불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법

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

서울대학교 컴퓨터공학부 [1] 삼성전자 DMC 연구소 [2]

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

Oversampling-Based Ensemble Learning Methods for Imbalanced Data

Kyung-Min 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

필기체 낱글자 인식을 위해서 사용되는 데이터는 대개 다수의 사용자들로부터 수집된 자연어 문장들을 이용하기 때문에 해당 언어의 언어적 특성에 따라서 낱글자의 종류별 개수 차이가 매우 큰 특징이 있다. 일반적인 기계학습 문제에서 학습데이터의 불균형 문제는 성능을 저하시키는 중요한 요인으로 작용하지만, 필기체 인식에서는 데이터 자체의 높은 분산과 비슷한 모양의 낱글자 등이 성능 저하의 주요 인이라 생각하기 때문에 이를 크게 고려하지 않고 있다. 본 논문에서는 이러한 데이터의 불균형 문제를 해결하기 위한 과표본화 기반의 앙상블 학습 기법을 제안하였다. 제안한 방법은 데이터의 불균형 문제를 고려하지 않은 방법보다 전체적으로 향상된 성능을 보일 뿐만 아니라 데이터의 개수가 부족한 낱글자들의 분류성능에 있어서도 향상된 결과를 보여주었다.

1. 서 론

일반적인 기계학습 기법들은 학습데이터가 클래스별로 비슷한 비율로 구성되어 있다고 가정하고 학습을 진행하게 된다. 그러나 많은 실세계 문제들이 불균형 데이터(imbalanced data) 문제에 속하게 되고 이러한 경우 소수 범주의 클래스들은 다수 범주의 클래스보다 잘 못 분류될 가능성이 높다 [1]. 이러한 부작용(side-effect)은 기계학습 알고리즘의 설계 특성상 각 클래스의 상대적인 분포를 고려하는 대신 전반적인 성능을 최적화 시키려하기 때문에 결정트리나 다층 퍼셉트론과 같은 분류기에서 흔히 나타난다 [1, 2].

필기체 인식의 경우도 언어적 특성에 따라서 글자별 데이터의 비율이 크게 다른 전형적인 불균형 데이터 문제로 볼 수가 있다. 예를 들면 자주 쓰여지는 알파벳인 a, o, i, e와 같은 소문자는 학습 데이터에서 차지하는 빈도가 높은 반면, Y, N, L과 같은 대문자는 자주 사용되지 않아 학습데이터에서 차지하는 빈도가 낮다. 이와 같이 데이터의 분포가 불균형한 상태에서 학습을 진행하게 되면 인식기는 훈련 데이터에서 차지하는 빈도가 높은 데이터에 과적응하게 된다.

그러나 필기체인식의 경우 이러한 데이터 불균형 문제보다 데이터 자체의 높은 분산과 유사한 모양의 글자들 간의 분류 문제 등이 전체적인 성능에 더 큰 영향을 미친다고 알려져있기 때문에 데이터의 불균형 문제를 크게 고려하지 않는다. 그러나 높은 빈도의 데이터에 대한 과적응은 학습 초기에 모델의 성능을 높이는데는 효율적일 수 있지만, 일정 정도 이상의 성능을 보이는 모델에서는 결국 성능 향상의 장애 요인으로 작용할 수 밖에 없다.

이러한 문제점을 해결하기 위해서 본 논문에서는 과표본화에 기반한 앙상블 학습 기법을 제안하고, 기법의 효과를 필기체 데이터를 이용하여 보여주었다.

본 논문의 구성은 다음과 같다. 2장에서는 과표본화 기반의 앙상블 학습 기법을 제안하고 3장에서는 실험 결과를 보이고 4장에서는 결론을 맺고 향후 연구방향을 모색한다.

2. 불균형 데이터 처리를 위한 앙상블 기법

2.1 과표본화 기법

과표본은 샘플링 기법의 한 방법으로 소수 범주의 집합 S_{min}에서 무작위로 데이터를 추출하여 집합 E를 만들고 이를 기존 집합 S에 더하는 과정으로 이뤄진다. 이러한 과정을 거쳐 S_{min}의 데이터 개수는 $|E|$만큼 증가하게 되고 집합 S의 클래스 분포는 그에 따라 조절이 된다 [3]. 이 방법은 모든 데이터를 사용할 수 있다는 장점이 있는 반면, 데이터의 수를 증가시켜

계산에 필요한 시간이 커지거나 복제되는 데이터에 분류기가 과적응 할 수 있다는 단점이 있다.

