온라인 필기인식을 위한 증가하는 데이터를 이용한 앙상블 기법

김태준 [1O] 장하영 [1] 박정완 [2] 황성택 [2] 장병탁 [1]

서울대학교 컴퓨터공학부 [1] 삼성전자 DMC 연구소 [2]

{tjkim, hyjang}@bi.snu.ac.kr {timothy.park, shwang}@samsung.com btzhang@bi.snu.ac.kr

Ensemble Methods with increasing data for online handwriting recognition

Tae-Jun kim[1] Ha-Young Jang[1] Jeongwan Park[2] Seongtaek Hwang[2] Byoung-Tak Zhang[1]

Department of Computer Science and Engineering, Seoul National University[1]
DMC R&D Center, Samsung Electronics Co., LTD. [2]

요 약

모바일 기기의 대중화와 함께 필기체 인식의 중요성은 더욱 커지고 있다. 필기 데이터는 데이터에 존재하는 분산(variance)이 매우 크기 때문에 인식기를 학습시키기도 어렵고 학습시간도 길어진다는 문제점이 있다. 본 논문에서는 이러한 문제점들을 해결하기 위한 앙상블 기법을 제시하였다. 제안한 방법론은 모바일 기기를 통해서 축적되는 필기데이터를 효율적으로 이용하기 위하여 일정량의 데이터가 모일 때마다 새로운 약분류기(weak learner)를 추가함으로써 앙상블 모델을 구축한다. 필기체 인식을 위해서 많이 사용되는 인공신경망은 필기 데이터의 크기가 커짐에 따라서 데이터 내의 분산도 같이 커지는 문제로 인하여 학습 시간이 급격히 증가하게 되는데 앙상블 기법을 이용한 점진적 학습을 통해서 빠른 시간 안에 보다 효율적인 학습이 가능하게 된다. 또한 앙상블 기법의 적용으로 인해서 분산이 큰 데이터에서 일반적으로 발생하는 데이터 집합의 변화에 따른 급격한 성능 변화 문제 또한 해결이 가능하다.

1. 서 론

필기체 인식은 손으로 쓴 글씨를 종이나 사진, 터치 인터페이스 등을 통해서 입력 받아 인식하는 기술이다. 모바일 기기의 대중화로 터치 인터페이스의 중요성이 커지면서 필기체 인식은 중요한 사용자 인터페이스의 하나로 자리잡게 되었고, 모바일 기기 상에서의 필기체 인식의 중요성은 더욱 커지고 있다. 그러나 그림 1에서 볼 수 있는 것처럼 필기 데이터는 작성자의 필기 습관과 방법에 따라서 같은 문자라도 그 모양이 크게 변화하게 되며, 동일한 작성자의 경우에도 글씨를 입력하는 환경 및 방법에 따라서 모양 및 크기 등에서 큰 차이가 발생하게 된다. 이러한 특성으로 인해 필기 데이터는 데이터에 존재하는 분산이 매우 크고, 학습에 많은 어려움이 발생하게 된다.

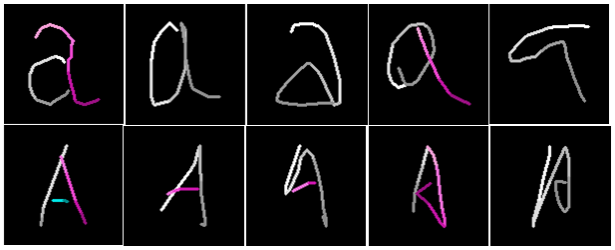


그림 1. 소문자 a와 대문자 A의 입력 예

이러한 문제점을 해결하기 위해서 최근 회귀형 신경망(recurrent neural network)[1]이나 심층학습(deep learning)[2]을 이용한 다양한 시도들이 좋은 결과를 보이고 있지만 막대한 계산능력과 오랜 학습 시간을 필요로 한다는 단점이 있다. 또한 기존의 방법들은 전체 데이터를 일괄처리(batch) 방식으로 학습하기 때문에 학습데이터의 변화에 적응하기 힘들다는 문제점이 있다.

