가상예제를 이용한 Naïve Bayes 분류기 성능 향상[1)]

이유정° 강병호 강재호 류광렬
부산대학교
{yjlee°, bhokang, jhkang, krryu}@pusan.ac.kr

Improving Performance for Naïve Bayes Classifier Using Virtual Examples

Yujung Lee,° Byoungho Kang, Jaeho Kang, Kwang Ryel Ryu
Busan, Pusan National University

요 약

기계학습에서 분류는 훈련 예제들로 학습하여 생성한 분류기를 활용하여 새로운 예제에 어느 한 범주를 부여하는 것을 말한다. 일반적으로 분류의 성능 즉 정확도의 향상은 학습 알고리즘을 개선하거나 훈련예제 집합을 변형시킴으로써 가능하다. 본 논문에서 소개하는 가상예제를 이용한 분류기 성능 향상 방안은 후자에 속한다. 실세계 분류문제에서 많은 수의 훈련예제들을 수집하는 일은 대상문제에 따라 비용이 많이 드는 경우가 있다. 또한 적은 수의 훈련예제를 학습해 생성한 분류기는 분류성능이 좋지 않을 수 있다. 본 논문에서는 이런 문제를 해결하기 위해서 가상예제를 생성해 훈련예제 집합에 추가하는 방안을 제안하고자 한다. 가상예제를 이용한 분류성능 향상방안이 naïve Bayes 학습 알고리즘 성능 개선에 효과가 있음을 실험을 통해 확인하였다.

1. 서 론

기계학습에서 분류란 훈련예제들로 학습하여 생성된 분류기를 활용하여 새로운 예제들에게 카테고리를 부여하는 것을 말한다.

일반적으로 분류성능을 향상시키는 방법은 크게 두 가지로 나눌 수 있다. 첫째는 학습 알고리즘을 개선하는 방법이다. 그 예로는 보다 성능이 좋은 학습알고리즘을 활용하는 방법, 각 학습 알고리즘 내에 있는 여러 가지 파라미터들을 조정하는 방법, 여러 학습 알고리즘을 섞어 쓰는 메타 분류기(meta classifier)를 활용하는 방법 등이 있다. 둘째는 훈련 예제집합을 변형시키는 방법이다. 그 예로는 잡음예제를 제거하는 방법, 수치속성을 이산화(discretization)하는 방법, 새로운 속성을 추가하는 방법(synthesis), 분류 성능에 도움이 되는 속성을 선택해서 학습하는 방법(attribute selection) 등이 있다. 가상예제를 이용한 분류 성능 향상 방안은 훈련 예제집합을 변형시키는 방법 중 하나이다. 본 논문에서 제안하는 방안을 간략히 설명하면, 먼저 후보 가상예제를 생성하고 훈련예제에 후보 가상예제를 추가했을 때 분류기의 성능 향상 정도를 추정한다. 이렇게 추정된 값을 바탕으로 분류 성능에 분류 성능에 도움이 되는 후보 가상예제를 선별한 후 훈련예제 집합에 추가하는 방법이다.

실세계 분류 문제에 있어서는 훈련예제 수집 양에 한계가 있고 대상문제에 따라 훈련예제를 수집하는 비용이 많이 드는 경우가 있다. 예를 들면 자동차 번호판 인식, 사람의 홍채, 지문 인식, 영상 인식 등이 있다. 앞의 문제들은 훈련예제들이 대부분 값 비싼 영상 데이터이거나 데이터 수집비용 자체가 크다. 훈련예제들을 충분히 수집하지 못한 경우, 생성된 분류기는 성능이 좋지 않을 수 있다. 본 논문에서 제안하는 가상예제를 이용한 분류 성능 향상방안은 훈련예제의 수가 제한적이거나 수집비용이 큰 분류 문제에서 분류 성능을 올리는 데에 효과가 있다.

