KIPS Tr. Software and Data Eng.
Vol.5, No.10 pp.483~488 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.10.483

Prototype-Based Classification Using Class Hyperspheres

Hyun-Jong Lee† · Doosung Hwang††

ABSTRACT

In this paper, we propose a prototype-based classification learning by using the nearest-neighbor rule. The nearest-neighbor is applied to segment the class area of all the training data with hyperspheres, and a hypersphere must cover the data from the same class. The radius of a hypersphere is computed by the mid point of the two distances to the farthest same class point and the nearest other class point. And we transform the prototype selection problem into a set covering problem in order to determine the smallest set of prototypes that cover all the training data. The proposed prototype selection method is designed by a greedy algorithm and applicable to process a large-scale training set in parallel. The prediction rule is the nearest-neighbor rule and the new training data is the set of prototypes. In experiments, the generalization performance of the proposed method is superior to existing methods.

Keywords : Prototype Selection, Nearest-Neighbor Rule, Set Covering Optimization, Greedy Algorithm

클래스 초월구를 이용한 프로토타입 기반 분류

이 현 종† · 황 두 성††

요 약

본 논문은 최근접 이웃 규칙을 이용한 프로토타입을 이용하는 분류 학습을 제안한다. 훈련 데이터가 대표하는 클래스 영역을 초월구로 분할하는데 최근접 이웃규칙을 적용시키며, 초월구는 동일 클래스 데이터들만 포함시킨다. 초월구의 반지름은 가장 인접한 다른 클래스 데이터와 가장 먼 동일 클래스 데이터의 중간 거리 값으로 결정한다. 그리고 전체 훈련 데이터를 대표하는 최소의 프로토타입 집합을 선택하기 위해 집합 덮개 최적화를 이용한다. 제안하는 선택 방법은 클래스 별 프로토타입을 선택하는 그리디 알고리즘으로 설계되며, 대규모 훈련 데이터에 대한 병렬처리가 가능하다. 분류 예측은 최근접 이웃 규칙을 이용하며, 새로운 훈련 데이터는 프로토타입 집합이다. 실험에서 제안하는 방법은 기 연구된 학습 방법에 비해 일반화 성능이 우수하다.

키워드 : 프로토타입 선택, 최근접 이웃 규칙, 집합 덮개 최적화, 그리디 알고리즘

1. 서 론

최근접 이웃 규칙(nearest-neighbor rule)은 테스트 데이터와 가장 가까운 데이터의 클래스로 분류하는 학습 규칙으로 구현이 단순하나 활용이 높은 기계 학습 알고리즘이다[1]. 그러나 대량의 데이터로 구성되는 분류 문제에 적용 시 저장 공간, 데이터 간의 유사도 계산과 정렬시키는 계산량 등이 급격히 증가하는 단점이 나타난다[2]. 프로토타입(prototype)은 기계 학습에서 훈련 데이터의 클래스 내 데이터를 대표할 수 있는 적은 수의 데이터로 정의되며, 이러한 프로토타입의 선택은 최근접 이웃 알고리즘의 앞서 언급된 단점들을 보완할 수 있는 전략으로 이용되고 있다[3]. 프로토타입 선택(prototype selection) 기법으로 중복 또는 잡음 데이터의 제거, 샘플링(sampling)[3, 4], 클래스 간 경계(inter-class boundary) 정보[5, 6], 동일 클래스 내 데이터 분포[7], 클래스를 대표할 수 있는 새로운 데이터 생성(new data editing)[3], 집합 덮개 최적화(set covering optimization)[8, 9] 등을 이용한 알고리즘들이 제안되었다.

프로토타입 기반 분류 학습은 대표 데이터의 선택과 분류 학습의 두 단계로 구성된다. 첫 번째 단계에서 훈련 데이터 간의 유사도와 클래스 정보를 이용하여 프로토타입을 선택한다. 선택된 프로토타입들은 준비된 훈련 데이터의 수보다 적은 수로 구성되며, 각 클래스를 대표하여 학습 분류 알고리즘에 적용가능하다고 가정한다. 두 번째로 선택된 분류 알고리즘은 프로토타입들로 구성된 데이터 집합에 대해 분류 규칙을 학습하고, 테스트 데이터에 대해 분류 예측을 수행한다. 이러한 학습 전략은 어떤 학습 알고리즘에 적용될 수 있으

† 준 회 원 : 단국대학교 컴퓨터과학과 학사과정
†† 종신회원 : 단국대학교 소프트웨어학과 교수
Manuscript Received : March 30, 2016
First Revision : June 3, 2016
Accepted : June 3, 2016
* Corresponding Author : Doosung Hwang(dshwang@dankook.ac.kr)

며, 선택된 소수의 대표 학습 데이터를 이용하여 학습 과정은 낮은 데이터 저장 공간과 계산량을 보장하게 된다[3].

