클래스 불균형 문제를 해결하기 위한 개선된 집중 샘플링

김 만 선[†] · 양 형 정[††] · 김 수 형[†††] · 챠 위 핑[††††]

요 약

실세계의 문제에서 많은 기계학습의 알고리즘들은 데이터의 클래스 불균형 문제에 어려움을 겪는다. 이러한 클래스 불균형 문제를 해결하기 위하여 데이터의 비율을 변경하거나 좀 더 나은 샘플링 전략으로 극복하려는 연구들이 제안되었다. 그러나 데이터의 비율을 변경하는 연구에서는 전체 데이터 분포의 특성을 고려하지 못하고, 샘플링 전략을 제안하는 연구에서는 여러 가지 제한 조건을 고려해야만 한다. 본 논문에서는 위의 두가지 방법의 장점을 모두 포함하는 개선된 집중 샘플링 방법을 제안한다. 제안된 방법에서는 클래스 불균형 문제를 해결하기 위해 학습에 유용한 데이터들을 샘플링하는데 스코어링에 기반한 데이터 분할 방법을 이용한다. 즉, 입력 데이터들에 대해 SOM(Self Organizing Map)의 학습 결과로 얻은 BMU(Best Matching Unit)와의 거리를 계산하고, 이 거리를 스코어라 한다. 측정된 스코어는 오름차순으로 정렬되며, 이 과정에서 입력 데이터의 분포가 재 표현되고, 재 표현된 분포는 전체 데이터의 특성을 대표하게 된다. 그 결과로 얻은 데이터들 중에서 유용하지 못한 데이터들에 대해 제거하는 과정을 수행하여 새로운 학습 데이터 셋을 얻는다. 새로운 학습 데이터 생성 과정에서는 재 표현된 분포의 결과를 두 구간(upper, lower)으로 분할하는데, 두 구간 사이의 데이터들은 유용하지 못한 패턴들로 간주되어 학습에 이용되지 않는다. 본 논문에서 제안한 방법은 클래스 불균형의 비율 감소, 훈련 데이터의 크기 감소, 과적합의 방지 등 몇 가지 장점을 보인다. 제안한 방법으로 샘플링된 데이터에 kNN 을 적용하여, 분류 실험한 결과 심한 불균형이 있는 ecoli 데이터의 분류 성능이 최대 2.27배 향상되었다.

키워드 : 비 감독 학습, 자기조직화지도, 베스트 매칭 유니트, 집중샘플링

Improved Focused Sampling for Class Imbalance Problem

Man-Sun Kim[†] · Hyung-Jeong Yang[††] · Soo-Hyung Kim[†††] · Wooi Ping Cheah[††††]

ABSTRACT

Many classification algorithms for real world data suffer from a data class imbalance problem. To solve this problem, various methods have been proposed such as altering the training balance and designing better sampling strategies. The previous methods are not satisfy in the distribution of the input data and the constraint. In this paper, we propose a focused sampling method which is more superior than previous methods. To solve the problem, we must select some useful data set from all training sets. To get useful data set, the proposed method devide the region according to scores which are computed based on the distribution of SOM over the input data. The scores are sorted in ascending order. They represent the distribution of the input data, which may in turn represent the characteristics of the whole data. A new training dataset is obtained by eliminating unuseful data which are located in the region between an upper bound and a lower bound. The proposed method gives a better or at least similar performance compare to classification accuracy of previous approaches. Besides, it also gives several benefits : ratio reduction of class imbalance; size reduction of training sets; prevention of over-fitting. The proposed method has been tested with kNN classifier. An experimental result in ecoli data set shows that this method achieves the precision up to 2.27 times than the other methods.

Key Words : Unsupervised Learning, SOM(Self Organizing Map), BMU(Best Matching Unit), Focused Sampling

1. 서 론

실세계 데이터에서 한 범주에 속하는 패턴의 수가 다른 범주에 속하는 패턴의 수보다 적거나 많은 경우에 클래스 불균형 문제가 발생한다. 클래스 불균형은 응답모델링, 리모트 센싱, 영상 장면 분류 등에서도 쉽게 나타나는 현상으로 예를 들어, 휴대폰 이용 대상자 중 사기 행위를 하는 사용자를 찾는 사기탐지(fraud detection)[1]에서 정상 사용자에 비해 사기 사용자의 수가 적게 나타나는 클래스 불균형이 발생한다[2-4,17,19,20].

클래스 불균형 문제를 내포하는 데이터 집합의 경우 소수

※ 이 논문은 2006년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2006-003-D00511).
† 정 회 원 : 전남대학교 연구원
†† 종신회원 : 전남대학교 전자컴퓨터공학부 조교수
††† 정 회 원 : 전남대학교 전자컴퓨터공학부 교수
†††† 준 회 원 : 전남대학교 전산학과 재학중 (박사과정)
논문접수 : 2007년 3월 28일, 심사완료 : 2007년 7월 12일

범주의 데이터에 비하여 다수 범주에 속하는 데이터의 수가 과도하게 분포한다. 이런 경우 다수가 소수 범주의 영역을 침범하게 되어 분류 알고리즘의 성능에 좋지 못한 영향을 미친다. 따라서 클래스 불균형 문제를 해결하는 것은 분류 성능을 향상 시키는데 있어 필수적인 과정이다[13].

불균형 데이터에 접근하기 위해 제안된 것은 샘플링 전략에 따라 랜덤 샘플링(Random Sampling)[9,10,18]과 집중 샘플링(Focused Sampling)[5,6,7,8,11]이 있다. 랜덤 샘플링이란 훈련에 추가되거나 제거되어야 할 데이터를 임의로 추출하는 것이다. 랜덤 샘플링은 샘플링 방법에 따라 오버 샘플링과 언더 샘플링으로 나눈다.

오버 샘플링(Over Sampling)[9,10]은 정해진 규칙에 따라 다수 범주의 수만큼 소수 범주에서 그 데이터를 생성한다. 이 방법은 모든 데이터의 정보를 사용할 수 있다는 장점이 있지만 데이터의 수를 증가시켜 시간복잡도가 증가한다는 단점이 있다. Chawla et al.[9]은 kNN(k Nearest Neighbor) 기법을 사용하여 소수 범주 주변에 인공적으로 데이터를 생성하는 오버 샘플링으로 불균형을 해소하였다. 그러나 소수 범주에 속한 데이터 중에서 이상치(outlier)가 존재할 경우 부정적인 영향을 끼칠 수 있다는 단점을 갖는다.

