문서 분류를 위한 특징 선택

진 훈[0] 김인철
경 기 대 학 교 대 학 원 전 자 계 산 학 과
{jinun, ickim}@kyonggi.ac.kr

Feature Selection for Document Classification

Hoon Jin[0] In-Cheol Kim
Dept. of Computer Science, Kyonggi University

요 약

본 논문은 텍스트 형태로 존재하는 문서가 특정 범주에 속하는 지를 판별하는데 있어서 그 문서를 표현하고 있는 특징을 어떻게 선택할 것인가와 얼마나 선택할 것인가가 미치는 영향을 실험을 통하여 측정하였다. 우리는 실험을 통하여 특징 선택 방법이 분류 성능에 미치는 영향을 알아보고자 하였고, 특징의 개수와 분류 성능과의 상관관계, 그리고 범주의 개수와 특징의 개수와의 관계를 규명하고자 하였다. 결과를 통하여 우리는 뉴스 그룹 문서의 경우 그 분포상황의 특이성에 기인하여 정보획득 방법이 가장 좋은 성능을 냄을 알 수 있었고, 문서의 특징의 개수에 따라 성능에 있어서 커다란 차이가 있음도 알게 되었다. 또한 정보획득 방법과 나이브 베이지안 분류방법을 이용했을 때 가장 좋은 성능을 도출하는 특징의 개수가 범주의 개수에 비례함을 알 수 있었다.

1. 서론

형태적인 면에서 하나의 문서는 그 문서에서 발생하는 단어의 유무, 단어의 종류, 단어의 위치, 단어의 빈도수 등의 특징(feature)으로 표현될 수 있다. 또한 이렇게 단어를 표현하는 방법에 있어서도 단어 구성면에서 단일어인지, 구문적 어구(syntactic phrase)인지, 시소러스(thesaurus) 인지에 따라 다르게 표현될 수 있다. 하지만 문서는 위와 같은 문서의 특징이 될 수 있는 요소들이 대단히 많다[1,2,3,4].

문서를 분류한다 함은 미리 정의되어 있는 여러 범주에 각각의 문서들을 할당하는 것이다. 하지만 문서의 수가 증가할수록 각각의 문서를 효과적으로 검색 및 색인화(indexing)하고, 내용 요약(summarization)과 같은 작업을 수행할 때 많은 어려움을 겪게 된다. 이를 해결하기 위해 각 문서들을 범주 별로 귀속시키는 작업을 수행하며, 휴리스틱(heuristic)을 이용하는 방법대신 컴퓨터를 이용하는 자동화된 기계학습 기술이 이용되었다. 대표적인 문서 분류 기법으로는 최근접 이웃분류(nearest neighbor classification), 베이지안 확률분류(Bayes probabilistic classification), 결정트리(decision trees), 신경망(neural networks), 결정 규칙(decision rules), 그리고 지지벡터기계(support vector machine)들이 있다[1,4]. 이러한 분류 알고리즘들은 문서의 특징을 선택하는 여러 방법들과 함께, 최근 활발한 연구가 진행중인 문서 분류를 위한 특징 선택 연구에 많이 적용되고 있다. 그러나 지금까지의 연구들은 특징 부분집합(feature subset)을 결정하기 위한 평가, 특징 선택 방법 별 단순 성능 비교, 특징 선택 시 구문적 어구를 선택함으로 인한 효과 등에 초점이 맞추어져 있었다. 이에, 본 연구에서는 위 연구들을 바탕으로 하여 특정한 분류 방법 또는 특징 선택 방법에 의존하지 않으면서도 좋은 분류 성능을 내는 개별적인 특징과 범주와의 상관관계를 밝히고자 하였다.

