method using the Impurity of Words

Jinbeom Kang[0] JaeYoung Yang Joongmin Choi
IS Laboratory Dept. of Computer Science & Engineering, Hanyang University

요 약

효과적인 문서 분류를 위해 학습 하고자 하는 클래스와 관련된 많은 특징들이 필요하다. 하지만 학습하고자 하는 개념과 관련이 없거나 중복된 정보가 수집된 정보 속에 존재한다. 학습 과정에서 정확한 지식 습득을 하기 위해 특징 선택 방법을 사용하였다. 본 논문에서는 클래스에 대한 단어의 불순도를 이용한 특징 선택 방법을 제안한다. 기존의 특징 선택 방법과 비교 분석하여 기존 특징 선택 방법의 문제점을 파악하고 개선된 기법을 보인다.

1. 서 론

특정 의미를 가진 단어들이 모여 문장을 이루고 이런 문장들이 모여 하나의 문서를 구성하고 있다. 기계 학습 분야에서 특정 문서를 대표할 수 있는 단어들 및 패턴들이 문서를 나타낼 수 있는 특징(feature)이 된다. 기계 학습을 이용한 문서 분류에서는 예제 집합(example set)으로부터 클래스를 표현할 수 있는 지식을 습득한다. 생성된 지식은 알려지지 않은 새로운 자료에 대한 클래스로 분류하기 위해 제공될 수 있다. 기계 학습은 두 단계로 이루어진다. 학습(learning) 단계에서는 특징들과 클래스 간의 관계나 규칙성을 찾기 위한 시도를 하고, 분류(classification) 단계에서는 학습 단계에서 추론된 학습 모델을 이용하여 결과를 예측할 수 없는 새로운 예제(example)에 대한 클래스를 찾는다.

효과적인 문서 분류를 위해서는 학습하고자 하는 클래스와 관련된 많은 특징들이 필요하다. 하지만 수집된 많은 정보들 중에는 학습하고자 하는 개념(concept)과 관련이 없거나(irrelevant) 중복된(redundant) 정보를 가진 경우도 있다. 또한 자료 자체에 노이즈가 있기도 하다. 이와 같이 학습 모델 생성을 위해 수집된 정보가 신뢰할 수 없다면, 학습 과정에서도 정확한 지식의 습득이 어렵다[4].

특징 선택(feature selection)의 과정은 학습할 클래스와 관련이 없거나 중복된 정보를 학습 모델 생성 이전에 제거함으로써 학습 알고리즘의 성능을 향상시키기 위해 학습 알고리즘이 수행되기 전의 전처리 과정으로 사용된다. 이러한 과정을 통해서 많은 자료들 중 실제 분류 성능에 영향을 줄 수 있는 특징들을 검증해 낼 수 있다. 또한 학습 모델 생성에 사용될 자료의 수를 줄임으로써 학습 알고리즘이 좀 더 빠르고 효과적으로 동작할 수 있으며 생성된 학습 모델의 크기도 줄일 수 있다.

본 논문에서는 클래스를 잘 표현할 수 있는 특징에 대해 탐구하고 향상된 특징 선택 방법에 대해 제안한다. 특징 선택을 하기 위해 두 가지의 가정을 두었다. 하나는 특징이 하나의 클래스에만 나타나면 그 특징은 다른 클래스의 간섭을 받지 않고 특정 클래스를 잘 나타낼 수 있는 좋은 특징이라 가정하였다. 또한 학습을 하기 위해 수집하는 예제 집합들이 균등(balance)하게 분포가 되지 않는다는 단점을 고려하는 가정을 가지고 있다.

2. 관련 연구

특징 선택은 기계학습, 통계 데이터 마이닝 그리고 패턴 인식 분야에서 활발히 연구되고 있다. 초기에는 특징의 중복성 및 관련성이 적은 특징들을 제거하는 것이 목표였다. 이와 같이 예제 집합으로부터 불필요한 특징들을 제거함으로써 학습 모델 생성 시 발생하는 계산 시간이나 많은 자료의 수집 및 관리에 드는 비용을 줄일 수 있다. 또한 만들어진 학습 모델로부터 생성되는 규칙들을 보다 쉽게 이해할 수 있다.

