n-gram 파티클을 이용한 베이지안 필터링 기법

장하영○ 장병탁

서울대학교 전기컴퓨터공학부

hyjang@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Bayesian Filtering Method using n-gram Particle

Ha-Young Jang○ Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

베이지안 필터링(Bayesian Filtering)은 관측데이터와 연관된 확률을 이용하여 관찰된 데이터를 설명할 수 있는 은닉 변수를 설정하여 마코프 연쇄(Markov Process)를 따르는 은닉 변수의 값을 추정하는 모델로 칼만 필터(Kalman Filter)나 파티클 필터(Particle Filter) 등의 방법이 대표적이다. 본 논문에서는 다자간의 대화에서 발생하는 이야기 흐름의 전환을 분석하기 위해서 대담 형식으로 구성된 데이터를 임의보행(Random Walk) 확률과정을 따르는 시계열 데이터로 간주하고, 이야기 흐름의 전환이 마코프 연쇄에 의해 결정된다고 가정하여 베이지안 필터링 기법을 이용한 대화 분석 기법을 제시하였다. 제안한 방법론은 n-gram 언어모델과 베이지안 필터링 기법을 결합하여 말뭉치로부터 n-gram 언어모델을 구축하여 이를 초기 분포로 이용하고, 이 n-gram 들을 파티클로 이용하여 이야기 흐름의 전환을 예측하게 된다. 일반적으로 언어 데이터는 그 특성상 베이지안 필터링 기법의 폭넓은 적용이 어려운데 본 논문에서 제시한 n-gram 언어모델과 베이지안 필터링 기법을 결합을 통해서 언어처리에 있어서 베이지안 필터링 기법의 보다 넓은 적용이 가능할 것으로 기대 된다.

1. 서론

오차가 존재하는 관측값으로부터 실제값을 추정해내기 위해 사용되는 베이지안 필터링은 특정 조건을 만족할 경우에 최적해를 보장하는 방법이지만, 언어처리의 경우에는 n-gram의 길이가 길어질 경우 데이터의 희소성이 급격히 증가하는 특성 때문에 베이지언 필터링 기법의 적용에 어려움이 있다. 따라서 베이지언 필터링 기법이 언어처리에 적용될 경우에는 유니그램을 사용하는 스팸 필터링[4]이나 단어 분리(Word Segmentation)[3] 등에 주로 사용된다. 또한 일반적인 언어처리 문제의 경우에는 데이터의 시간적 특성을 고려하지 않고 구성된 말뭉치를 이용하여 일괄처리 방식으로 모델을 구축하기 때문에 언어 데이터에 들어 있는 시간 정보를 활용하기 힘들다는 단점이 있다.

본 논문에서는 이를 해결하기 위해서 n-gram 모델과 베이지안 필터링 기법을 결합하여 언어 데이터의 시간 정보를 활용할 수 있는 방법을 제시하였다. 제안된 방법론은 기존의 n-gram 모델이 말뭉치 전체의 빈도수를 이용[5]함으로써 문장의 선후관계나 대화에서의 시간 정보 등을 활용할 수 없는 것 과는 달리 대화문으로 구성된 말뭉치로부터 구축된 n-gram을 파티클로 이용함으로써 데이터에 존재하는 시간 정보를 활용할 수 있다는 장점이 있다. 이로 인해서 기존의 n-gram 언어모델에서 처리할 수 없었던 텍스트 스트림의 감지(Detection), 구분(Segmentation), 인식(Recognition) 등의 다양한 분석 방법에 적용이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 n-gram 파티클을 이용한 베이지안 필터링 방법에 대해서 설명하고, 3장에서는 제안한 방법론을 여러 명의 출연자가 이야기하는 토크쇼 데이터에 적용하여 이야기 흐름에 변화가 있는 부분을 감지한 실험 결과를 보인 후에 결론을 맺는다.