언더 샘플링(Under Sampling)[10]은 소수 범주의 수만큼 다수 범주에서 데이터를 추출한다. 이 기법은 시간복잡도가 감소한다는 장점이 있으나, 다수 범주 데이터를 많이 버려야 하기 때문에 다수 범주의 분포 및 특징을 대표 할 수 없다. Liu[10]는 언더 샘플링을 위하여 비감독학습 기반의 기법인 EasyEnsemble과 감독 학습 기반의 기법인 BalanceCascade을 제안하였다. 이 연구에서는 기존의 언더 샘플링 결함을 극복하기 위하여 여러 개의 서브셋을 랜덤 생성하고 학습하여 각각의 결과를 조합하였다. 그러나 랜덤하게 구성된 서브셋이 다수 범주의 전체적인 특징을 대표할 수 있는 서브셋으로 볼 수 없다.

이러한 랜덤 샘플링의 문제점을 해결하고자 최근 집중 샘플링에 관한 연구가 활발히 수행되고 있다. 주로 데이터의 분포와 속성에 의해 집중 샘플링(Focused Sampling)이 수행되고 있으며, 최근 연구로는 Japkowicz의 연구[5], 마할라노비스 거리(mahalanobis distance)를 이용한 Foody의 연구[6], 앙상블 네트워크의 출력값이 갖는 편기(bias)와 분산(variance)을 이용한 Shin의 연구[7], SHRINK 알고리즘을 이용한 Kubat et al.연구[8], 결정 경계(decision boundary) 근처의 데이터만을 활용한 Shin과 Cho[11]연구가 있다. 그러나 이들 연구들은 제한 조건을 만족해야 하거나, 계산시간이 많이 소요되는 문제점 그리고 다수 범주의 데이터 특성을 대표하지 못하는 문제점을 안고 있다.

본 논문에서는 이러한 문제점을 해결하기 위하여 유용한 데이터만을 선택하는 방법을 제안한다. 제안하는 방법에서는 비감독 학습에 의해 클래스 분포의 중심에 기초한 결과를 얻고, 각 데이터의 거리를 계산하여 스코어를 측정한 후, 이를 기반으로 두 단계 샘플링을 수행한다. 이 방법은 데이터 분포를 이용하여 전체 데이터의 특성을 반영할 수 있는 장점이 있고 샘플링을 이용하여 데이터의 크기를 감소시키는 결과도 얻을 수 있다. 입력 데이터의 분포를 그대로 유지하면서 유용한 데이터만을 샘플링 한 결과에서도 분류 정확률은 최고 2.27배까지 향상되었다.

본 논문의 구성은 다음과 같다. 2장에서 클래스 불균형 문제를 해결하기 위한 집중 샘플링과 관련된 연구들을 살펴보고, 3장에서는 제안하는 방법인 데이터 분포에 기반한, 개선된 집중 샘플링 기법을 기술한다. 4장에서는 실험 결과와 관련 연구와의 비교 분석을 하고, 5장에서 결론을 맺는다.

2. 관련 연구

집중 샘플링은 특정 전략을 적용하여 샘플링을 하는 방법이다. 흔히 능동 학습 데이터 선택 방법이라고도 한다. 클래스 불균형 문제를 해결하기 위해 집중 샘플링을 적용한 연구는 다음과 같이 정리해 볼 수 있다.

Japkowicz[5]는 선형적으로 구분이 가능한 경우에는 불균형 데이터의 절대적인 수가 성능에 영향을 끼치지 않는다고 하였다. 그러나 각 범주간의 데이터 비율의 차이가 크고, 선형적으로 구분이 불가능한 경우에는 성능에 큰 영향을 미치는 것으로 나타났다.

Foody[6]는 분류 경계에 인접한 학습 데이터들을 찾기 위하여 클래스 중심과의 마할라노비스 거리를 이용하였다. 마할라노비스 거리는 변수들 사이의 표준편차와 상관관계를 고려하여 만들어진 거리이다. 양분된 두 개의 학습 셋에서 클래스 경계 부근 패턴 셋과 클래스 중심부근 패턴 셋에 대한 성능 비교 실험 결과 클래스 경계 부근의 패턴 셋이 우수한 분류 결과를 보였다. 그러나 변수들과의 관계가 반드시 비독립적이라고 보장할 수 없고, 모든 변수가 표준화되어 있고 서로 독립적인 관계를 가지고 있다면, 마할라노비스 거리는 유클리디안(Euclidean) 거리와 차이가 없다.

Shin[7]은 앙상블 네트워크의 출력값이 갖는 편기(bias)와 분산(variance)을 이용하여 유용한 패턴을 선택하였다. 다양한 구조 및 학습 파라미터를 가진 여러 개의 네트워크들로부터, 각 데이터마다 출력값 분포를 얻을 수 있었다. 분류 문제에서 편기(bias)가 작고 분산이 높은 패턴이 높은 효용지수(Utility Index)를 얻어서 이를 유용한 데이터로 선택한다. 그러나 이 연구에서는 데이터의 사이즈가 작은 이진 분류 문제만을 다루었다.

Kubat et al.[8]은 SHRINK 알고리즘을 통하여 다수 범주와 소수 범주의 분포가 혼재하는 영역은 소수 범주로 분류되어야 한다고 하였다. SHRINK 알고리즘이란 가장 높은 성능을 낼 수 있는 소수 범주의 영역을 찾아내는 것으로, 데이터의 속성(attributes)들 중에서 의미 없는 속성들을 하나씩 제거해 감으로써 그 영역을 찾는다. 그러나 이 방법은 다수 범주의 데이터가 소수 범주의 영역을 침범한다는 전제조건을 만족해야 한다.

Shin과 Cho[11]는 결정 경계(decision boundary) 근처의 데이터만을 활용하는 전략으로 SVM (Support Vector

Machine)을 적용하여 이진 분류 문제를 다루었다. 하지만, 실세계의 클래스 불균형 문제는 이진 분류뿐만 아니라 다중 분류 문제를 해결할 수 있어야 하는데 그렇지 못한 점이 문제로 남아 있다.

