Journal of The Korea Society of Computer and Information
Vol. 20, No. 2, February 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.2.021

클래스 영역의 다차원 구 생성에 의한 프로토타입 기반 분류

심 세 용*, 황 두 성**

Prototype based Classification by Generating Multidimensional Spheres per Class Area

Seyong Shim*, Doosung Hwang**

요 약

본 논문에서는 최근접 이웃 규칙을 이용한 프로토타입 선택 기반 분류 학습을 제안하였다. 각 훈련 데이터가 대표하는 클래스 영역을 구(sphere)로 분할하는데 최근접 이웃 규칙을 적용시키며, 구의 내부는 동일 클래스 데이터들만 포함하도록 한다. 프로토타입은 구의 중심점이며 프로토타입의 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 이용하여 프로토타입 선택 문제를 변형시켰다. 제안하는 프로토타입 선택 방법은 클래스 별 적용이 가능한 그리디 알고리즘으로 설계되었다. 제안하는 방법은 계산 복잡도가 높지 않으며, 대규모 훈련 데이터에 대한 병렬처리의 가능성이 높다. 프로토타입 기반 분류 학습은 선택된 프로토타입 집합을 새로운 훈련 데이터 집합으로 사용하고 최근접 이웃 규칙을 적용하여 테스트 데이터의 클래스를 예측한다. 실험에서 제안하는 프로토타입 기반 분류기는 최근접 이웃 학습, 베이지안 분류 학습과 다른 프로토타입 분류기에 비해 일반화 성능이 우수하였다.

▸ Keywords : 프로토타입 선택, 최근접 이웃 규칙, 분류학습, 집합 덮개 최적화, 그리디 알고리즘

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 into spheres within which the data exist from the same class. Prototypes are the center of spheres and their radii are computed by the mid-point of the two distances to the farthest same class point and the nearest another class point. And we transform the prototype selection problem into a set covering problem in order to

•제1저자 : 심세용 •교신저자 : 황두성
•투고일 : 2014. 10. 22, 심사일 : 2014. 12. 21, 게재확정일 : 2015. 1. 26.
* 단국대학교 컴퓨터과학과(Dept. of Computer Science, Dankook University)
** 단국대학교 운동의과학과(Dept. of Kinesiologic Medical Science & Computer Science, Dankook University)

determine the smallest set of prototypes that include all the training data. The proposed prototype selection method is based on a greedy algorithm that is applicable to the training data per class. The complexity of the proposed method is not complicated and the possibility of its parallel implementation is high. The prototype-based classification learning takes up the set of prototypes and predicts the class of test data by the nearest neighbor rule. In experiments, the generalization performance of our prototype classifier is superior to those of the nearest neighbor, Bayes classifier, and another prototype classifier.

▸ Keywords : Prototype selection, Nearest-neighbor rule, Classification learning, Set covering optimization, Greedy algorithm

I. 서 론

최근접 이웃 규칙(nearest-neighbor rule)은 테스트 데이터와 가장 근접한 훈련 데이터의 클래스로 분류하는 알고리즘으로써 구현이 단순하나 실용적 활용이 높은 기계 학습 방법을 제공한다[1]. 하지만 대량의 데이터로 구성된 학습 데이터 또는 중복 데이터의 비율이 높은 데이터의 처리는 데이터의 저장 공간, 데이터 간의 비유사도 계산량 그리고 정렬시키는 계산량 등이 급격히 증가하게 된다[2]. 프로토타입(prototype)은 기계학습에서 훈련 데이터의 클래스 내 데이터를 대표할 수 있는 적은 수의 데이터로 정의되며, 이러한 프로토타입의 선택은 최근접 이웃 알고리즘의 단점을 보완할 수 있는 학습 전략으로 이용되고 있다[3]. 프로토타입 선택(prototype selection) 전략으로는 중복 데이터(duplicate data) 또는 잡음 데이터(noisy data)의 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 데이터간의 비유사도와 클래스 정보를 이용하여 대표 프로토타입을 선택한다. 선택된 프로토타입들은 준비된 훈련 데이터의 수보다 적은 수로 구성되고 각 클래스 데이터를 대표하며 학습 분류 알고리즘에 적용가능하다고 가정한다. 두 번째로 선택된 대표 프로토타입들로 구성된 훈련 데이터 집합들로 학습하고 테스트 데이터에 대해 분류 예측을 한다. 이러한 학습 전략은 어떤 분류 학습기에도 적용될 수 있으며 선택된 소수의 대표 학습 데이터를 이용하여 학습을 하고 낮은 데이터 저장공간과 계산량을 보장하는 분류 학습이 된다[3].