2. n-gram을 이용한 베이지안 필터링

베이지안 필터링은 마코프 연쇄를 따르는 시스템의 상태를 은닉변수로 가정하고 이로부터 관측되는 관측값을 이용하여 시간에 따른 시스템의 상태를 나타내는 확률 밀도 함수를 재귀적(recursive)으로 추정하는 기법이다[2]. 제안한 방법론은 베이지안 필터의 posterior를 가중치 w_t를 가지는 m개의 n-gram으로 이루어진 파티클의 집합 S_t를 이용하여 표현한다.

$$S_t = \left\{ \langle x_t^{(i)}, w_t^{(i)} \rangle \middle| i = 1, \ldots, n \right\} \quad (1)$$

제안한 방법론을 이용하여 파티클의 분포를 추정하는 방법은 크게 3단계로 나누어 생각할 수 있다. 먼저 말뭉치의 확률 분포를 사전분포(prior distribution)으로

이용하여 파티클 집합 S_t를 초기화 한다. 이 때 S_t는 말뭉치에 들어 있는 문장에서 균등(uniform)하게 추출된 n-gram 들로 구성된다. 이 파티클 집합과 매 시간 간격마다 입력으로 들어오는 문장에서 관측되는 n-gram들을 이용하여 예측(prediction)과 가중치의 수정(update)가 반복되게 되는데 시간 t일때의 파티클의 분포에 대한 예측은 시간 t-1일때의 파티클의 분포 S_{t-1}와 시간 t-1에서의 입력문장에서 관측되는 n-gram을 이용하여 수행되는데 이때 중요한 점은 파티클 필터에서 발생하는 퇴보(degeneracy) 현상을 피하기 위해서 연속된 단어로 구성된 n-gram만을 이용하는 것이 아니라 하이퍼에지로 구성된 비연속적인 단어로 구성된 n-gram을 사용 한다는 것이다. 하이퍼그래프를 이용한 언어모델은 3장에서 설명할 것이다.

마지막으로 가중치의 수정은 파티클 필터에서의 임포턴스 샘플링(importance sampling)과 동일한 방법을 사용하는데, 목적분포 $p(x)$의 추정치(estimation)는 아래와 같이 정의된다.

$$\hat{p}(x_{0:k} \mid Z_k) = \sum_{i=1}^{N} w_k^i \delta\,(x_{0:k} - x_{0:k}^i) \tag{2}$$

이 때, δ는 dirac delta 함수이고, w_k^i는 다음과 같이 정의된다.

$$w_k^i = \frac{p(x_{0:k}^i \mid Z_k)}{q(x_{0:k}^i \mid Z_k)} \tag{3}$$

S_t의 분포 $q(x_{0:k}|Z_k)$를 분해(factorize)하면 아래와 같이 되고,

$$q(x_{0:k} \mid Z_k) = q(x_0)\prod_{j=1}^{k} q(x_j \mid x_{0:j-1}, Z_j) \tag{4}$$

이 때 은닉변수가 마코프 연쇄를 따르기 때문에 아래와 같이 가중치를 추정할 수 있다.

$$w_k = w_{k-1}\frac{p(y_k \mid x_k)p(x_k \mid x_{k-1})}{q(x_k \mid x_{0:k-1}, Z_k)} \tag{5}$$

w_k의 분산(variance)를 최소화 하기 위해서 $q(x)$를 아래와 같이 선택하면,

$$q(x_k \mid x_{0:k-1}, Z_k) = p(x_k \mid x_{0:k-1}, Z_k) \tag{6}$$

가중치는 아래와 같이 정의된다.

$$w_k^i = w_{k-1}^i \frac{p(z_k \mid x_k^i)p(x_k^i \mid x_{k-1}^i)}{q(x_k^i \mid x_{0:k-1}^i, D_k)} = w_{k-1}^i p(z_k \mid x_k^i) \tag{7}$$

3. 하이퍼그래프 언어모델

하이퍼그래프 모델에서는 말뭉치에 존재하는 단어들의 조합으로 구성된 하이퍼에지와 하이퍼에지의 출현빈도를 표현하는 가중치로 확률분포를 표현[1]하게 되는데 이때 하이퍼그래프 모델의 에너지는 다음과 같이 정의된다[6].