Oh와 Jang[15]은 불균형 데이터 문제를 해결하기 위해 퍼셉트론에 기초한 부스팅 기법을 제안하였다. 부스팅 기법은 학습을 어렵게 하는 데이터에 집중하여 앙상블 머신을 구축하는 기법이다. 이 연구는 각각의 학습 예제에 서로 다른 중요도를 부여하는 방법을 도입하였다. 부스팅 기법에서는 약학습기(weak learner)의 조건을 만족시키지 못하는 경우를 보완하기 위해 커널을 도입한 커널 퍼셉트론을 사용하여 학습기의 표현 능력을 높였다. 실험 결과 부스팅은 결정 경계 근처의 데이터에 집중한다는 것을 알 수 있었다. 그러나 불균형의 비율이 (+1 데이터 비율) 4.2%, 4.7%에 해당하는 데이터들은 만족할만한 F1의 결과를 얻었으나, +1 데이터 비율이 32.4% 을 갖는 실험에서는 MLP보다 성능이 떨어졌다.

Sun et al[16]의 연구에서는 클래스 불균형 문제를 해결하기 위하여 cost-sensitive 알고리즘을 제안하였다. 그들은 클래스 불균형 문제를 해결하고자 하는 연구를 크게 data-level approaches, algorithm-level approaches, cost-sensitive learning 3가지로 구분하였다. data-level approaches는 데이터를 재-균등(rebalance)하게 데이터 공간의 표본을 재추출하는 방법에 의한 접근법이다. algorithm-level approaches는 소수 범주에 더 강력하게 분류기를 적응적으로 알고리즘의 변형을 시도하는 방법이다. 마지막으로 cost-sensitive learning 은 위 두 가지를 모두 조합한 방법으로 오분류된 데이터에 더 많은 관심을 기울일 수 있는 방법이라고 소개했다. 즉, 비용 에러(cost error)를 소수 범주(rare class)에서 최소화되는 지점을 찾는 연구로서 AdaBoost 알고리즘에서 가중치 업데이트 수식(weight update formula)을 변형하여 3가지 변형된 식(AdaC1, AdaC2, AdaC3)을 완성했다. 이 연구에서 제안된 방법은 기존의 AdaBoost 알고리즘을 적용한 결과보다 우수하지만, 이러한 장점에도 불구하고, 다수 범주와 소수범주의 비율을 찾는 문제는 해결하지 못하고 있다. 즉, 다양한 실험을 통해서만 최적의(다수 범주 대 소수 범주) 비율을 찾는 것이 여전히 문제로 남아 있다.

이상과 같이 이들 연구들은 제한 조건을 만족해야 하거나, 계산시간이 많이 소요되는 문제점과 다수 범주의 데이터 특성을 대표하지 못하는 문제점을 안고 있다. 본 논문에서는 위와 같은 기존 연구들의 문제점을 해결하기 위해 개선된 집중 샘플링 방법을 이용한다.

3. 데이터의 분포에 기반한 개선된 집중 샘플링

집중 샘플링을 위해 본 연구에서는 비감독 학습(unsupervised learning)인 SOM을 이용하여 데이터 분포를 분석하고 이를 활용해 스코어링을 수행한다. 데이터 분포의 스코어링 결과는 각 입력 데이터가 학습에 의해 새롭게 얻어진 분포에서 클래스 중심에 얼마나 근접한지를 보여준다. 스코어링 결과는 오름차순으로 정렬되며, 정렬된 결과는 스코어를 기반으로 두 영역으로 나누어져 샘플링이 수행된다. 두 구간을 분할하는 전략으로서 학습에 유용한 데이터가 집중되어 있는 상한(upper)구간과 하한(lower)간의 데이터만을 샘플링한다.

3.1 SOM 학습과 스코어링

본 논문에서는 스코어를 얻는 방법에 SOM[21]을 이용하여 데이터의 분포를 분석한다. SOM은 고차원 데이터의 가시화를 위한 효율적인 도구로서 클러스터링에 많이 이용된다. 이 방법은 복잡하고 비선형적인 통계적 연관성을 저차원 표현법으로 단순한 기하학적 연관성으로 표현한다. SOM은 위상적 연관성을 보전하면서 데이터에 존재하는 정보를 압축하기 때문에 주어진 데이터에 대한 요약된 정보로 간주할 수 있다. 여기에서 유사한 성질을 가지는 데이터들은 서로 인접하도록 조정되고 클래스 중심 분포로부터 가깝거나 매우 먼 데이터들만 학습에 유용한 패턴이 될 수 있다. 반면에 두 구간 사이의 데이터들은 학습 성능을 저하시키지는 않으나 학습에 대한 기여도가 상대적으로 낮다. 즉, 불필요한 패턴들로서 이러한 데이터들은 학습에 방해가 되므로 제거된다.

스코어를 얻기 위한 거리 측정은 자기 조직화 학습에 의해 갱신된 연결강도와 각 입력 데이터와의 거리 계산으로 수행된다. 가장 가까운 거리의 뉴런을 승자뉴런이라 하는데 승자 뉴런만이 출력을 보낼 수 있는 유일한 뉴런이 되며, 승자와 그 이웃의 뉴런들만이 학습과정에서 연결강도를 갱신한다. 이러한 경쟁학습을 통하여 자연스럽게 클래스의 중심이 선정되고, 그 중심을 기준으로 이웃 데이터들은 점점 가깝게 조정된다. 그 결과 데이터의 분포를 얻을 수 있다. 학습 결과 얻어진 스코어는 각 입력 데이터가 새롭게 얻어진 분포에 의해 클래스 중심 분포에 얼마나 근접한지 여부

① 초기 연결 가중치를 스스로 결정하고, 네트워크의 MAP 사이즈를 자기 조직화하여 구성한다.
② 모든 입력 데이터에 대해 학습이 끝나면, 클래스 중심 부근으로 판단할 수 있는 BMU(Best Matching Units)와의 거리를 수식 3.1과 같이 구한다.

$$d_j = \sum_{i=0}^{n-1}(x_i(t)-w_{ij}(t))^2 \qquad \text{(수식 3.1)}$$

여기서 d_j :입력과 출력 뉴런 j사이의 거리, $x_i(t)$: 시각 t에서의 i번째 입력 벡터, $w_{ij}(t)$: i번째 입력벡터와 j번째 출력 뉴런 사이의 연결 강도
③ 모든 데이터에 대한 거리를 스코어로 정의하고 정렬한다. BMU 로부터 가까운 데이터들은 낮은 스코어를 갖게 되고, 멀리 떨어진 데이터들은 높은 스코어를 갖게 된다. 얻어진 스코어는 유용한 데이터를 샘플링하기 위하여 오름차순으로 정렬된다.

