Naïve Bayes 문서 분류기를 위한 점진적 학습 모델 연구

김제욱, 김한준, 이상구
서울대학교 컴퓨터공학부

{jerry, hjkim, sglee}@europa.snu.ac.kr

A Study on Incremental Learning Model for Naïve Bayes Text Classifier

Je-uk Kim, Han-joon Kim, Sang-goo Lee
School of Computer Science and Engineering, Seoul National University

요약

본 논문에서는 Naïve Bayes 문서 분류기를 위한 새로운 학습모델을 제안한다. 이 모델에서는 라벨이 없는 문서들의 집합으로부터 선택한 적은 수의 학습 문서들을 이용하여 문서 분류기를 재학습한다. 본 논문에서는 이러한 학습 방법을 따를 경우 작은 비용으로도 문서 분류기의 정확도가 크게 향상될 수 있다는 사실을 보인다. 이와 같이, 알고리즘을 통해 라벨이 없는 문서들의 집합으로부터 정보량이 큰 문서를 선택한 후, 전문가가 이 문서에 라벨을 부여하는 방식으로 학습문서를 결정하는 것을 selective sampling이라 한다. 본 논문에서는 이러한 selective sampling 문제를 Naïve Bayes 문서 분류기에 적용한다. 제안한 학습 방법에서는 라벨이 없는 문서들의 집합으로부터 재학습 문서를 선택하는 기준 측정치로서 평균절대편차(Mean Absolute Deviation), 엔트로피 측정치를 사용한다. 실험을 통해서 제안한 학습 방법이 기존의 방법인 신뢰도(Confidence measure)를 이용한 학습 방법보다 Naïve Bayes 문서 분류기의 성능을 더 많이 향상시킨다는 사실을 보인다.

Keywords: Naïve Bayes 문서 분류기, Selective sampling, uncertainty, incremental learning

1. 서론

문서의 분류는 클래스의 소속이 알려지지 않은 문서를 미리 정해진 클래스로 할당하는 것을 말한다. 문서 분류기는 문서를 입력으로 받아 미리 학습된 정보를 이용하여 문서를 자동으로 분류한다. 문서 분류기의 성능은 학습에 의해 크게 좌우된다. 일반적으로 문서 분류기를 학습하기 위해, 라벨이 있는 학습문서를 얻는 것은 비용이 크다. 여기서 라벨이 있는 문서란, 사람에 의해 소속 클래스가 밝혀진 문서를 의미한다. 라벨이 있는 학습 문서를 얻기 위해서는 사람이 많은 시간을 들여 문서의 라벨을 부여해

야 한다. 따라서 적은 수의 문서만을 선택하여 적절한 학습의 효과를 얻는 것은 여러 학습 방법의 중요 목표가 되었다. 임의로 문서를 선택하여 라벨을 부여하고 이를 학습 문서로 채택하기 보다는 정보량이 큰 문서만을 골라서 학습문서로 채택한다면, 보다 적은 수의 학습 문서로도 적절한 학습의 효과를 낼 수 있을 것이다.

보통 많은 양의 라벨이 없는 문서들은 쉽게 얻을 수 있다. Selective sampling 학습 방법에서는 이들 라벨이 없는 문서들의 집합으로부터 정보량이 큰 문서를 선택하여, 이 문서의 라벨을 부여하는 방식으로 학습을 진행한다. 이 때 전문가가 라벨을 부여해야 하는 문서의 수가 크게 줄어들기 때문에 학습 비용을 크게 줄일 수 있다.

정보량이 큰 문서를 선택하는 방법으로서 여기서는 uncertainty 개념을 이용한다. 이 방법에서는 현재의 문서 분류기가 불명확하게 분류하는 문서를 정보량이 큰 문서라고 간주한다. 본 논문에서는 문서 분류에 있어서의 uncertainty 개념을 일반적으로 정의하고, 이것을 수치로 나타낸 측정치인 평균 절대편차와 엔트로피 측정치를 제안한다. 그리고 이 측정치를 이용하여 학습한 문서 분류기 성능과 기존의 방법인 confidence 측정치를 사용한 문서 분류기의 성능을 비교해 본다. 이러한 학습 방법을 적용할 문서 분류기로 이 논문에서는 Naïve Bayes 문서 분류기를 사용한다. Naïve Bayes 문서 분류기는 해당 문서가 클래스에 속할 확률을 이용하여 문서를 분류하는데, 본 논문에서는 이러한 특성을 이용하여 uncertainty를 측정한다.