$$\varepsilon(s^{(n)}; W) = -\sum_{i=1}^{|E|} w_{i_1 i_2 \ldots i_{|E_i|}} x_{i_1}^{(n)} x_{i_2}^{(n)} \ldots x_{i_{|E_i|}}^{(n)} \tag{3}$$

$x_{i_1}^{(n)} x_{i_2}^{(n)} \ldots x_{i_{|E_i|}}^{(n)}$ 는 말뭉치 내에 존재하는 단어들의 조합으로 구성된 하이퍼에지를 의미하고 W는 하이퍼에지의 가중치를 의미한다. 즉, n-gram 방식의 언어모델에서 사용하는 단어들의 출현빈도를 가중치의 형태로 표현하여 이를 이용하여 하이퍼그래프 모델을 표현하는 것이다.

이렇게 만들어진 하이퍼그래프 모델에서 문장 $s^{(n)}$이 나타날 확률은 다음과 같이 깁스 분포에 의해서 주어지게 되고,

$$P(s^{(n)}|W) = \frac{1}{Z(W)} exp\{-\varepsilon(s^{(n)}; W\} \tag{4}$$

분할함수(partition function) $Z(W)$는 다음과 같이 정의된다.

$$Z(W) = \sum_{x^{(m)}} exp\{-\varepsilon(s^{(m)}; W\} \tag{5}$$

4. 실험 및 결과

제안한 방법론은 기존의 언어모델들과는 달리 시간 정보가 있는 데이터에의 적용이 가능하다. 실험에서는 이를 잘 보여주기 위해서 한명의 호스트와 한명 이상의 게스트가 함께 이야기를 나누는 CNN의 Larry King Live show 대본을 말뭉치로 이용하여 대화에 대한 방청객의 반응을 예측하는 실험을 진행하였다. 말뭉치는 총 77,447개의 문장으로 구성된 1,000개의 대본에서 수집하였고, 말뭉치 내에는 방청객이 웃음을 터트린 1,123개의 문장과 그렇지 않은 76,324개의 문장이 들어 있다.

일반적인 분류 문제에서는 데이터가 한쪽으로 치우쳐져 있을 경우에 학습에 어려움이 생기고 대개의 경우에 정확도(precision)나 재현률(recall)에 문제가 생기기 때문에 이를 해결하기 위한 별도의 방법이 필요하게 된다. 그러나 사실 이러한 문제는 이미 앞에서 이야기 한 언어 데이터에 존재하는 시간 정보를 무시하기 때문에 발생하는 현상으로 순차적인 대화의 과정을 고려할 수 있다면 보다 쉽게 해결할 수 있는 문제에서 순서를 제외함으로써 문제 자체를 보다 어렵게 만들어 버리게 되는 것이다. 이를 해결하기 위해서 다음과 같이 문제를 설명하는 모델을 정의하였다.

제안한 방법론은 시간 t에서 n-gram 으로 정의되는 관찰 데이터는 해당 데이터에 대한 은닉 변수 X_t에서 생성되며, 동일한 분위기의 대화를 나타내는 은닉 변수 θ를 추정하는 것이 목적이다. 이 때 은닉 변수 X_t는 관찰 데이터 Z_t를 생성해 내는 문장으로 생각할 수 있다.

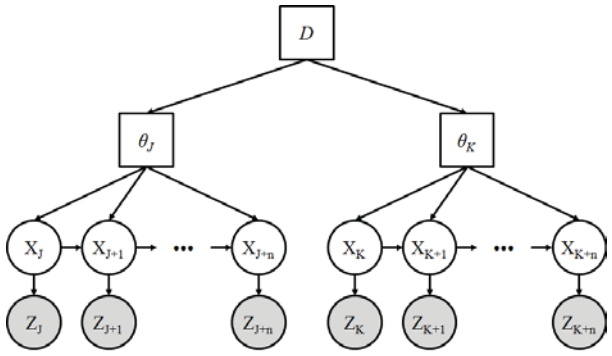

그림 1. 대화 분위기 감지를 위한 베이지안 필터링 모델

제안한 방법론의 효과를 보여주기 위해서 말뭉치에 존재하는 시간 정보를 사용하지 않고 unigram의 빈도수만을 이용하여 Naïve Bayes Classifier, Support Vector Machine (SVM), Decision Tree 알고리즘을 적용하여 데이터를 분류해 보았다.