(그림 1) SOM 학습과 스코어링 과정

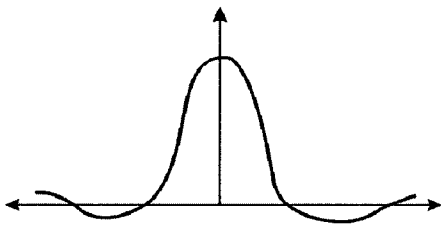
를 알기 위한 척도로 반영된다. 스코어링은 학습 데이터들의 특정 클래스 중심에 가까운 정도를 점수화하는 과정이다. 스코어들은 구간별 효율적인 분할을 하기 위해 오름차순 정렬된다. 위 과정을 정리하면 아래의 (그림 1)과 같다. 스코어링의 결과로서 전체 데이터 분포의 중심부에 해당하는 클래스와 해당하지 않는 클래스에 대한 정보, 즉 요약된 정보로 재 표현된 결과를 얻을 수 있다.

3.2 샘플링

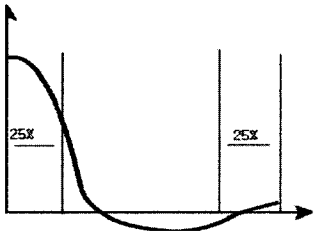
3.1의 결과는 (그림 1)의 데이터 분포 형태를 측면(lateral)에서 보는 것과 같이 구간을 선정하여 세분화할 수 있다. 즉, 10%, 15%, 20%, 25%, 30%만큼 구간을 선정하면 전체 학습 데이터를 모두 사용하지 않고 전체 데이터 비율 중 20%~60% 만 샘플링 된다. 분할 구간을 이렇게 선정한 이유는 데이터의 크기를 감소시키는 것이 샘플링의 목적이지만 훈련 데이터의 크기가 매우 작거나 많다면 전체 데이터의 특성을 대표할 수 없으므로 샘플링의 의미가 없게 된다. 따라서 전체 데이터 비율 중 20%~60% 구간에 포함되지 않는 데이터들은 삭제한다.

클래스 분포의 중심이 되는 BMU(Best Matching Unit)에 근접한 구간을 상한 구간, 상대적으로 반대편에 위치하는 구간을 하한 구간이라 정의한다. 만약 분류를 위하여 상한 구간의 데이터만을 이용한다면 n개의 클래스로 나눠지는 패턴에서 n보다 적은 수의 클래스로 구성된 데이터들만 얻게 된다. 이러한 결과는 실험 4.2.2의 <표 4>에서 자세히 보인다. 이러한 데이터들은 모델을 학습시키기에 충분한 데이터라고 볼 수 없는데 그 이유는 검증단계에서 올바른 분류 결과를 얻을 수 없기 때문이다.

SOM의 결과를 측면에서 살펴보면 (그림 2)의 a)과 같이 멕시칸 모자와 유사하다. 예를 들어, 학습에 유용한 부분만 선택하는 전략에 맞도록 전체 데이터의 50%만 학습에 이용하고자 한다면, (그림 1)의 b)와 같이 스코어의 오름차순 정렬된 결과에 상한구간 25%, 하한구간 25%로 분할하면 된다. 분리 구간을 넓게 설정할수록 많은 데이터를 학습에 이용할 수 있지만, 반드시 학습에 좋은 결과를 보장할 수 없다는 실험 결과를 얻었다. 이에 대한 연구는 실험4.2.2의 <표 3>에서 구간별 데이터 구성에 따른 실험 비교에서 보인다.



(그림 2) a) 멕시칸 모자 모양의 분포 형태



b) 25% 양쪽 구간 설정

```
Procedure Data_Selection_Using SOM

Input:
 D : training patterns
 ND : redefine of D by SOM
 S_ND=sorting
 UR : upper bound of S_ND
 LR : lower bound of S_ND

Ouput:
S_ND' : new S_ND
begin
step 1 //initalization
 step 1.1 normalize D
 step 1.2 determine map size
step 2  //training the map
 step 2.1 ND : training using SOM
step 3  //scoring
 step 3.1 computing distance from BMU
 step 3.2 sort distance in ascending order
step 4  //remove
 step 4.1 UR=find upper bound of S_ND
 step 4.2 LR=find lower bound of S_ND
 step 4.3 remove X  where UR<X<LR

end
```

(그림 3) 의사코드(Pseudo code)

따라서, 클래스 분포의 중심과 가까운 데이터들과 중심에서 먼 데이터들만이 학습에 필요한 유용한 데이터가 된다. 즉, 상한 구간과 하한 구간 사이의 데이터들은 학습에 사용하지 않게 되는데, 두 구간 사이의 데이터들은 유용하지 못한 패턴들로 간주되어 제거된다. 제안하는 방법의 의사코드는 다음의 (그림 3)과 같다.

4. 실험 결과 및 논의

4.1 실험 환경 및 데이터

실험 환경은 펜티엄4 3.39Ghz, 1GB RAM, 윈도우 XP 환경에서 MATLAB 7.0을 사용하였다.

제안하는 방법의 실험을 위하여 웹상에 공개되어 있는 UCI Repository의 데이터[12]를 사용한다. 유방암 데이터(WDBC :

〈표 1〉 실험 데이터 요약

클래스 당 데이터의 개수		평균	표준편차
br	(0,444) (1,239)	339.5	142.1
pima	(1,268) (-1,500)	384	164.0
yeast	(1,244) (2,429) (3,463) (4,44) (5,35) (6,51) (7,163) (8,30) (9,20) (10,5)	148.4	173.6
ecoli	(1,143) (2,77) (3,2) (4,2) (5,35) (6,20) (7,5) (8,52)	42, 48	48.70
glass	(1,70) (2,76) (3,17) (5,13) (6,9) (7,29)	35.6	29.7