본 논문에서는 이러한 문제점을 해결하기 위해서 앙상블 모델의 점진적인 학습을 통해서 증가하는 학습데이터를 효율적으로 이용할 수 있는 방법을 제안한다. 또한 제안하는 방법을 이용하여 분산이 큰 필기 데이터에 대해서도 안정적인 성능을 보이는 앙상블 모델을 빠른 시간 안에 학습할 수 있음을 실험 결과를 통해서 제시한다.

2. 점진적 학습을 위한 앙상블 기법

2.1 배깅(Bagging)

Bagging은 Breiman이 처음 제안한 방법으로 분류기의 예측 성능을 향상시킬 수 있는 방법 중 하나이다[3][4]. 전체 데이터 T에서 복원추출을 통하여 n개의 데이터 집합 T_1, T_2, ..., T_n을 생성한 후 이를 학습 데이터로 사용하여 약분류기 h_1, h_2, ..., h_n을 만든다. 이렇게 만들어진 약분류기들의 출력결과를 취합한 후 최종

결과를 도출하게 된다. 그림 2에 bagging 알고리즘의 수행과정이 설명되어 있다.

Bagging (T, L_b, M)
For each $m = 1,2,\ldots, M$
T_m = *Sample_With_Replacement*(T, N)
$h_m = L_b(T_m)$
Return $h_{fin}(x) = argmax_{y \in Y} \sum_m^M I(h_m(x) = y)$
Sample_With_Replacement (T, N)
$S = \varphi$
For $i = 1,2,\ldots, N$
r = *random_integer*$(1, N)$
Add T[r] to S
Return S

T : *original training set of* N *examples*
M : *# of base models to be learned*
L_b : *base model learning algorithm*
$I(A)$: *indicator function that returns 1 if event A is true and 0 otherwise*

그림 2. Bagging Algorithm

2.2 증가하는 데이터를 이용한 점진적 학습

일괄처리 방식으로 학습이 이루어지는 일반적인 bagging 알고리즘은 학습데이터가 지속적으로 증가하는 상황에는 적용하기가 쉽지 않다. 일괄처리 방식이 아닌 온라인 학습을 이용한 온라인 bagging 기법이 Oza[5]에 의해서 제안되었지만 Oza가 제안한 방법 역시 고정된 학습데이터를 이용하여 포아송 분포에 기반한 복원추출을 통해서 새로운 데이터 집합을 만든 후에 학습만 온라인 방식으로 진행하였다. 복원 추출을 통해서 새로운 데이터 집합을 만드는 일반적인 bagging 기법과 달리 본 논문에서 제안한 방법은 축적되는 데이터가 일정 크기에 도달했을 때, 이 축적된 데이터를 학습 데이터 집합으로 이용하여 새로운 약분류기를 만든다. 각 약분류기들은 테스트 데이터에 대해 인식한 결과를 내고, 최종 결과는 다수결(Majority Voting)을 통해 결정(Decision)한다.

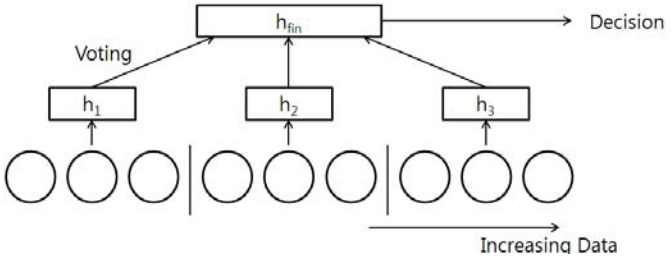


그림 3. 점진적 학습을 통한 앙상블 모델 구축

그림 3에 제안한 방법론을 이용한 앙상블 모델의 구축 과정이 설명되어 있다. 제안한 방법론은 데이터가 지속적으로 증가하는 온라인 환경에서 추가되는 데이터들을 효율적으로 이용할 수 있을 뿐 아니라 전체 모델을 새로 학습하지 않기 때문에 빠른 시간에 학습이 가능하다. 또한 서로 다른 분포를 가진 학습데이터를 통해 만들어진 약분류기들을 이용하는 앙상블 기법의 특성으로 인해서 과적합(overfitting) 문제가 발생하지 않고, 분산이 큰 데이터에 대해서도 좋은 성능을 보인다는 장점을 가지고 있다.

