자질선정에 따른 Naive Bayesian 분류기의 성능 비교

Performance Evaluation of a Naive Bayesian Classifier using various Feature Selection Methods

국민상, 정영미
연세대학교 문헌정보학과

Kook Min-Sang and Chung Young-Mee
Department of Library and Information Science, Yonsei University

베이즈 확률을 이용한 분류기는 자동분류 초기부터 사용되어 아직까지 이 분야에서 가장 많이 사용되는 분류기 중 하나이다. 본 논문에서는 KTSET 문서에서 임의로 추출한 198건의 정보과학회 관련 논문의 제목 및 초록을 대상으로 베이즈 확률을 이용한 문서의 자동분류 실험을 수행하였으며, 더불어 Naive Bayesian 분류기에 가장 적합한 자질선정 방법을 찾고자 카이제곱 통계량, 상호정보량 및 기대상호정보량, 정보획득량, 역문헌빈도, 역카테고리빈도 등 6가지의 자질선정 기준을 실험하였다. 실험 결과는 카이제곱 통계량을 이용한 분류 실험의 성능이 가장 좋았고, 기대상호정보량과 정보획득량, 역카테고리빈도 또한 자질수에 큰 영향을 받지 않고 비교적 안정적인 성능을 보였다.

1. 서론

문서의 자동분류(Automatic Classification)란 분류작업 이전에 분류체계가 만들어져 있는 상태에서 각 문서를 가장 적합한 클래스에 배정함으로써 문서들을 집단화하는 것으로, 현재는 텍스트 범주화(text categorization)란 용어로 더 많이 사용되고 있다.

현재까지 많은 통계적 분류방법과 기계학습 방법이 텍스트 범주화에 적용되어 왔지만 그 중 베이즈 확률을 이용한 텍스트 범주화는 문서의 자동분류 초기부터 사용되어 아직도 가장 많이 사용되는 방법 중 하나이다.

Naive Bayesian 분류기(이하 NB)는 간단하면서도 매우 실용적인 분류기 중 하나이다. NB는 범주가 발생하는 사전확률을 기반으로 하여 범주 C_i가 문서에 할당될 확률과, 특정 단어가 문서에서 발생할 조건확률을 측정한다. 계산을 간단히 하기 위해 단어 발생 확률은 서로 독립적이라 가정한다.

텍스트 범주화에 있어서 학습집단에서 사용되는 자질을 선정하는 것은 매우 중요한 부분이다. 특히 베이즈 확률을 사용하는 분류기에서는 더욱 그러하다. 보통의 데이터 집단에서 자질공간은 수십에서 수백만개의 단어로 이루어져 있다. 예를 들어 베이지안 네트워크는 독립성 가설이 전제되지 않는다면 계산 자체가 어려워질 것이다. 따라서 자동분류에 있어 자

질공간의 축소는 계산시간 복잡도만의 문제가 아니라 분류기 자체의 성능에도 영향을 미치는 것이다.

자질공간의 축소는 정확도(accuracy)를 감소시키지 않는 범위 내에서 이루어져야 한다. 또한 수작업의 개입 없이 자동적으로 이루어져야 한다.

이에 따라 본 논문에서는 확률적 접근방법을 사용한 Naive Bayesian 분류기에서 어떤 자질선정 방법을 사용하는 것이 보다 더 나은 성능을 보이는가를 살펴보도록 한다. 본 실험에서는 카이제곱 통계량(CHI), 상호정보량(MI) 및 기대상호정보량(EMI), 정보획득량(IG), 역문헌빈도(IDF), 역카테고리빈도(ICF) 등 6가지의 자질선정 기준을 사용하였다.

2. Naive Bayesian 분류기

전술했다시피 NB 분류기는 문서의 분류를 위해 베이즈 이론을 사용한다.