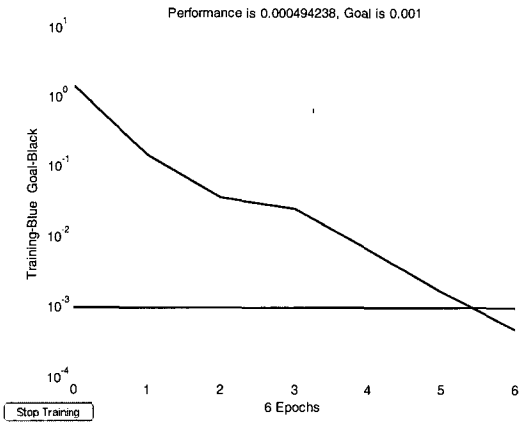
Wisconsin Diagnostic Breast Cancer Dataset, 이하 br)는 각 사람별(총569명)로 종괴에서의 핵의 모습(radius, texture, perimeter, area, smoothness, concavity, compactness, concave points, symmetry, fractaldimension)의 10개 속성들의 평균값, 표준 편차값, 최악 또는 최대값이며 양성(benign)/악성(malignant) 여부를 나타내고 있다. pima 데이터는 8개의 속성, 768개의 샘플수를 갖는다. 이 데이터의 불균형 비율은 65.1%(500개의 샘플: negative), 34.9%(268개의 샘플: positive)와 같다. yeast 데이터는 1484개의 데이터수와 10가지 속성으로 구성되어 있으며 10개의 클래스들(CYT, NUC, MIT, ME3, ME2, ME1, EXC, VAC, POX, ERL)로 나누어진다. ecoli의 8개의 클래스는 성분에 따라서 클래스의 유형(cp, im, pp, imU, om, omL, imL, imS)이 결정된다. ecoli 데이터는 336개의 데이터수와 8가지 속성으로 구성된다. glass 데이터는 214개의 샘플수와 10가지 속성으로 구성되어 있다. glass의 성분에 따라서 glass의 유형(building_windows_float_processed, building_windows_non_float_processed, vehicle_windows_float_ processed, containers, tableware, head lamps)이 결정된다. 클래스 당 데이터의 개수는 다음과 같으며, 클래스 불균형 문제가 존재한다. 아래의 <표 1>에 실험 데이터의 클래스 당 데이터의 개수, 평균, 표준편차를 요약하였다.

이 실험에서는 제안된 방법의 정확한 실험 결과 측정을 위해 5-fold hold out method cross-validation 방법을 사용하였다. 이 실험 방법은 실험 데이터를 5개의 부분 데이터 집합으로 나눈다. 하나의 부분 데이터 집합은 검증에 이용되고 나머지 4개의 부분 데이터 집합을 이용해 학습을 한다.

4.2 실험 결과

4.2.1 타 연구와의 성능 평가

첫 번째, 비교 실험은 기존 연구[7]와 동일한 실험 조건으로 유방암 데이터에 대한 MSE(Mean Squared Error)의 성능을 비교하였다. 네트워크의 구조는 1개의 은닉층(hidden layer)과 15개의 은닉 노드(hidden node)를 같은 MLP구조로 구성하였다. 기존 연구에서 유방암 데이터의 MSE 값은 단일(single) MLP 적용시 0.0614, 배깅(bagging) MLP 적용시 0.0061을 얻었으나, 제안한 방법으로 실험한 결과 0.0004를 얻었다. MSE는 0에 가까울수록 충분히 망(network)이 학습



(그림 4) 제안한 방법을 적용한 유방암 데이터의 실험 결과

<표 2> MSE 값의 결과 비교

실험데이터 / 데이터 선택	yeast	ecoli	glass	br
패턴 선정 전	2.2790	1.6746	0.0316	0.00222
10% 구간(패턴 선정)	1.3348	0.4854	0.0140	0.00085
15% 구간(패턴 선정)	1.4615	0.4072	0.0093	0.00080
20% 구간(패턴 선정)	1.4506	0.0581	0.0009	0.00084
25% 구간(패턴 선정)	1.7380	1.1673	0.0216	0.00099
30% 구간(패턴 선정)	2.1331	0.8835	0.0232	0.00096

되었다는 것을 의미하며, 제안된 방법은 기존 연구보다 훨씬 0에 가깝기 때문에 성능 향상이 있음을 알 수 있다. 아래의 (그림 4)는 제안한 방법을 적용한 실험 결과로 유방암 데이터의 수렴 과정을 보여준다.

다층 퍼셉트론 (MLP:Multi Layer Perceptron)에 훈련 성능 평가를 한 결과 <표 2>와 같이 MSE(Mean Squared Error)의 출력값이 데이터 선택 이전보다 점차 줄어드는 현상을 확인할 수 있었다. 상한 및 하한 구간 사이의 데이터들은 학습에 방해가 되는 불필요한 데이터이므로 이것을 제거하였고, 과적합을 방지할 수 있었다.

두 번째, 비교실험은 기존 연구[16]와 동일한 데이터인 유방암 데이터(br)와 pima(pima indian data)에 대한 F1의 성능을 비교하였다. 비교 실험은 앙상블 학습을 적용하였고, 기본 분류기로서 모두 C4.5를 사용하였다. <표 3>의 실험결과는 F1 값의 결과를 비교한 것으로 (a)는 제안하는 방법에 의한 F1 결과이며, (b)는 [16]의 실험결과인 F1이다.

두 가지 데이터의 실험 결과는 단순히 AdaBoost만을 적용한 결과보다(F1:0.4260, 0.6561) 모두 우수한 결과를 나타냈다. AdaC1, AdaC2, AdaC3, AdaCost, CSB2와 같은 기존 연구는 각기 다른 불균형 데이터의 비율을 결정하고 실험을 통하여 가장 높은 F1의 결과를 보이는 지점을 찾았다. 유방암 데이터는 1:0.6의 비율을 갖고 AdaC3을 적용한 실험에서, pima 데이터는 1:0.9의 비율을 갖고 AdaC2를 적용한 실험에서 가장 우수한 결과를 보였다. 하지만 제안하는 방법에 의한 실험에서는 각각의 데이터에 15%, 10% 데이터를 선택한 경우 0.7708, 0.8400으로 보다 높은 결과를 나타냈다.