일반적으로 가장 많이 사용하는 특징 선택 알고리즘은 filter 접근법[1,5]이다. 학습 알고리즘과 독립적으로 동작하는 방법으로 자료들의 일반적인 특징을 기반으로 한 휴리스틱 함수를 이용하여 선택된 특징 집합을 평가한다. 이 방법은 빠르고 많은 예제 자료 집합을 사용하기에 효율적이다.

분류 알고리즘들을 실세계에 존재하는 거대 문서 자료 집합에 적용하기 위해서는 반드시 문서를 표현하는데 이용되는 특징의 수를 축소시켜야 한다[3]. 이를 위해 [1]에서는 문서 집합과 하부 특징 집합과의 관련성을 비교하여 특징을 추출한 후 ID3, C4.5를 이용한 분류 실험에서 특징들의 차원 축소(dimensionality reduction)가 성능향상에 미치는 효과를 입증하였다. 또한 [2,3]에서는 통계적인 기반의 문서 분류 학습 기법을 이용하였다. [3]에서는 로이터(Reuters) 자료에 대해 여러 가지의 특징 선택 방법과 k-NN과 LLSF(Linear Least Squares Fit

mapping) 알고리즘을 적용한 분류 실험에서 원래 문서의 98%를 제거한 특징만을 가지고도 더욱 정확한 분류를 해낼 수 있음을 측정하였다.

3. 불순도를 이용한 특징 선택

클래스를 대표할 수 있는 특징이 좋은 특징들이다. 이것은 다른 클래스에는 나타나지 않고 오직 한 클래스에만 나타나는 특징을 말한다. 이런 가정을 기반으로 한 특징 선택 방법이 정보 획득 방법이다. 정보 획득은 엔트로피(entropy)를 이용하여 불순도를 측정한다. 하지만 더욱 간단한 수식으로 불순도를 측정할 수 있다.

결정트리(decision tree)를 이용한 데이터 마이닝에서 분류 지점(split point)을 찾기 위해 *Gini* 인덱스를 이용한다. *Gini* 인덱스는 불순도를 통해 좋은 분류 지점을 찾고 이를 통해 트리의 구조를 단순화 시킨다.

본 논문에서는 *Gini* 인덱스를 특징 선택에 사용하기 위해 새롭게 정의 하였다.

$$Gini(w,c)=[1-P(d|w)(1-P(d|w)]$$

특정 단어가 클래스에 대해 나타날 확률과 나타나지 않을 확률을 곱함으로써 불순도를 측정할 수 있다. 클래스 c에 대해서 불순도가 크다는 것은 그 클래스를 대표할 수 있는 좋은 특징이 아님을 알 수 있다. 그래서 1에서 그 수를 뺌으로써 클래스 c에 대한 순수도를 측정할 수 있다.

모든 클래스에 대해 수집한 예제 집합이 균등하게 분포되어 있지 않다. 이로 인해 특정 클래스에 대해 많은 예제가 집중 될 수 있다. 특정 클래스에 집중된 예제들은 많은 특징들을 보유하여 상대적으로 다른 클래스의 특징들이 누락 될 수 있다. 더불어 기존의 특징 선택 방법은 이 점을 고려하지 않는다. *Gini* 인덱스 방법에 학습 시 클래스에 배당되는 예제들을 고려하여 가중치를 주었다.

$$Gini^{weight}(w,c)=Gini(w,c)P(w|c)$$

이 가중치는 클래스 c에서 단어 w가 나타날 확률로 클래스 c의 학습 문서 수에 대한 단어 w가 나타난 문서의 비율을 의미한다.

$Gini^{weight}$ 값이 크다는 것은 다른 클래스에 나타나지 않고 오직 해당 클래스에만 나타난 좋은 단어임을 나타낸다. 따라서 $Gini^{weight}$의 값이 큰 집합을 분류 작업 시 사용하게 된다.

