ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol 20, No 10, pp 549-554, 2014 10
http //dx doi org/10 5626/KTCP 2014 20 10 549

불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법

(Oversampling-Based Ensemble Learning Methods for Imbalanced Data)

김 경 민 † 장 하 영 † 장 병 탁 ††
(Kyung Min Kim) (Ha Young Jang) (Byoung Tak Zhang)

요 약 필기체 낱글자 인식을 위해서 사용되는 데이터는 일반적으로 다수의 사용자들로부터 수집된 자연언어 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 주요인이라 생각하기 때문에 이를 크게 고려하지 않고 있다. 본 논문에서는 이러한 데이터의 불균형 문제를 고려하여 필기체 인식기의 성능을 향상시킬 수 있는 과표본화 기반의 앙상블 학습 기법을 제안한다. 제안한 방법은 데이터의 불균형 문제를 고려하지 않은 방법보다 전체적으로 향상된 성능을 보일 뿐만 아니라 데이터의 개수가 부족한 낱글자들의 분류성능에 있어서도 향상된 결과를 보여준다.

키워드: 앙상블, 필기체 인식, 불균형 데이터, 표본화 기법

Abstract Handwritten character recognition data is usually imbalanced because it is collected from the natural language sentences written by different writers. The imbalanced data can cause seriously negative effect on the performance of most of machine learning algorithms. But this problem is typically ignored in handwritten character recognition, because it is considered that most of difficulties in handwritten character recognition is caused by the high variance in data set and similar shapes between characters. We propose the oversampling based ensemble learning methods to solve imbalanced data problem in handwritten character recognition and to improve the recognition accuracy. Also we show that proposed method achieved improvements in recognition accuracy of minor classes as well as overall recognition accuracy empirically.

Keywords: ensemble method, handwritten character recognition, imbalanced data, sampling method

· 본 연구는 삼성전자와 한국연구재단의 지원(NRF 2010 0017734)을 일부 받았음
· 이 논문은 제40회 추계학술발표회에서 '불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울대학교 컴퓨터공학과
kmkim@bi snu ac kr
hyjang@bi snu ac kr
†† 종신회원 : 서울대학교 컴퓨터공학과 교수(Seoul National Univ)
btzhang@bi snu ac kr
(Corresponding author임)

논문접수 : 2014년 1월 23일
(Received 23 January 2014)
논문수정 : 2014년 9월 3일
(Revised 3 September 2014)
심사완료 : 2014년 9월 9일
(Accepted 9 September 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제10호(2014 10)

1 서 론

일반적인 기계학습 기법들은 학습데이터가 범주 별로 비슷한 비율로 구성되어 있다고 가정하고 학습을 진행하게 된다. 그러나 많은 실세계 문제들이 불균형 데이터(imbalanced data) 문제에 속하게 되고 이러한 경우 소수 범주에 속한 데이터들은 다수 범주에 속한 데이터보다 잘못 분류될 가능성이 높다[1]. 이러한 부작용(side effect)은 기계학습 알고리즘의 설계 특성상 각 범주의 상대적인 분포를 고려하는 대신 전반적인 성능을 최적화 시키려하기 때문에 발생하는 것으로 결정트리(decision tree)나 다층 퍼셉트론(multilayer perceptron)과 같은 분류기에서 흔히 나타난다[1,2].

필기체 인식의 경우도 언어적 특성에 따라서 범주 별(글자 별) 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 과적응(overfitting)하게 되는 문제가 발생하게 된다.

그러나 필기체 인식의 경우 이러한 데이터 불균형 문제보다 데이터 자체의 높은 분산과 유사한 모양의 글자들 간의 분류 문제 등이 전체적인 성능에 더 큰 영향을 미친다고 알려져 있기 때문에 데이터의 불균형 문제를 크게 고려하지 않는다. 그러나 높은 빈도의 데이터에 대한 과적응은 학습 초기에 모델의 성능을 빨리 향상시키는 데에는 효율적일 수 있지만, 일정 정도 이상의 성능을 보이는 모델에서는 결국 최종적인 성능 향상의 장애요인으로 작용할 수 밖에 없다. 이러한 문제점을 해결하기 위해서 본 논문에서는 과표본화에 기반한 앙상블 학습 기법을 제안하고, 이를 이용한 필기체 인식기의 성능 향상을 실험적으로 보여주었다.