본 논문은 2장에서 이전의 연구 중에서 가상예제와 관련한 연구들을 언급하고 3장에서 가상예제 추가방안의 전체적인 흐름, 가상예제의 생성방안, 그리고 가상예제의 평가방안을 서술하고 있다. 4장에서는 실제 이러한 가상예제 분류성능 향상에 기여하는지를 실험으로 확인하고 마지막으로 5장에서는 결론과 향후 연구에 대해 기술한다.

1) 이 논문은 한국과학재단 국가지정연구실사업의 지원으로 이루어진 것임.(Contract number: M10400000279-05J0000-27910)

2. 관련 연구

가상예제와 관련된 기존 연구로는 잡음 추가를 통한 신경망의 일반화 향상이 있다.[1][2] 잡음 추가란 신경망 학습시 사용되는 훈련예제에 고의로 잡음을 주는 방안으로 과부합을 방지하고 신경망을 일반화하는 데에 도움을 준다는 것이 입증되었다. 또한 가상예제을 이용하여 신경망의 일반화 능력을 향상시키는 연구가 있다. 이 연구는 예제가 희박한 영역에 가상예제을 생성시킴으로써 신경망의 일반화 능력을 향상시켰다.[3][4][5] 좋은 가상예제를 생성하기 위한 방안으로써 단순 가상예제 생성, bootstrap 가상예제 생성, 선별 가상예제 생성, 그리고 검증 가상예제 생성방안이 제안되어 비교 실험되었다.[5]

영상문제에서 가상예제에 대한 연구는 얼굴 인식 성능 향상을 위한 간단한 하이브리드 분류기를 가상의 훈련예제와 함께 사용하는 방안이다.[6] 이 연구에서는 실제 이미지로부터 추출된 거울 얼굴이미지(mirror face image)를 통해 가상예제를 생성하고 또는 동일한 카테고리 내의 실제 훈련예제로부터 무작위로 두 개의 예제를 선택한 후에 그 예제들의 평균을 구함으로써 가상 예제를 생성한다.

기존 연구들은 주로 신경망 분류 알고리즘의 일반화에 가상예제를 적용하여 분류기의 성능이 향상됨을 보였다. 그리고 예제의 속성이 수치 속성으로만 표현된 문제에만 적용되었다. 이전 연구들과는 달리 본 논문에서는 수치속성뿐 아니라 문자속성으로 표현된 문제에도 가상예제을 이용한 방안을 적용하고 naïve Bayes 분류 성능 향상 방안에 대해서 연구했다.

3. 가상예제 생성 및 평가방안

가상예제 추가방안이 포함된 전체 학습 과정은 그림 1과 같다. 먼저 입력으로 훈련예제 집합과 가상예제의 최대 개수가 주어진다. 먼저 후보 가상예제를 만들고 그 후보 가상예제에 대한 평가를 수행한다. 평가가 끝난 후 평가값이 0 이상이면 가상예제를 훈련예제 집합에 추가한다. 0과 같거나 0이하일 경우에는 가상예제를 훈련예제 집합에 추가하지 않고 다른 후보 가상예제를 생성하고 평가하는 과정을 반복한다. 이렇게 가상예제 한 개를 훈련예제 집합에 추가한 뒤에 학습을 통해 분류기를 생성한다. 훈련예제 집합에 추가된 가상예제의 개수가 입력으로 받은 가상예제 최대 개수가 될 때까지 이 과정을 반복한다.

```
function VIRTUAL-EXAMPLE-LEARNING returns classifier
    Inputs: T, v_maxcount
    v ; a candidate virtual example
    score ; evaluation score for a candidate virtual example
    T+ ; original training set including a virtual example
    T+ = T
    For j=0 until j< v_maxcount do
        v ← VIRTUAL-EXAMPLE-GENERATION(T)
        score ← VIRTUAL-EXAMPLE-EVALUATION(v,T)
        if score > 0
          then T+ ← T+ U {v}
              classifier ← NAÏVE-BAYES-LEARNING(T+)
    return classifier
```

그림 1. 가상예제 추가 방안이 적용된 학습과정

3.1 후보 가상예제 생성방안

본 절에서는 가상예제의 생성방안에 대해서 설명한다. 먼저 후보 가상예제의 카테고리가 결정된다. 그 다음 해당 카테고리의 훈련예제들이 훈련예제 집합 안에서 가지는 각 속성값 범위 내에서 무작위로 각 속성의 값을 결정한다.