일반적으로 여러 데이터 중 프로토타입의 선택은 데이터 간 유사도를 계산하며, 일정 거리내 중복 또는 근접한 데이터 등을 대표하는 데이터를 선택한다. Tomek Link는 유사도와 분류정보를 같이 사용하여 클래스 분류 경계면에 위치한 프로토타입을 선택하는데 이용되었다. 유사도의 계산방법은 유클리디안 거리(Euclidean distance), 맨하탄 거리(Manhattan distance), 하우스도르프 거리(Hausdorff distance), Tomek link 등이 있다[5-7].

본 논문에서는 초월구(hypersphere)를 이용하는 프로토타입 선택과 학습을 제안한다. 제안하는 프로토타입 선택 방법은 훈련 데이터 간의 유사도와 클래스 정보를 이용한 동일 클래스 내 데이터로만 구성되는 초월구를 계산하고 클래스를 대표하는 적은 수의 프로토타입을 결정한다. 2장에서는 관련 연구에 대해서 토의하고, 3장에서는 집합 덮개 최적화 기반 프로토타입 선택 알고리즘을 기술한다. 4장에서는 알려진 분류 학습 문제에서 제안하는 학습전략의 실험결과를 제시하고 제안된 프로토타입을 이용한 학습결과와 베이지안(Bayesian), k-최근접 이웃(k-nearest neighbors, NN), 지지벡터기계(support vector machine, SVM)의 실험 결과를 비교한다. 마지막으로 5장에서는 제안하는 프로토타입 선택 알고리즘의 문제점과 개선 방향에 대해서 논의한다.

2. 관련 연구

학습 알고리즘의 선택에 따라 새로운 분류 규칙을 찾는 과정은 학습데이터 집합의 크기, 계산 시간 그리고 저장 공간의 크기에 영향을 미친다. 샘플링을 이용한 대표 데이터 선택은 사전에 정한 비율에 따라 프로토타입의 수가 결정된다. 그러나 데이터의 분포를 반영하지 않는 임의 선택(random selection) 방법의 경우에는 분류 예측률이 낮게 보고되었다[3]. 데이터 간의 유사도를 이용한 중복 데이터 또는 잡음 데이터의 제거는 학습데이터의 크기를 줄여 학습시간을 단축시킬 수 있는 전략이다[4, 7].

프로토타입 선택 방법에서 최근접 이웃 규칙은 데이터를 중심으로 상수 거리에 위치한 학습 데이터를 포함하는 영역을 구별하는데 사용되었다[8, 10, 11]. 클래스 영역 내 위치한 데이터 간의 유사도를 계산하여 고정 상수 거리 내에 포함된 데이터로 구성되는 집합들은 초월구 또는 초월사각형(hyperrectangle)들로 구성시켜, 클래스 영역을 분할한다. 대표 학습 데이터 선택 방법은 상수 반지름 거리 내에 동일 클래스 데이터를 가장 많이 포함하는 데이터가 프로토타입이 된다. 이렇게 선택된 프로토타입들의 집합은 새로운 학습데이터가 되며, 원래 학습데이터의 크기보다 적은 수로 구성된다.

IPS(interpretable prototype selection)은 데이터 간 유사도와 고정거리 반지름을 이용하여 클래스 영역을 분할하는 초월구로 구성시키고, 가능한 모든 학습데이터를 포함하는 소수의 프로토타입을 선택하는 최적화 기법을 사용하였다[8]. 프로토타입 선택 문제를 집합 덮개 최적화 문제로 변형시켜 독립된 클래스마다 프로토타입을 선택하는 단계적 알고리즘으로 설계되었다. 프로토타입이 포함하는 데이터 집합은 상이한 클래스에 속한 데이터들도 포함될 수 있으며 사전 실험을 이용하여 구의 반지름을 선택해야 한다.