〈표 3〉 F1 값의 결과 비교

(a) 제안하는 방법에 의한 F1 결과

실험데이타	F1	실험데이타	F1
br_10	0.4494	pima_10	0.8400
br_15	0.7708	pima_15	0.6938
br_20	0.6956	pima_20	0.6222
br_25	0.7166	pima_25	0.6176
br_30	0.6629	pima_30	0.7292

(b) 기존 연구[16]의 F1 결과

	AdaBoost	AdaC1	AdaC2	AdaC3	AdaCost	CSB2
		1:0.6	1:0.6	1:0.6	1:0.4	1:0.1
br	0.4260	0.4477	0.4822	0.4981	0.4872	0.4831
		1:0.6	1:0.9	1:0.9	1:0.3	1:0.7
pima	0.6561	0.6617	0.6869	0.6829	0.6777	0.6498

이러한 결과는 cost 만을 달리 적용하는 실험(cost sensitive approaches)보다 데이터의 분포(data level approaches)를 전처리하여 적용하는 방법이 훨씬 문제를 해결하는데 도움이 된다는 것을 보이고 있다. 그러나 <표 3>의 (a), (b) 실험 결과는 아직 해결하지 못한 문제가 남아있다. 그 문제는 (a)의 실험에서 데이터 선택비율 선정방식과, (b) 실험에서 cost를 설정하는 방법을 자동적으로 찾지 못한다는 것이다.

4.2.2 제안한 방법의 특성 평가

효율적인 분할구간의 비율을 찾기 위하여 상한구간, 하한구간을 각각 10%, 15%, 20%, 25%, 30%로 선정하여 실험하였다. <표 4>는 제안한 방법에 의한 세 가지 학습 데이터의 실험 결과이다. 유용한 데이터로 선정된 결과는 학습에 미치는 영향을 알아보기 위하여 kNN(k Nearest Neighbor) 분류기로 분류한다. kNN 분류기는 지금까지 개발된 분류 알고리즘들 중 가장 간단하면서도 비교적 좋은 성능을 보이는 것으로 평가되고 있다. kNN 분류기는 근접 이웃의 개수인 k값의 변화에 따라서 결과가 달라질 수 있기 때문에 실험을 통하여 분류 정확도가 가장 우수한 k값을 사용하였다. k값을 3, 5, 7, 9로 변화하며 실험한 결과에서는 k값이 5에서 가장 높은 분류 정확도를 얻었다. 데이터 선택 전, 후의 분류 정확도와 사용된 샘플의 수를 비교하였다. 상한 구간 10%, 하한 구간 10% 만을 선택하여 분류하여도 전체 데이터를 사용한 결과보다 우수한 결과를 보였다.

yeast 데이터의 분류 결과는 구분선의 위치가 10%~15%인 지점인 영역에서 우수한 결과를 보였다. 전체 데이터수로 학습한 경우(1484개)와 선택된 샘플만으로 학습한 경우를 비교한 결과 444개, 296개, 742개, 592개, 890개 선택한 순서로 높은 정확도를 보였다. ecoli 데이터의 분류 결과는 구분선의 위치가 20%~25%인 지점인 영역에서 우수한 결과를 보였다. 제안한 방법은 최대 2.27배 패턴 선정 전 보다 성능 향상을 보였다. 이 실험에서는 전체 데이터수로 학습한 경우(336개)와 선택된 샘플만으로 학습한 경우를 비교한 결과 168개, 134개, 200개, 66개, 112개 선택한 순서로 높은

〈표 4〉 제안한 방법에 의한 분류 결과 비교

실험데이터 / 데이터 선택	yeast		ecoli		glass	
	정확도	데이터 수	정확도	데이터 수	정확도	데이터 수
패턴 선정 전	0.4410	1484개	0.3869	336개	0.8411	214개
10% 구간 (패턴 선정 후)	0.5473	296개	0.7727	66개	0.8725	42개
15% 구간 (패턴 선정 후)	0.5630	444개	0.5571	112개	0.8525	64개
20% 구간 (패턴 선정 후)	0.4917	592개	0.8508	134개	0.9148	84개
25% 구간 (패턴 선정 후)	0.5470	742개	0.8809	168개	0.8681	106개
30% 구간 (패턴 선정 후)	0.4898	890개	0.7848	200개	0.8588	128개

〈표 5〉 ecoli 데이터의 구간별 데이터 구성에 따른 실험 결과

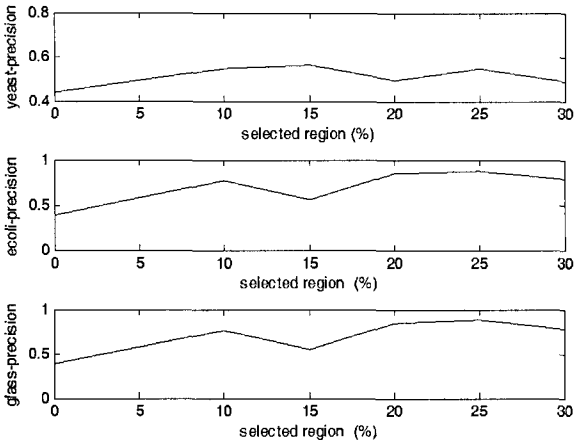
실험 데이터 / 데이터 선택	ecoli(상한구간)		ecoli(상한+하한구간)	
	정확도	클래스수	정확도	클래스수
10% 구간(패턴 선정 후)	0.0000	2	0.7727	6
15% 구간(패턴 선정 후)	0.6588	3	0.5571	6
20% 구간(패턴 선정 후)	0.5994	3	0.8508	7
25% 구간(패턴 선정 후)	0.6904	5	0.8809	6
30% 구간(패턴 선정 후)	0.7700	4	0.7848	7

정확도를 보였다. 매우 괄목할만한 실험 결과로서 클래스 불균형 문제가 심각한 데이터에 매우 효율적이라는 결과를 보여주고 있다. glass 데이터의 분류 결과에서는 구분선의 위치가 25%인 지점에서 0.9148로 매우 우수한 결과를 보였다. 이 실험에서는 전체 데이터 수의 39.35%의 재구성만으로 패턴 선정 전 보다 8.7% 성능 향상을 보였다.