본문의 구성은 다음과 같다. 먼저 2절에서는 문서 분류와 학습을 위한 배경 이론에 대하여 알아본다. 3절에서는 uncertainty의 정의를 내려보고 이 정의를 바탕으로 새로운 측정치들을 제안한다. 4절에서는 uncertainty 측정치를 이용한 점진적 학습 방법을 Naïve Bayes 문서 분류기에 적용해본다. 5절에서는 제안한 학습방법을 기존의 방법과 비교하기 위한 실험을 제시한다. 끝으로 6절에서는 결론을 내린다.

1. D ← $\{\langle x_1, f(x_1)\rangle, \ldots, \langle x_n, f(x_n)\rangle\}$
2. $h \leftarrow L(D)$
3. While stop-condition is not satisfied do:
   a) Apply S_L and get the next example,
      $x \leftarrow S_L(X, D)$
   b) Ask the teacher to label x, $w \leftarrow f(x)$
   c) Update the labeled examples set,
      $D \leftarrow D \cup \{\langle x, w\rangle\}$
   d) Update the classifier, $h \leftarrow L(D)$
4. Return classifier h

<그림1> Selective Sampling 을 이용한 학습 알고리즘

2. 배경 이론

2.1 Selective Sampling

Selective sampling에서는 라벨이 없는 문서의 클래스 소속값(membership value)을 현재의 문서 분류기에게 질의하는 방식으로 해당 문서의 정보량을 측정한다. 이러한 방법은 기존의 여러 학습 알고리즘에서 제시되었다. [9]에서는 Nearest Neighbor 문서 분류기를 위한 학습 방법을 제시하였고, [11]은 Query by committee 알고리즘을 이용한 selective sampling 학습 방법을 제안하였다.

<그림1>은 selective sampling 학습 방법을 알고리즘화 한 것이다.[9] X를 라벨이 없는 문서들의 집합이라 가정하자. 그리고 f는 전문가라고 가정하자. 즉, f는 문서의 라벨을 부여하는 함수라 할 수 있다. (f: x → $\{C_1, C_2, ..., C_k\}$, k는 클래스의 개수) D는 X의 부분 집합으로서, 전문가에 의해 라벨이 부여된 문서들의 집합이다. D의 원소인, $\langle x_i, f(x_i)\rangle$는 x_i문서의 라벨이 전문가에 의해 $f(x_i)$로 할당되었음을 의미한다. S_L은 문서 집합 X로부터 정보량이 가장 큰 문서 x를 선택하는 함수이다. 이 함수는 selective sampling 알고리즘의 핵심이라 할 수 있다. L은 라벨이 있는 문서들로부터 문서 분류기를 학습하는 알고리즘이다. 이는 문서 분류기를 구현하는 모델에 따라 달라질 수 있다. h는 D로부터 만들어진 문서 분류기의 함수이다. f와 마찬가지로 h도 특정 문서를 클래스와 관련시키는 함수이다. (h: x → $\{C_1, C_2, ..., C_k\}$)

Selective sampling 알고리즘은 우선, 초기 문서 분류기를 만드는 것으로 시작한다.(line 1) 그 후, 종결조건(stop condition)이 만족될 때까지 점진적으로 학습을 진행한다.(line 3) 라벨이 없는 문서의 집합으로부터 정보량이 큰 문서를 선택한 다음(line a), 이 문서에 라벨을 부여한다.(line b) 이 문서를 D에 추가하고(line c), 현재의 문서 분류기를 재학습한다.(line d) 종결 조건이 만족되면 최종 문서 분류기가 만들어지면서 알고리즘은 종료된다.(line 4) 여기서 종결조건은 전문가가 문서에 라벨을 부여하는 것을 그만두는 시점을 의미한다.