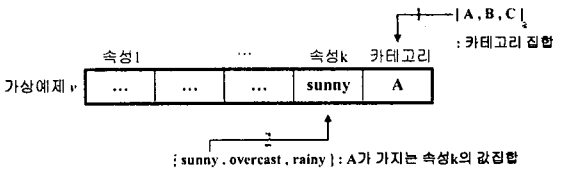

그림 2. 후보 가상예제 생성방안 예

그림 2에서 만약 A, B, C가 카테고리 집합이라고 하면, 카테고리가 A인 후보 가상예제를 한 번 생성하고 그 다음 생성 때에는 B인 후보 가상예제를 생성한다. A가 후보 가상예제의 카테고리로 결정되면, 후보 가상예제의 k번째 속성값은 카테고리가 A인 훈련예제들이 훈련예제 집합에서 가지는 k번째 속성값 집합인 sunny, overcast, rainy 중에서 무작위로 선택된다. 속성값을 무작위로 생성하는 이유는 좀 더 다양한 가상예제를 생성하여 분류에 도움이 되는지를 평가하기 위함이다. 다른 속성들도 동일한 방식으로 속성값을 결정한다.

3.2 후보 가상예제 평가방안

본 절에서는 가상예제의 평가함수를 정의하고 평가함수를 통해 후보 가상예제가 분류 성능 향상에 기여하는가를 판단하는 과정을 설명한다. 생성된 후보 가상예제는 훈련예제에 바로 추가되지 않고 후보 가상예제가 훈련예제에 추가되었을 때의 성능을 추정한 값을 통해 분류 성능 향상에 도움이 되는지를 평가한 후 훈련예제 집합에 추가한다.

그림 3에서 h_i는 i번째 훈련예제 t_i를 훈련예제 집합에서 빼고 학습한 결과 나온 분류기이고 h_i'는 t_i를 뺀 훈련예제 집합에 가상예제 v를 더하여 학습한 결과 나온 분류기이다. 두 개의 분류기를 t_i로 테스트 하면 분류기에서 t_i의 실제 카테고리에 대한 확률 p_i와 p_i'가 나온다. p_i'가 p_i보다 크면 가상예제를 훈련예제 집합에 추가해 학습한 분류기 h_i'의 성능이 t_i에 대해 가상예제를 추가하기 전보다 좋아졌다고 말할 수 있다. 이런 과정을 모든 훈련예제에 대해 수행한 후, p_i'와 p_i 차이의 평균을 취한 값이 후보 가상예제에 대한 평가점수가 된다. 결과적으로 이 평가점수가 0보다 크다면 후보 가상예제 v는 전체적으로 분류성능에 도움이 된다고 판단한다.

수식 (1)은 앞에서 설명한 가상예제 평가방안을 수식으로 정의해 놓았다. 수식에서 T는 훈련예제 집합이고 T_i는 훈련예제 집합 T에서 i번째 훈련예제를 뺀 훈련예제 집합이다, T_{i+v}는 T_i에 가상예제 v를 추가한 훈련예제 집합이다. c_i는 t_i의 실제 카테고리를 의미한다. 그리고