2. 관련연구

분류 알고리즘들을 실사회에 존재하는 거대 문서 데이터 집합에 적용하기 위해서는 반드시 문서를 표현하는데 이용되는 특징의 수를 축소시켜야 한다[4]. 이를 위해 George H. John, Ron Kohavi, Karl Pfleger은 [2]에서 문서의 하부 특징(feature subset) 집합을 구별하는 수단으로서 포장적 모델을 이용했다. 그리고 문서 집합과 하부 특징 집합과의 관련성을 비교하여 관련성이 높은 것들과 낮은 것, 약한 관련성을 갖는

특징을 추출한 후 ID3, C4.5를 이용한 분류 실험에서 특징들의 차원 축소(dimensionality reduction)가 성능향상에 미치는 효과를 입증하였다. 또한 Yiming Yang은 [4]에서 통계적인 기반의 문서 분류 학습 기법을 이용하여 로이터(Reuters) 데이터에 대해 여러 가지의 특징선택방법과 k-NN과 LLSF(Linear Least Squares Fit mapping) 알고리즘을 적용한 분류실험에서 원래 문서의 98%를 제거한 특징만을 가지고도 더욱 정확한 분류를 해낼 수 있음을 측정하였다. 그리고 Dunja Mladenic은 [1]에서 Yahoo사이트에 존재하는 계층적인 웹 문서들에 대해 여러 특징선택 방법들과 나이브 베이지안(Naï ve Bayesian) 분류 알고리즘을 이용하여 어떤 특징선택 방법이 좀 더 나은 성능을 내는지를 실험하였다.

3. 특징 선택 방법

특징을 선택하는 방법은 크게 세 가지로 나누어 생각해 볼 수 있다. 첫째는 (가)분류 방법에 의존하지 않고 특징을 선택하는 방법과 (나)의존하는 방법이고 [1,2,3], 둘째는 (다)전체의 특징들을 개별적으로 이용하는 방법과 (라)부분집합을 취하여 이용하는 방법, 그리고 셋째는 전체의 특징들을 2차원의 공간 상에 위치해 있다고 가정하고 위에서부터 하나씩 특징을 더해가며 최종의 특징 집합을 구하는 방법과 반대로 아래에서부터 전체의 특징 중에서 하나씩 제거해 가면서 적절한 특징만을 취하는 방법이며 약술하면 다음과 같다.

가) 분류방법에 비의존적인 방법

계획에 의존하지 않는 접근(Scheme-independent approach) 방식이다.하부 특징 선택은 선처리 작업으로 수행되며 추론 알고리즘을 작동하는데 있어서 선택된 하부 특징 집합의 영향을 고려하지 않는다는 단점을 지닌다. 각각의 특징들을 측정(evaluation)하는 작업은 특징 선택을 위해 문서를 학습하는 과정 동안에 이루어진다. 일반적으로 가장 많이 사용되는 특징 선택 방법들로는 문서 빈도수 임계값(Document Frequency thres-holding), 단어 빈도수 임계값(Term Frequency thresholding), 정보 획득(Information Gain), χ_2 통계치(CHI) 등이 있으며[1,3,4] 본 실험에서는 정보 획득 방법과 문서 빈도수 임계값, 그리고 단어 빈도수 임계값을 이용하는 방법이 사용되었다.

[식.1]에서 F는 단어 W를 표현하는 특징이며, P(W)는 단어 W가 발생하는 확률, $\bar{W}$ 는 단어 W가 발생하지 않은 확률을 나타낸다. $P(C_i)$ 는 i번째 범주일 확률이고 $P(C_i|W)$ 는 단어 W가 발생했을 때 i번째 범주에 속하는 조건적인 확률이다.

정보획득 방법은 문서에 나타나는 단어의 자유도(Entropy)의 변화에 따르는 차를 측정하여 그 문서가 어느 범주에 속하는 지를 알아내는 방법이고 χ_2 통계치는 특정 단어와 특정 범주와의 관련성을 측정한 평균값을, 자유도를 갖는 χ_2 분포값과 비교하여 측정한다.

나) 분류방법에 의존적인 방법

모델생성 시 추론 알고리즘을 이용하여 질적으로 우수한(good) 하부 집합을 탐색하는 과정을 수행한다. 두 가지 과정이 존재하는데 하나는 하부 집합을 평가하는 과정이고, 다른 하나는 평가 결과에 의해서 하부집합 공간을 탐색하는 과정이다.