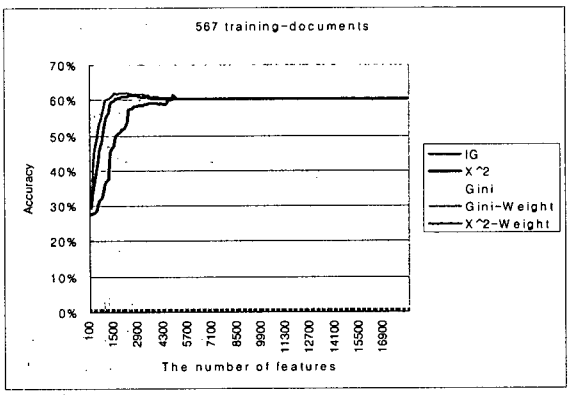
$$Gini_{\max}^{weight}(w)=\max_{i=1}^{m}\{Gini^{weight}(w,c_i)\}$$

모든 클래스에 걸쳐 높은 $Gini^{weight}$ 값을 가진 특징들은 분류 작업 시 클래스를 나타낼 수 있는 좋은 특징일 뿐만 아니라 모든 클래스에 대해 균등한 수의 특징을 선택할 수 있다. 이렇게 선택된 특징은 적은 수로도 새로운 자료에 대한 분류 성능을 높일 수 있다.

4. 성능 평가

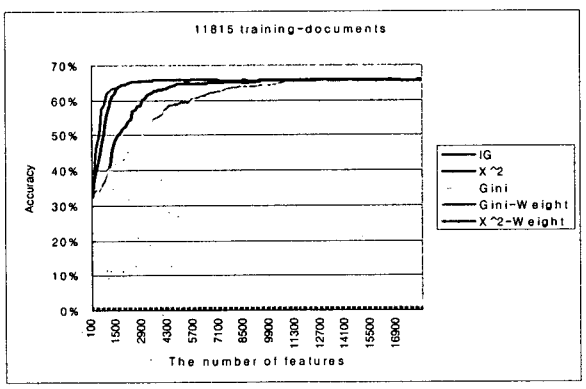
본 논문에서 제안한 특징 선택 방법의 성능 평가를 위해 Reuters21578 자료 집합에 대해 10-폴드 교차 타당성 평가를 하였다. Reuters21578 자료 집합은 119개의 클래스와 21578개의 문서로 구성되어 있다.

나이브 베이지안을 이용하여 5가지의 특징 선택 방법을 평가 하였다. 일반적으로 잘 알려진 정보 획득, χ^2 방법과 본 논문에서 제안한 $Gini_{\max}$, $Gini^{weight}_{\max}$ 방법을 비교 실험 하였다. *Gini* 인덱스에 곱해진 가중치가 특징 선택에서 어떤 영향을 주는지 알기 위해 χ^2에 제안한 가중치를 부여하여 비교 실험 하였다.



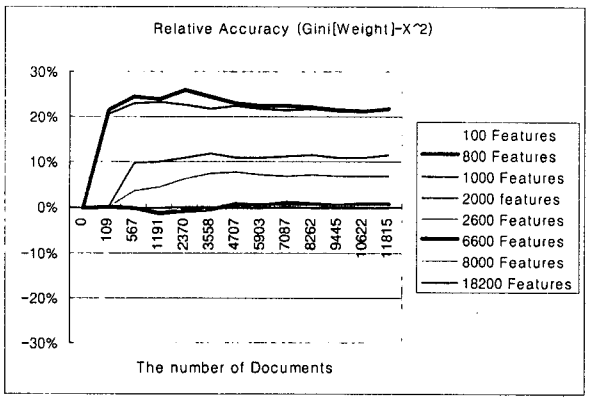
[그림 1] 567개 학습 문서에서 특징 수에 따른 비교 그래프

[그림 1]의 실험은 567개의 문서에서 특징 수에 따른 비교 그래프이다. 정보획득 방법은 불순도를 측정하는 $Gini_{\max}$와 같은 특성을 가지고 있어 그래프의 변화가 없음을 알 수 있다. χ^2 방법으로 선택된 적은 수의 특징들은 성능이 좋지 못함을 알 수 있다. 반면, $Gini^{weight}_{\max}$ 방법은 적은 특징만으로도 높은 성능을 보여줌을 알 수 있다. 가중치를 곱한 χ^2 방법은 높은 성능을 보이긴 하지만 본 논문에서 제안한 $Gini^{weight}_{\max}$ 방법보다 좋지 못함을 알 수 있다.



