Competitive Unit을 사용한 Helmholtz Machine에 의한 문서 클러스터링

장 정 호° 장 병 탁

서 울 대 학 교 컴 퓨 터 공 학 부

{jhchang, btzhang}@scai.snu.ac.kr

Topical Clustering of Documents using Helmholtz Machines with Competitive Units

Jeong-Ho Chang° Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

문서 클러스터링은 정보검색 시스템에서 검색 과정의 효율성을 향상시키기 위해서 많이 사용된다. 기존의 K-means 클러스터링과 같은 거리-기반 접근 방법은 거리에 대한 척도를 정해야 하는 문제가 있고, 또한 전체 자질 공간에서 지역적 특성에 민감하기 때문에 문서 내에 노이즈가 존재할 경우 만족스러운 결과를 내지 못할 수 있다. 그리고 기본적으로 문서 데이터는 희소성(sparseness)을 갖기 때문에 정규 분포를 가정한 mixture 모델을 적용하기에도 어려움이 있다. 본 논문에서는 Helmoholtz machine에 의한 문서 클러스터링 방법을 제안한다. 제안되는 방법에서는 하나의 문서를 어떤 내재적인 요인(factor)들의 다양한 결합에 의한 결과로 가정하는데, 이 때의 요인은 주제어 집합 또는 적어도 의미적으로 유사한 단어들의 집합이다. 그리고 기본적으로 Helmholtz machine은 이진 데이터를 다루는데, 텍스트 문서에 나타나는 단어들의 빈도를 고려하기 위해 수정된 Helmholtz machine을 제시한다. TREC-8 adhoc 데이터와 20 Newsgroup 문서 집합에 대한 클러스터링 실험 결과, 제안된 방법이 K-means 알고리즘에 비해 우수한 성능을 보였으며 주제어 추출을 통해 문서 집합의 전체 내용 파악을 용이하게 하는 특성이 있었다.

1. 서론

인터넷의 발달과 이에 따른 정보량의 폭발적 증가로 온라인 텍스트나 전자화된 텍스트 문서의 양이 크게 증대되고 있다. 하지만 이러한 정보의 폭주 자체가 곧바로 사용자들로 하여금 필요한 정보를 쉽게 얻을 수 있도록 하는 것은 아니다. 왜냐하면 다양한 주제에 관련된 텍스트를 모두 검색하고 조직화하는 것은 너무도 많은 시간을 필요로 하는 작업이기 때문이다. 이에 따라 컴퓨터를 이용한 문서의 자동 분석에 대한 요구가 증대되고 있는데, 문서 클러스터링은 그러한 요구를 충족시키기 위한 도구 중의 하나라고 볼 수 있다. 본 논문에서는 다양한 주제에 관련된 문서 집합으로부터 주제어를 추출하고 이를 통해 문서를 주제별로 클러스터링하는 방법으로서 확률 그래프를 이용한다.

지난 몇 년간 확률 그래프 모델은 표현력과 계산 가능성 면에서 많은 발전을 하였으며 이에 따라 확률적 결정 모델링에 대한 많은 관심과 연구가 진행되어 왔다. 일반적으로 그래프 모델이 복잡할 경우, 모델 내에서의 정확한 확률적 추론은 그 계산 비용이 너무 크다. 따라서 이에 대한 근사 방법에 대한 연구가 진행되어 왔다. 본 논문에서는 학습 및 추론시 사용되는 근사 방법으로서 Helmholtz machine을 이용한다.

Helmholtz machine[1, 2]은 기본적으로 이진 데이터에 대해 적용되는 모델이다. 하지만 문서 데이터의 경우, 이진 데이터 즉 단어의 존재 유무만으로는 문서 데이터에 포함된 특징을 직절히 표현할 수 없다. 논문에서는 Helmholtz machine학습의 기본 골격을 그대로 유지하면서, 빈도수 데이터에 대한 분석을 진행할 수 있도록 하였다. 실험 결과 제안된 방법이 문서 집합으로부터의 주제어 추출이나 클러스터링 면에서 기존의 이진 데이터만을 사용하는 Helmholtz machine보다 성능 향상을 보임을 알 수 있었다.

2 장에서는 Helmholtz machine을 간략하게 소개하고 3 장에서는 빈도수 데이터를 고려하기 위해 고안된 수정된 Helmholtz machine을 소개한다. 4 장에서는 3 개의 실세계 문서 데이터에 대한 주제어 추출 및 클러스터링 실험 결과를 보이고 이를 분석하며 5 장에서는 결론과 앞으로의 연구 방향에 대해 서술한다.