일반적으로 두 데이터 중 어떤 데이터가 프로토타입으로 선택되는지는 데이터간 비유사도를 측정하여 중복 또는 근접한 데이터 중 대표 데이터를 선택한다. Tomek links는 비유사도와 분류정보를 같이 사용하여 클래스 분류 경계면에 위치한 프로토타입을 선택하는데 이용되었다. 비유사도의 계산방법은 유클리디안 거리(Euclidean distance), 맨하탄 거리(Manhattan distance), 하우스도르프 거리(Hausdorff distance), Tomek links 등이 있다[5, 6, 7].

본 논문에서는 집합 덮개 최적화(set covering optimization) 기반 프로토타입 선택 알고리즘을 제안한다. 제안하는 프로토타입 선택 방법은 훈련 데이터 간의 비유사도와 클래스 정보를 이용한 동일 클래스 내 데이터로만 구성된 덮개 집합들(covering set)로부터 클래스 대표 데이터를 선택하는 집합 덮개 최적화 문제로 변형시킨다. 변형된 집합 덮개 최적화 문제의 해를 구하기 위한 그리디 알고리즘(greedy algorithm)을 제안하였으며 분류 실험을 진행하였다. 2장에서는 관련 연구에 대해서 토의하고 3장에서는 집합 덮개 최적화 기반 프로토타입 선택 알고리즘을 기술한다. 4장에서는 알려진 분류 학습 문제에서 제안하는 학습전략의 실험결과를 제시하고 최근접 이웃 알고리즘, 베이지안(Baysian) 알고리즘, 고정된 원의 반지름을 이용한 프로토타입 선택 방법[8]의 실험 결과를 비교한다. 마지막으로 5장에서는 제안하는 프로토타입 선택 알고리즘의 문제점과 개선 방향에 대해서 논의한다.

II. 관련 연구

학습 알고리즘의 선택에 따라 새로운 분류 규칙(classification rule)을 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 않은 임의 선택(random selection)으로 인해 분류 예측율이 낮게 보고되었다. 데이터 간의 비유사도를 이용한 중복 데이터 또는 잡음 데이터의 제거는 학습데이터의 크기를 줄여 학습시간을 단축시킬 수 있는 전략이다[4, 7].

최근접 이웃 규칙은 테스트 데이터의 클래스를 예측하기 위하여 이미 분류된 학습 데이터와 최소 비유사도 값을 갖는 학습 데이터의 클래스를 분류 예측한다. 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 비유사도를 계산하여 특정 상수 거리 내에 포함된 데이터로 구성되는 집합들은 다차원 공간의 구(sphere)들로 구성되며 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.

Bien 등은 데이터 간 비유사도와 고정거리 반지름을 이용하여 잠재적 프로토타입이 포함하는 데이터 영역을 구로 구분하고, 가능한 모든 학습데이터를 포함하는 소수의 프로토타입을 선택하는 알고리즘을 제안하였다. 또한 프로토타입 선택 문제를 집합 덮개 최적화 문제로 정형화시켜 독립된 클래스마다 프로토타입을 선택하는 단계적 알고리즘을 설계하였다. 그러나 잠재적 프로토타입이 포함하는 데이터 집합은 상이한 클래스에 속한 데이터들도 포함될 수 있으며 사전 실험을 이용하여 구의 반지름을 선택해야 하는 단점이 있다[8].