사건 E와 C_i가 있을 때, E가 주어졌을 때 C_i가 발생할 확률은 다음과 같다.

$$P(C_i \mid E) = \frac{P(E \mid C_i) \times P(C_i)}{P(E)}$$

이를 문서분류에 적용해서, 분류하려는 문서에 단어 $W_1, W_2, W_3, \ldots, W_n$이 출현한 경우를 사건 E, 문서가 범주 c_i에 분류되는 것을 사건 C_i라고 하면, 이 문서가 범주 C_i에 분류될 확률은 다음과 같다.

$$P(C_j \mid W_1, \ldots, W_n) = \frac{P(W_1, \ldots, W_n \mid C_j) P(C_j)}{P(W_1, \ldots, W_n)}$$

NB에서는 한 문서가 특정 카테고리에 포함될 때 그 문서에서 나타나는 단어의 출현은 독립적이라고 가정하고, 범주의 할당도 상호 배타적이라고 가정한다. 이러한 가정 하에 위의 공식을 다시 표현하면

$$P(C_j \mid W_1, \ldots, W_n) = P(C_j) \times \prod_i \frac{P(W_i \mid C_j)}{P(W_i)}$$

가 된다. 마론은 이 공식을 다음과 같이 간단한 형식으로 표현하였다.

$$P(C_j \mid D_m) = k \times P(C_j) \times \prod_i P(W_i \mid C_j)$$

이때 $P(C_j \mid D_m)$는 문서 D가 범주 C에 분류될 확률이고, k는 $\sum_j P(C_j \mid D_m) = 1$, 즉 한 문서가 전체 범주에 속할 각 확률을 더한 값이 1이 되도록 하는 비례상수이다.

$P(C_j)$와 $P(W_i \mid C_j)$는 실험집단으로부터 다음과 같이 계산할 수 있다.

$$P(C_j) = \frac{C_j\text{에 할당된 학습문서수}}{\text{모든 학습문서수}}$$

$$P(W_i \mid C_j) = \frac{C_j\text{에 할당된 문서에서 } W_i\text{가 발생한 횟수}}{C_j\text{에 할당된 문서에 나타난 모든 단어의 발생 횟수}}$$

3. 자질선정 기준

3.1 카이제곱 통계량 (χ^2 statistic : CHI)

카이제곱 통계량은 용어 t와 범주 c간의 의존성을 측정하는 것이다. t와 c에 대한 2×2 분할표에서 A가 t와 c가 동시에 발생한 횟수, B는 t는 발생했지만 c는 발생하지 않은 횟수, C는 t는 발생하지 않고 c만 발생한 횟수, D는 t와 c 모두 발생하지 않은 횟수라 하고, N이 총 문서수라 할 때, 카이제곱 통계량은 다음과 같이 구해진다[2].

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

자질의 범주구분능력은 주어진 자질에 대한 범주의 카이제곱 통계량의 최대값으로 설정된다.

$$\chi^2_{max}(t) = MAX_{i=1}^{m} \{\chi^2(t, c_i)\}.$$

3.2 상호정보량 (Mutual Information : MI)

상호정보량은 단어 연관성에 대한 통계적 언어 모델링에서 주로 사용되는 기준이다. 용어 t와 범주 c에 대한 2×2 분할표에서 A가 t와 c

가 동시에 발생한 횟수, B는 t는 발생했지만 c는 발생하지 않은 횟수, C는 t는 발생하지 않고 c만 발생한 횟수라 하고, N이 총 문서수라 할 때, 상호정보량은 다음과 같이 구해진다[2].

$$I(t,c) = \log P(t|c) - \log P(t) \approx \log \frac{(A \times N)}{(A+C) \times (A+B)}$$

자질의 범주구분능력은 주어진 자질에 대한 범주의 상호정보량의 최대값으로 설정된다.