2.2 Naïve Bayes 문서 분류기

Naïve Bayes 문서 분류기는 하나의 문서가 입력으로 받아 그것이 각 클래스에 할당될 수 있는 확률을 계산하는 방법으로 문서를 분류한다. 특정 문서가 특정 클래스에 속하는 확률을 계산하기 위하여 다음과 같은 Bayesian 이론을 사용한다.

$$P(c_j \mid d) = \frac{P(C_j)P(d \mid C_j)}{P(d)} \quad (2.1)$$

여기에서, d는 임의의 문서를 의미하고, C_j는 임의의 클래스를 의미한다. P(d)는 모든 클래스에 대하여 같은 값을 가지므로 확률을 계산하는데 있어 고려하지 않아도 된다. 따라서 $P(C_j)$와 $P(d \mid C_j)$만 알면 d가 C_j에 할당될 확률을 계산할 수 있다. $P(C_j)$는 모든 학습 문서들의 수(n)와 C_j 클래스에 속하는 학습 문서들의 수(n_{c_j})의 비율로 구할 수 있다. 따라서 다음과 같은 식이 성립한다.

$$p(C_j) = \frac{n_{c_j}}{n} \quad (2.2)$$

Naïve Bayes 문서 분류기는 $P(d \mid C_j)$의 계산을 좀 더 쉽게 하기 위해, 문서 내에 존재하는 모든 단어들은 서로 독립적이라고 가정한다. 이러한 가정을 따르면 $P(d \mid C_j)$는 다음과 같은 식으로 변형이 될 수 있다.

$$P(d \mid C_j) = \prod_{w \in d} P(w \mid C_j) \quad (2.3)$$

f를 C_j 클래스에 출현하는 모든 단어들의 빈도수의 합이라 하고, f_k를 C_j 클래스에

출현하는 w_k 단어의 빈도수라 할 때, $P(w_k|C_j)$ 의 최대 추정치는 $\frac{f_k}{f}$이라 할 수 있다. 그러나 이 추정치를 확률식에 그대로 적용하면, (2.1) 전체 식의 값을 0으로 만들 확률이 높다. 왜냐하면, 특정 문서에 존재하는 단어가 확률을 계산하려는 클래스 내에 존재하지 않을 수도 있기 때문이다. 이러한 문제를 해결하기 위해서 일반적으로 다음과 같이 m-estimate 개념을 응용한 기법을 이용한다.[10]

$$P(w_k|C_j) = \frac{n_k + 1}{n + |Vocabulary|} \quad (2.4)$$

여기서 |Vocabulary|는 모든 학습문서 내에 포함되어 있는 서로 다른 단어들의 수이다.

3. 문서 분류에 있어서의 Uncertainty의 정의와 측정

위에서 selective sampling 알고리즘을 소개하였다. 위 알고리즘에서 S_L 함수는 문서 집합에서 정보량이 가장 큰 문서를 선택하는 함수라고 정의하였다. [1], [2], [8], [11] 등의 연구에서는 이 함수를 uncertainty 개념을 이용하여 구현하였다. 여기서는 uncertainty를 일반적으로 정의해 보고, 이 정의를 만족하는 새로운 측정치 두개를 제안한다.

3.1 Uncertainty의 정의

Uncertainty란 특정 문서가 분류기에 의해 분류될 경우, 불명확하게 분류되는 정도를 말한다. 즉, uncertainty가 클수록 해당 문서를 문서 분류기가 클래스로 분류하는 확신이 작은 것이다. 일반적으로 uncertainty는 문서의 클래스를 예측하고, 이 예측에 대한 확신을 수치로 나타낼 수 있는 문서 분류기에서는 모두 정의가 가능하다.

예를 들어, 문서와 클래스 간의 거리를 이용하여 문서를 분류하는 문서 분류기를 생각해보자. 문서 d와 클래스 C_1, C_2, ..., C_k 를 가정하자. d와 가장 가까운 클래스를 C_i라 하고, 두 번째로 가까운 클래스를 C_j라 하자. 또한 dist(d, C_i)를 d와 C_i 간의 거리, dist(d, C_j)를 d와 C_i간의 거리라고 가정하자. 이때, dist(d, C_i)와 dist(d, C_j)의 차이가 작을수록 현재 문서 분류기가 문서 분류에 대한 확신을 작게 가진다고 할 수 있으므로 uncertainty가 크고, 차이가 클수록 uncertainty가 작다.

