최대 엔트로피 기반 문서 분류기의 학습

장정호, 장병탁, 김영택
서울대학교 컴퓨터공학과
jhchang@scai.snu.ac.kr, {btzhang, ytkim}@comp.snu.ac.kr

Text Categorization Based on the Maximum Entropy Principle

Jeongho Chang, Byoung-Tak Zhang, Yung Taek Kim
Dept. of Computer Engineering
Seoul National University

요 약

본 논문에서는 최대 엔트로피 원리에 기반한 문서 분류기의 학습을 제안한다. 최대 엔트로피 기법은 자연언어처리에서 언어 모델링(Language Modelling), 품사 태깅 (Part-of-Speech Tagging) 등에 널리 사용되는 방법 중의 하나이다. 최대 엔트로피 모델의 효율성을 위해서는 자질 선정이 중요한데, 본 논문에서는 자질 집합의 선택을 위한 기준으로 chi-square test, log-likelihood ratio, information gain, mutual information 등의 방법을 이용하여 실험하고, 전체 후보 자질에 대한 실험 결과와 비교해 보았다. 데이터 집합으로는 Reuters-21578을 사용하였으며, 각 클래스에 대한 이진 분류 실험을 수행하였다.

1. 서 론

문서 분류(text categorization)는 임의의 텍스트 문서를 이미 정해진 범주에 따라 분류하는 문제이다. 전산화된 문서의 양이 점점 더 증가하고 있고, 이에 따른 정보의 분류 문제 역시 중요한 문제로 제기되고 있다. 지금까지 최근접 이웃 방법 (nearest neighbor method), naïve Bayes 방법, 신경망 등의 여러 통계적 방법이나 기계 학습 기법들이 문서 분류 문제에 적용되었다. 본 논문에서는 최대 엔트로피(Maximum Entropy)에 기반한 문서 분류기를 제안하고자 한다.

최대 엔트로피 기법은 최근 자연언어처리 분야에서 언어의 통계적 모델링[1], 단어의 품사 태깅[2] 등에 성공적으로 적용되고 있는 기법들 중의 하나이다. 자연언어처리에서 다루는 많은 문제들은 확률 분포에 기반한 분류(classification)문제로 재 정의될 수 있으며, 최대 엔트로피 기법은 주어진 문맥(context)에 따른 조건부 확률 분포를 생성하고, 이에 따라 분류 작업을 진행한다.

문서 분류 문제에서는 하나의 문서가 주어질 때, 문서를 구성하는 단어나 구(phrase) 등을 분류를 위한 문맥으로 볼 수 있다. 이런 입장에서 주어진 문맥에 기반한 조건부 확률 분포를 구성하여 이미 정의된 문서 범주들 각각에 대한 확률을 추정할 수 있으며, 이 확률 값에 기초하여 각 문서에 적절한 범주를 지정할 수 있다.

최대 엔트로피 모델 구축은 그 계산 과정에서 복잡도가 크기 때문에, 효율적 학습을 위해서는 적절한 자질 정보의 선택이 중요하다. 이를 위하여 chi-square test(CHI)[8], log-likelihood ratio(LL)[6], information gain(IG)[8], mutual information(MI)[8] 등의 방법에 의한 자질 집합을 구성하고, 문서의 이진 분류 문제에 대해 그 실험 결과를 비교한다.

2절에서는 먼저 최대 엔트로피 원리에 대해 간략히 설명하고, 3절에서는 문서 분류기의 학습과 구축을 위해 최대 엔트로피 원리를 어떻게 적용할 수 있는지에 대해 서술한다. 4절에서는 Reuters-21578 문서 집합에 대한 실험 결과를 제시한다.

2. 최대 엔트로피 원리

최대 엔트로피 원리는 통계 물리(statistical physics) 연구에서 유래한 것으로서, 미지의 사실에 대한 가정들은 배제하고 알려진 부분적인 사실만 지식 획득의 원천으로 보는 기법이다. 이에 기반한 최대 엔트로피 모델은 미리 정의된 제한 조건(constraint)들을 만족하면서 그 이외의 사항에 대해서는 균일 분포(uniform distribution)를 갖는 확률 모델이다.