제안된 방법에서 모든 클래스를 학습에 반영하는지 여부를 알아보기 위해 구간별 데이터 구성에 따른 분류 결과를 비교하였다. 아래의 <표 5>에서 상한 구간만을 샘플링하여 학습에 이용하면 데이터 분포의 특성을 잘 반영하지 못하여 효율적으로 모델을 학습 시킬 수 없다. ecoli 데이터는 8가지 클래스 중에서 (3,2) (4,2) 데이터만이 매우 극소수인 특징을 갖는다. ecoli 데이터의 실험 결과는 <표 5>와 같이 상한 구간만을 선정했을 경우 2~5개 정도의 클래스가 선정되었지만, 상한 및 하한 구간을 혼합하여 구성한 데이터에서는 1개의 클래스를 제외한 나머지 클래스들이 모두 학습에 이용되었다.

세 가지 데이터의 실험 결과를 종합하여 (그림 5)와 같이 도표를 나타낸다. 이 도표는 선택된 영역의 범위를 증가시키며 분류결과를 그래프로 표현한 것이다. 분류 정확도는 선택된 영역의 범위가 넓어질수록 상승과 하강을 반복하고 있다.

불균형의 비율 감소를 비교하기 위한 실험에서 주어진 데이터의 불균형의 비율은 <표 1>과 같다. 패턴 선정 이전의 3가지 데이터에서 각 클래스에 해당하는 데이터의 수와의 표준편차는 각각 173.646, 48.738, 29.743 이 나왔다. 표준 편차는 자료의 흩어진 정도를 나타내는 척도로서 이 값이 작



(그림 5) 실세계 데이터의 분류 정확도 비교

〈표 6〉 각 데이터 비율 간 표준 편차

실험 데이터 / 데이터 선택	yeast	ecoli	glass
패턴 선정 전	173.646	48.738	29.743
10% 구간(패턴 선정)	31.230	12.790	26.969
15% 구간(패턴 선정)	49.441	16.954	24.687
20% 구간(패턴 선정)	63.874	25.517	23.200
25% 구간(패턴 선정)	74.429	25.067	21.655
30% 구간(패턴 선정)	90.376	26.216	20.153

다는 것은 평균 데이터 수와 비슷한 비율을 유지한다는 의미를 갖는다. 제안하는 방법에 의하여 각각 31~90, 12~26, 20~26 까지 표준 편차를 줄임으로써 불균형의 비율을 감소시켰다. <표 6>에서는 각 데이터 비율의 표준 편차를 보여준다.

4.3 논의

본 논문에서 클래스 분포의 중심을 재 표현 하는데 SOM을 사용하였다. 이것은 외부의 명시적 지시 없이 주어진 입력 데이터들을 학습하여 그 안에 내재해 있는 특성을 공간상에 놓여 있는 특정 출력 뉴런으로 표출시켜 줄 수 있는 특징을 갖는다. 따라서 그 결과를 클래스 불균형 문제에 해결하는 전처리 과정에 활용한다면 다음과 같은 장점을 얻을 수 있다.

첫째, 클래스 불균형 데이터의 비율(ratio)을 감소시킬 수 있다. 전체 데이터의 분포를 요약된 정보로 재 표현하여 특정 영역안의 데이터만 샘플링 하므로 불균형의 비율을 현저히 감소시킬 수 있다. 즉, 특정 영역안의 데이터들은 클래스의 중심 근처에 몰려있는 데이터와 상대적으로 먼 데이터들로 구성되어 있고, 같은 비율만큼 양쪽 구간에서 샘플링 되어 모든 클래스에 속하는 데이터들을 사용할 수 있다. 결과적으로는 클래스 불균형의 비율을 낮추는 효과를 얻게 된다.

둘째, 데이터의 크기를 감소시켜 학습에 소요되는 메모리 및 계산 복잡도를 감소시킨다. <표 2>의 실험결과와 같이 전체 데이터 yeast 데이터 집합의 1484개의 데이터, ecoli 데이터 집합의 336개, glass 데이터 집합의 214개에서 각각 444개와, 168개로 전체 대비 70%, 50%, 60% 감축된 데이터만으로 향상된 결과를 얻을 수 있다. 따라서 대용량 데이터가 주어졌을 때 여러 모델들 간의 성능을 평가할 경우 전처리 방법으로 활용할 수 있다.

셋째, 학습 파라미터가 과도하게 설정되었을 경우 과적합을 방지한다. 문제의 복잡도를 사전에 알기 어렵기 때문에 이에 적합한 모델의 구조나 학습 파라미터를 설정하는 일은 대부분 경험적으로 이뤄지는데 이는 많은 시행착오를 통해서 이루어지므로 많은 시간이 소요되는 단점이 있다. 또한 이상 패턴이 포함되었을 경우 과적합이 발생하는데, 비감독 학습 방법인 SOM은 이를 방지한다. 제안하는 집중 샘플링 방법은 훈련에 유용한 데이터만을 정련하여 선정하는 방법이므로 분류기로 사용될 모델의 종류나 복잡도에 상관없이 과적합에 대한 대응 필요성이 없게 된다. 신경망과 같은 모델에서는 학습 데이터 셋의 변화에 매우 민감하기 때문에 매 학습시마다 출력 값의 변동이 크면 모델의 신뢰도가 떨어진다. 이러한 경우에도 학습에 유용한 데이터만을 선택한다면 모델의 출력 값 변동이 안정된다.

5. 결 론

본 연구에서는 클래스 불균형 문제를 해소하기 위하여 효율적인 데이터 선택 방법을 제안하였다. 수행과정으로는 첫째, 비감독 학습의 SOM을 적용하여 재 표현된 데이터 분포를 얻고, 클래스 분포의 중심값을 기반으로 거리를 계산하여 스코어를 산정하였다. 둘째, 특정 구간을 설정하고 특정 구간의 상한, 하한 구간으로 분리된 구간 사이의 데이터들은 클래스의 구분을 명확히 할 수 없으므로 제거하였다. 즉, 유용한 데이터만을 이용함으로써 클래스 불균형 문제를 해결하였다. 개선된 집중 샘플링방법은 다음과 같은 장점을 얻었다. 클래스 불균형의 비율 감소, 데이터의 크기 감소, 과적합 방지, 불안정한 모델의 안정화 그리고 메모리 및 계산 복잡도를 감소시켰다. 제안하는 방법을 실제 문제에 대하여 적용한 결과 실험 결과 최대 2.27배까지 분류 정확도의 성능 향상의 결과를 얻었다.