3. 실험 및 결과

제안한 방법론의 성능을 평가하기 위해서 다수의 사용자로부터 수집된 138,084개의 온라인 필기 데이터를 학습 데이터로 사용하였고, 99,353개의 UNIPEN[6] Train-R01/V07 데이터를 테스트 데이터로 사용하였다. 각각의 데이터는 대소문자의 모양이 동일한 일부 경우와 동일한 형태의 알파벳이 존재하는 숫자 0,1을 제외한 숫자와 알파벳 대소문자, 그리고 ?!@ 3종류의 특수문자로 이루어진 57개의 클래스로 구성되어 있다. 표 1에 UNIPEN 데이터에 대한 보다 자세한 설명과 각각의 데이터에 대한 기존 연구의 성능이 나와있다(1a: 숫자, 1b: 대문자, 1c: 소문자, 1d: 특수문자, 2: 혼합, 3: 혼합).[7]

표 1. UNIPEN 데이터 구성 및 성능

Category	# of Data	Accuracy (%)
1a	15953	96.4
1b	28069	91.3
1c	61360	81.2
1d	17286	73.6
2	122668	**72.6**
3	67352	**72.6**

본 논문에서는 데이터가 지속적으로 축적되는 상황을 가정하기 위하여 전체 학습데이터를 다섯 개로 나누어서 각각에 대하여 약분류기를 학습시킨 후에 이를 이용하여 앙상블 모델을 구축하고 그 결과를 확인해 보았다. 앙상블 모델을 위한 약분류기로는 인공신경망을 사용하였다. 다섯 개로 나눈 각각의 데이터집합에 대한 분석 결과가 그림 4에 나와 있다. 사각형점이 있는 실선이 학습에러이고, 테스트데이터1은 전체 학습데이터를 이용한 테스트 결과, 테스트데이터2는 UNIPEN 데이터를 이용한 테스트결과이다. 전체 데이터집합을 분류하여 새로운 데이터 집합을 만들어내는 경우에는 각 데이터집합의 분포를 균등하게 유지하는 것이 일반적인 방법이지만 본 실험에서는 다양한 분포를 가진 학습데이터가 추가되는 상황에서도 제안한 방법이 잘 동작하는지 확인하기 위하여 학습데이터의 분포에 차이를 두었다.

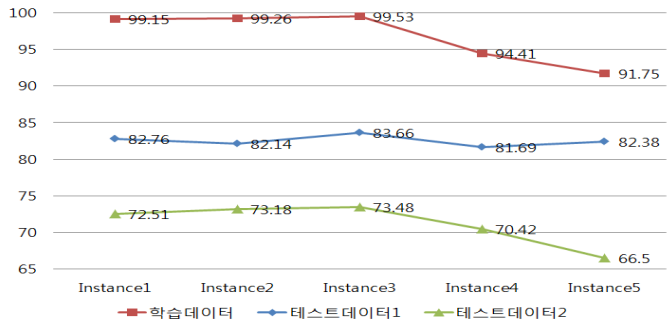
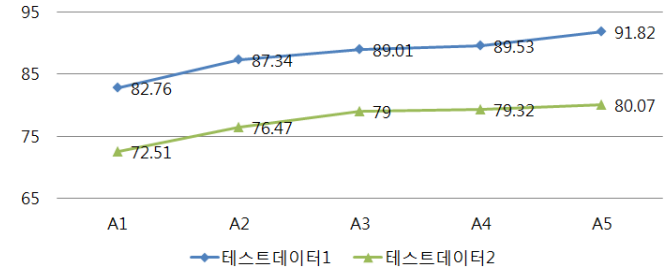
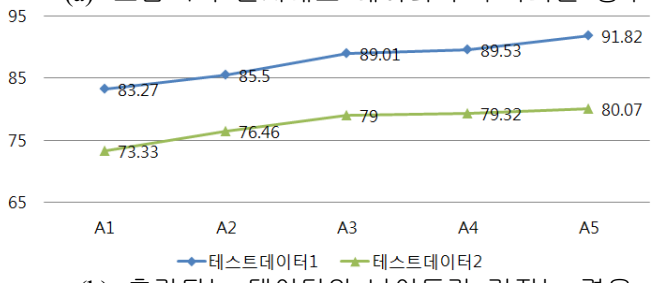