데이터를 단순 복제하는 대신 지능적으로 과표본화 기법을 사용한 대표적 연구로 Chawla가 제안한 Synthetic Minority Oversampling Technique(SMOTE)가 있다 [4]. SMOTE는 기존에 있는 데이터를 복제하는 대신 소수 범주 클래스의 데이터들을 서로 보간하여 새로운 인공적인 데이터를 합성했다. 이 기법은 먼저 k 근접 이웃(k-nearest neighbor) 알고리즘을 사용해 소수 범주 클래스의 데이터들과 가장 가까운 데이터들을 찾은 뒤 새로 합성되는 데이터가 그 성향을 반영하도록 했다.

Hui Han은 SMOTE를 수정한 기법인 borderline-SMOTE(BSM)을 제안했다 [5]. SMOTE가 소수 범주 클래스의 모든 데이터를 대상으로 기법을 적용했던 반면, BSM은 클래스의 결정 영역(decision region)에 있는 데이터들에만 기법을 적용시켰다.

다양한 샘플링 기법들의 성능을 비교해본 결과 이러한 지능적인 기법들을 사용한 [4,5]보다 오히려 단순 복제를 사용한 과표본화 기법이 더 좋은 분류 성능을 내는 경우가 많다는 연구 결과도 있다 [6]. 또한 [6]은 분류기의 성능을 높이기 위해서는 샘플링이 매우 중요한 요소 중 하나임을 확인했다.

2.2 과표본화 기반 앙상블 학습 기법

불균형 데이터를 사용한 학습 과정에서는 일반적으로 관측수가 많은 클래스의 데이터가 지배적인 영향을 미치기 때문에 학습된 모델의 성능 저하가 발생하게 된다. 이를 해결하기 위해 사용하는 과표본화 기법의 경우에는 데이터의 불균형 정도에 따라서 표본화된 데이터 크기의 급격한 증가로 인해 학습에 어려움이 발생한다는 문제점과 함께 표본화된 데이터의 분포가 원래 데이터의 분포와 달라진다는 문제점이 있다. 이를 해결하기 위해서 본 논문에서는 과표본화에 기반한 앙상블 학습 기법을 제안하였다.

제안한 방법은 각각의 클래스에서 동일한 횟수만큼 복원 추출하여 만들어진 전체 데이터의 부분집합들을 이용하여 앙상블 모델을 구축함으로써 기존의 과표본화 기법에서 발생할 수 있는 복제된 데이터에 대한 과적응 문제의 해결이 가능하다. 또한 과표본화로 인한 전체 데이터의 크기 증가로 인한 학습시간의 증가 문제도 앙상블 모델을 이용함으로써 해결이 가능하다 [7]. 또한 앙상블 모델의 구축을 위하여 전체데이터의 부분집합을 생성하는 과정은 언더샘플링(undersampling)의 경우와 유사하게 다수 클래스의 데이터 일부만을 사용하지만, 이를 이용하여 학습된 약분류기들의 조합으로 앙상블 모델을 구축하기 때문에 전체 모델의 관점에서는 모든 데이터를 사용한 것과 같은 효과를 얻을 수 있어 언더샘플링 과정에서 흔히 발생하는 데이터의 정보 손실 문제가 발생하지 않는다.

다시 말하면 언더샘플링된 데이터의 앙상블로 과표본

Oversampling_Based_Ensemble_Learning(T, L_b, M)

For each $m = 1,2,...,M$

$T_m = Oversampling(T, N)$

$h_m = L_b(T_m)$

Return $h_{fin}(x) = argmax_{y \in Y} \sum_m^M I(h_m(x) = y)$

Oversampling(T, N)

$S = \varphi$

For each class in T

For $i = 1,2,...,N$

$r = random\_integer(1, N)$

Add $T[r]$ *to* S

Return S

T : *original training set*

N : *# of sampling*

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 1. 과표본화 기반 앙상블 학습 기법

를 구현함으로써 언더샘플링에서 발생하는 데이터의 손실을 피할 수 있을 뿐만 아니라 과표본화로 인해서 발생하는 과적응이나 학습시간의 증가 등과 같은 학습의 어려움도 피할 수 있다. 제안한 방법론의 수행과정이 그림 1에 나타나 있다.

3. 실험 및 결과

3.1 데이터

제안한 방법론의 성능을 평가하기 위해 다수의 사용자로부터 수집된 238,450개의 필기체 데이터를 훈련 데이터로 사용하였고, 9만여개의 UNIPEN[8] Train-R01/V07 데이터를 테스트 데이터로 사용하였다. 데이터의 클래스는 모두 50개로 소문자 알파벳 26개와 대문자 알파벳 16개, 숫자 8개이다. 클래스당 평균 데이터개수는 4,769개이고 데이터를 가장 많이 포함하고 있는 클래스 상위 10개와 가장 적게 포함하고 있는 클래스 하위 10개가 표 1, 2에 나타나 있다.