Marchette는 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, 임의의 선택에 따라 모든 학습데이터를 포함시키는 프로토타입 집합을 구성시켰다[10]. 한편, Younsi 등은 프로토타입 영역 내 포함되는 데이터 수를 고려하여 잠재적 분류 경계면에 위치한 잡음 데이터를 조절시켰다[11].

Tomek links와 비유사도를 이용하여 클래스 분리 경계에 위치한 학습 데이터들로 구성된 새로운 학습 데이터를 생성시켜 분류 예측을 수행하는 프로토타입 선택 알고리즘이 제안되었다[5, 6]. 이 프로토타입 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 매우 적은 수의 지지벡터(support vector)로 구성되는 SVM(support vector machine)에는 적합하나, 데이터 분포를 가정하는 베이지안, 가우시안(Gaussian) 알고리즘 등의 통계 기반 분류 학습에는 적합하지 않다.

III. 프로토타입 선택 알고리즘

주어진 분류 문제 $\chi = \{(x_i, y_i) \mid i = 1, \cdots, n\}$에서 각 x_i는 d차원의 벡터($x_i \in R^d$)이며 $y_i \in \{1, \cdots, C\}$이다. χ는 C개의 훈련 데이터 집합으로 구성되어 $\chi = \chi^1 \cup \chi^2 \cup \cdots \cup \chi^C$가 되며 $\chi^c = \{(x_i, c) \mid i = 1, \cdots, n_c\}$는 클래스 c의 훈련 데이터이다. 그리고 $n = \sum_{c=1}^{C} n_c$가 된다. 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 가정되어 영역내 위치하는 테스트 데이터의 분류는 가장 가까운 프로토타입의 클래스로 예측된다.

분류 문제의 각 데이터가 대표하는 동일 클래스 내 데이터의 부분 집합은 비유사도 거리를 이용하여 계산한다. 클래스 c의 데이터 x로부터 거리 r_x내에 위치한 동일 클래스 데이터 집합은 x가 대표하는 데이터들을 포함한다. 거리 r_x는 모든 데이터와 거리를 구하여 동일 클래스 내 가장 큰 거리값 r_1과 다른 클래스 중 가장 작은 값 r_2을 갖는 거리의 중간값으로 결정한다. 클래스 c의 데이터 x와 그의 상수 거리 r_x가 대표하는 집합 $S(x)$는 다음과 같다.

$$S(x) = \{z \mid d(x,z) < r_x, l(z) = l(x)\} \quad \cdots \quad (1)$$

$$r_x = \frac{r_1 + r_2}{2}$$

$$d(x, x_{i_1}) \leq d(x, x_{i_2}) \leq \cdots \leq d(x, x_{i_n})$$

$$r_1 = \max_{l(x)=l(z)} d(x,z)$$

$$r_2 = \min_{l(x)\neq l(z)} d(x,z)$$

여기서 $l(x)$는 x의 클래스이다.

$$\min_{\alpha} \sum_{j=1}^{n} \alpha_j \quad \cdots \quad (2)$$

$$s.t \sum_{j=1, x_i \in S(x_j)}^{n} \alpha_j \geq 1, \forall x_i \in \chi$$

$$\alpha_i \in \{0,1\}$$

주어진 훈련데이터의 대표 집합에서 최소의 클래스 프로토타입들을 선택하는 문제는 수식 (2)와 같이 새로운 집합 덮개 최적화의 해가 된다[13]. 수식 (2)의 α_i는 훈련 데이터 (x_i, c)의 프로토타입 선택 여부를 나타내는 변수이며 0 또는 1이다. 수식 (2)의 해를 구하는 최적화 문제는 NP-hard 문제로써 대량의 데이터 처리 시에는 높은 복잡도로 인해 해를 구하는데 많은 시간이 요구된다[13].