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

2 불균형 데이터 처리를 위한 앙상블 기법

2 1 과표본화 기법

과표본화는 샘플링 기법의 한 방법으로 소수 범주의 집합 S_{minor}에서 무작위로 데이터를 추출하여 집합 E를 만들고 이를 기존 집합 S에 더하는 과정으로 이뤄진다. 이러한 과정을 거쳐 S_{minor}의 데이터 개수는 $|E|$만큼 증가하게 되고 집합 S의 범주 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 과적응할 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 범주의 데이터들을 서로 보간하여 새로운 인공적인 데이터를 합성하였다. 이 기법은 먼저 k 근접 이웃(k nearest neighbor) 알고리즘을 사용해 소수 범주의 데이터들과 가장 가까운 데이터들을 찾은 뒤 새로 합성된 데이터가 그 성향을 반영하도록 하였다.

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

SMOTE 계열 외에 다른 과표본화 기법으로는 ADASYN(Adaptive Synthetic Sampling)이 있다. ADASYN은 데이터들의 밀도 분포인 Γ를 계산하여 범주마다 다른 양의 데이터를 합성했다[6].

다양한 표본화 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[7]. 또한 [7]은 분류기의 성능을 높이기 위해서는 표본화가 매우 중요한 요소 중 하나임을 확인하였다.

2 2 언더샘플링 기법

언더샘플링 기법은 다수 범주의 집합 S_{major}에서 무작위로 데이터를 추출하여 $|E| < |S|$를 만족하는 집합 E를 만든 뒤 이를 기존 집합 S에서 제거하는 방식으로 이뤄진다. 언더샘플링 결과 집합 S의 크기는 집합E의 크기, $|E|$ 만큼 줄어들게 된다. 언더샘플링 기법은 데이터의 크기가 매우 클 때 효과적이지만 데이터의 일부를 버림으로써 정보가 손실된다는 단점이 있다.

과표본화 기법과 관련된 연구와 마찬가지로 지능적으로 언더샘플링 기법을 사용한 연구들이 있다. 대표적인 예로 EasyEnsemble과 BalanceCascade가 있는데 두 기법의 목적은 정보 손실의 단점을 해소하기 위한 것이다[8]. EasyEnsemble은 다수 범주 S_{major}에서 부분집합 N_1, N_2, …, N_T를 독립적으로 샘플링한 뒤 N_i와 소수 범주 S_{minor}를 학습한 분류기 H_i를 T개 만들어 결과를 취합한다. EasyEnsemble가 무감독 학습방식을 통해 제거할 다수 범주 집합의 데이터를 탐색하는 반면, BalanceCascade에서는 감독 학습방식을 취한다. BalanceCascade

는 S_{major}에서 부분집합 N_1을 우선 한 번 샘플링하여 N_1과 S_{minor}를 학습한 분류기 H_1을 만들고 $x \in S_{major}$인 x가 H_1에 의해 정확하게 분류되면 x는 충분히 많다고 판단하여 x를 S_{major}에서 제거한다. 이 과정을 T번 순차 반복하여 마지막으로 하나의 분류기 H가 만들어지게 된다.

다른 방식의 언더샘플링 기법으로는 one sided selection (OSS)가 있다. OSS에서는 S_{major}의 데이터를 4개의 그룹(노이즈 데이터, 범주 경계선 근처 데이터, 중복된 데이터, 안전한 데이터)으로 나누고 S_{major}에서 노이즈 데이터, 범주 경계선 근처 데이터, 중복된 데이터가 제거된 부분집합 E를 만들었다. 이를 S_{minor}와 더해 집합 N, $N = \{E \cup S_{minor}\}$을 학습하였다[9].