GSC(greedy sphere covering)는 데이터로부터 동일한 클래스들만으로 구성시킬 수 있는 반지름 계산에 최근접 이웃 규칙을 이용하였다. 최단 거리에 위치한 상이한 클래스까지의 거리를 계산하여 프로토타입이 대표할 수 있는 공간 영역으로 간주하였으며, 가능한 많은 수의 데이터를 포함하는 학습데이터를 프로토타입으로 선택한다[10]. RSC(randomised sphere cover)는 GSC와 동일하게 데이터가 커버하는 클래스 영역의 설정은 같으나, 프로토타입 영역 내 포함되는 데이터 수를 고려하며 분류 경계면에 위치한 잡음 데이터를 조절한다[11]. 그리고, 새로운 프로토타입 전략은 이미 선택된 프로토타입 집합이 포함하지 않는 임의 선택된 데이터의 초월구 내 위치한 동일 클래스 데이터의 수가 높은 경우가 된다.

초월사각형을 이용한 다차원 데이터 영역의 분할은 데이터와 차원 축과 평행하는 최소와 최대 벡터를 계산하여 구성된다[12]. 각 초월사각형내 나타나는 데이터가 모두 동일 클래스에 속하면, 새로운 프로토타입은 초월사각형내 나타나는 동일 데이터의 평균이다. 그러나 초월사각형내 다른 초월사각형내 다른 클래스 데이터는 분리 경계면에 위치할 가능성이 높아 유지시켜 데이터의 분포를 반영시키고, 테스트 데이터의 분류예측은 최근접 이웃규칙을 이용한다[13].

Tomek link와 유사도를 이용하여 클래스 분리 경계에 위치한 학습 데이터들로 구성된 새로운 학습 데이터를 생성시켜 분류 예측을 수행하는 프로토타입 선택 알고리즘이 제안하였다[5, 6]. 선택 알고리즘은 분리 경계 영역에 위치한 데이터들을 구별하며, 이미 선택된 데이터, 클래스와 거리 관계를 분석한 정보를 이용하여 프로토타입 집합에 추가할 것인지 여부를 결정한다. 이러한 프로토타입 선택 방법은 클래스 영역을 지배하는 대표 데이터를 선택할 가능성이 낮아 분류 경계면에 위치한 적은 수의 지지벡터(support vector)로 구성되는 SVM(support vector machine)에는 적합하나, 데이터 분포를 가정하는 최근접 이웃, 베이지안, 가우시안(Gaussian) 알고리즘 등의 통계 분류 학습에는 적합하지 않다.

3. 프로토타입 선택 알고리즘

c개 클래스로부터 생성된 분류 문제를 $\chi=\{\chi^1 \cup \chi^2 \cup \cdots \cup \chi^c\}$를 가정하자. 여기서 $\chi^l=\{(x_i, l) \mid i=1, \cdots, n_l\}$는 클래스 l에 속하며 각 x_i는 d-차원의 벡터($x_i \in R^d$)이며 클래스 $l \in \{1, 2, \cdots, c\}$이다. 제안하는 프로토타입 선택 방법은 패턴 분류 문제 χ로부터 각 클래스 내 데이터를 대표할 수 있는 적은 수를 가지는 데이터 집합인 프로토타입 $P=\{P^1 \cup P^2 \cup \cdots \cup P^c\}$를 선택한다. 최근접 이웃 알고리즘의 사용 시 선택된 프로토타입은 클래스의 상수 영역을 대표하는 클래스 데이터로 가

정되어 영역 내 위치하는 테스트 데이터의 분류는 가장 가까운 프로토타입의 클래스로 예측된다.

주어진 문제의 각 데이터가 포함 가능한 동일 클래스 내 데이터의 집합은 유사도 거리를 이용하여 계산한다. 데이터 (x,l)로부터 거리 r_x내에 위치한 동일 클래스 데이터 집합은 (x,l)가 대표하는 데이터들을 포함한다. 거리 r_x는 모든 데이터와 거리를 구하여 동일 클래스 내 가장 큰 거리값과 다른 클래스 중 가장 작은 값을 갖는 거리의 중간값으로 결정한다. 훈련 데이터 (x,l)와 그의 상수 거리 r_x가 대표하는 집합 $S(x)$는 다음과 같다.

$$S(x)=\{z\,|\,d(x,z)<r_x, l(z)=l(x)\} \tag{1}$$

여기서 $l(x)$는 x의 클래스이다.