	Naïve Bayes	SVM	Decision Tree
Accuracy	95.82	95.32	97.42
Recall	2.01	2.38	0.92

표 2. 치우친 데이터를 사용한 실험 결과

실험결과 세가지 알고리즘 모두 95% 이상의 매우 높은 성능을 보이고 있으나 재현율이 모두 매우 낮은 것을 확인할 수 있다. 이 결과가 보여주는 것은 치우친 데이터의 처리를 위한 별도의 작업이 없을 경우에는 일반적으로 모두 한 쪽으로 분류를 해 버리는 경우가 많기 때문에 위와 같은 결과가 나온 것으로 판단된다.

다음으로는 제안한 방법론의 성능을 확인하기 위해서 균형 잡힌 데이터를 사용한 실험 결과와 비교해 보았다. 이전 실험과 마찬가지로 unigram의 빈도수 만을 이용하였고, 데이터는 1,123개의 positive 데이터를 그대로 복사하여 총 5,615개의 positive 데이터를 만들고 76,324개의 negative 데이터 중에서 랜덤하게 5,615개의 데이터를 선택하여 총 11,230개의 데이터를 사용하였다.

	n-gram Filtering	Naïve Bayes	SVM	Decision Tree
Accuracy	79.87	72.19	72.30	70.97

표 2. 제안 방법과의 성능 비교

5. 결론

본 논문에서는 n-gram을 이용한 베이지안 필터링을 이용하여 언어 데이터에 들어 있는 시간 정보를 활용할 수 있는 기법을 제시하였다. 많은 언어데이터들이 시간과 연관된 정보를 가지고 있고, 시간적인 분석을 필요로 하지만 기존의 언어모델이 시간 정보를 무시한 대용량의 말뭉치만을 이용한다는 한계를 가지고 있는 것과는 달리 제안된 방법론은 시간 정보의 활용을 위한 언어 데이터의 분석에의 가능성을 보여주었다. 이러한 결과는 순차적인 대화로 구성된 말뭉치를 이용한 실험에서도 잘 보여지고 있다.

또한 단순히 시간 정보의 활용만이 아니라 데이터의 희소성으로 인해서 언어처리 분야에의 적용이 많지 않았던 베이지안 필터링 기법을 n-gram 모델과 결합함으로써 언어처리 분야에의 적용 가능성을 보여줬다는 것에 본 논문의 의의가 있다. 이를 위해서는 분포 추정 및 학습 과정에 대한 보다 명확한 정의가 필요하리라 생각된다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. 2012-0005643, Videome), 정부(지식경제부)의 재원으로 한국산업기술평가관리원의 지원(10035348, mLife) 및 교육과학기술부의 BK21-IT 프로그램에서 일부 지원되었음.

참고문헌

[1] 장하영, 장병탁, 자동 스토리텔링을 위한 하이퍼그래프 언어모델 기반의 문장의 전후 관계 분석, *한국정보과학회 가을학술발표 논문집*, 제38권 2(B), pp. 271-274, 2011.11

[2] Arulampalam, M. S., Maskell, S., Gordon, N., A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing* 50:174–188, 2002.

[3] Goldwater, S. and Griffiths, T.L. and Johnson, M., A Bayesian framework for word segmentation: Exploring the effects of context, *Cognition*, 112(1):21-54, 2009.

[4] Graham , P., Better Bayesian Filtering, 2003.

[5] Song, F. and Croft, W.B., A general language model for information retrieval, *Proceedings of the eighth international conference on Information and knowledge management*, pp. 316-321, 1999.

[6] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3) pp. 49-63, 2008.