♦TFreq(F) = TF(W)

♦DFreq(F) = DF(W)

♦InfoGain(F)=

$$P(W)\sum_i P(Ci|W)\log\frac{P(Ci|W)}{P(Ci)} + P(\bar{W})\sum_i P(Ci|\bar{W})\log\frac{P(Ci|\bar{W})}{P(Ci)}$$

♦ $\chi^2 avg(F) = \sum_{i=1}^{m} \Pr(Ci)\chi^2(F,ci)$ $\chi^2 \max(F) = \max_{i-1}^{m}\{\chi^2(F,ci)\}$

식. 1 특징 선택 알고리즘

다) 개별적인 특징을 취하는 방법

문서의 경우와 같이 선택 가능한 특징의 수가 많을 때 사용하기에 적합한 방법이며, 각각의 요소들을 취하여 비교, 평가하여 특징으로 삼는다.

라) 부분집합을 취하여 이용하는 방법

일반 데이터 집합의 경우에는 고려할 수 있는 특징의 개수가 작기 때문에 이용가능하지만 문서집합의 경우는 불가능하다.

4. 실험

4.1 실험 목표와 방법

본 논문에서의 실험 목표는 다음과 같다.

- 특징 선택 방법이 어떻게 문서의 분류 성능에 영향을 미치는가?
- 특징의 개수가 분류 성능 방법에 영향을 미치는가?
- 분류 성능에 있어서 클래스의 개수와 특징의 개수와의 상관관계가 존재하는가

본 실험에서는 특징 선택을 위해 3절에서 언급한 (가),(다)의 방법을 이용하였으며 그 중에서도 정보획득, 문서빈도수, 단어빈도수를 사용하였다. 문서 데이터 집합은 20개의 유즈넷(Usenet) 뉴스 그룹 문서이며, 각 범주마다 1,000여 개의 뉴스 문서를 포함하므로, 전체 문서 수는 20,000여 개이며, 특징이 될 수 있는 전체 단어의 개수는 약 97,000개이다. 실험을 위해 10-fold 교차 검증 방법을 사용하였으며, 검증 과정을 각 경우별로 1회 시행하였다. 우리는 Dual Pentium II 300Mhz CPU, 256M Memory, Linux(Kernel version 2.2.14) OS 사양의 서버를 사용하였고, 10Mbps의 네트워크 환경의 클라이언트에서 실험하였다. 또한 k-NN 분류 실험의 경우 k값을 10으로 사용하였고, 모든 경우에 대하여 분류시간 문제로 인해 측정시간 단축을 위해 훈련문서 집합의 범주 당 10개의 문서만을 추출하여 검증과정에 이용하였다.

각 범주별 1,000개의 문서를 가진 문서 집합은 분류방

법에 의존하지 않는 방법과 개별적인 특징을 이용하는 접근 방식을 이용하여 인덱싱(indexing)된다. 인덱싱된 결과는 분류 방법을 이용하여 분류되며, 각각의 경우에 시간과 정확도를 측정하였다. k-NN, 나이브베이지안, TF/IDF 각각의 경우에 특징선택 방법인 DC(Document Counts), IG(Information Gain), OC(word Occur Counts)의 특징개수를 10, 50, 100, 1000, 10,000 인 경우에 각각 측정하였다. 또한 범주의 수를 임의의 10개로 줄인 후 같은 실험을 반복하였고, 포장적 접근방식을 사용하지 않은 기본 모델만을 가지고 분류실험을 진행하기도 하였다.

4.2 실험 결과

지금까지의 모든 실험 및 [Fig.1-3]을 종합하여 다음과 같은 결과를 확인할 수가 있었다.

(가) 분류시간 및 정확도는 NB, TF/IDF 분류기법을 사용할 때 IG의 경우에는 문서에 대한 특징의 개수가 50개, k-NN, IG를 의 경우에는 35개일 경우 가장 적은 시간이 소요되면서 가장 좋은 성능을 기록하였다.

(나) DC 나 OC 의 경우에는 NB, TF/IDF 분류기법을 사용하였을 때, 특징의 개수가 1,000개일 경우까지는 지속적으로 증가하나 그 이상의 경우에는 비슷한 성능을 유지함을 알 수 있으며, k-NN의 경우에는 전체적으로 큰 변화 없이 동일한 수행성능을 나타냈다.

(다) 계획에 의거하여 생성한 문서모델을 가지고 실험한 분류성능이 기본문서모델을 가지고 실험한 것보다 근소하게 우수한 성능을 나타냈다.