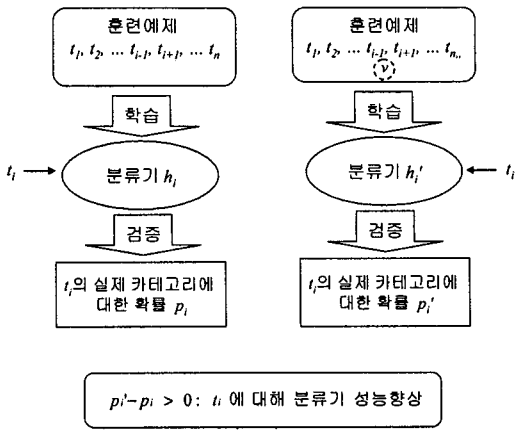

그림 3. 훈련예제 t_i를 통한 후보 가상예제 평가방안

h_{Ti}는 T_i를 학습한 분류기이다. Score(v)는 후보 가상예제

$$\text{Score}(v) = \frac{\sum_{i=1}^{|T|} \{p(t_i \in c_i \mid h_{Ti+v}) - p(t_i \in c_i \mid h_{Ti})\}}{|T|} \quad (1)$$

v가 분류 성능을 얼마만큼 향상시키는가를 추정한 평가점수이다. Score(v)는 그림 3과 동일하게 가상예제를 추가한 후의 $p(t_i \in c_i | h_{Ti+v})$와 추가하기 전의 $p(t_i \in c_i | h_{Ti})$의 차이를 분류 성능 향상정도로 보고 있다. 다시 말해 t_i의 실제 카테고리 c_i에 대한 확률이 높게 나오는 naïve Bayes 분류기가 성능이 더 좋다고 판단한다. 모든 훈련예제에 대해서 $p(t_i \in c_i | h_{Ti+v})$와 $p(t_i \in c_i | h_{Ti})$간 차이의 평균을 구하면 후보 가상예제 v가 분류 성능을 얼마만큼 향상시키는 지에 대한 추정값을 구할 수 있다. 추정된 Score(v) 점수가 0이상이면 후보 가상예제 v가 전체적으로 분류 성능 향상에 도움이 된다고 판단하고, 훈련예제 집합에 후보 가상예제 v를 추가한다. 하지만 Score(v) 점수가 0과 같거나 0 보다 작으면 분류성능 향상에 도움이 되지 않는다고 판단해 훈련예제에 후보 가상예제를 추가하지 않는다.

4. 실험 결과 및 분석

본 장에서 제안한 가상예제를 이용한 분류기 성능 향상 방안을 학습 알고리즘 naïve Bayes에 대해 실제 적용해 실험한 결과를 정리하고 분석한다. 가상예제를 이용한 naïve Bayes 분류기 성능 향상 방안에 대한 실험환경은 다음과 같다.

- naïve Bayes algorithm(NB)을 학습 알고리즘으로 사용 (supervised entropy-based discretization 적용)
- UCI Repository 26개 데이터에서 실험
- ten-fold cross validation 으로 10번 실험

표 1에서 보면 UCI 26개 데이터에 대한 데이터명, 예제 개수, 속성 개수(수치속성 수/문자속성 수), 클래스 개수를 표기하였다. 여기서 기본 비교대상으로 가상예제를 추가하지 않은 naïve Bayes (NB)를 사용하고 본 논문에서 제안한 가상예제를 이용한 naïve Bayes 분류기 성능 향상 방안을 NBVE라고 한다. 추가할 최대 가상예제 수를 훈련예제의 500%으로 가정하였다. NB는 naïve bayes에서 각 데이터별 분류 정확도이고, NBVE는 50~500%의 가상예제를 추가해 생성된 분류기의 정확도를 표기하였다. 마지막 Best는 50~500% 중 가장 좋았던 정확도를 표기하였다. 표에서 mean은 전체적인 향상정도를 보기 위해 모든 데이터들의 정확도를 평균한 것이다. 그리고 win/loss에서 win는 NBVE가 naïve Bayes보다 좋은 경우의 수이고 loss는 NBVE가 naïve Bayes보다 좋지 않은 경우의 수이다.