GLPK[12]와 같은 소프트웨어를 사용하면 집합 덮개 최적화의 해를 구할 수 있으나 시간 복잡도 때문에 그리디 알고리즘(greedy algorithm), 임의의 라운딩(randomized rounding), 근접 알고리즘(approxiation algorithm) 등을 사용하여 해를 구하는 접근 방법이 일반적이다[8, 11, 13]. 프로토타입 선택 과정에서 수식 (2)의 해는 훈련 데이터에서 각 클래스를 대표하는 멤버 집합이 결정되면 각 클래스별로 독립적으로 구할 수 있다. 그러므로 수식 (2)는 C개의 소규모 집합 덮개 최적화 문제로 축소된다. 각 클래스로부터 프로토타입 집합의 선택은 가장 적은 대표 데이터로 클래스 내 모든 훈련 데이터를 포함시킨다.

그림 1은 분류문제 χ, 각 데이터의 대표 집합 S를 입력으로 하여 각 클래스 단위의 프로토타입 선택을 하기 위해 제안된 그리디 알고리즘이다. C는 클래스 수이며 τ는 선택된 프로토타입이 대표하는 데이터의 수를 조절하는 상수이다. $\tau = m$이면 선택될 프로토타입은 최소 m개 이상의 동일 클래스 데이터를 대표할 수 있어야 한다. 알고리즘의 출력은 클래스 별 선택된 프로토타입 집합 $P = P^1 \cup P^2 \cup \cdots \cup P^C$이며 $|P^c| << |\chi^c|$이다.

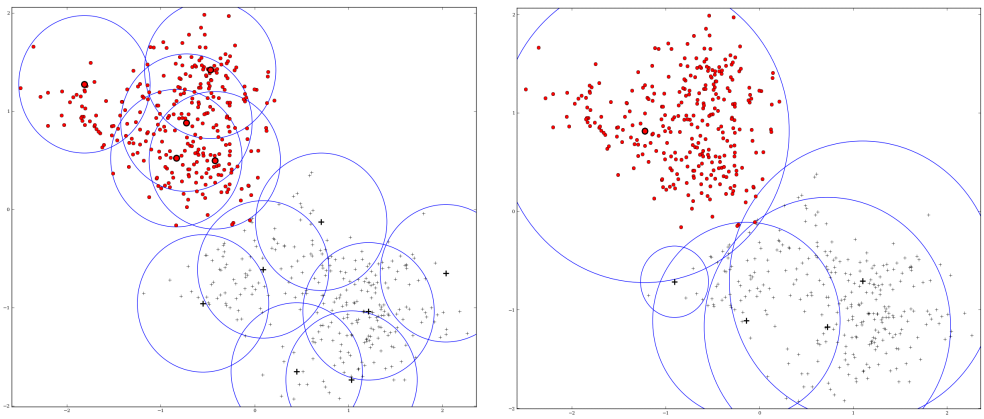
$$\triangle obj(x_j) = |\chi^c \cap S(x_j) \setminus \cup_{x_i \in P^c} S(x_i)| \quad \cdots \quad (3)$$

수식 (3)의 $\triangle obj\ (x_j)$는 $x_j \in \chi^c$가 프로토타입으로 선택 시 클래스 c영역을 대표할 수 있는 데이터의 크기이다. 잠재적 프로토타입은 $\triangle obj\ (x_j)$를 최대로 하는 (x_j, c)이다.

그림 2는 반지름이 고정된 프로토타입 선택 방법[8]과 제

```
Prototype selection(χ, S, C, τ)
// χ = {(x_i, c) | i = 1, ..., n and c ∈ {1, ..., C}}
// S(x) = {z | d(x_j, z) ≤ r_x and l(x_j) = l(z)}
// C : 클래스 수
// τ : 프로토타입 선택 임계값
// P^c, c = 1,2, ..., C
P = Φ
for c = 1 to C do
  P^c = Φ;  χ^c = { x_i | (x_i, c) ∈ χ }
  while △obj (x_j) > 0 and |S(x_j)| ≥ τ do
    x_j = argmax_{x_i ∈ χ^c} △obj (x_i)
    P^c = P^c ∪ {x_j}
  end while
  P = P ∪ P^c
end for
return P
```