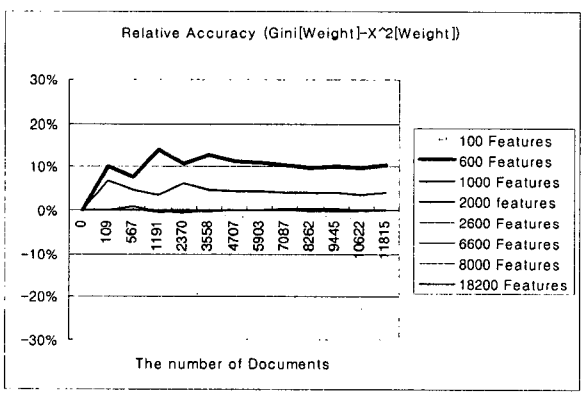
[그림 2] 11816개 학습 문서에서 특징 수에 따른 비교 그래프

[그림 2]는 11816개 문서에서 비교 실험한 그래프이다. $Gini_{max}$ 와 정보 획득 방법은 특징의 수가 적을 때 낮은 성능을 보이지만 학습 문서가 많아짐에 또한 따라 χ^2 방법은 더욱 좋은 성능을 보임을 알 수 있다. 가중치가 곱해진 $Gini^{weight}_{max}$와 χ^2 가중치 방법이 매우 좋은 성능을 보임을 알 수 있다. 하지만 특징의 불순도에 따른 특징 선택 방법이 더 좋음을 알 수 있다.



[그림 3] $Gini^{weight}_{max}$ 방법과 χ^2 방법의 상대적 정확도

[그림 3]은 $Gini^{weight}_{max}$ 방법의 정확도에서 χ^2 방법의 정확도를 뺀 상대적인 정확도를 나타내는 그래프이다. 그래프에서 800개 특징을 사용한 경우 χ^2 방법보다 최대 26% 높은 성능을 보였다. χ^2 방법의 경우 최대 1% 높게 나타남을 알 수 있다.



[그림 4] $Gini^{weight}_{max}$ 방법과 χ^2 가중치 방법의 상대적 정확도

실험에서 가장 좋은 성능을 보였던 $Gini^{weight}_{max}$와 χ^2 가중치 방법의 상대적 정확도를 [그림 4]에서 볼 수 있다. 800개의 특징을 이용한 분류 실험에서 $Gini^{weight}_{max}$ 방법이 최대 14% 높은 성능이 나타났다.

5. 결론 및 향후 과제

본 논문에서 클래스에 대한 단어의 불순도를 고려하고 클래스별 학습 문서 비율을 가중치로 부여함으로써 새로운 특징 선택 방법을 제안하였다. 제안한 방법이 기존의 방법들보다 높은 성능을 보임을 알 수 있었다.

적은 학습 문서에서 χ^2 방법은 선택된 특징이 좋은 분류 결과를 주지 못하였다. 하지만, 학습 문서가 많아짐에 따라 높은 성능을 보였다. 정보 획득 방법은 적은 문서에서 좋은 특징 들이 선택 되었지만, 문서가 많아짐에 따라 좋은 특징을 선택하지 못하였다.

본 논문에서 제안한 $Gini^{weight}_{max}$ 방법은 실험을 통해 χ^2 방법과 정보 획득 방법의 문제점을 보완함을 알 수 있었다.

χ^2 가중치 방법과 비교 실험을 통해 예제 집합의 구성에 대한 중요성을 파악할 수 있었다.

향후 제안한 특징 선택 방법을 문서 분류뿐만 아니라 음성 및 영상의 특징 선택에 적용하여 연구할 수 있을 것이다.

참고문헌

[1] G. H. John, R. Kohavi, K. Rfleger, "Irrelevant Features and the Subset Selection Problem", Proc. of ICML94, 121-129, Morgan Kaufmann Pulishers, San Francisco, CA, 1994

[2] I. H. Witten and Eibe, Frank, Data Mining, Morgan Kaufmann Publishers, 2002

[3] Y. Yang, J. O. Pedersenm "A Comparative study on Feature Selection in Text Categorization", Proc. of ICML97, pp 412-420, 1997

[4] M. A. Hall. "Correlation-based Feature Selection for Machine Learning", Ph. D diss. Hamilton, NZ: Wailkato University, Department of Computer Science.

[5] K. kira and L. A. Rendell. "The feature selection problem: Traditional methods and a new algorithm." In 10th National Conference on Artificial Intelligence, pp 129-134. MIT Press 1992.