KIPS Trans. Softw. and Data Eng.
Vol.9, No.3 pp.83~90 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.3.83

Hyper-Rectangle Based Prototype Selection Algorithm Preserving Class Regions

Byunghyun Baek† · Seongyul Euh†† · Doosung Hwang†††

ABSTRACT

Prototype selection offers the advantage of ensuring low learning time and storage space by selecting the minimum data representative of in-class partitions from the training data. This paper designs a new training data generation method using hyper-rectangles that can be applied to general classification algorithms. Hyper-rectangular regions do not contain different class data and divide the same class space. The median value of the data within a hyper-rectangle is selected as a prototype to form new training data, and the size of the hyper-rectangle is adjusted to reflect the data distribution in the class area. A set cover optimization algorithm is proposed to select the minimum prototype set that represents the whole training data. The proposed method reduces the time complexity that requires the polynomial time of the set cover optimization algorithm by using the greedy algorithm and the distance equation without multiplication. In experimented comparison with hyper-sphere prototype selections, the proposed method is superior in terms of prototype rate and generalization performance.

Keywords : Prototype Selection, Prototype, Hyper-Rectangle, Set Cover Optimization Algorithm

클래스 영역을 보존하는 초월 사각형에 의한 프로토타입 선택 알고리즘

백 병 현† · 어 성 율†† · 황 두 성†††

요 약

프로토타입 선택은 훈련 데이터로부터 클래스 영역을 대표하는 최소 데이터를 선택하여 낮은 학습 시간 및 저장 공간을 보장하는 장점을 제공한다. 본 논문은 모든 분류 알고리즘에 적용할 수 있는 초월 사각형을 이용한 새로운 훈련 데이터의 생성 방법을 설계한다. 초월 사각형 영역은 서로 다른 클래스 데이터를 포함하지 않으며 클래스 공간을 분할한다. 선택된 초월 사각형 내 데이터의 중간값은 프로토타입이 되어 새로운 훈련 데이터를 구성하고, 초월 사각형의 크기는 클래스 영역의 데이터 분포를 반영하여 조절된다. 전체 훈련 데이터를 대표하는 최소의 프로토타입 집합 선택을 위해 집합 덮개 최적화 알고리즘을 설계했다. 제안하는 방법에서는 탐욕 알고리즘과 곱셈 연산을 포함하지 않은 거리 계산식을 이용하여 집합 덮개 최적화 알고리즘의 다항 시간을 요구하는 시간 복잡도 문제를 해결한다. 실험에서는 분류 성능의 비교를 위해 최근접 이웃 규칙과 의사 결정 트리 알고리즘을 이용하며 제안하는 방법이 초월 구를 이용한 프로토타입 선택 방법보다 우수하다.

키워드 : 프로토타입 선택, 프로토타입, 초월 사각형, 집합 덮개 최적화 알고리즘

1. 서 론

클래스 영역의 분포를 반영하여 수집된 훈련 데이터는 높은 일반화 성능을 보장하는 모델 학습이 가능하다. 대용량 훈련 데이터의 경우 학습 모델의 복잡도를 증가시키는 중복 및 잡음 데이터 등이 존재할 가능성이 높으며, 많은 학습 시간이 요구된다[1, 2]. 훈련 데이터를 대표하는 소수의 프로토타입 선택은 불필요한 데이터의 제거, 학습 시간의 단축 그리고 기존 훈련 데이터와 유사한 분류 성능을 얻을 수 있으며 높은 일반화 성능을 갖는다는 장점이 있다[3-8].

소수의 프로토타입 선택은 훈련 데이터 사이의 유사도 및 클래스 정보를 이용하며, 데이터 간 유사도 계산 방법은 유클리디안 거리(Euclidean distance), 맨하탄 거리(Manhattan distance), 마할라노비스 거리(Mahalanobis distance) 등이 있다[9]. 곱셈 연산이 포함된 유클리디안 거리를 이용하는 경우, 대용량 데이터의 유사도 계산에 많은 시간이 소요되어 낮은 계산 복잡도를 갖는 유사도 계산 방법이 필요하다.

† 준 회 원 : 단국대학교 소프트웨어학과 석사과정
†† 정 회 원 : 단국대학교 소프트웨어학과 박사과정
††† 종신회원 : 단국대학교 소프트웨어학과 교수
Manuscript Received : August 20, 2019
First Revision : October 15, 2019
Accepted : November 27, 2019
* Corresponding Author : Doosung Hwang(dshwang@dankook.ac.kr)

프로토타입은 상수 거리의 영역 내 위치한 동일 클래스 데이터들을 대표한다. 상수 거리의 영역을 다차원 공간의 초월구(hyper-sphere)들로 구성하고 클래스 영역을 분할하는 연구가 진행되었다[10-13]. 선택된 프로토타입은 영역에 동일 클래스 데이터를 가장 많이 포함하는 데이터이며, 선택된 프로토타입들은 새로운 학습 데이터로 구성된다.