$$I_{\max}(t) = MAX_{i=1}^{m} \{I(t,c_i)\}$$

하지만 상호정보량의 최대값은 범주의 분포를 고려하지 않기 때문에, 모든 범주를 대상으로 하기 위해 상호정보량에 표준편차와 변이계수를 사용할 수 있다. 아래 식에서 I_{SD}는 상호정보량의 표준편차, I_{AVG}는 범주에 대한 상호정보량의 평균, C는 범주수이다.

$$I_{SD} = \sqrt{\frac{\sum (I - I_{AVG})^2}{C}} \quad , \quad I_{CV} = \frac{I_{SD}}{I_{AVG}}$$

3.3 기대상호정보량(Expected MI : EMI)

상호정보량이 사건의 발생정보만 이용하는 자질의 적합성 척도인 반면 기대상호정보량은 발생하지 않은 사건에 관한 정보까지도 고려하는 척도이다. 자질의 범주구분 능력은 주어진 자질에 대한 범주의 기대상호정보량의 최대값으로 설정된다.

$$EMI(t,C_j) = \sum_{i=0}^{1} \sum_{k=0}^{1} P(t=i, C_j=k) MI(t=i; C_j=k)$$

$$EMI_{\max}(t) = MAX_{i=1}^{m} \{EMI(t,c_i)\}$$

3.4 정보획득량 (Information Gain : IG)

정보획득량은 한 문서에서 어떤 용어의 존재유무를 앎으로써 범주 예측을 위해 얻어진 정보의 비트수를 구하는 것이다. $\{c_i\}_{i=1}^{m}$가 범주 집합을 나타낸다고 하면 용어 t에 대한 정보획득량은 다음과 같이 계산된다[2].

$$G(t) = -\sum P(c_i) \log P(c_i) + P(t) \sum P(c_i|t) \log P(c_i|t) + P(\bar{t}) \sum P(c_i|\bar{t}) \log P(c_i|\bar{t})$$

3.5 역문헌빈도 (IDF)

역문헌빈도란 문서간의 분리도가 높은 단어에 높은 가중치를 주는 것으로, 보통 이 역문헌빈도에 용어빈도를 곱하여 자질에 가중치를 부여한다. 자질 W_i의 문서 j에서의 빈도수를 $freq_{ij}$, 총 문서의 개수를 N, 단어 W_i를 포함하는 문서의 개수를 DF_i라고 할 때 자질의 가중치는 다음과 같이 계산된다.

$$W_i = freq_{ij} \times IDF_i = freq_{ij} \times (\log(N) - \log(DF_i) + 1)$$

3.6 역카테고리빈도 (ICF)

역카테고리빈도란 문서의 분류를 위해 범주의 분리 능력이 우수한 색인어에 높은 가중치를 주는 방법이다. 자질 W_i의 범주 j에서의 빈도수를 $freq_{ij}$, 총 범주의 개수를 M, 단어 W_i를 포함하는 범주의 개수를 CF_i라고 할 때 자질의 가중치는 다음과 같이 계산된다[3].

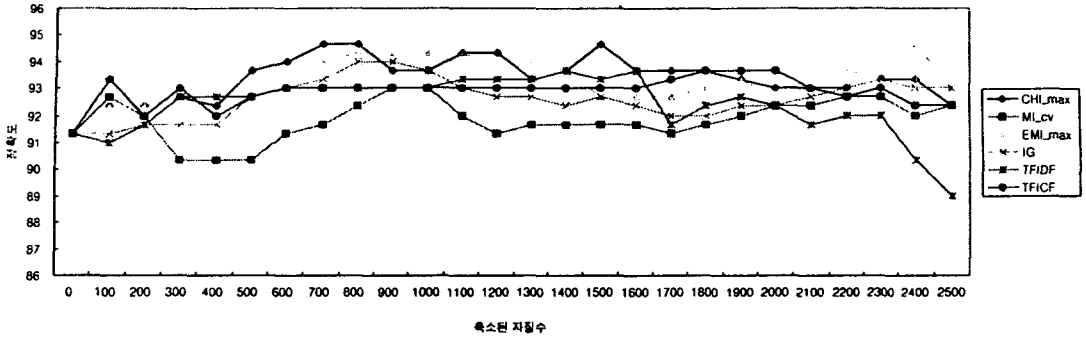
$$W_i = freq_{ij} \times ICF_i = freq_{ij} \times (\log(M) - \log(CF_i) + 1)$$