표에서 가상예제 비율이 50%인 win/loss/draw를 보면 20개의 데이

터에서 NBVE가 성능 향상을 보였다. 다른 가상예제 비율에 있어서도 많은 데이터에서 NBVE가 naïve Bayes보다 더 좋은 성능을 보였다. 종합적으로 Best를 표에서 보면 2가지 데이터를 제외하고는 대부분의 데이터에서 naïve Bayes의 분류 성능 개선이 일어났다. 또한 이산화된 수치속성과 문자속성에서 모두 성능향상을 보이는 것으로 보아 가상예제 추가방안은 수치속성과 문자속성의 구분 없이 적용이 가능하다는 것을 알았다.

5. 결론 및 향후 연구

실험을 통해 본 논문에서 제안된 가상예제를 이용한 naïve Bayes 성능향상이 있음을 UCI 데이터들을 대상으로 확인하였다. 그리고 문자속성과 수치속성에 관계없이 분류성능을 향상시키는 것을 알았다. 또한 본 논문에서 제안한 가상예제 추가방안은 가상예제를 생성하고 평가하는 데에 많은 비용이 들지 않는다. 이러한 이점을 훈련예제의 수집비용이 큰 분류 문제에 적용한다면 분류기 성능향상에 있어 비용절감의 효과를 기대할 수 있다.

가상예제에 대한 향후 연구로는 크게 3가지가 있다. 첫째로 가상예제 후보 생성방안에 대한 연구이다. 본 논문에서는 후보 가상예제를 무작위로 생성했다. 하지만 분류 성능 향상에 도움이 되는 후보 가상예제를 효율적으로 생성하는 방안에 대한 연구가 필요하다. 예를 들면 실제 훈련예제 집합의 예제분포를 기초로 하여 가상예제를 생성하는 방안 등이 있다. 둘째, 본 논문에서 주장한 평가방안을 다른 여러 학습알고리즘에도 적용하는 연구가 필요하다. naive Bayes와 같이 확률이 나오지는 않지만 확률과 비슷한 신뢰도가 나오는 여러 알고리즘에 본 논문에서 제안한 평가방안이 적용가능하다. 마지막으로 본 논문에서는 추가되는 가상예제 비율을 미리 가정했지만 각 데이터 별로 얼마만큼의 가상예제가 필요한지를 자동적으로 정할 수 있는 방안을 연구하는 것도 중요한 향후 연구 과제이다.

참고 문헌

[1] G. An. "The effects of adding noise during backpropagation training on a generalization performance" *Neural Computation*, 7(2):613-674, 1996.

[2] C.M. Bishop. "Training with noise is equivalent to tikhonov regularization." *Neural Computation*, 7(1):108-116, 1995.

[3] S. Cho. and K. Cha. "Evolution of Neural network training set through addition of virtual samples". In International Conference on Evolutionary Computation, page 685-688, Nagoya, Japen, 1996.

[4] S. Cho. M Jang, and S Chang "Virtual sample generation using a population of networks". Neural Processing Letters 5(2):83-89, 1997.

[5] 권유화, 조성준. "가상샘플 데이터를 이용한 신경망의 일반화 능력제고와 그 응용". 정보과학회논문지 제25권 제8호, 1998.

[6] Yeon-Sik, Ryu and Se-Young Oh, "SIMPLE Hybrid Classifier for Face Recognition with Generated Virtual Data", Pattern Recognition, 2002.