본 논문에서 제안하는 방법은 초월 사각형을 이용한 프로토타입 선택 알고리즘을 설계한다. 훈련 데이터 사이의 유사도와 클래스 정보를 이용하여 동일 클래스 데이터만을 포함하는 초월 사각형(hyper-rectangle)을 구성하고, 각 클래스를 대표하는 적은 수의 프로토타입을 결정한다. 2장에서는 관련 연구에 대해 토의하고, 3장에서는 제안하는 방법의 알고리즘을 기술한다. 4장에서는 제안하는 방법과 초월 구를 이용한 프로토타입 선택방법을 비교하기 위해 최근접 이웃 규칙, 의사 결정 트리 알고리즘을 적용한 실험 결과를 제시한다. 5장에서는 개선 방향에 대해서 논의한다.

2. 관련 연구

일반적으로 지도학습 알고리즘은 학습 데이터의 크기와 분류 규칙의 복잡도에 따라 학습 단계에서 높은 계산량이 요구된다. 샘플링(sampling)을 이용한 프로토타입의 선택은 사전에 정의된 비율에 따라 프로토타입의 수가 결정된다. 그러나 데이터의 분포를 반영하지 않는 임의 선택(random selection) 방법의 경우, 분류 예측률이 낮다는 연구 결과가 보고되었다[14].

IPS(interpretable prototype selection, [10])는 데이터 간 유사도와 고정 거리 반지름을 이용하여 클래스 영역을 분할하는 초월 구를 구성하며, 가능한 모든 학습 데이터를 포함하는 소수의 프로토타입을 선택하기 위한 최적화 기법을 제안하였다. 이 기법은 프로토타입 선택 문제를 집합 덮개 최적화 문제로 변형시켜 클래스마다 독립적으로 프로토타입을 선택하는 단계적 알고리즘이 설계되었다. 그러나 프로토타입이 포함하는 데이터 집합에 다른 클래스의 데이터들도 포함될 수 있으며, 사전 실험을 통해 초월 구의 반지름을 선택하는 비용이 발생한다.

GSC(greedy sphere covering, [11])는 최근접 이웃 규칙을 이용하여 동일 클래스의 데이터들만 포함하는 초월 구의 반지름을 구한다. 최단 거리에 위치한 다른 클래스 데이터까지의 거리를 계산된 값은 프로토타입이 대표할 수 있는 공간 영역으로 간주하며, 가능한 많은 수의 데이터를 포함하는 학습 데이터가 프로토타입으로 선택된다.

RSC(randomized sphere cover, [12])는 GSC와 동일하게 데이터가 커버하는 클래스 영역의 설정은 같으나, 프로토타입 영역 내 포함되는 데이터의 수를 고려하며 분류 경계면에 위치한 잡음 데이터를 조절한다. 선택된 프로토타입 집합에 포함되지 않은 데이터 중 임의의 데이터가 구성한 초월 구 내 동일 클래스의 데이터 수가 많은 경우, 새로운 프로토타입을 생성한다.

PBL(prototype based learning, [13])은 프로토타입 영역 내 포함되는 데이터를 고려한다는 측면에서 RSC와 유사하다. 그러나 가장 근접한 서로 다른 클래스 데이터의 거리와 가장 떨어져 있는 동일 클래스 데이터 거리의 중간을 반지름으로 설정하여 초월구를 구성한다.

Tomek link와 유사도를 이용하여 클래스 분리 경계에 위치한 학습 데이터들로 구성된 새로운 학습 데이터를 생성시켜 분류 예측을 수행하는 프로토타입 선택 알고리즘이 제안되었다[15, 16]. 제안된 방법은 분리 경계 영역에 위치한 데이터들을 구별하며, 이미 선택된 데이터 및 클래스와의 거리 정보를 이용하여 프로토타입 집합에 추가할 것인지 결정한다. 이러한 프로토타입 선택 방법은 클래스 영역을 지배하는 대표 데이터를 선택할 가능성이 낮아 분류 경계면에 위치한 적은 수의 지지 벡터(support vector)로 구성되는 SVM(support vector machine)에는 적합하나, 데이터 분포를 가정하는 최근접 이웃, 베이지안, 가우시안(Gaussian) 등의 알고리즘에서 높은 일반화 성능을 보장할 수 없다.