표1. 데이터가 개수가 가장 많은 낱글자 10개

클래스	l	h	i	r	s	n	t	o	a	e
데이터 개수	8842	9267	10120	11601	11655	12427	13199	14834	18417	19240

표2. 데이터가 개수가 가장 적은 낱글자 10개

클래스	2	3	9	7	6	5	8	4	Y	N
데이터 개수	446	642	670	843	845	872	888	917	1283	1317

3.2 실험결과

본 논문에서 제안하는 과표본화 기반 앙상블 학습 기법의 성능을 측정하기 위해 제안한 방법을 가지고 학습한 모델의 성능과 전체 데이터를 가지고 학습한 모델과 비교해 보았다. 학습을 위한 분류기는 인공신경망을 이용하였고, 앙상블 모델은 배깅(Bagging)을 이용하였다 [7]. 과표본화 기반 앙상블 학습 기법을 적용 했을 때 분류기의 평균 정확도는 81.79%였고 단일 분류기의 평균 정확도는 77.564%였다. 제안한 기법을 적용했을 경우 기존의 모델에 비해서 4% 이상 향상되었음을 확인할 수 있다.

또한 표3에서 볼 수 있듯이 데이터의 개수가 많은 상위 10개 클래스에서는 일부를 제외하고 성능이 향상되거나 약간 저하되었지만, 데이터의 개수가 적은 하위 10개 낱글자에서는 모든 경우에 성능이 큰 폭으로 향상되었음을 확인할 수 있다. 클래스 별 분류 정확도의 상승 폭을 보면 하위 낱글자의 대부분 낱글자가 10%이상 정확도가 향상되었고 낱글자 '4'의 경우에는 28.2%까지 향상되었다. 위와 같은 결과를 통해서 데이터의 개수가 부족한 클래스의 분류 성능이 향상될 뿐만이 아니라 전체적인 분류 성능도 향상됨을 확인할 수 있다.

표3. 낱글자 별 정확도

(a) 상위 10개 낱글자의 정확도

클래스	단일분류기	과표본화기반 앙상블
L	61.63	**76.36**
H	**88.20**	88.00
I	78.27	**82.74**
R	76.02	**78.81**
S	95.42	**97.07**
n	**76.99**	61.57
t	84.42	**85.91**
o	**89.27**	88.39
a	**87.97**	80.18
e	**94.04**	92.76

(b). 하위 10개 낱글자의 정확도

클래스	단일분류기	과표본화기반 앙상블
2	47.84	**57.01**
3	85.33	**97.15**
9	30.89	**51.29**
7	42.83	**64.32**
6	67.76	**81.93**
5	50.82	**66.68**
8	46.06	**74.91**
4	50.03	**78.22**
Y	28.05	**39.64**
N	71.13	**80.44**

4. 결론 및 향후 연구

본 논문에서는 필기체 데이터에서 데이터의 분포가 불균형한 문제를 해결하기 위해 과표본화 기반 앙상블 학습 기법을 제안하였다. 이 기법을 적용한 결과 데이터에서 개수가 부족한 낱글자들의 분류 성능을 올릴 수 있었고 전체적인 평균 분류 성능도 향상될 수 있음을 확인했다. 제안한 방법론은 앙상블 모델을 이용한 과표본화 기법을 구현함으로써 표본화 기법들 간의 단점을 배제한체 각각의 장점만을 구현할 수 있는 기법으로써 보다 다양한 불균형 데이터에 적용이 가능할 것으로 예상된다.

감사의 글

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

참 고 문 헌

[1] S. Ertekin, J. Huang, L. Bottou, L. Giles, Learning on the border: active learning in imbalanced data classification, *CIKM '07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 127–136, 2007.

[2] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[3] H. He, E. A. Garcia, Learning from Imbalanced Data, *IEEE Transactions on knowledge and data engineering*, vol. 21, No 9, pp. 1236-1284, 2009.

[4] N. V. Chawla, L. O. Hall, K. W. Bowyer, W. P. Kegelmeyer, Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research 16*, 321–357, 2002.

[5] H. Han, W. Wang, B. Mao, Borderlinesmote: A new over-sampling method in imbalanced data sets learning. *In International Conference on Intelligent Computing*, pp. 878–887, 2005.

[6] J. V. Hulse, T. M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, *In Proc. of the 24th International Conference on Machine Learning*, pp.935-942, 2007.

[7] 김태준, 장하영, 박정완, 황성택, 장병탁, 온라인 필기 인식을 위한 증가하는 데이터를 이용한 앙상블 기법, 2013 한국컴퓨터종합학술대회(KCC2013)논문집, pp. 1396-1398, 2013.

[8] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, 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.