2.3 앙상블 기법

앙상블은 각각 다양한 가설 공간에서 약분류기 H_1, H_2, …, H_n을 만든 뒤 이들의 결과를 조합해 하나의 강분류기를 만드는 기법이다. 앙상블 기법에는 대표적으로 배깅(Bagging, Bootstrap Aggregating)과 부스팅(Boosting)이 있다. 배깅은 Breiman이 제안한 기법으로 결정트리나 신경망과 같은 기본 모델의 분산을 줄이는 기법이다[10]. 배깅의 진행 과정은 다음과 같다. 훈련 데이터에서 k개의 부분집합을 무작위로 복원추출한 뒤 각각의 부분집합을 결정트리나 신경망과 같은 기본 모델로 학습하여 k개의 약분류기를 만든다. 그리고 이들의 결과를 취합하여 하나의 강분류기의 결과를 낸다.

Adaptive Boost(w, m)

For each i = 1,…, N

Initialize $w_i = 1 / N$

For each t = 1, 2,…, T

Classify (x, y) with classifier $f_t(x) \in \{-1, 1\}$ *and* w_i

$$err_t = \frac{\sum_{i=1}^{n} w_i I(y_i \neq f_t(x_i))}{\sum_{i=1}^{n} w_i}$$

$$c_t = \log\left(\frac{1 - err_t}{err_t}\right)$$

Update $w_i = w_i \exp(c_t I(y_i \neq f_t(x_i)))$

Construct a strong classifier $\operatorname{sign}\left(\sum_{i=1}^{T} c_t f_t(x)\right)$

(x, y) : labeled data

w : weight vector

N : # of labled data

T : # of iteration

그림 1 Adaptive Boost
Fig. 1 Adaptive Boost

부스팅은 Schapire와 Freund에 의한 제안되었다[11]. 배깅이 병렬적인 앙상블 기법인 반면에 부스팅은 순차적인 앙상블 기법이다. 부스팅에서는 각 데이터 인스턴스마다 가중치를 갖는다. 인스턴스가 높은 가중치를 가질수록 학습된 분류기에 더 많은 영향을 준다. 각 단계 t마다 가중치 w_t와 주어진 데이터에 의해 분류기 f_t가 만들어지고 f_t의 오차 err_t에 따라 가중치 벡터 w는 조정된다. 인스턴스가 잘못 분류될수록 가중치는 증가한다. 마지막 T단계에서는 T개의 약분류기들의 결과를 조합해 하나의 강분류기가 만들어지게 되고 결과 취합 방식은 각 약분류기의 성능의 함수로 나타나게 된다. 부스팅 기법의 한 예로 AdaBoost(Adaptive Boost)의 알고리즘이 그림 1에 나타나 있다. 본 논문에서 제안한 과표본화 기반 앙상블 기법은 일반적인 부스팅 기법에서 각각의 데이터 인스턴스들이 갖는 가중치를 범주 별 데이터의 개수에 기반하여 표본화를 통해 결정하는 변형된 부스팅 기법이라고 생각할 수 있으며, 앙상블 모델의 구축과정이 일반적인 부스팅 기법과는 다르게 배깅에 기반한 병렬적인 앙상블 기법이라는 특징을 가지고 있다.

2.4 과표본화 기반 앙상블 학습 기법

불균형 데이터를 사용한 학습 과정에서는 일반적으로 관측수가 많은 범주의 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 범주에서 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 가능하게 된다 [12]. 또한 앙상블 모델의 구축을 위하여 전체데이터의 부분집합을 생성하는 과정은 언더샘플링(undersampling)의 경우와 유사하게 다수 범주의 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.