기 연구된 프로토타입 선택 방법은 데이터의 유사도를 계산하여 반지름을 구하고, 초월 구를 구성하여 클래스 영역을 분할한다. 모든 데이터를 포함하는 소수의 프로토타입 선택 방법이 제안되었으며, 파라미터를 통해 클래스 경계면에 위치한 잡음 데이터를 조절한다. 그러나 대용량 데이터의 경우 유사도 계산에 많은 시간이 필요하고, 클래스 경계면에 소수의 데이터를 포함하는 프로토타입이 빈번하게 선택되어 높은 프로토타입 선택율을 갖는다는 단점이 있다.

3. 프로토타입 선택 알고리즘

분류 문제 $\mathrm{X}=\{(x_i, y_i)|i=1,\ldots,n\}$는 c개의 클래스 $y_i \in \{1,\ldots,c\}$를 가지며 d차원의 입력 벡터 $x^i \in \mathrm{R}^{\mathrm{d}}$로 가정한다. 특정 클래스 l에 나타나는 데이터 집합은 $\mathrm{X_l}=\{(\mathrm{x_i},\mathrm{l})|\mathrm{i}=1,\ldots,\mathrm{n_l}\}$이며, X_l의 크기는 $n_l=|\mathrm{X}_l|$이다. 따라서, 전체 훈련 데이터의 크기는 $\Sigma_{l=1}^{c} n_l = n$이 된다. 초월 사각형 기반 프로토타입 선택 알고리즘을 훈련 데이터 X에 적용함으로써 프로토타입 집합 $\mathrm{P}=\{\mathrm{P_l}|\mathrm{l}=1,\ldots,\mathrm{c}\}$이 새로운 훈련 데이터가 되며, 각 클래스 l에 대해서 $|\mathrm{P}_l| \ll n_l$이 된다.

3.1 초월 사각형

다차원 훈련 데이터 공간을 구분하는 초월 사각형 $\mathrm{hr}=(\mathrm{min},\mathrm{max})$은 최소 정점 min과 최대 정점 max로 정의되며, 동일 클래스의 데이터만 포함한다. 데이터 $x=(x_1, \cdots, x_d) \in \mathrm{R}^d$ 가 초월 사각형의 영역 내 포함 여부와 초월 사각형의 대각선 길이 계산은 다음 거리 계산식 (1)을 이용한다. Equation (1)은 초월 사각형의 중심인 mid와 초월 사각형의 대각선 길이 r, 그리고 각 차원에서 나타나는 좌표의 차를 통해 거리를 계산하기 때문에 유클리디안 거리 식과 같은 곱셈 연산의 거리 계산보다 시간 복잡도가 감소된다.

$$\mathrm{dist}(x,\mathrm{hr}) = \mathrm{Maximum}_{i\in\{1,\ldots,d\}}\left\{|x-\mathrm{mid}| - \frac{\mathrm{r}}{2}\right\} \quad (1)$$
$$\mathrm{mid} = \frac{\max+\min}{2},\quad r = \max-\min$$

Fig. 1은 hr이 최소정점 $\min=(-0.15,-0.15)$과 최대정점 $\max=(0.15,0.15)$이라고 가정했을 때, 범위 $[-0.5,0.5]$에서 나타나는 Equation (1)의 등고선 그래프를 보여준다. hr의 내부 데이터는 0 또는 음수 값을 가지며, 외부 데이터의 경우는 양수 값으로 계산된다. $\mathrm{dist}(x,\mathrm{hr})\le 0$의 조건을 만족하는 경우 데이터 x가 초월 사각형 hr에 포함되며, $\mathrm{dist}(x,\mathrm{hr})>0$인 경우 x는 hr에 포함되지 않는다. 즉, 초월 사각형을 이용한 프로토타입 $\mathrm{hr}(x)$은 다음과 같이 문제 X의 동일 클래스 데이터를 포함한다.

$$\mathrm{hr}(x) = \{z|\mathrm{dist}(x,z)\le 0\}$$

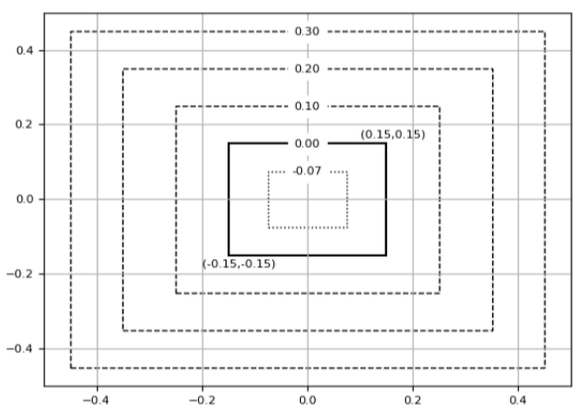


Fig. 1. Contour Plot for Distance Measure Equation (1)