2. Helmholtz Machine

Helmholtz machine [1, 2]은 확률 생성 모델에서의 학습과 추론 과정을 보다 용이하게 하기 위한 근사적 방법의 일종이다. Helmholtz machine은 하나의 생성 네트워크(generative network)와 인식 네트워크(recognition network)의 쌍으로 구성된다. 인식 모델은 하나의 데이터 또는 패턴이 주어질 때, 그 데이터에 내재된 특성들의 확률 분포를 추정하는데 이용되며, 생성 모델은 그 내부적 표현으로부터 입력 데이터를 추정함으로써 이러한 인식 모델을 학습시키는데 사용된다. 이러한 점에서 Helmholtz machine은 자기감독(self-supervised) 학습의 한 형태로 파악할 수 있다.

임의의 데이터 집합 $\mathbf{D}=\{d^{(1)}, d^{(2)}, \ldots, d^{(m)}\}$에 대해 생성된 모델의 적합성을 측정하기 위한 하나의 척도로 흔히 모델의 데이터에 대한 유사도(likelihood)가 사용된다. 모델의 파라미터 집합이 Θ로 주어지고 집합 $\mathbf{D}$에 속한 각 데이터 $d^{(i)}$가 서로 독립이고 동일한 분포를 따른다면, 유사도에 로그를 취한 수치, 즉 로그 유사도는 다음과 같다.

$$\log(\mathbf{D}|\boldsymbol{\Theta}) = \sum_{i=1}^{I} \log\left[\sum_{\alpha^{(i)}} P(d^{(i)}, \alpha^{(i)}|\boldsymbol{\Theta})\right] \quad (1)$$

위 식에서 $\alpha^{(i)}$는 데이터 $d^{(i)}$를 생성하는 은닉 요인을 의미하는데 모든 α에 대한 분포 P를 계산하는 작업은 은닉 요인의 수에 따라 그 복잡도가 지수적으로 증가하기 때문에 위 식에 대한 로그 유사도 최대화 계산 역시 아주 어렵게 된다. 그래서 보다 쉬운 형태의 분포를 도입함으로써 위 문제를 근사적으로 해결할 필요가 있는데 Jensen의 부등식을 적용하면 아래와 같은 부등식을 얻을 수 있다.

$$\begin{aligned}\log(D|\boldsymbol{\Theta}) &= \sum_{i=1}^{I} \log\left[\sum_{\alpha^{(i)}} P(d^{(i)}, \alpha^{(i)}|\boldsymbol{\Theta})\right] \\ &= \sum_{i=1}^{I} \log\left[\sum_{\alpha^{(i)}} Q(\alpha^{(i)}) \frac{P(d^{(i)}, \alpha^{(i)}|\boldsymbol{\Theta})}{Q(\alpha^{(i)})}\right] \\ &\geq \sum_{i=1}^{I} \sum_{\alpha^{(i)}} Q(\alpha^{(i)}) \log \frac{P(d^{(i)}, \alpha^{(i)}|\boldsymbol{\Theta})}{Q(\alpha^{(i)})} \quad (2)\end{aligned}$$

따라서 위 식 오른쪽의 하한값을 최대화시키면 로그 유사도에 대해서도 역시 최대화 작업을 수행할 수 있다. 이 방식은 *generalized EM*과 유사한 작업이라고 할 수 있다. 로그 유사도에 대한 하한값 추정을 가능하게 하는 Q를 학습하기 위한 것으로 인식네트워크가 사용된다. 그림은 Helmholtz machine의 간단한 예를 보인다.

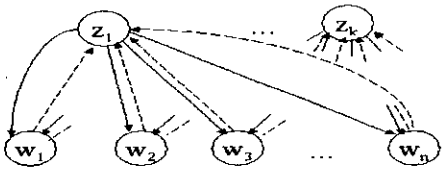

그림 1 기본적인 Helmholtz machine

3. 문서 데이터에 적용된 Helmholtz machine

기본적으로 Helmholtz machine의 입력 노드는 확률적 이진 노드이며, 생성 네트워크나 인식 네트워크 모두 각 층의 노드가 1이 될 확률 P는 식 (3)과 같이 sigmoid 함수에 의해 결정된다.

$$P(s_i = 1) = \frac{1}{1 + \exp(-\sum w_{ij} s_j)} \quad (3)$$

여기에서 s_j는 노드 j의 활성값이며, w_{ji}는 노드 j로부터노드 i에 이르는 연결선의 가중치를 의미한다. 하지만 일반적으로 생성 네트워크와 인식네트워크에서의 각 노드의 활성 함수가 반드시 같을 필요는 없다. 본 논문에서는 competitive 함수[3]를 생성 네트워크상에서의 입력 노드에 대한 활성함수로 사용하며 나머지에 대해서는 sigmoid함수를 취한다. Competitive 함수에 의한 확률 P는 아래와 같이 주어진다.

$$P(s_i = 1) = 1 - \frac{1}{1 + \sum s_j \exp(w_{ji})} \quad (4)$$

식 (4)에서 $\exp(w_{ji})$는 s_j=1 일때 s_i의 값 역시 1이 되는 경우에 대한 *odds ratio*로 해석된다. 그리고, 입력값으로 문서 내에 존재하는 단어들의 빈도수를 사용하기 위해, [9]에서 이용된 방법을 이용하여 입력노드는 그 복사본을 여러 개 가질 수 있는 것으로 가정하였다. 이를 그래프 구조로 표현하면 그림과 같다.