주어진 훈련데이터로부터 최소의 클래스 프로토타입들을 선택하는 집합 덮개 최적화 문제는 NP-hard 문제로써 대량의 데이터 처리 시에는 높은 계산 복잡도로 인해 해를 구하는데 많은 시간이 요구된다[14]. GLPK[15]와 같은 소프트웨어를 사용하면 집합 덮개 최적화의 해를 구할 수 있으나 복잡도 때문에 그리디 알고리즘(greedy algorithm), 임의의 라운딩(randomized rounding), 근접 알고리즘(approximation algorithm) 등을 사용하여 해를 구하는 접근 방법이 일반적이다[8, 11, 14].

훈련 데이터의 상수 거리내 위치한 멤버들이 결정되면 프로토타입 선택 과정은 각 클래스별로 독립적으로 구할 수 있다. 그러므로 X로부터 P를 구하는 문제는 c개의 소규모 집합 덮개 최적화 문제의 해이다. 클래스 l의 커버 집합 $S(x)$들로부터 최소의 프로토타입 집합을 선택하는 문제는 다음 Equation (2)와 같이 집합 덮개 최적화의 해가 된다.

$$\min_{\alpha^{(l)}} \sum_{j=1}^{n_l} \alpha_j^{(l)} \tag{2}$$
$$s.t \sum_{j=1,\, x_i \in S(x_j)}^{n_l} \alpha_j^{(l)} \geq 1,\ \forall x_i \in \chi^l$$
$$\alpha_j^{(l)} \in \{0,1\}$$

Equation (2)의 $\alpha_j^{(l)}, j=1,\cdots,n_l$는 라그랑쥐(Lagrange) 변수로 $\alpha_j^{(l)}=1$이면 (x_j,l)은 프로토타입이 되어 동일 클래스 내 부분 영역을 커버하는 새로운 훈련 데이터가 된다. 각 클래스 l의 집합 덮개 문제의 해는 가장 적은 프로토타입들의 선택이 클래스 데이터를 포함시킬 수 있는 경우이다. Equation (3)의 $\triangle obj\,(x)$를 최대로 하는 클래스 데이터 x가 선택될 프로토타입이 된다.

$$\triangle obj(x)=|\,\chi^l \cap S(x) \setminus \cup_{x_i \in P_l} S(x_i)| \tag{3}$$

$\triangle obj\,(x)$는 $x \in \chi^l$가 프로토타입으로 선택될 때 이미 선택된 프로토타입들이 포함시키지 않은 클래스내 데이터의 수가 되며, 추가되는 프로토타입은 (x,l)이다.

```
procedure selectPrototypes(X, S, c):
// χ = {(x_i, l) | i = 1, ..., n and l ∈ {1, ..., c}}
// S(x) = {z | d(x, z) ≤ r_x and l(x) = l(z)}
// c: 클래스 수
// P_l, l = 1,2,...,c
P = Φ
for l = 1
  P_l = Φ;   χ^l = { x_i | (x_i, l) ∈ χ }
  do {
    x = max_{(x,l)∈x^l} △obj(x)
    P_l = P_l ∪ {x}
  } while( △obj (x) > 0 )
  P = P ∪ P_l
end for
return P
```

Fig. 1. Prototype Selection Algorithm

Fig. 1은 분류문제 χ, 각 데이터의 대표 집합 S를 입력으로 하여 각 클래스 단위의 프로토타입 선택 알고리즘이다. c는 클래스 수이며 S는 데이터가 커버하는 동일 클래스 데이터의 집합이다. 제안하는 프로토타입 선택 방법의 출력은 클래스별 선택된 프로토타입 집합 $P=\{P^1 \cup P^2 \cup \cdots \cup P^c\}$이며 $|P^l| << |\chi^l|$이다. 훈련데이터를 대표하는 프로토타입 집합 P는 학습 알고리즘의 새로운 훈련 데이터이나 적은 수의 데이터로 구성된다. Fig. 2는 테스트 데이터 x의 최근접 이웃 프로토타입을 이용한 분류 함수이다.

```
procedure predict(x, P):
// x: 테스트 데이터
// P: 프로토타입 집합
y = min_{(x_i,y)∈P} d(x, x_i)
return y
```