엔트로피는 확률 변수의 불확실성(uncertainty)에 대한 수학적 척도로서, 이산 확률 변수 X의 엔트로피는 다음과 같이 정의된다.[4]

$$H(X) = -\sum_{\forall x} p(x)\log p(x)$$

일반적으로 이산 확률 변수 X가 균일 분포를 따르는 경우 그 엔트로피 값은 최대가 된다. 따라서 최대 엔트로피 모델이 미지의 사실에 대해서 균일 분포를 갖는다는 것은 미지의 내용에 대해서는 모델의

엔트로피가 최대가 된다는 것을 의미한다. 주어진 두 사건에 대해 그 둘을 명확히 구분할 정보가 존재하지 않을 때의 가장 좋은 해결책은 그 두 사건이 똑같은 확률로 발생한다고 간주하는 것이며, 이는 불확실성의 최대화, 즉 엔트로피의 최대화를 의미한다. 따라서 생성된 모델은 알려진 부분 정보에만 부합하고 미확인 정보에 대해서는 그 어떤 가정도 포함하지 않는 최선의 모델이 된다[1, 2] 이러한 내용에 따른 최대 엔트로피 원리의 정의는 다음과 같다.

최대 엔트로피 원리 (Maximum Entropy Principle)
주어진 제한 조건들을 만족하는 확률 분포들의 집합 P에서, 최선의 확률 분포는 엔트로피 값이 최대인 확률 분포이다.

$$p^* = \arg\max_{p \in P} H(p)$$

3. 문서 분류기의 설계 및 구현

문서 분류기는 최대 엔트로피 원리에 기반하여 통계적 확률 모델로 구축된다. 이러한 모델의 생성은 크게 두 단계로 구성된다. 첫 번째는 문서 분류에 있어 중요한 정보를 제공하는 자질(feature)의 선택이고, 두 번째는 선택된 각 자질에 연관되는 가중치(weight) 변수의 값을 추정(parameter estimation)하는 단계이다.

3-1. 문서 분류를 위한 자질

문서 분류 문제에서는 문서를 구성하는 특정 단어, 구(phrase), 또는 단어들의 집합들을 하나의 자질로 구성하는 것이 보통이다. 이 논문에서는 개별적인 단어에 대한 자질들만 생성한다. 구축하고자 하는 확률모델을 구성할 수 있는 자질들은 잠재적으로 모든 단어 집합과 모든 문서 범주 집합의 순서쌍에 대해 생성될 수 있다.

$$F_{candiate} = \{ f_{w,c} \mid w \in W,\ c \in C \}$$

C: 미리 정의된 문서 범주들의 집합
W: 문서를 구성하는 단어들의 집합

그리고 각각의 자질은 다음과 같은 이진 함수의 형태로 표현될 수 있다.

$$f_{w,c}(d,y) = \begin{cases} 1 & \text{if } w \in d,\ c = y \\ 0 & otherwise \end{cases}$$

w: 단어 $\quad c, y$: 문서 범주 $\quad d$: 문서

즉, 자질 함수는 문서를 구성하는 특정 단어와 그에 따른 문서의 범주에 대한 조건을 검사하고 단순히 참(1), 거짓(0)의 값을 반환하는 함수이다.

3-2. 모델 생성과 가중치 추정

문서 분류를 위한 통계적 확률 모델은 주어진 자질들을 기반으로 최대 엔트로피 원리에 의하여 구축된다. 학습 문서 집합 D에 대해, 자질 f의 경험적 확률 분포 $\tilde{p}(d,y)$에 대한 기대값은 다음과 같이 정의된다.

$$E_{\tilde{p}}(f) = \sum_{d,y} \tilde{p}(d,y) f(d,y),$$

$\tilde{p}(d,y) = N(d,y)/N(D)$, $N(D)$: 문서의 총 개수
$N(d,y)$: 문서 d가 범주 y인 빈도

이 값은 학습 데이터에서 자질 f가 나타나는 정규화된 빈도 수를 의미한다. 그리고, 학습하고자 하는 확률 분포 $p(d,y)$에 대한 f의 기대값은 다음과 같다.

$$E_p(f) = \sum_{d,y} p(d,y) f(d,y)$$

최대 엔트로피 원리에 의하면, 확률 분포 $p(d,y)$는 다음과 같은 제약 조건을 만족하여야 한다.

$$E_p(f) = E_{\tilde{p}}(f)$$

하지만, 모든 가능한 (d,y)에 대한 결합 확률 분포 $p(d,y)$를 구할 수는 없기 때문에 다음과 같은 근사적 기대값을 사용한다.

$$E_p(f) = \sum_{d,y} p(d,y) f(d,y) \approx \sum_{d,y} \tilde{p}(d)\, p(y|d) f(d,y)$$