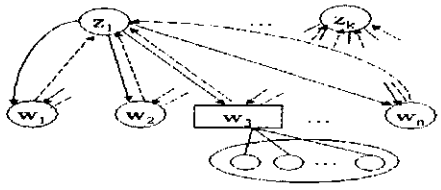

그림 2 변형된 Helmholtz machine의 구조

각 복사본들은 이진 노드이며 위 그림에서 w_3에 대한 연결선의 가중치는 각 복사본(replica)들이 서로 공유하게 된다. 만약 w_3의 출현 빈도가 3 이라면 3 개의 복사본이 생성된다. 이러한 구조에서 Helmholtz machine의 학습알고리즘인 wake-sleep 알고리즘 [5] 중 sleep 단계는 기존과 똑같이 수행되며, 다만 wake 단계에서 인식 네트워크를 이용하여 은닉 노드의 상태를 확률적으로 결정할 때, w_3가 해당 문서에 나타난 빈도수가 계산과정에 포함된다.

4. 실험

이 절에서는 Helmholtz machine을 이용하여 문서 집합을 주제에 따라 클러스터링하는 실험을 하고 그 결과를 분석한다. 특히 입력 노드가 이산값을 가지는 경우 기존의 이진 값을 가지는 경우에 대해 어떤 성능 향상을 보일 수 있는지를 분석한다.

4.1 문서 데이터

TREC-8 adhoc 데이터 집합은 총 50 개의 주제 및 주제에 대한 설명과 해당 문서 집합으로 구성되어 있다. 이 실험에서는 관련 문서가 비교적 많은 4 개의 주제를 선택하였으며, 문서의 총 개수는 1,069 개이다.

20 Newsgroup 데이터는 Lang [6]에 의해 수집된 데이터로서, 20 개의 주제에 대한 전자 뉴스 그룹에 등록된 문서들의 집합이다. 각 주제에는 1,000 개씩, 총 20,000 개의 문서로 구성되어 있다. 이 실험에서는 그 중 science와 recreation에 관련된 뉴스 그룹 4 개씩을 선택하여 각 뉴스 그룹당 500 개의 문서를 임의로 추출하여 실험하였다. 표 1은 각 데이터의 구성을 보여준다. 괄호 안의 숫자는 각 문서 집합을 구성하는 분서의 개수를 의미한다.

표 1 문서 집합의 구성

TREC (1069)	*Foreign minorities, Germany* *Estonia, economy* *Invention, scientific discoveries* *King Husayn, peace*
Science (2000)	*Sci.crypt, sci.electronics, sci.med, sci.space*
Recreation (2000)	*Rec.autos, rec.motorcycles* *rec.sport.baseball, rec.sport.hockey*

4.2 실험

각 문서 집합에 포함된 문서는 'bag-of-words' [7] 또는 'Boolean' 방식에 의해 벡터 공간상에서 표현된다. 이 때, 문서에 포함된 단어 중에서 숫자는 제외되며 알파벳이 아닌 단어 또한 제거되었다. 그리고 기본적인 불용어(stop-list)에 포함된 단어 역시 제외하였으며, 스테밍은 적용되지 않았다. 마지막으로, 식 (5)에 제시된 자질 선택 과정[8]을 통해 상위 2000 개의 단어만 문서의 해당 벡터를 구성하도록 하였다.

각 데이터 집합에 대해 Helmholtz machine은 하나의 은닉층과

입력층으로 구성하였다.

4.3 실험 결과

표 2에 세 개의 문서 집합에 대한 클러스터링 결과가 요약되어 있다. 괄호 안의 숫자는 Helmholtz machine의 경우 은닉노드의 수를 의미하며, K-means 알고리즘의 경우 에는 설정된 클러스터의 개수를 의미한다.

$$l(w) = p(d)\sum_{d\in D} p(w|d)\log\frac{p(w|d)}{p(w)},$$
$$p(d) = \frac{1}{m}, \quad p(w) = \frac{\sum_d n(w|d)}{\sum_{w'}\sum_d n(w'|d)},$$
$$p(w|d) = \frac{n(w|d)}{\sum_{w'} n(w'|d)} \qquad (5)$$

표 2 문서 집합에 대한 클러스터링 결과 (오류 %)

	Helmholtz machines		*K*-means
	Numeric	Binary	
TREC-8 adhoc	2.2 (4)	7.7 (4)	20.2 (8)
Science	14.9 (4) 13.6 (6)	16.2 (6)	32.1 (8).
Recreation	9.7 (5)	11.6 (5)	29.2 (8)