즉, 제안된 방법론은 언더샘플링된 데이터의 앙상블로

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 2 과표본화 기반 앙상블 학습 기법
Fig. 2 Oversampling Based Ensemble Learning

과표본화를 구현함으로써 언더샘플링에서 발생하는 데이터의 손실을 피할 수 있을 뿐만 아니라 과표본화로 인해서 발생하는 과적응이나 학습시간의 증가 등과 같은 학습의 어려움도 피할 수 있는 방법으로 샘플링 기법과 앙상블 기법의 결합으로 인하여 기존의 샘플링 기법들이 가지고 있는 단점을 극복하고 장점만을 이용할 수 있도록 하였다. 제안한 방법론의 수행과정이 그림 2에 나타나 있다.

3 실험 및 결과

3 1 데이터

제안한 방법론의 성능을 평가하기 위해 다수의 사용자로부터 수집된 238,450개의 필기체 데이터를 훈련 데이터로 사용하여 실험을 진행하였다. 훈련 데이터의 범주는 모두 50개로 소문자 알파벳 26개와 대문자 알파벳 16개, 숫자 8개이다. 대문자의 개수가 소문자의 개수보다 적은 이유는 인식과정에서 대문자와 소문자의 모양이 같은 C, K, O, P, S, U, V, W, X, Z를 소문자로 간주하여 인식했기 때문이고 숫자 0과 1도 소문자 o와 l로 간주하여 인식했다. 일부 훈련 데이터의 예가 그림 3에 나와 있다. 그림 3과 같은 원시자료(raw data)로부터 획의 필기순서에 따른 온라인 특징 256차원을 추출하였고 글자를 구성하는 획 (x,y) 좌표 벡터에 따라 오프라인 특징 377차원을 추출하여 인식기의 학습에 사용하였다.

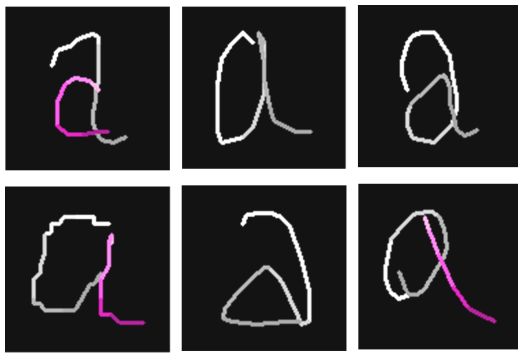


그림 3 훈련 데이터 a의 예
Fig. 3 Examples of character 'a'

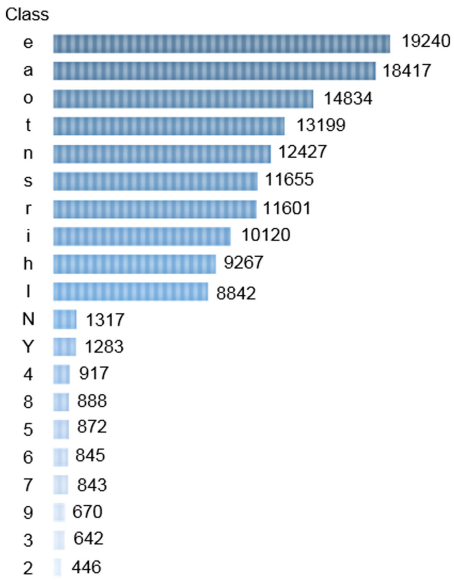


그림 4 데이터가 개수가 가장 많은 낱글자 10개와 가장 적은 낱글자 10개
Fig. 4 10 most frequent characters and 10 most rare characters in training data set

범주당 평균 데이터개수는 4,769개이고 데이터를 가장 많이 포함하고 있는 범주 상위 10개와 가장 포함하고 있는 범주 하위 10개가 그림 4에 나타나 있다. 테스트 데이터로 9만여개의 UNIPEN Train R01 /V07 데이터가 사용되었다[13].