위의 식을 따르면, 관심이 되는 확률 분포는 결합확률분포가 아닌 조건부 확률 분포 $p(y|d)$가 된다. 자질 집합 F에 대해 가능한 모든 확률 분포의 집합 P 중에서, 모든 자질들에 대해 위의 조건을 만족하는 확률 분포의 집합 C는 다음과 같이 정의된다.

$$C = \{ p \in P \mid E_p(f) = E_{\tilde{p}}(f),\ \text{for } all\ f \in F \}$$

최대 엔트로피 원리에 의하면, 집합 C에 속하는 확률 분포 중에서 최선의 확률 분포는 다음과 같이 표현된다.

$$p^* = argmax_p H(p)$$

위의 문제는 주어진 제약 조건을 만족해야 하는 제한적 최적화(constrained optimization)문제이다. 이 문제는 라그랑제 승수(Lagrange multipliers)방법을 적용하여 비제한적 최적화(unconstrained optimization)문제로 변형하여 풀 수 있다. 그 결과, 구하고자 하는 조건부 확률 분포는 다음과 같은 지수 또는 로그 선형 확률 분포 형태로 표현된다.

$$Q = \left\{ p \mid p_\Lambda(y|d) = \frac{1}{Z_\Lambda(d)} \exp\left(\sum_i \lambda_i f_i(d,y)\right) \right\}$$

여기에서 $Z_\Lambda(d)$는 정규화 상수로서 다음과 같이 정의된다.

$$Z_\Lambda(d) = \sum_y \exp\left(\sum_i \lambda_i f_i(d,y)\right)$$

결국 확률 모델 $p(y|d)$를 생성하는 것은 각 자질의 가중치 값인 λ_i를 구하는 문제가 된다. 일반적으로 최대 엔트로피 기법에서는 GIS(Generalized Iterative Scaling)와 같은 반복적 최적화 방법을 이용하는데, 이 논문에서는 보다 더 일반적인 문제에 적용 가능한 IIS(Improved Iterative Scaling)방법을 사용하여 가중치 값들을 근사적으로 계산한다. IIS방법은 반복적인 최적화 과정이기 때문에 어느 시점에서 반복을 중단할지를 결정해야 한다. 이 논문에서는 매 반복마다 생성된 모델의 경험적 분포에 대한 로그-유사도(log-likelihood)를 계산하여 그 값의 변화량을 척도로 삼았다. 모델의 로그-유사도(log-likelihood) $L_{\tilde{p}}(p)$는 다음과 같이 정의된다.

$$L_{\tilde{p}}(p) = \log \Pi_{d,y} p(y|d)^{\tilde{p}(d,y)} = \sum_{d,y} \tilde{p}(d,y) \log p(y|d)$$

최대 엔트로피를 가지는 지수 모델을 구하는 것은 최대 로그-유사도를 가지는 모델을 구하는 문제와 같다.[1, 3] 즉,

$$p^* = argmax_{p \in C} H(p) = argmax_{q \in Q} L_p(q)$$

4. 실험 및 결과

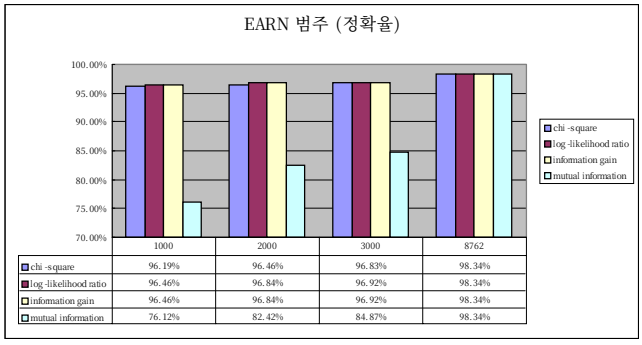
최대 엔트로피 기법에 의해 학습, 구축된 문서 분류기를 위한 실험 문서 집합으로는 Reuters-21578 문서 집합[1]을 사용하였다. 이 문서 집합은 크게 5개의 범주 집합으로 구성되어 있는데, 이 논문에서는 그 중 TOPICS 범주 집합에 포함된 3개의 범주에 대해 이진 분류 실험을 진행하였다. 학습 문서 집합과 검증 문서 집합의 구성은 ModApte 분리를 따랐으며, 빈 문서를 제외한 문서의 개수는 각각 8,762, 3,009개이다. 3개 범주에 대한 학습 문서 집합의 구성은 다음과 같다.