Fig. 2. Prototype Based Classification

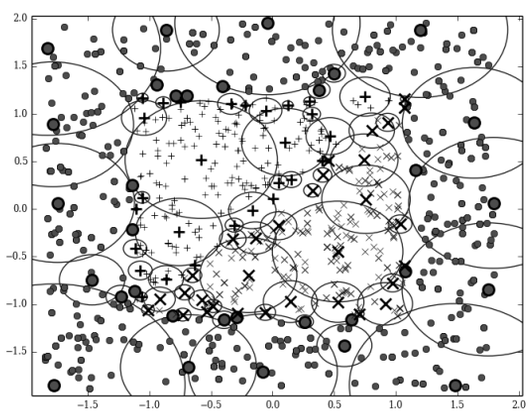


Fig. 3. An Example of PBL

Fig. 3은 프로토타입 선택 예이다. 3-클래스 분류문제의 학습 데이터는 임의로 발생시켰으며, 클래스 별 400개의 데이터를 발생시켜 1,200개 이차원 벡터로 구성된다. 91개의 프로토타입이 선택되어 7.6%의 데이터 선택율을 보이고 있으며, 클래스 분리 경계 영역에서 프로토타입의 선택이 높게 나타났다. 클래스 분리 경계 영역에서는 서로 다른 클래스 영역을 대표하는 작은 반지름의 초월구들의 선택 비율이 높기 때문이다.

4. 실험 결과

제안하는 알고리즘의 일반화 성능 비교를 위해 k-최근접 이웃 알고리즘(k-NN), 베이지안(Bayesian), 지지벡터기계(SVM), IPS, GSC, 그리고 PBL(Prototype based learning)의 분류 예측율을 측정하였다. PBL은 제안하는 방법으로 프로토타입을 선택하고 최근접 이웃 규칙을 기반으로 분류한다. 분류 예측율은 UCI로부터 선택된 벤치마크 분류 문제를 대상으로 하였다[16]. k-NN, 베이지안, SVM의 실험은 원래 문제의 학습데이터가 사용되며, IPL, GSC, 그리고 PBL은 선택된 프로토타입에 대해 분류예측을 수행한다.

선택된 문제는 Table 1에 제시되었다. 문제는 2~26클래스를 갖는 다중 분류 문제들이며 6~180개의 속성들이 있다. TAB digits은 기타 악보를 구성하는 프렛(fret) 숫자로써 초보 연주자들이 많이 사용하는 13개의 클래식 곡과 18개의 외국곡, 그리고 22개의 한국 곡들로부터 준비되었다[17]. 추출된 프렛 번호는 0~15까지이며, 클래스별 약 100~300개의 데이터로 구성된다. 각각의 실험은 데이터를 5번 5-fold 교차 검증으로 훈련 데이터와 테스트 데이터를 나누어 실험하였다.

Table 1. Selected Benchmark Classification Problems

분류문제	크기	속성	Num. 속성	Nom. 속성	클래스
A1A	1,605	14	14	0	2
Abalone(ABA)	4,177	8	7	1	8
Australian credit approval(AUS)	690	14	6	8	2
Breast Cancer(BC)	699	10	10	0	2
Car evaluation(CAR)	1,728	6	0	6	4
DNA	2,000	180	0	180	3
Four classes(FC)	862	2	2	0	2
German(GER)	1,000	20	7	13	2
Glass(GLA)	214	9	9	0	7
Letter(LTT)	20,000	16	16	0	26
Mammograhic(MA)	961	6	6	0	2
Mushroom(MU)	8,124	21	0	21	2
Pendigits(PD)	10,992	16	16	0	10
Pima(PIM)	768	8	8	0	2
Satimage(SAT)	6,435	36	36	0	7
Segment(SEG)	2,310	19	19	0	7
Svmguide1(SG)	3,089	4	4	0	2
TAB digits(TAB)	4,784	118	0	118	16
USPS(US)	9,298	256	256	0	10
Vehicle(VE)	946	18	18	0	4

오류율(error rate)은 분류평가에 일반적으로 사용되나, 클래스에 나타나는 데이터의 수에 큰 차이가 보이는 경우에는 객관적 평가 도구로 이용되기 어렵다. 이러한 클래스 불균형 문제(class imbalance problem)에서 오류율은 다수 클래스 데이터의 수에 크게 의존되기 때문이다. 객관적인 평가를 위해 클래스 불균형 문제에 덜 민감한 균형 오류율(balanced error rate) $berr$을 평가도구로 이용한다.