그림 1. 프로토타입 선택 알고리즘
Fig. 1. Prototype selection algorithm



(a) 반지름이 고정된 프로토타입 선택
(a) Prototype selection with a fixed radius

(b) 제안하는 프로토타입 선택
(b) Proposed prototype selection

그림 2. 2개의 클래스에서 프로토타입을 선택한 예
Fig. 2. Examples of the selected prototypes for a 2-class problem

표 1. 선택된 벤치마크 분류 문제
Table 1. Selected benchmark classification problems

데이터	크기	속성	Numeric 속성	Nominal 속성	클래스	출처
DNA	2,000	180	0	180	3	statlog
Glass	214	9	9	0	7	UCI
Liver disorder	345	6	6	0	2	UCI
Svmguide2	391	20	20	0	3	libsvm
USPS	7,291	256	256	0	10	libsvm
Vehicle	846	18	18	0	4	statlog
Wine	178	13	13	0	3	UCI
Abalone	4,177	8	7	1	8	UCI
Ringnorm	7,400	20	20	0	2	DELVE
Twonorm	7,400	20	20	0	2	DELVE

표 2. 훈련 데이터 성능 비교
Table 2. Perfomance comparison of the train data

데이터	3-NN		고정 반지름		베이지안		제안하는 방법		반지름
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차	
DNA	0.88	0.010	0.60	0.020	0.93	0.016	0.93	0.000	0.3
Glass	0.81	0.020	0.56	0.010	0.51	0.076	0.83	0.020	0.3
Liver disorder	0.83	0.020	0.58	0.030	0.61	0.038	0.86	0.020	0.2
Svmguide2	0.87	0.013	0.62	0.060	0.84	0.015	0.94	0.008	0.1
USPS	0.99	0.000	0.80	0.020	0.80	0.005	0.97	0.000	1.0
Vehicle	0.85	0.010	0.65	0.019	0.48	0.019	0.86	0.000	0.5
Wine	0.98	0.010	0.92	0.030	0.98	0.008	0.96	0.010	1.0
Abalone	0.74	0.006	0.49	0.021	0.52	0.006	0.84	0.005	1.0
Ringnorm	0.81	0.002	0.62	0.018	0.99	0.001	0.85	0.004	3.0
Twonorm	0.98	0.001	0.93	0.008	0.98	0.001	0.98	0.001	2.7