3.2 기존 방법에서의 Uncertainty 측정치

3.2.1 신뢰도 측정치

[8]에서는 현재의 문서 분류기가 분류 결과에 대한 확신을 얼마나 갖고 있나를 수치로 표현하기위해 신뢰도 측정치를 제안하였다. 문서 d가 클래스 C_i로 할당되는 경우, 신뢰도(confidence)는 다음과 같이 정의된다.

$$\text{confidence}(d, C_i) = \frac{\text{sim}(d, C_i) - \text{sim}(d, C_j)}{\text{sim}(d, C_i)} \quad (3.1)$$

여기에서, C_i는 문서 d와 가장 가까운 클래스이고, C_j는 문서 d와 두 번째로 가까운

클래스이다. 그리고 sim(d, C_i)는 문서 d와 클래스 C_i 간의 similarity를 말한다. sim(d, C_i) 대신에, 문서 d가 클래스 C_i 에 속할 확률을 사용해도 위에서 정의한 신뢰도 측정치와 같은 효과를 얻을 수 있다.

신뢰도를 이용하여 uncertainty를 측정할 수 있다. 신뢰도가 크다는 것은 현재 문서 분류기가 분류 결과에 대한 확신을 크게 갖고 있다는 의미이므로, uncertainty가 작다는 의미로 해석할 수 있다. 그 반대의 경우는 uncertainty가 크다는 것을 의미한다. 신뢰도가 클수록 현재 문서 분류기의 확신은 큰 것이고, 따라서 문서를 정확하게 분류할 가능성도 커지는 것이다.[8]

3.3 새로운 Uncertainty의 측정치 제안

3.3.1 평균절대편차(MAD) 측정치

3.1절에서 uncertainty에 대한 정의를 내렸었는데, 여기서는 이 정의를 바탕으로 새로운 측정치를 제시해 본다.

어떤 문서 분류기이든지 간에, 문서와 클래스간의 관련성에 대한 수치나 확률을 계산하는 방법으로 분류를 할 것이다. 이러한 수치나 확률의 분포가 골고루 퍼져있냐 아니면 한쪽으로 치우쳐 있냐는 uncertainty의 중요한 단서가 된다. 이 개념을 바탕으로 다음과 같이 MAD를 이용한 uncertainty 측정치인 $U_{MAD}(d)$를 제안한다.

$$U_{MAD}(d) = \frac{1}{|C|}\sum_{i=1}^{|C|}(p_{c_i} - \mu) \qquad (3.2)$$

여기서, P는 P={ p_{c_1} , ... , p_{c_2} , $p_{c_{|C|}}$ }로 정의되며 이는 문서 d가 각 클래스 C={C_1, C_2, ..., $C_{|C|}$} 에 속할 확률을 의미한다. 또한 μ는 각 확률들의 평균으로서 다음과 같이 정의된다.

$$\mu = \frac{1}{|C|}\sum_{i=1}^{|C|} p_{c_i} \qquad (3.3)$$

MAD는 각 확률이 평균으로부터 떨어진 평균거리를 말한다. 따라서 MAD가 작을수록 확률이 고르게 분포되어 있는 것이며, 그만큼 uncertainty가 크다고 할 수 있다.

3.3.2 엔트로피 측정치

여기서는 엔트로피 개념을 응용하여 uncertainty 측정치를 제안한다. 이를 위해서 정보이론에서 사용되는 정보량 개념을 도입한다.

x_d를 클래스집합 C={C_1, C_2, ..., $C_{|C|}$}의 원소 중 하나의 값을 P={p_1, p_2, ..., $p_{|C|}$}의 확률로 가지는 확률 변수라 정의하자. 여기서 x_d=C_i이 의미하는 것은 문서 집합 D로부터 선택된 임의의 문서 d가 클래스 C_i에 분류됨을 의미한다. 여기서 $P(x_d=C_i)=p_i$, $\sum_{C_i \in C} P(x_d = C_i)$ = 1이다. 이 때, 특정 클래스 C_i가 갖는 정보의 양인 Info(x_d=C_i)을 다음과 같이 정의할 수 있다. [3]

$$\text{Info}(x_d=C_i) = \log_2 \frac{1}{p(x_d = C_i)} \qquad (3.4)$$

따라서, 확률분포 P의 정보량은 다음과 같이 정의된다.

$$\text{Info}(P) = \sum_{C_i \in C} \log_2 \frac{1}{p(x_d = C_i)} \qquad (3.5)$$

x_d	C_1	C_2	C_3
$P(x_d=C_i)$	0.50	0.40	0.10

<표1> x_d의 확률분포 예제

x_d	C_1	C_2	C_3
$P(x_{d'}=C_i)$	0.90	0.07	0.03

<표2> $x_{d'}$의 확률분포 예제

식 (3.2)에서 정의한 P의 정보량은 결국 확률 P의 엔트로피 정의와 같다. 확률분포 P의 엔트로피는 이 분포의 uncertainty를 의미한다.[3] 따라서 여기서 다음과 같은 정의를 내릴 수 있다.

$$U_{entropy}(d) = Info(P) = Entropy(P) \qquad (3.6)$$

위 식에서 $U_{entropy}$ (d)는 엔트로피를 이용한 문서 d의 uncertainty 측정치를 의미한다.