뉴스 그룹의 'Recreation' 데이터에 대한 실험에서 은닉 노드가 4 개일 때는 'autos'와 'motorcycle', 또는 'baseball'과 'hockey'에 대해 구분을 잘 하지 못하는 경향을 보였다. 그래서 은닉 노드 하나를 추가하였으며, 표에 나타난 결과는 이와 같이 은닉노드가 5 개로 설정했을 때의 결과를 보인 것이다. 실제 클러스터링 작업을 진행할 때, 5 개의 노드 중 하나는 제외된다. 그 이유는 아래 표를 참고하면 쉽게 알 수 있다.

표 3은 각 은닉노드에 의해 추출된 주제어를 보인 것이다. 해당 단어들은 각 은닉노드에 대한 확률값, 즉 $P(w|z)$가 높은 순서대로 나열되었다. 5 번째 항목의 경우, 나머지 항목과 비교하여 볼 때, 특별한 주제를 의미하지는 않음을 알 수 있다. 다만 뉴스 그룹 기사에 공통적으로 나타나는 부분만을 나타낼 뿐이다. 이러한 것은 Helmholtz machine의 Multiple cause mixture model[5]로서의 특성을 잘 나타내며, 기존의 다른 클러스터링 알고리즘과 다른 점이다.

표 3 'Recreation' 데이터에 대한 주제어 집합

	$P(w\|z)$에 따른 상위 10 개의 단어
1	bike, ride, good, riding, motorcycle, bmw, bikes, ama, road, rider
2	game, time, year, baseball, play, good, games, league, season, team
3	team, hockey, season, year, nhl, game, pittsburgh, toronto, play, fan
4	car, engine, good, cars, drive, people, speed, ford, make, price
5	**world, time, people, good, mail, make, canada, real, read, post**

'Science' 데이터 집합의 경우, 문서에 대해 단어의 존재 유무만을 고려하면 은닉 노드가 4 개, 5 개가 되어도 클러스터링 성능이 아주 좋지 않았다. 은닉 노드가 4 개, 5 개인 경우 정확율(accuracy)는 각각 48.5%, 58.6%에 불과하였다. 은닉노드를 6 개로 주었을 때, 비로소 어느 정도의 성능을 보임을 알 수 있었다. 이 경우에도 2 개의 은닉 노드는 뉴스 기사들에 일반적으로 사용되는 단어들을 표현하고 있었다. 하지만, 문서 내에 존재하는 단어의 빈도수를 고려하면 4 개의 은닉노드만으로도 충분히 주제어를 추출할 수 있었고 어느 정도의 클러스터링 성능을 보임을 알 수 있다. 따라서 문서 데이터에 대해 기존의 Helmholtz machine에서처럼 이진 값만 사용하는 것보다는 단어의 빈도수도 고려함으로써 성능 향상을 이룰 수 있음을 알 수 있다.

5. 결론

본 논문에서는 Multiple cause model을 이용하여 텍스트 문서로부터 주제어를 추출하고 문서들을 클러스터링하는 접근법을 제시하였으며, 기존의 Helmholtz machine에 빈도수 데이터를 이용할 수 있는 하나의 방안을 제시하였다. 향후 생성 네트워크에 대해 다양한 확률 모델을 연구하고 각 확률 모델이 어떤 특징을 가질 수 있는지에 대해 검토가 필요할 것이다.

감사의 글

본 연구는 과학기술부 뇌연구개발사업(BR-2-1-G-06)과 교육부 BK21 사업에 의하여 일부 지원되었음

참고 문헌

[1] P. Dayan and G. E. Hinton, Varieties of Helmholtz machine, *Neural Networks*, 9:1385-1403, 1996.
[2] P. Dayan, G.E. Hinton, R. M. Neal, and R. S. Zemel, The Helmholtz machine, *Neural Computation*, 7:889-904, 1995.
[3] P. Dayan and R. S Zemel, Competition and multiple cause models, *Neural Computations*, 7:889-904, 1995
[4] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, The MIT Press, 1998.
[5] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, The wake-sleep algorithm for unsupervised neural networks, *Science*, 268:1158-1161, 1995.
[6] K. Lang, Learning to filter netnews, In *Proceedings of the 12th International Conference on Machine Learning*, pages 331-339, 1995.
[7] M. Sahami, Using machine learning to improve information access, PhD thesis, Stanford University, 1998.
[8] N. Slonim and N. Tishby, Document clustering using word clusters via the information bottleneck method, In *Proceedings of the 23th International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 208-215, ACM Press, 2000.
[9] Y. W. The and G. E. Hinton, Rate-coded restricted Boltzmann machines for face recognition, *NIPS 2000*, 2000.