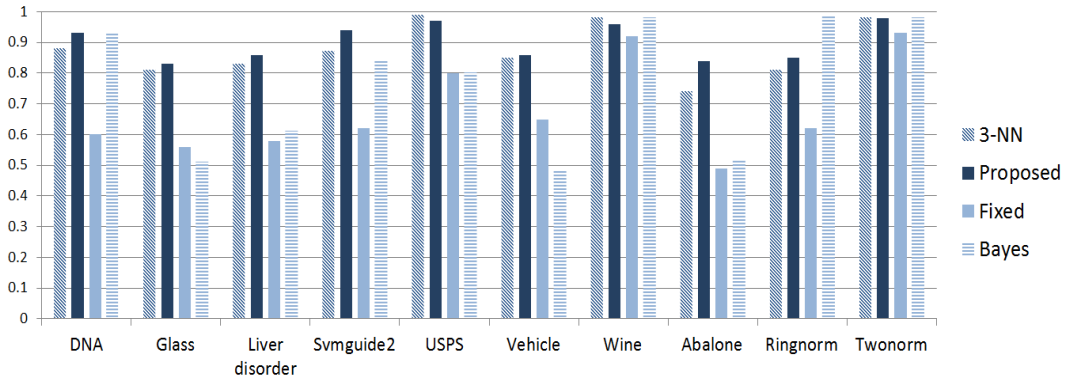


그림 3. 훈련 데이터 성능 비교
Fig. 3 Perfomance comparison of the train data

표 3. 테스트 데이터 성능 비교
Table 3. Perfomance comparison of the test data

데이터	3-NN		고정 반지름		베이지안		제안하는 방법	
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
DNA	0.76	0.020	0.60	0.040	0.93	0.021	0.93	0.000
Glass	0.66	0.070	0.56	0.100	0.43	0.094	0.83	0.050
Liver disorder	0.62	0.040	0.60	0.044	0.59	0.072	0.85	0.050
Svmguide2	0.87	0.011	0.61	0.078	0.79	0.032	0.94	0.026
USPS	0.97	0.000	0.80	0.020	0.80	0.009	0.97	0.000
Vehicle	0.70	0.030	0.65	0.043	0.45	0.028	0.86	0.030
Wine	0.96	0.030	0.92	0.050	0.97	0.024	0.96	0.030
Abalone	0.74	0.006	0.48	0.025	0.52	0.011	0.84	0.009
Ringnorm	0.81	0.002	0.62	0.018	0.99	0.002	0.85	0.007
Twonorm	0.98	0.001	0.93	0.010	0.98	0.003	0.98	0.003

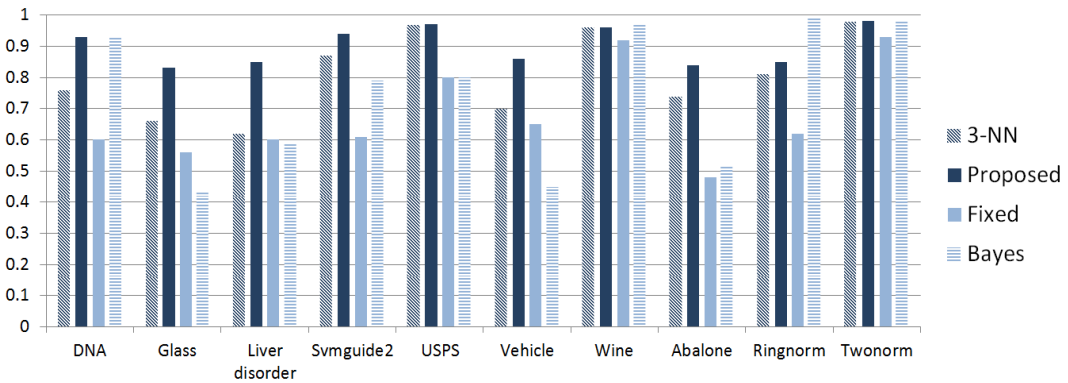


그림 4. 테스트 데이터 성능 비교
Fig 4. Perfomance comparison of the test data

안하는 방법의 프로토타입이 구성되어 있는 예이다. 학습 데이터는 임의로 발생시켰으며 이진 클래스 문제에 대한 실험결과이다. 고정된 원의 반지름은 0.7이다. 고정된 반지름을 이용한 프로토타입 선택 방법 (a)는 12개의 데이터가 선택되었으며 클래스 분리 경계 영역의 프로토타입은 서로 다른 클래스의 데이터가 포함되어 있다. (b)는 제안하는 그리디 알고리

즘의 결과이며, 5개의 프로토타입을 선택하였으며 각 프로토타입 반지름의 선택은 필요하지 않다.

IV. 실험결과

제안하는 알고리즘의 일반화 성능 비교를 위해 K-최근접 이웃 알고리즘(K-NN), 베이지안(Bayesian), 그리고 고정된 원의 반지름을 이용한 프로토타입 선택 알고리즘의 분류 예측율을 측정하였다. 분류 예측율은 선택된 벤치마크 분류 문제에 대해 테스트를 진행하였고 선택된 문제는 표 1에 제시되었다[14, 15 ,16]. 문제는 2~10클래스를 갖는 다중 분류 문제들이며 6~180개의 속성을 갖고 있다. 각각의 실험은 데이터를 10-겹 교차 검증(10-fold cross validation)으로 훈련 데이터와 테스트 데이터를 나누어 실험하였고 그 실험을 5번씩 실행한 결과에 대해서 평균값을 계산하였다. 표 2와 3은 학습과 테스트에 대한 실험 결과이다. 고정된 원의 반지름 값은 사전 실험을 통해 각 문제마다 분류 예측율을 측정하여 가장 높게 측정된 반지름을 선택하였다.