초월 사각형의 크기는 두 정점 min과 max의 길이로 결정되며, 초월 사각형의 최대 크기를 조정하기 위해 파라미터 θ를 정의한다. 초월 사각형의 대각선 길이는 $\theta\ge \mathrm{r}$를 만족함과 동시에 서로 다른 클래스 데이터를 포함하지 않기 때문에 가변적이다.

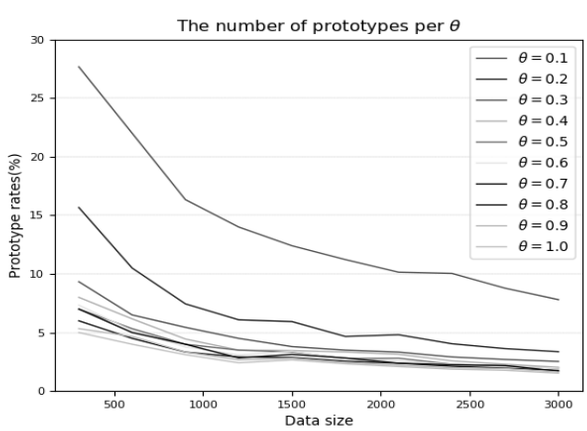


Fig. 2. The Number of Prototypes per θ

Fig. 2는 θ에 따라 선택되는 프로토타입의 수를 보여준다. 사용된 데이터는 3개 클래스의 2차원 데이터이며 300~3000개까지 무작위로 발생시켰다. θ가 큰 값으로 설정된 경우, 초월 사각형이 포함하는 영역이 확대되어 프로토타입 선택 수가 적어진다. 또한, 선택된 프로토타입은 클래스 경계면에서 떨어져 분포되어 있을 가능성이 높다. 작은 θ가 설정된 경우, 초월 사각형이 포함하는 영역이 축소되어 프로토타입 선택 수가 많아지며 클래스 경계면에 선택된 프로토타입이 위치할 가능성이 높아질 수 있다. 따라서, 일반화 성능을 높이기 위해 학습 알고리즘 및 데이터의 분포에 따라 적절한 θ값을 찾는 것이 필요하다.

3.2 프로토타입 선택 알고리즘

훈련 데이터 $x_i\in$X를 포함하는 최소의 프로토타입 집합을 찾기 위해 가변수(dummy variable) $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)$를 설정한다. 데이터 x_i가 프로토타입 P에 포함되는 경우 α_i의 값은 1이 되며, 포함되지 않으면 0이다. 따라서, $\Sigma_{j,x_i\in\mathrm{hr}(\hat{x}_j)}\alpha_j$는 초월 사각형 $\mathrm{hr}(\hat{x}_j)$의 영역에 포함되는 데이터 x_i의 수이다. Equation (2)를 통해 모든 훈련 데이터를 포함하는 최소의 프로토타입을 선택한다.

$$\begin{aligned}&\mathrm{Min}_{\alpha}\ \Sigma_{j=1}^{n}\alpha_j \\ &\text{s.t}\quad \Sigma_{j,x_i\in\mathrm{hr}(\hat{x}_j)}\alpha_j\ge 1\ \forall x_i\in\mathrm{X}\end{aligned} \quad (2)$$

Equation (2)는 집합 덮개 최적화 문제이며 다항 시간(polynomial time)의 시간 복잡도를 갖는 NP-hard 문제이다[14, 17]. 이 문제는 높은 계산 복잡도를 갖기 때문에 데이터의 수가 증가함에 따라 해를 구하기 위해 많은 시간이 소요된다. 초월 사각형 $\mathrm{hr}(\hat{x}_j)$은 서로 다른 클래스 데이터를 포함하지 않기 때문에, 클래스 단위의 분리된 집합 덮개 최적화 문제 전략을 이용하여 병렬 처리를 통해 시간을 단축할 수 있다. $\Sigma_{j,x_i^{(l)}\in\mathrm{hr}(\hat{x}_j^{(l)})}\alpha_j^{(l)}$는 클래스 l을 대표하는 초월 사각형 $\mathrm{hr}(\hat{x}_j^{(l)})$의 영역에 포함되는 데이터 $x_i^{(l)}$의 수가 된다. Equation (3)을 통해 각 클래스의 훈련 데이터를 대표하는 프로토타입을 독립적으로 선택한다.