예제 1: <표1>, <표2>는 각각 에서는 문서 d와 d'을 현재 문서 분류기로 분류했을 때 생기는 확률분포를 나타낸다. 먼저 d의 uncertainty를 위의 정의를 이용하여 구해보자.

$$U_{entropy}(d) = Info(P) = 0.50*\log_2(1/0.50) + 0.40*\log_2(1/0.40) + 0.10*\log_2(1/0.10) \fallingdotseq 1.36$$

위와 같이 계산하면, $U_{entropy}(d') \fallingdotseq 0.56$ 이다. 문서 d의 uncertainty가 d'의 것보다 크므로, 현재의 문서 분류기는 d'를 d보다 더 확신 있게 분류한다고 할 수 있다. ■

4. Naïve Bayes 문서 분류기를 위한 점진적 uncertainty 학습 방법

1. D ← {<x1, f(x1)>, ..., <xn0, f(xn0)>
2. h ← L(D)
3. While stop-condition is not satisfied do:
   a) Apply U(X) and get the next example,
      x ← U(X)
   b) Ask the teacher to label x, w ← f(x)
   c) Update the labeled examples set,
      D ← D ∪ {<x, w>}
   d) Update the classifier, h ← L(D)
4. Return classifier h

<그림2> Uncertainty를 이용한 학습방법

$$U(X) \stackrel{def}{=} \arg\min_{d \in X} U_{MAD}(d)$$

$$\stackrel{def}{=} \arg\max_{d \in X} U_{entropy}(d)$$

$$\stackrel{def}{=} \arg\min_{d \in X} U_{confidence}(d)$$

<그림3> uncertainty를 이용한 selection function U(X)의 정의

여기서는 3.1절에서 정의한 uncertainty 개념을 Naïve Bayes 문서 분류기에 적용하는 방법을 제시한다. 식 (2.1)에 나타냈듯이 Naïve Bayes 문서 분류기는 각 문서와 클래스 간의 관련성을 확률로 나타낸다. 따라서, 분류하려는 문서와 클래스 간에는 다음과 같은 확률분포가 형성된다

$$P=\{P(C_1 | d), P(C_2 | d), ..., P(C_{|C|} | d)\} \qquad (3.7)$$

이 분포와 4절에서 정의한 uncertainty 측정치들을 이용하여, Naïve Bayes 문서 분류기에서의 uncertainty 측정치들을 다음과 같이 정의할 수 있다. U_{MAD}(d),

$U_{entropy}(d)$, $U_{confidence}(d)$는 각각 평균절대편차, 엔트로피, 신뢰도를 이용한 uncertainty 측정치를 의미한다.

$$U_{MAD}(d) = \frac{1}{|C|}\sum_{i=1}^{|C|}(p(C_i \mid d) - \mu)$$

$$, \ \mu = \frac{1}{|C|}\sum_{i=1}^{|C|} p(C_i \mid d) \qquad (3.8)$$

$$U_{entropy}(d) = P(C_1 \mid d)*\log_2(P(C_1 \mid d)) + P(C_2 \mid d)*\log_2(P(C_2 \mid d)) + \ldots + P(C_{|C|} \mid d)*\log_2(P(C_{|C|} \mid d)) \qquad (3.9)$$

$$U_{confidence}(d) = \frac{p(C_i \mid d) - p(C_j \mid d)}{p(C_i \mid d)},$$

(Ci, Cj는 각각 $C=\{C_1, C_2, \ldots, C_{|C|}\}$ 의 원소 중, $p(C|d)$가 가장 큰 클래스와 두 번째로 큰 클래스)

(3.10)

<그림2>에서는 uncertainty를 이용한 점진적 학습 방법을 소개한다. 여기서, U(X)는 라벨이 없는 문서 집합 X로부터 가장 uncertainty가 큰 문서를 골라내는 함수이다. 식 (3.8), (3.9), (3.10)에서 정의한 측정치를 이용하여 U(X)함수를 <그림3>에서 정의하였다.