실험 결과, 표 2에서 제안하는 알고리즘은 훈련 데이터보다 작은 크기의 프로토타입을 사용하지만 훈련 데이터전체를 사용하는 3-최근접 이웃 알고리즘의 분류 예측율과 유사하거나 더 우수한 결과를 보였다. 베이지안 알고리즘과 비교 시 Glass, Liver disorder, Vehicle, Abalone 등에서 제안하는 알고리즘의 분류 예측율은 매우 큰 차이를 보이고 있다. 고정된 원의 반지름을 사용하는 프로토타입 선택 알고리즘과 비교에서는 모든 문제에서 제안하는 학습의 일반화 성능이 높았다. 이러한 이유는 제안하는 알고리즘은 클래스 영역의 데이터 분포를 반영하는 프로토타입을 선택하기 때문이다.

제안하는 프로토타입 선택 전략은 고정 반지름을 사용하는 경우보다 선택되는 프로토타입 수가 많아질 가능성이 있다. 고정된 원의 반지름을 사용하는 프로토타입 선택 알고리즘과 제안하는 프로토타입 선택 알고리즘이 선택한 프로토타입 수를 표 4에서 비교하였다($\tau=1$). 실험 결과를 보면 제안한 알고리즘이 선택한 프로토타입 수는 전체 데이터의 25~75%이고 고정 반지름을 사용한 경우는 1~10%로 제안한 알고리즘이 선택한 프로토타입 수는 고정된 반지름의 경우보다 높게 나타났다. 선택되는 프로토타입 수는 τ에 의해 영향을 받을 수 있다. 그리고 고정된 반지름을 사용한 프로토타입 선택 방법은 프로토타입이 대표하는 클래스 영역에 상이한 클래스를 포함될 가능성이 높다. 이러한 문제는 최근접 이웃 분류 규칙을 이용 시 일반화 성능을 높이는데 장애가 될 수 있다.

표 4. 선택된 프로토타입 수 비교
Table 4. Comparison of the number of selected prototype

데이터	크기	제안하는 방법	고정 반지름
DNA	2,000	1,476.2(73.8%)	51.1(2.6%)
Glass	214	110.3(51.6%)	17.3(8.1%)
Liver disorder	345	226.0(65.5%)	28.1(8.1%)
Svmguide2	391	297.3(76.0%)	11.3(2.9%)
USPS	7,291	2,029.0(27.8%)	81.5(1.1%)
Vehicle	846	475.6(56.2%)	96.4(11.4%)
Wine	178	46.0(25.8%)	17.9(10.1%)
Abalone	4,177	2,864.6(68.6%)	50.3(1.2%)
Ringnorm	7,400	3,444.7(46.5%)	98.0(1.3%)
Twonorm	7,400	4,849.5(65.5%)	325.9(4.4%)

V. 결 론

본 논문에서는 최근접 이웃 분류 규칙을 이용한 프로토타입 기반 학습 전략을 제안하였다. 잠재적 프로토타입이 대표할 클래스 내 데이터 집합은 데이터 간의 비유사도를 사용하여 상수 거리 내에 위치한 동일 클래스 데이터들로 구성하고, 프로토타입 선택 문제를 집합 덮개 최적화 문제로 정형화시켰다. 구성된 프로토타입 집합들로부터 모든 데이터를 포함시키는 프로토타입 선택은 집합 덮개 최적화 알고리즘의 해가 된다. 변형된 선택 문제의 해를 구하는 문제를 해결하기 위해 그리디 방법을 이용한 프로토타입 선택 알고리즘을 제안하였다.