$$berr = \frac{1}{C}\sum_{l=1}^{C}\frac{1}{n_l}\sum_{i=1}^{n_l}1(y(x_i) \neq 1) \tag{4}$$

균형 오류율 berr평가에서 각 클래스에 속한 오류율은 분류 오류율과 동일하게 계산되나, 전체 오류율은 클래스 가중치를 반영하는 클래스 오류율의 평균이 된다. $1-berr$은 균형 정확률(balanced accuracy rate) $bacc$이다.

Table 2는 실험 결과이다. IPS의 반지름은 사전 실험을 통해 각 문제마다 분류 예측율을 측정하여 낮은 $berr$의 반지름을 선택하였다. 지지벡터기계 SVM은 OVO(one vs one) 학습전략을 사용하며 $C=10$이다. SVM1은 RBF 커널($\gamma=3$), SVM2는 POLY 커널($\gamma=5, c_0=0$)을 선택하였다. 학습 알고리즘의 오른쪽 열 %는 데이터 축소율이다. 한편, SVM의 %열은 선택된 지지벡터의 비율이다.

Friedman 순위 검정 케이스를 수행한 결과 전체 IPS, GSC, PBL 등 프로토타입 기반 학습이 1-NN, 3-NN 등과 성능에서 비교 경쟁력을 보이고 있다. 1-NN은 가장 낮은 균형 오류율을 보이고 있으며, IPS와 PBL의 오류율은 3-NN과 RBF 커널 SVM1과의 중간 범위에 나타났다. 베이지안 학습 결과가 가장 높은 오류율을 보였고, GSC는 SVM2보다 약간 낮았다. 전체적으로 IPS, GSC, PBL은 POLY 커널의 SVM보다 균형 오류율이 높았다. Fig. 4는 Table 2의 문제에 대한 균형 정확률 $bacc$을 비교한 그래프이다.

훈련데이터로부터 선택된 프로토타입 비율은 GSC, IPS, PBL 순으로 낮게 나타났다. SVM1의 지지벡터 비율 32.0%와 비교시 프로토타입의 비율은 16.0~17.0%에서 나타났다. 프로토타입의 비율은 문제에 따라 차이를 보이고 있으나 GSC 알고리즘이 선택한 비율이 약간 낮았다. IPS는 프로토타입이 대표하는 클래스 영역에 상이한 클래스를 포함될 가능성이 높아, 최근접 이웃 분류 규칙을 이용한 분류 오류율을 높일 가능성이 있다.

Table 2. Experimental Results

Problem	1-NN	3-NN	Bayes	SVM1	%	SVM2	%	IPS	%	GSC	%	PBL	%
A1A	0.16	0.19	0.44	0.30	38%	0.21	55%	0.23	19%	0.30	8%	0.19	29%
Abalone	0.23	0.34	0.46	0.44	64%	0.23	57%	0.40	11%	0.24	35%	0.19	42%
Australian	0.08	0.14	0.21	0.13	28%	0.18	55%	0.16	19%	0.18	16%	0.11	21%
Breast Cancer	0.02	0.04	0.10	0.03	10%	0.04	43%	0.06	15%	0.08	5%	0.05	6%
Car Evaluation	0.03	0.05	0.19	0.02	17%	0.02	54%	0.03	19%	0.25	3%	0.16	4%
DNA	0.13	0.20	0.22	0.03	45%	0.26	54%	0.21	19%	0.18	28%	0.08	51%
Fourclass	0.00	0.00	0.27	0.08	31%	0.00	13%	0.14	3%	0.18	2%	0.15	3%
German Numer	0.19	0.30	0.28	0.23	48%	0.24	56%	0.28	19%	0.20	26%	0.17	4%
Glass	0.16	0.40	0.35	0.35	71%	0.35	53%	0.52	18%	0.26	27%	0.27	36%
Letter	0.03	0.05	0.35	0.05	42%	0.35	53%	0.09	19%	0.13	10%	0.09	17%
Mammographic	0.11	0.17	0.22	0.16	34%	0.15	38%	0.19	13%	0.19	16%	0.19	19%
Mushroom	0.00	0.00	0.01	0.00	14%	0.24	55%	0.00	19%	0.06	0%	0.04	1%
Pendigits	0.00	0.00	0.13	0.00	14%	0.26	55%	0.01	20%	0.06	3%	0.04	4%
Pima	0.15	0.26	0.27	0.27	45%	0.24	55%	0.25	21%	0.22	23%	0.17	32%
Satimage	0.06	0.09	0.23	0.13	18%	0.36	38%	0.09	13%	0.13	7%	0.11	12%
Segment	0.02	0.05	0.21	0.05	28%	0.13	52%	0.06	16%	0.12	7%	0.10	10%
Svmguide1	0.02	0.03	0.06	0.03	21%	0.02	60%	0.06	10%	0.05	8%	0.05	12%
TAB Digits	0.00	0.00	0.11	0.00	25%	0.02	6%	0.05	4%	0.23	2%	0.18	2%
USPS	0.00	0.00	0.00	0.00	1%	0.21	13%	0.00	5%	0.02	0%	0.02	0%
Vehicle	0.15	0.20	0.54	0.20	51%	0.33	50%	0.28	16%	0.20	22%	0.17	31%
mean	0.08	0.13	0.23	0.13	32%	0.19	46%	0.16	15%	0.16	12%	0.13	17%
F-평균	1.6	3.5	6.6	5.0		5.5		4.7		5.2		3.8	
순위	1	2	8	5		6		4		7		3	