3 2 실험결과

본 논문에서 제안하는 과표본화 기반 앙상블 학습 기법의 성능을 측정하기 위해 제안한 방법으로 학습한 모델의 성능과 전체 데이터를 6등분한 뒤 앙상블을 적용한 모델의 성능, 전체 데이터를 한 번에 학습한 기본 모델의 성능을 비교해 보았다. 학습을 위한 분류기는 인공 신경

표 1 각 분류기의 정확도
Table 1 Accuracy of each classifier

(a) Overall accuracy of each classifier

Base Model	Bagging	Oversampling Based Ensemble Method
77.564	80.42	81.79

(b) Accuracy on 10 most frequent characters

Class	Base Model	Bagging	Oversampling Based Ensemble Method
L	61.63	85.05	76.36
H	88.20	79.83	88.00
I	78.27	47.62	82.74
R	76.02	91.59	78.81
S	95.42	98.85	97.07
n	76.99	70.19	61.57
t	84.42	93.62	85.91
o	89.27	92.19	88.39
a	87.97	90.70	80.18
e	94.04	95.13	92.76

(c) Accuracy on 10 most rare characters

Class	Base Model	Bagging	Oversampling Based Ensemble Method
2	47.84	56.34	57.01
3	85.33	90.83	97.15
9	30.89	43.52	51.29
7	42.83	50.58	64.32
6	67.76	83.85	81.93
5	50.82	51.75	66.68
8	46.06	60.29	74.91
4	50.03	68.51	78.22
Y	28.05	30.17	39.64
N	71.13	69.74	80.44

망을 이용하였고, 앙상블 모델은 배깅을 이용하였다. 표 1의 (a)에서 볼 수 있듯이 과표본화 기반 앙상블 학습 기법을 적용했을 때 N = 1000, M = 6인 경우, 분류기의 평균 정확도는 81.79%였고 배깅을 이용한 앙상블 모델이나 기본 모델에 비해서 성능이 향상되었음을 확인할 수 있었다. 특히 표 1의 (c)에서 나타나듯이 기본 모델은 전반적인 성능을 최적화하기 위해 데이터 개수가 적은 범주에 대한 성능을 희생시킨 반면, 제안한 기법에서는 그러한 성능 희생 없이 전반적인 성능이 개선되었다. 또한 하위 10개 낱글자에 대한 배깅의 성능과도 비교해보았을 때 제시한 기법이 배깅보다 데이터 개수가 적은 범주에 대한 정보손실이 더 적다는 점을 알 수 있었다.

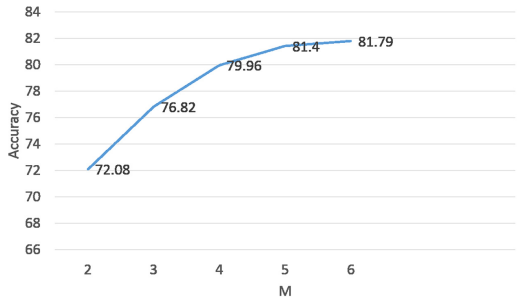


그림 5 약분류기 개수에 따른 정확도
Fig. 5 Accuracy of the different number of weak learners

과표본화 기반 앙상블 기법에서 약분류기의 개수가 추가됨에 따라서 보여지는 성능 변화가 그림 5에 나타나 있다. 실험 결과로부터 학습의 초기에 급격한 성능 향상을 보이다가 학습이 진행됨에 따라서 성능이 수렴하는 경향을 보이고, 또한 전체 데이터의 절반만을 이용하여 학습이 진행된 성능이 전체 데이터를 이용하여 학습한 신경망 단일 분류기의 경우에 크게 뒤지지 않음을 확인할 수 있다. 이러한 결과로부터 제안한 방법론이 학습 초기에 문제 공간을 효율적으로 탐색하여 빠른 성능 향상을 보임과 동시에, 최종적으로는 훈련 데이터에 있는 정보들을 빠짐없이 잘 활용하고 있다고 판단할 수 있다.

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 확인하였다. 제안한 방법론은 앙상블 모델을 이용한 과표본화 기법을 구현함으로써 표본화 기법들 간의 단점을 배제한체 각각의 장점만을 구현할 수 있는 기법으로써 필기체 인식 문제만이 아닌 다양한 불균형 데이터에 적용이 가능할 것으로 예상된다.