실험에서 제안하는 학습 전략은 고정된 원의 반지름을 이용한 학습, 베이지안 학습, 그리고 최근접 이웃 학습에 비해 일반화 성능이 우수하였다. 그러나 선택된 프로토타입 수는 제안하는 방법이 고정된 반지름을 이용한 방법보다 높게 나타났다. 제안하는 방법은 데이터의 분포를 고려하여 동일 클래스 데이터만으로 구성된 프로토타입을 구성하기 때문에 고정 반지름을 이용한 방법보다 프로토타입 수가 높게 나타날 수 있었다. 제안하는 방법의 성능을 유지하면서 프로토타입 수를 줄일 수 있는 연구가 필요하다. 그리고 제안하는 프로토타입 선택 전략은 병렬 처리의 가능성이 높아 향후 연구 방향으로 분류속도를 높이기 위한 연구로 진행할 수 있으며, 빅데이터 응용과 분석에서 활용될 가능성이 높다.

REFERENCES

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

〔2〕 T. Hastie, R. Tibshirani, and J. Friedman, "*The Elements of Statistical Learning: Data Mining*," Inference, and Prediction, Springer Series in Statistics, 2001.
〔3〕 J. Arturo Olvera-Lopez, J. Ariel Carrasco-Ochoa, J. Francisco Martinez Trinidad, and J. Kittler, "A review of instance selection methods," Artif. Intell. Rev Vol. 34, No. 2, pp. 133-143, Aug. 2010.
〔4〕 S. Garcia, J. Derrac, J. Cano, and F. Herrera, "Prototype Selection for Nearest Neighbor Classification : Taxonomy and Empirical Study," IEEE Transactions on Pattern Analysis and Machone Intelligence, Vol. 34, No. 3, pp. 417-435, Mar. 2012.
〔5〕 D. S. Hwang and D. W. Kim, "Near-boundary data selection for fast support vector machines," Malasian journal of Computer Science, Vol. 25(1), pp. 23-37, Mar. 2012
〔6〕 F. Angiulli, "Fast Nearest Neighbor Condensation for Large Data Sets Classification," IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 11, pp. 1450-1464, Nov. 2007.
〔7〕 D. R. Wilson, and T. R. Martinez, "Reduction Techniques for Instance-BasedLearning Algorithms," Machine Learning, Vol. 38, No. 3, pp. 257-286, Mar. 2000.
〔8〕 J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," The Annuals of Applied Statistics Vol. 5, No. 4, pp. 2403-2424, Dec, 2011.
〔9〕 I. Takigawa, M. Kudo, and A. Nakamura, "Convex sets as prototypes for classifying patterns," Engineering Applications of Artificial Intelligence, Vol. 22, No. 1, pp.101-108, Feb. 2009.
〔10〕 D. Marchette, "Class cover catch digraphs," Wiley Interdisciplinary Reviews : Computational Statistics Vol. 2, No. 2, pp. 171-177, Mar. 2010.
〔11〕 R. Younsi, and A. Bagnall, "An efficient randomised sphere cover classifier," Int. J. of Data Mining, Modelling and Management, Vol. 4, No. 2, pp.156-171, Jan. 2012.
〔12〕 GLPK, The GLPK Linear Programming Kit Package, *https://www.gnu.org/software/glpk/*
〔13〕 Vijay V. Vazirani, "*Approximation Algorithms*," Springer, 2001.
〔14〕 UCI Machine Learning Repository, *http://archive.ics.uci.edu/ml/*
〔15〕 The DELVE Manual, *http://www.cs.utoronto.ca/~delve/*
〔16〕 Stalog project, *http://www1.maths.leed.ac.uk/~charles/statlog/indexdos.html*
〔17〕 K. S. Kim and D. S. Hwang "Support Vector Machine Algorithm for Imbalanced Data Learning," Journal of the Korea Society of Computer and Information, Vol. 15, No. 7, pp. 11-17, July. 2010

저 자 소 개

심 세 용

2013: 단국대학교 멀티미디어학과 공학사
현 재: 단국대학교 전자계산학과 석사과정
관심분야: Machine Learning, Parallel Processing
Email : sheyong88@gmail.com

황 두 성

1986: 충남대학교 이학사
1990: 충남대학교 이학 석사
2003: Wayne State University 박사
현 재: 단국대학교 컴퓨터과학과, 운동의과학과 부교수
관심분야: Machine Learning, Parallel Processing, Semantic Web
Email : dshwang@dankook.ac.kr