	Acq	Crude	Earn
Positive	1,483	334	2,706
Negative	7,279	8,428	6,056

불용어 목록, 스테밍(stemming), 단어의 문서 빈도에 따른 제거 후의 문서의 크기는 8,754였다. 아래 표는 8,754개의 모든 단어를 자질로 사용할 때의 각 범주에 대한 이진 분류 결과를 보여준다. Acq나 Crude 범주에서 재현율이 낮은 것은 해당 범주의 데이터가 Earn 범주에 비해서는 상대적으로 적기 때문인 것으로 간주된다.

	재현율(recall)	정확율(precision)
Acq	86.88%	96.03%
Crude	64.74%	84.87%
Earn	96.64%	98.34%

그리고, [그림 1]과 [그림 2]에서는 각 자질 선정 방법들을 Earn 범주에 대해 실험한 결과이다. MI에 의한 것을 제외하고는 비슷한 결과를 보였다. 이 점에서는 [6],[8]과 같은 결과를 보인다. 하지만 자질의 수를 줄일 경우, 학습 시간은 많이 줄지만 정확율[5] 면에서 성능 향상이 보이지 않았다. 재현율[5] 면에서는 거의 비슷한 결과를 보였다. 나머지 두 범주 역시 Earn 범주에 대한 실험과 비슷했다.

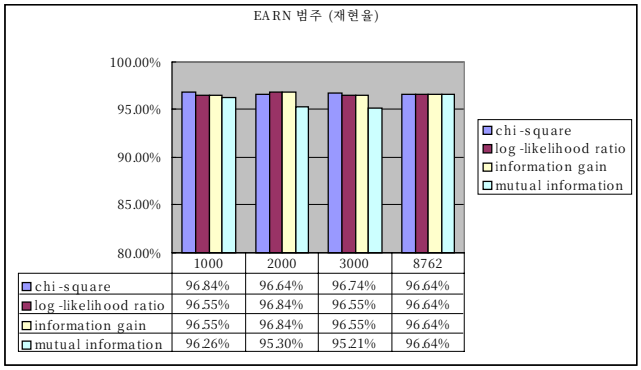


	1000	2000	3000	8762
chi-square	96.19%	96.46%	96.83%	98.34%
log-likelihood ratio	96.46%	96.84%	96.92%	98.34%
information gain	96.46%	96.84%	96.92%	98.34%
mutual information	76.12%	82.42%	84.87%	98.34%

[그림 1]

5. 결론

이 논문에서는 하나의 단어만으로 구성된 자질을 이용하여 최대 엔트로피 기법에 의해 문서 분류기를 학습하였다. 둘 이상의 단어나 구(phrase)로 구성된 자질과, 이진 함수 형태가 아닌 실수 값을 가지



	1000	2000	3000	8762
chi-square	96.84%	96.64%	96.74%	96.64%
log-likelihood ratio	96.55%	96.84%	96.55%	96.64%
information gain	96.55%	96.84%	96.55%	96.64%
mutual information	96.26%	95.30%	95.21%	96.64%

[그림 2]

는 자질 함수가 문서 분류기의 성능향상에 도움이 될 수 있는지의 여부에 대한 연구가 필요하다. 그리고 이 논문에서는 자질을 미리 선정하고 최대 엔트로피 모델을 학습하였지만, 점진적 자질 선정에 대해서도 연구가 필요하다.

감사의 글

본 연구는 대학 기초 연구 기술 지원 사업 (c1-98-006800)에 의해 지원되었음.

참고문헌

[1] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, pp. 39-72.

[2] A. Ratnaparkhi. 1996. A Maximum Entropy Model for Part-of-Speech Tagging, *Proceedings of Conference on Empirical Methods in Natural Language Processing.*

[3] Adwait Ratnaparkhi. 1997. A Simple Introduction to Maximum Entropy Models for Natural Language Processing, *Technical Report 97-08*, Institute for Research in Cognitive Science, University of Pennsylvania.

[4] Cover, T and Thomas, J. 1991. *Elements of Information Theory*. John Wiley & Sons.

[5] David D. Lewis. 1991. Evaluating Text Categorization. *Proceedings of the Speech and Natural Lanugage Workshop.* pp. 312-317.

[6] T. Dunning. 1993. Accurate Methods for Statistics of Surprise and Coincidence. *Computational Linguistics*. Vol . 19. pp. 61-74

[7] William B. Frakes and Richard Baeze-Yates. 1997. *Information Retrieval Data Structures & Algorithms.* Prentice-Hall, Inc.

[8] Yang, Y., Pedersen J.P. 1997. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning.*

1) http://www.research.att.com/~lewis/reuters21578.html