5. 실험 및 결과

5.1 데이터 집합

본 논문에서는 Reuters-21578 문서 집합을 이용하여 실험을 하였다. <표3>은 실험에서 사용한 7개의 클래스들을 나열한 것이다. 각 클래스의 초기 학습문서의 수는

클래스	초기 학습 문서의 수	테스트 문서의 수
acq	20	50
crude	20	50
earn	20	50
Interest	20	50
money-fx	20	50
ship	20	50
trade	20	50

<표3> 실험에서 사용한 클래스 설명

각각 20개로 하였다. 또한 문서 분류기의 성능측정에 이용될 테스트 문서는 각 클래스 당 50개, 총 350개로 하였다. 또한 라벨이 없는 문서 집합을 구성하기 위해, 각 클래스로부터 임의로 200개의 문서를 선택하여, 총 1400개의 문서로 구성된 집합을 만들었다.

5.2 Feature Selection

문서 분류의 성능을 높이고 분류 계산시간을 줄이기 위하여 본 실험에서는 feature selection을 실시하였다. [12]에서는 feature selection의 대표적인 방법인 document frequency, information gain, χ^2-statistics, mutual information 그리고 Term strength 방법의 성능을 비교하였다. 그 결과 앞의 세 방법이 상당히 효과적임을 밝혔다. 여기서는 document frequency 방법을 사용하였다. 이 방법들을 Naïve Bayes 문서 분류기에 적용한 결과, [12]에서와 같이 앞의 세 방법의 성능이 비슷하게 효과적으로 밝혀졌다. 그런데 실행 시간면에서 document frequency 방법이 이 세가지 중에서 가장 우수한 이유로 이를 선택하

였다.

특정 단어의 document frequency는 해당 단어가 출현하는 학습 문서의 수를 의미한다. 희귀한 단어는 문서 분류를 하는데 있어 정보를 거의 제공하지 못한다는 것이 이 방법의 기본적인 가정이다. 본 논문에서는 학습 문서 집합 내의 각 단어들에 대하여 document frequency를 계산한 후에, 이 수치로 단어들의 순위를 부여하여 상위 30%인 것들만 feature로 선택하는 방법을 사용한다.

5.3 성능 평가 측정치

여러 개의 클래스 중 하나의 클래스에 입력 문서를 할당하는 문서 분류기의 경우, 각 클래스와 각 테스트 문서에 대하여 특정 문서가 특정 클래스로 분류되었는지 여부에 따라 다음과 같은 측정치를 계산할 수 있다.

- a: 해당 클래스에 정확하게 분류된 문서의 수
- b: 해당 클래스에 틀리게 분류된 문서의 수
- c: 해당 클래스에 속하지만 이 클래스로 분류되지 않은 문서의 수 (incorrectly rejected)
- d: 해당 클래스에 속하지 않고, 이 클래스로 분류되지 않은 문서의 수 (correctly rejected)

이를 합쳐서 해석하면, (a+c)는 해당 클래스에 속하는 모든 문서의 수이고, (a+b)는 해당 클래스에 실제로 할당된 문서의 수이다. 이를 통해서, 이제 recall 과 precision을 정의해보자.

$$\text{recall} = \frac{a}{a+c} \tag{5.1}$$

$$\text{precision} = \frac{a}{a+b} \tag{5.2}$$

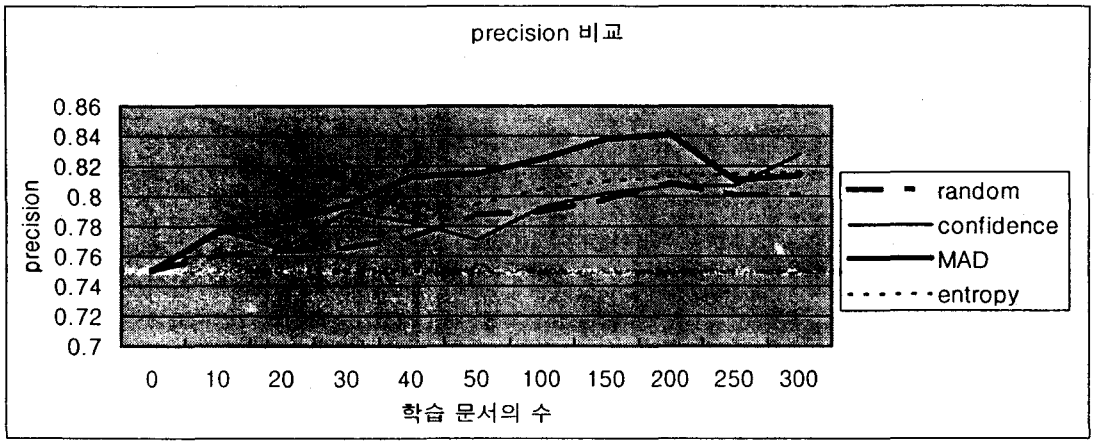
recall과 precision은 일반적으로 문서 분류기의 성능을 평가하는 측정치로서 많이 사용된다. 이 두 측정치를 합성한 측정치도 생각할 수 있는데, F1 측정치가 그 대표적인 것이다. 다음은 F1 측정치를 위한 식이다.