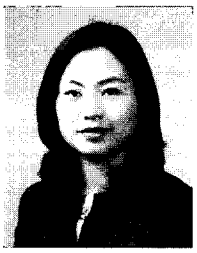
향후 연구 과제로는 상한, 하한 구간을 설정하는 기준을 등 분할에 의존하는 휴리스틱한 방법보다는 최적화 기법을 도입하여 구간을 설정하는 연구가 이루어져야 할 것이다. 최적화 문제들을 풀기 위한 방법으로써 유전자 알고리즘(GA)을 적용하고, 국부최적화를 극복하기 위하여 지역 탐색 연산을 단순 유전자 알고리즘(SGA)에 결합한 혼합 유전자 알고리즘(HGA)에 대한 연구가 필요하다.

참 고 문 헌

[1] T. Fawcett, F. Provost, "Adaptive Fraud Detection, Data Mining and Knowledge Discovery," Vol.1, No.3, pp. 291-316, 1997.

[2] S. Cho, H. Shin, E. Yu, K. Ha, and D. MacLachlan, "Data Mining Problems and Solutions for Response Modeling in CRM," Entrue Journal of Information Technology, Vol.5, No.1, pp.55-64, 2006.

[3] L. Bruzzone, D. Fernàndez Prieto, "A Combined Supervised and Unsupervised Approach to Classification of Multi Temporal Remote Sensing Images," In Proceedings of the IEEE 2000 International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, Hawaii, 24-28, Vol. 1, pp. 162- 164, July, 2000.
[4] R. Yan, Y. Liu, R. Jin, A. Hauptmann, "On Predicting Rare Classes With SVM Ensembles In Scene Classification," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21-24, April, 2003.
[5] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies," in Proceedings of the 2000 International Conference on Artificial Intelligence, pp. 111-117, 2000.
[6] G. M, Foody, "The Significance of Border Training Patterns in Classification By A Feedforword Neural Network Using Back Propagation Learning," International Journal of Remote Sensing, Vol.20, No.18, pp. 3549-3562, 1999.
[7] 신 현정, 조 성준, "신경망 앙상블의 편기와 분산을 이용한 "분류" 패턴 선택," 한국정보과학회 추계학술대회, 2001.
[8] M. Kubat, S. Matwin, "Addressing the Curse of Imbalanced Data Sets: One-Sided Sampling," Proceedings of the Fourteenth International Conference on Machine Learning , pp. 179-186, 1997.
[9] N. Chawla, N. Japkowicz, A. Kolcz, Special Issue on Class Imbalances, SIGKDD Explorations 6(1), June 2004.
[10] X. Liu, J. Wu, Z. Zhou, "Exploratory Under-Sampling for Class-Imbalance Learning," International Conference on Data Mining(ICDM) pp. 965-969, 2006.
[11] H. Shin and S. Cho, "Fast Pattern Selection for Support Vector Classifiers," 7th Pacific-Asia Conference, PAKDD 2003, Seoul, Korea, April 30 - May 2, 2003.
[12] http://www.ics.uci.edu/~mlearn/databases/
[13] Foster Provost, "Machine Learning from Imbalanced Data Sets 101," Learning from Imbalanced Data Sets Papers from the AAAI Workshop, 2005.
[14] Mixture of Expert Agents for Handling Imbalanced Data Sets, annals of mathematics, computing & teleinformatics, Vol 1, no 1, pp. 46-55, 2003.
[15] 오장민, 장병탁, "불균형 데이터의 효과적 학습을 위한 커널 퍼셉트론 부스팅 기법," 한국정보과학회 2001년도 봄 학술발표논문집 제28권 제1호(B), pp. 304-306, 2001.
[16] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong and Yang Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognition, In Press, Corrected Proof, Available online 5 May 2007.
[17] Guobin Ou and Yi Lu Murphey, "Multi-class pattern classification using neural networks," Pattern Recognition, Vol 40, Issue 1, pp. 4-18, 2007.
[18] Jigang Xie and Zhengding Qiu, "The effect of imbalanced data sets on LDA: A theoretical and empirical analysis," Pattern Recognition, Vol 40, Issue 2, pp. 557-562, 2007.
[19] Vicenc Soler, Jesus Cerquides, Josep Sabria, Jordi Roig, Marta Prim, Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), pp. 330-336, 2006.
[20] Yang Liu, Nitesh V. Chawla, Mary P. Harper, Elizabeth Shriberg and Andreas Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," Computer Speech & Language, Vol 20, Issue 4, pp. 468-494, 2006.
[21] Teuvo Kohonen, Self-Organizind Maps:Second Edition, Springer, 1997.

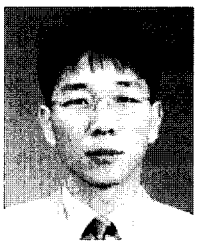


김 만 선

e-mail : kms9688004@nate.com
2000년 홍익대학교 전자전기공학부(학사)
2002년 공주대학교 전산학과(석사)
2005년 공주대학교 컴퓨터공학과(박사)
2006~현재 전남대학교 연구원
관심분야 : 데이터 마이닝, 기계학습, 에이전트, 인공지능, 패턴인식

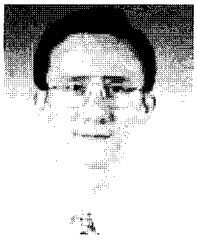
양 형 정

e-mail : hjyang@chonnam.ac.kr
1991년 전북대학교 전산통계학과(학사)
1993년 전북대학교 전산통계학과(석사)
1998년 전북대학교 전산통계학과(박사)
2003-2005 카네기멜런 대학교 연구원
2005~현재 전남대학교 전자컴퓨터공학부 조교수
관심분야 : e-Design, 데이터 마이닝, e-Learning



김 수 형

e-mail : shkim@chonnam.ac.kr
1986년 서울대학교 컴퓨터공학과(학사)
1988년 한국과학기술원 전산학과(공학석사)
1993년 한국과학기술원 전산학과(공학박사)
1993년~1996년 삼성전자 멀티미디어연구소 선임연구원
1997년~현재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 인공지능, 패턴인식, 유비쿼터스컴퓨팅



Wooi Ping Cheah

e-mail : cheahwooiping@gmail.com
1986년 CampbellUniversity, USA(학사)
1993년 University of ScienceMalaysia (석사)
2006년~현재 전남대학교 전산학과 재학중 (박사과정)
1996~2006 Faculty of Information Science and Technology, Multimedia University, Malaysia (lecturer)
관심분야 : software and knowledge engineering, decision support systems, and datamining