References

[1] S. Ertekin, J. Huang, L. Bottou, L. Giles, "Learning on the border: active learning in imbalanced data classification," *Proc. of ACM conference on Confer ence on Information and Knowledge Management*, pp. 127 136, 2007.

[2] A. Estabrooks, T. Jo, and N. Japkowicz, "A mul tiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, Vol. 20, No. 1, pp. 18 36, 2004.

[3] H. He, and E. A. Garcia, "Learning from Imba lanced Data," *IEEE Transactions on knowledge*

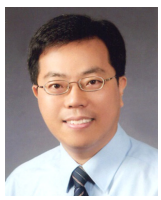
and data engineering, Vol. 21, No. 9, pp. 1236 1284, 2009.
[4] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "Smote: Synthetic minority oversam pling technique," *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321 357, 2002.
[5] H. Han, W. Wang, B. Mao, "Borderlinesmote: A new over sampling method in imbalanced data sets learning," *Proc. of International Conference on Intelligent Computing*, pp. 878 887, 2005.
[6] H. He, Y. Bai, E.A. Garcia, S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imba lanced Learning," *Proc. of International Joint Con ference on Neural Networks*, pp. 1322 1328, 2008.
[7] J. V. Hulse, T. M. Khoshgoftaar, A. Napolitano, "Experimental perspectives on learning from imba lanced data," *Proc. of International Conference on Machine Learning*, pp. 935 942, 2007.
[8] X. Y. Liu, J. Wu, Z. H. Zhou, "Exploratory Under Sampling for Class Imbalance Learning," *Proc. of International Conference on Data Mining*, pp. 965 969, 2006.
[9] M. Kubat, S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection," *Proc. of International Conference on Machine Lear ning*, pp. 179 186, 1997.
[10] L. Breiman, "Bagging predictors," Machine Learning, Vol. 24, No. 2, pp. 123 140, 1996.
[11] Y. Freund and R. E. Schapire, "A decision theoretic generalization of on line learning and an application to boosting," *Journal of Computer and System Sci ences*, Vol. 55, No. 1. pp. 119 139, 1997.
[12] T. J. Kim, H. Y. Jang, J. W. Park, S. T. Hwang, B. T. Zhang, "Ensemble Methods with increasing data for online handwriting recognition," *Proc. of the KIISE Korea Computer Congress 2013*, pp. 1396 1398, 2013.
[13] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, "UNIPEN project of on line data exchange and recognizer benchmarks," Proc. of International Conferences on Pattern Recognition, pp. 29 33, 1994.

김 경 민

2013년 홍익대학교 컴퓨터공학과 학사 2013년~현재 서울대학교 컴퓨터공학부 석박사통합과정. 관심분야는 기계학습, Com putational Intelligence, 멀터미디어마이닝, 인지과학

장 하 영

2002년 연세대학교 컴퓨터과학과 공학사 2004년 서울대학교 컴퓨터공학과 공학석사. 2004년 현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 기계학습, 진화연산, 확률그래프 모델



장 병 탁

1986년 서울대 컴퓨터공학과 학사. 1988년 서울대 컴퓨터공학과 석사. 1992년 독일 Bonn 대학교 컴퓨터과학 박사. 1992년~1995년 독일국립정보기술연구소(GMD, 현 Fraunhofer Institutes) 연구원. 1997년~현재 서울대 컴퓨터공학부 교수 및 인지과학, 뇌과학, 생물정보학 협동과정 겸임교수. 2003년~2004년 MIT 인공지능연구소(CSAIL) 및 뇌인지과학과(BCS) 객원 교수. 2007년~2008년 삼성종합기술연구원(SAIT) 객원 교수. 현재 서울대 인지과학연구소 소장, Applied Intelligence, BioSystems, Journal of Cognitive Science 등 국제저널 편집위원. 관심분야는 바이오지능, 인지기계학습, 분자진화 컴퓨팅기반 뇌인지 정보처리 모델링