$$F_1 = \frac{2 * recall * precision}{recall + precision} \tag{5.3}$$

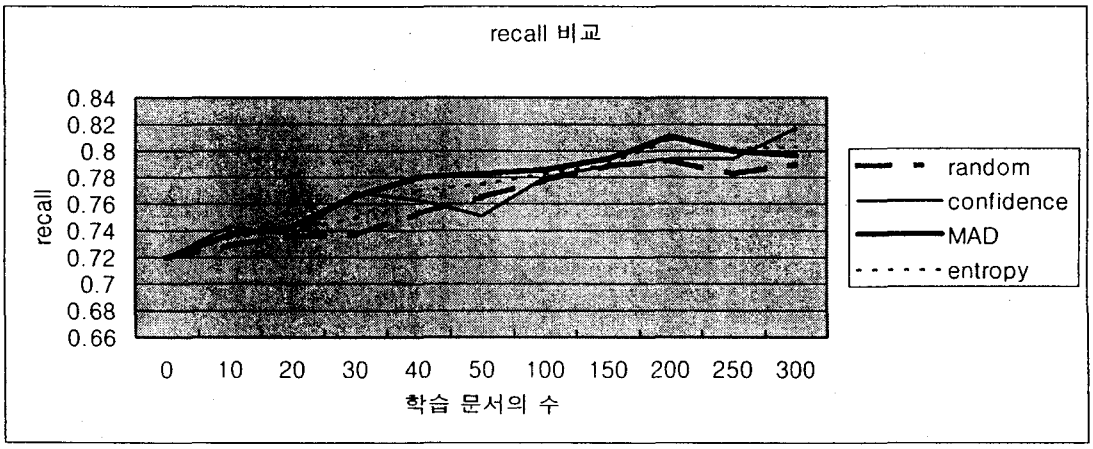
위에서 설명한 recall과 precision, F1 측정치는 각 클래스의 성능을 개별적으로 평가하는 것이다. 모든 클래스에 대한 평균적인 성능을 평가하기 위해, 여기서는 macro-averaging 방법을 이용한다. 이 방식에서는 각 카테고리 별로 recall, precision, F1 측정치 등을 계산하고 이들의 평균을 계산하여 전체적인 문서 분류기의 성능을 평가한다.

5.4 실험 결과

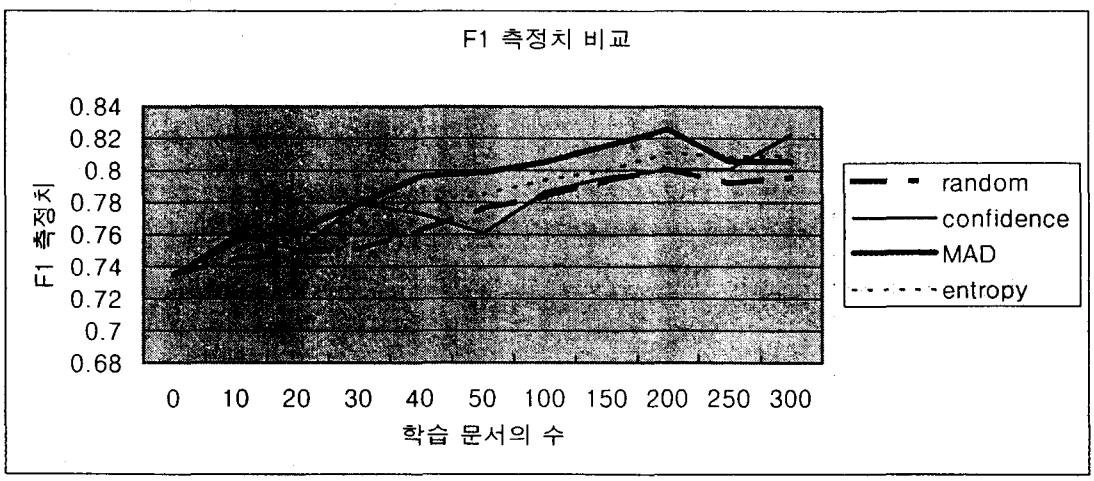
5.1절에서 설명한 데이터 집합을 이용하여 실험을 실시하였다. 본 실험에서는 위에서 제시한 학습 방법을 비교하였다. 신뢰도, 평균절대편차, 엔트로피를 이용한 학습 방법의 성능을 precision, recall, F1 측정치를 통해 비교하였다. 또한 임의로 문서