데이터	예제	속성 (수/문)	클래스	NB	NBVE (%)										
					50%	100%	150%	200%	250%	300%	350%	400%	450%	500%	Best
autos	205	16/10	7	64.88	70.02	72.06	72.73	73.62	74.35	74.70	75.76	76	76.59	**76.93**	**76.93**
vote	435	0/16	2	90.11	92.25	93.28	94.73	**96.44**	96.42	95.3	95.21	95.19	95.19	95.17	**96.44**
audiology	226	0/69	24	73.45	76.48	**78.27**	77.74	77.74	77.52	77.69	77.69	77.60	77.56	77.52	**78.27**
kr-vs-kp	3196	0/36	2	87.89	**93.97**	92.29	92.34	92.34	92.34	92.29	92.29	92.29	92.39	92.86	**93.97**
monks-2	601	0/6	2	56.80	57.97	58.89	59.47	59.93	60.34	60.41	60.35	60.52	60.35	**60.52**	**60.52**
labor	57	8/8	2	85.96	**89.02**	88.06	86.4	85.33	85.33	85.16	84.63	84.3	83.96	84.3	**89.02**
zoo	101	0/17	7	93.07	94.36	96.35	96.35	96.25	96.45	96.55	**97.25**	96.55	96.55	96.55	**97.25**
heart-stat	270	10/0	7	81.11	84	83.81	84	84.04	**85.14**	84.18	84.11	83.88	83.96	83.96	**85.14**
ionosphere	351	13/0	2	89.17	90.54	90.74	91.25	91.62	91.68	91.85	91.99	92.05	**93.14**	92.88	**93.14**
monks-1	556	0/6	2	77.42	76.33	76.57	77.89	78.89	78.45	78.23	79.46	79.99	80.78	**80.99**	**80.99**
hepatitis	155	6/13	4	83.23	**85.04**	84.59	83.95	83.16	82.46	81.87	81.53	81.47	81.09	81.1	**85.04**
lymph	148	3/15	4	84.46	84.89	**86.21**	85.98	86.03	86.02	86	85.68	85.43	84.35	84.57	**86.21**
glass	214	10/0	7	70.56	**72.07**	70.25	68.53	68.80	68.81	68.80	68.90	69.17	69.31	69.41	**72.07**
anneal	798	9/29	6	96.43	98.40	98.66	98.77	98.78	98.84	98.91	98.90	98.92	98.92	**98.95**	**98.95**
diabetes	768	8/0	2	74.35	75.27	75.44	75.45	**75.53**	75.47	75.32	75.30	75.31	75.32	75.35	**75.53**
credit-g	1000	7/13	2	76	73.69	73.78	74.05	74.41	74.36	75.54	**76.93**	76.57	75.49	74.83	**76.93**
iris	150	4/0	3	92.67	93.33	93.6	93.6	93.53	93.53	93.66	93.73	93.8	93.8	**93.8**	**93.8**
balance	625	4/0	3	72.32	72.35	**73.20**	72.19	71.46	71.41	71.27	71.22	71.35	71.25	71.17	**73.20**
breast-c	286	0/9	2	71.68	**72.44**	71.54	71.61	71.75	71.92	71.68	71.51	71.68	71.92	71.82	**72.44**
colic	368	7/15	2	79.62	**80.08**	78.93	78.23	77.15	75.74	75.36	74.70	73.86	73.29	72.69	**80.08**
credit-a	690	6/9	2	86.52	86.57	**87.01**	86.31	86.11	86	85.97	86.07	85.97	85.95	85.89	**87.01**
soybean	683	0/35	19	92.97	93.42	93.60	93.57	93.64	**93.76**	93.65	93.65	93.62	93.56	93.62	**93.76**
heart-h	900	6/7	5	84.01	83.89	**84.35**	83.55	83.38	83.08	83.11	82.77	82.80	82.83	82.70	**84.35**
breast-w	699	9/0	2	97.14	**96.91**	96.88	96.75	96.71	96.65	96.56	96.55	96.53	96.53	96.49	**96.91**
primary	339	0/17	25	50.15	**47.19**	46.57	45.84	44.93	44.16	44.34	43.87	44.16	43.54	43.51	**47.19**
monks-3	554	0/6	2	93.44	**93.44**	93.44	93.44	93.44	93.44	93.44	93.44	93.44	93.44	93.44	**93.44**
mean				80.97	82.05	**82.16**	82.10	82.12	82.03	81.99	82.02	82.02	81.92	81.93	**84**
win/loss					20/5	17/8	16/9	15/10	13/12	13/11	14/11	16/8	14/11	17/8	**23/2**

표 1. 가상예제 추가방안을 Naive Bayes 학습에 적용했을 때의 정확도