그림 4. 데이터 집합 별 성능

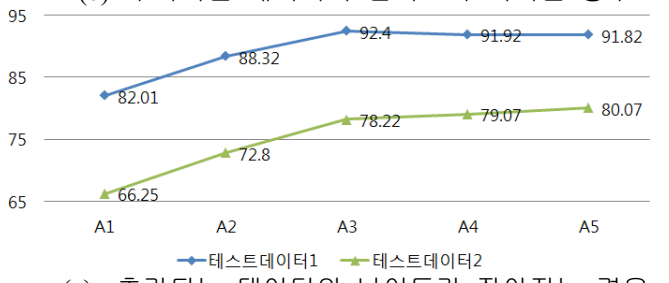
그림 5에 제안한 앙상블 모델을 이용한 성능이 나와 있다. 제안한 모델이 데이터 분포의 변화에 영향을 받지 않고, 일정한 성능을 보인다는 것을 확인하기 위해서 실험은 그림에서와 같이 다섯 개로 나누어진 데이터의 순서를 바꾸어서 진행해 보았다. 순서를 바꾼 세가지 실험 모두에서 새로운 약분류기가 추가됨에 따라서 성능이 향상되는 모습을 확인할 수 있다.



(a) 그림 4의 순서대로 데이터가 추가되는 경우



(b) 추가되는 데이터의 난이도가 커지는 경우



(c) 추가되는 데이터의 난이도가 작아지는 경우

그림 5. 앙상블 모델의 성능.

또한 전체 학습데이터를 사용한 인공신경망이 제안한 모델과 같은 정도의 학습에러를 얻기 위해서 7일 정도의 학습시간이 걸리고 테스트 데이터에 대한 정확도가 76% 정도밖에 나오지 않는데 반해, 제안한 모델은 각각의 분류기를 학습하는데 4-5시간 정도밖에 걸리지 않고 테스트 데이터에 대한 정확도도 80% 이상의 성능을 보이고 있다. 표 1의 내용에서도 알 수 있듯이 기존 연구결과가 숫자와 대소문자, 특수문자 등이 혼재되어 있는 경우에 72% 정도의 성능밖에 나오지 않는 것에 비해서 제안된 모델은 10% 이상의 향상된 성능을 보이고 있음을 확인할 수 있다

4. 결론

본 논문에서는 앙상블 모델의 점진적인 학습을 통해서 증가하는 학습데이터를 효율적으로 이용할 수 있는 방법을 제안하였다. 제안된 모델은 분산이 큰 실제 필기 데이터를 이용한 실험에서 단일 분류기보다 학습시간과 성능 모두에서 좋은 결과를 보여주었을 뿐만 아니라 불균형 데이터인 필기체 데이터의 단점을 극복할 수 있는 부스팅(boosting) 기법과의 결합으로 더욱 좋은 성능을 낼 수 있을 것이라 기대된다.

감사의 글

이 논문은 삼성전자와 한국연구재단의 지원(NRF-2010-0017734, NRF-2013M3B5A2035921)을 일부 받았음.

참고문헌

[1] Graves, A., Schmidhuber, J., Offline handwriting recognition with multidimensional recurrent neural networks, *Advances in Neural Information Processing Systems*, 21:545-552, 2009.

[2] Ciresan, D., Meier, U., Schmidhuber, J., Multi-column deep neural networks for image classification, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on,* p. 3642-3649, 2012.

[3] Quinlan, J. R., Bagging, boosting, and C4. 5, *Proceedings of the National Conference on Artificial Intelligence*, p.725-730, 1996.

[4] Buhlmann, P., Yu, B., Analyzing bagging, *The Annals of Statistics*, 30(4):927-961, 2002.

[5] Oza, N. C., Online bagging and boosting, *Systems, man and cybernetics, 2005 IEEE international conference on*, p.2340-2345, 2005.

[6] Guyon, I., et al., UNIPEN project of on-line data exchange and recognizer benchmarks, *Pattern Recognition,Vol.2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, p.29-33, 1994.

[7] Ratzlaff, E. H., Methods, reports and survey for the comparison of diverse isolated character recognition results on the UNIPEN database, *Document Analysis and Recognition, Proceedings, Seventh International Conference on*, p. 623-628, 2003.