4. 실험문헌 집단

본 실험에서는 KTSET의 전체 1000개의 문서에서 정보과학회 관련 논문(884건) 중 한글초록이 있는 문서(880건)를 대상으로 하여 10개 이상의 문서를 갖는 범주와 그 범주 내에 있는 문서를 실험집단(198건)으로 설정하고, 이러한 198개의 실험집단 중 각 범주마다 임의로 75%의 문서를 추출하여 학습집단을 구성하고(148건), 나머지 25%(50건)를 검증집단으로 하였다. 여러 개의 분류기호를 갖는 문서의 경우에는 KTSET 상에서 첫 번째 범주만을 대상으로 하였다.

문서에서의 색인어 추출에는 한성대학교에서 개발한 형태소 분석기 HAM을 사용하였으며, 추출된 전체 자질수는 2542개였다.

실험의 성능평가 척도로는 정확도(accuracy)를 사용하였는데, 범주 C_j에 대한 2×2 분할



<그림 1> 각 자질선정 기준에서의 자질수의 변화에 따른 정확도

표에서 A, B, C 및 D가 다음과 같을 때 정확도는 아래 공식과 같이 구해진다.

A : 시스템이 C_i에 올바르게 할당한 문서수
B : 시스템이 C_i에 잘못 할당한 문서수
C : C_i에 할당되어야 하는 문서 중 할당되지 않은 문서수
D : C_i에 할당되지 않은 것이 옳은 문서수

$$Accuracy = \frac{A+D}{A+B+C+D}$$

5. 실험 결과

그림에서 볼 수 있듯이 전반적으로 카이제곱 통계량의 성능이 가장 좋았고, 기대상호정보량과 정보획득량, 역문헌빈도, 역카테고리빈도는 비슷한 수준을 보였다. 특히 카이제곱 통계량과 기대상호정보량, 정보획득량, 역카테고리빈도의 경우 전체 단어의 98%(전체 자질수 2542개 중 2500)까지 제거해도 꾸준히 안정적인 성능을 보였다. 변이계수를 사용한 상호정보량은 최대값을 사용할때보다 성능이 나아지긴 했지만, 여전히 5가지의 자질선정 기준 중 가장 낮은 성능을 보였다.

위의 결과로 볼 때 NB와 같은 확률을 사용한 분류기는 카이제곱 통계량이나 기대상호정보량, 정보획득량과 같이 모든 범주의 빈도 분포를 고려하고, 출현하지 않은 단어나 범주 또한 고려하는 자질선정 방법이 좋은 성능을 나타냈다고 할 수 있다.

6. 결론

점차 온라인상에서 텍스트 문서가 많아지고 인간의 수작업으로 모든 문서를 처리할 수 없게 된 현 상황에서 자동분류, 즉 텍스트 범주화는 여전히 중요한 위치를 차지하고 있고, 또한 그에 따라 자질선정이나 자질추출, 자질생성은 텍스트 범주화 작업에서 더욱 커다란 역할을 하게 되었다. 앞으로도 여러 텍스트 범주화 방법과 알고리즘에 적합한 자질선정 방법을 찾는 연구가 꾸준히 계속되어야 할 것이다.

참 고 문 헌

[1] David D. Lewis. Representation and Learning in Information Retrieval, PhD thesis, Department of Computer Science, Univ. of Massachusetts, 1992.

[2] Yiming Yang and Jan O. Pederson, "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 1997.

[3] 조광제, 김준태, "역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류", 한국정보과학회 학술발표논문집, 1997.