<그림4> 학습 알고리즘의 precision 비교



<그림5> 학습 알고리즘의 recall 비교



<그림6> 학습 알고리즘의 F1 측정치 비교

집단에서 선택한 문서들을 한꺼번에 학습 문서로 채택하는 임의 학습법의 성능도 함께 측정하였다.

<그림4>를 보자. Uncertainty 개념을 이용한 학습 방법에서는 약 0.79 정도의 precision을 얻기 위해 **약 30개의 라벨**이 있는 학습 문서가 필요한 반면, 임의 학습 방법에서는 약 150개의 학습 문서가 필요하다라는 것을 알 수 있다. 또한 본 논문에서 제시한 두 측정치(평균절대 편차, 엔트로피)를 이용한 학습 방법이 기존의 측정치인 신뢰도를 이용한 학습 방법 보다 평균적으로 성능이 우수함을 알 수 있다. 이러한 결과는 precision뿐만 아니라, recall과 F1 측정치의 경우도 마찬가지이다.

6. 결론

정확도가 큰 문서 분류기를 구현하기 위해서는 적절한 학습이 필수적이다. 그런데 일반적으로 전문가가 문서에 라벨을 부여하는 방식으로 학습 문서를 채택하는 것은 비용이 많이 든다. 본 논문에서는 이 비용을 줄이기 위한 학습 방법을 제안하였다. 본 논문에서는 uncertainty 개념을 이용한 점진적 학습 방법을 Naïve Bayes 문서 분류기에 적용한 알고리즘을 제시하였으며, 제시한 학습 방법의 우수성을 실험을 통해 입증하였다.

참고문헌

[1] David D. Lewis and Jason Catlett, Heterogeneous Uncertainty Sampling for Supervised Learning, *Machine Learning: Proceeding of the 11th international Conference*, pp. 148-156, 1994

[2] David D. Lewis and William A. Gale, A Sequential Algorithm for Training Text Classifiers, *Proceeding of 17h Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 3-12, 1994

[3] David J.C. MacKay, "Information Theory, Inference and Learning Algorithms", *available at http://wol.ra.phy.cam.ac.uk/mackay/itprnn/book.html*, pp. 27-51, 2001

[4] H. S. Seung and M.Opper and H.Sompolinsky, Query by Committee, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (ACM, New York)*, pp. 287-294, 1992

[5] Ion Muslea and Steven Minton and Craig A. Knoblock, Selective Sampling With Co-Testing, *The CRM Workshop on "Combining and Selecting Multiple Models With Machine Learning"*, 2000

[6] Ion Muslea and Steven Minton and Craig A. Knoblock, Selective Sampling With Redundant Views, *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-2000)*, pp. 621-626, 2000

[7] K. Nigam and A. McCallum and S. Thrun and T. Mitchell, Learning to classify text from labeled and unlabeled documents, *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998

[8] M. Trensh and N. Palmer and A. Luniewski, Type Classification of Semi-structured Documents, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1995

[9] M. Lindenbaum, S. Markovitch and D. Rusakov, Selective sampling for nearest neighbor classifiers, *American Association for Artificial Intelligence*, 1999

[10] Tom M. Mitchell, Machine Learning, *McGraw-Hill International Editions*, chapter 6, 1997

[11] Y. Freund and H. S. Seung and E. Shamir and N. Tishby, Selective sampling using the query by committee algorithm, *Machine Learning, 28(2-3)*, pp. 133-168, 1997

[12] Y. Yang and J. O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, *ICML 1997*, pp. 42-420, 1997