(라) 주어진 범주의 수를 절반으로 줄인 후 실험을 했을 때 NB, IG를 사용한 경우, 이전의 경우(가)와 비례하여 특징의 개수가 25개일 때 최고의 정확도 및 빠른 분류시간을 나타냈으며 DC의 경우에는 이전의 경우와 비슷한 결과가 도출되었다.

5. 결론

본 논문에서는 특징의 개수와 여러 가지의 특징선택 방법, 및 분류방법과의 상관성에 대하여 실험하고 그 결과를 기술하였다. 이를 살펴보면 최근 가장 많이 사용하는 특징선택 방법인 정보획득 방법을 사용하였을 경우, 분류방법에 의존한 모델을 가지고 실험했음에도 불구하고 특징의 개수가 범주의 개수에 비례하여 수렴함을 알 수 있다. 즉, 20개의 범주(약 20,000개의 문서)일때 50개, 10개의 범주(약 10,000개의 문서)일 경우 25개일 때 가장 좋은 성능을 나타낸 것이다. 또한 문서빈도수, 단어빈도수를 가지고 실험을 했을 경우에는 식 자체의 확률적인 특성으로 인하여 역시 특징의 개수가 많을수록 더 우수한 정확도를 보임을 알 수 있었다. 다만 이 경우에도 본 실험에서 사용되는 총 특징수가 97,000여개 임에 반해, 특징수가 1,000개를 넘어가게 되면 정확도 측면에서 큰 차이가 없어, 도리어 시간문제를 고려할 때 전체적인 성능이 떨어짐을 알 수 있다. 향후 연구로는 다른 문서 데이터 집합에 대하여 실험한 후 결과를 종합하여, 본 실험에서의 측정결과와의 차이를 확인하고 좀 더 구체적인 상관관계를 밝히고자 한다.

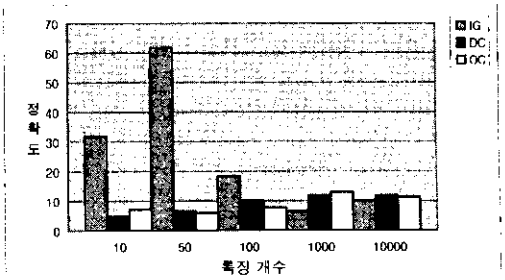

Fig. 1 k-NN분류기를 이용한 정확도 비교

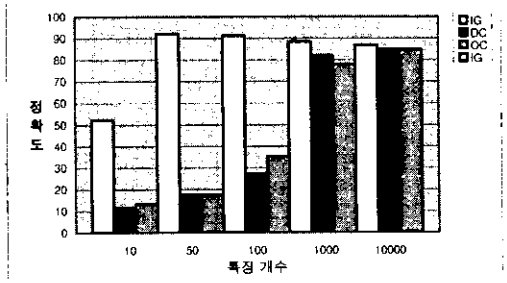

Fig. 2 나이브베이지안 분류기를 이용한 정확도 비교

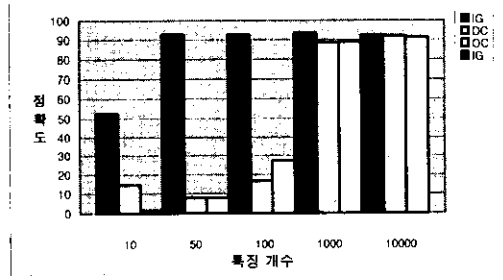

Fig. 3 TF/IDF분류기를 이용한 정확도 비교

6. 참고 문헌

[1]. Dunja Mladenic, Marko Grobelnik, " Feature selection for classification based on text hierarchy" , *Proc. of the Workshop on Learning from Text and the Web*, Pittsburgh, USA, 1998

[2]. George H. John, Ron Kohavi, Karl Rfleger, " Irrelevant Features and the Subset Selection Problem" , *Proc. of ICML94*, 121-129, Morgan Kaufmann Publishers, San Francisco, CA, 1994

[3]. Ian H. Witten and Eibe Frank, Data Mining, Morgan Kaufmann Publishers, Inc., 2000

[4]. Yiming Yang, Jan O. Pedersen, " A Comparative Study on Feature Selection in Text Categorization" , *Proc. of ICML97*, pp. 412-420, 1997