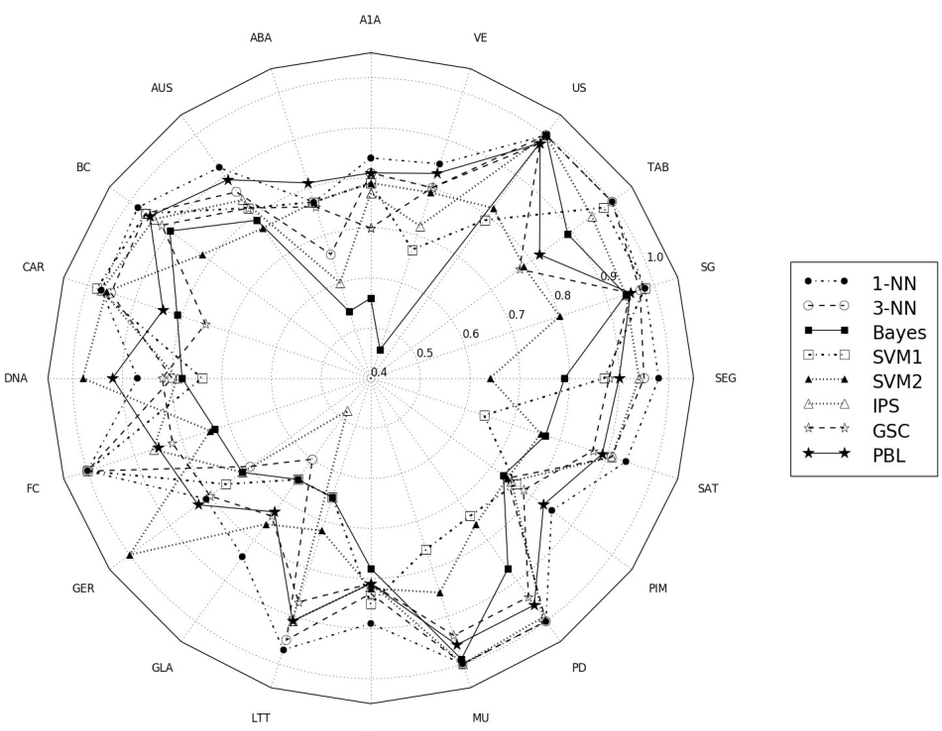


Fig. 4. Comparison of Balanced Accuracy

5. 결 론

본 논문에서는 최근접 이웃 분류 규칙을 이용한 프로토타입 기반 학습 전략을 제안하였다. 프로토타입이 대표할 클래스 내 데이터 집합은 데이터 간의 유사도를 사용하여 상수 거리 내에 위치한 동일 클래스 데이터들로 구성하고, 프로토타입 선택 문제를 집합 덮개 최적화 문제로 변형시켰다. 프로토타입이 대표하는 클래스 영역의 반지름은 자동으

로 결정된다. 변형된 선택 문제의 해를 구하는 문제를 해결하기 위해 그리디 전략을 이용한 프로토타입 선택 알고리즘이 제안되었다.

제안하는 방법은 전체 학습 데이터로부터 클래스 단위로 프로토타입을 선택한다. 클래스 단위별 프로토타입 선택은 계산량이 적으며 동시 처리가 가능하다. 실험에서 제안하는 학습 전략은 기 연구된 초월구 기반 프로토타입 선택 방법과 비교하여 일반화 성능이 우수하였다. 프로토타입을 이용한 분류 학습 전략은 병렬 처리 가능성이 높으며, 전체 훈련데이터를 요약 또는 정리할 수 있는 장점이 있다. 앞으로 처리속도를 높이기 위한 연구, 빅데이터 분석 및 활용, 기계학습 방법의 전처리 등에서 이용될 수 있다.

References

[1] X. Wu et al., "The top ten algorithms in data mining," CRC Press, 2009.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The elements of statistical learning: data mining," *Inference, and Prediction, Springer Series in Statistics*, 2001.

[3] Jose Arturo Olvera-Lopez, Jesus Ariel Carrasco-Ochoa, Jose Francisco Martinez Trinidad, and Josef Kittler, "A review of instance selection methods," *Artif. Intell. Rev*, Vol.34, No.2, pp.133-143, 2010.

[4] S. Garcia, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification : taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.34, No.3, pp.417-435, 2012.

[5] D. S. Hwang and D. W. Kim, "Near-boundary data selection for fast support vector machines," *Malaysian Journal of Computer Science*, Vol.25, No.1, pp.23-37, 2012.

[6] Fabrizio Angiulli, "Fast nearest neighbor condensation for large data sets classification," *IEEE Transactions on Knowledge and Data Engineering*, Vol.19, No.11, pp.1450-1464, 2007.

[7] D. Randall Wilson and Tony R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, Vol.38, No.3, pp.257-286, 2000.

[8] Jacob Bien and Robert Tibshirani, "Prototype selection for interpretable classification," *The Annuals of Applied Statistics*, Vol.5, No.4, pp.2403-2424, 2011.

[9] Ichigaku Takigawa, Mineichi Kudo, and Atsuyoshi Nakamura, "Convex sets as prototypes for classifying patterns," *Engineering Applications of Artificial Intelligence*, Vol.22, No.1, pp.101-108, 2009.

[10] David Marchette, "Class cover catch digraphs," *Wiley Interdisciplinary Reviews : Computational Statistics*, Vol.2, No.2, pp.171-177, 2010.

[11] Reda Younsi and Anthony Bagnall, "An efficient randomised sphere cover classifier," *International Journal of Data Mining, Modelling and Management*, Vol.4, No.2, pp.156-171, 2012.

[12] S. Salzberg, "A nearest hyperrectangle learning method," *Machine Learning*, Vol.6, pp.251-276, 1991.

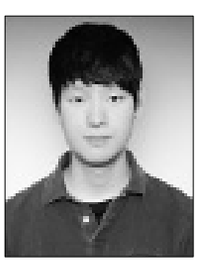
[13] J. Hamidzadeh, R. Monsefi, and H. S. Yazdi, "IRAHC: instance reduction algorithm using hyperrectangle clustering," *Pattern Recognition*, Vol.48, No.5, pp.1878-1889, 2015.

[14] Vijay V. Vazirani, *Approximation Algorithms*, Berlin: Springer, New York, 2001.

[15] GLPK, "The GLPK Linear Programming Kit Package."

[16] K. Bache and M. Lichman, "UCI Machine Learning Repository," University of California, School of Information and Computer Science., 2013.

[17] B. H. Baek et al., "System design and implementation for recognizing and playing guitar tab chords," *Korea Information Processing Society*, Vol.22, No.2, pp.119-112. 2015.



이 현 종

e-mail : guswhd321@naver.com
2013년~현 재 단국대학교
컴퓨터과학과 학사과정
관심분야 : Machine Learning, Image Processing

황 두 성

e-mail : dshwang@dankook.ac.kr
1986년 충남대학교(이학사)
1990년 충남대학교(이학석사)
2003년 Wayne State University(박사)
2003년 단국대학교 컴퓨터과학과 교수
2016년~현 재 단국대학교
소프트웨어학과 교수
관심분야 : Machine Learning, Image Processing, Parallel Processing