$$\begin{aligned}&\mathrm{Min}_{\alpha^{(l)}}\ \Sigma_{j=1}^{n_l}\alpha_j^{(l)} \\ &\text{s.t}\quad \Sigma_{j=1}^{n_l}\Sigma_{i=1}^{n_l}\alpha_j^{(l)}1\left(x_i^{(l)}\in\mathrm{hr}\left(\hat{x}_j^{(l)}\right)\right)\ge 1\end{aligned} \quad (3)$$

Equation (3)의 해는 비교적 낮은 계산이 필요한 탐욕 알고리즘(greedy algorithm)을 이용한다. 탐욕 알고리즘은 새롭게 포함되는 같은 클래스 데이터의 비율을 통해 프로토타입을 선택하며, 선택된 프로토타입은 각 클래스의 모든 데이터를 포함한다. $\triangle\mathrm{obj}(\hat{x}_j^{(l)})$는 $\hat{x}_j^{(l)}$이 프로토타입으로 선택되었을 때, 다른 프로토타입에 속하지 않은 같은 클래스 데이터를 영역 내 포함하는 데이터 수이다.

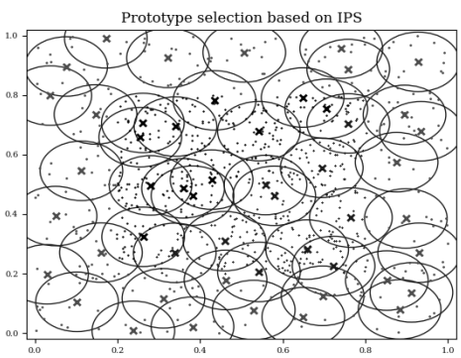
$$\triangle \mathrm{obj}\left(\hat{x}_j^{(l)}\right) = \left|\left(\mathrm{X}^{(l)} \cap \mathrm{hr}\left(\hat{x}_j^{(l)}\right)\right) \setminus \left(\cup_{x_i^{(l)} \in \mathrm{P}_l} \mathrm{hr}\left(\hat{x}_j^{(l)}\right)\right)\right|$$

새로운 훈련 데이터 P는 각 클래스 l을 대표하는 프로토타입 집합 P_l로 구성된다. 선택되는 프로토타입은 $\triangle \mathrm{obj}\left(\hat{x}_j^{(l)}\right)$의 값이 최대가 되는 $\hat{x}_j^{(l)}$이 되며, 새로운 훈련 데이터에 포함된다.

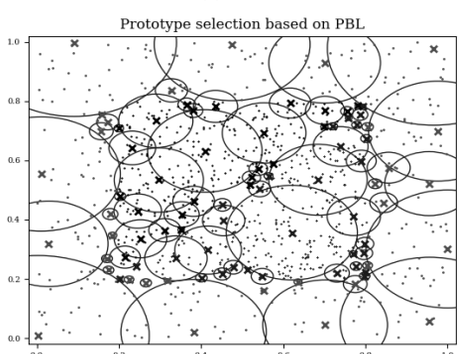
$$\mathrm{P} = \mathrm{P}_1 \cup \cdots \cup (\mathrm{P}_l \cup \{\hat{x}_j^{(l)}\}) \cup \cdots \cup \mathrm{P}_c$$

3.3 알고리즘 비교

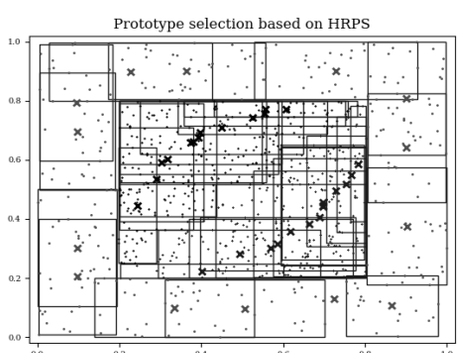
Fig. 3은 프로토타입 선택 알고리즘별 영역 구성과 프로토타입 선택 예이다. 사용된 데이터는 무작위로 생성한 900개의 2차원 데이터이며 3개의 클래스를 갖는다.



(a) IPS



(b) PBL



(c) HRPS

Fig. 3. Prototype Selection Examples of a Toy Problem

(a)는 사전에 정의된 고정 반지름의 초월 구를 이용한 프로토타입 선택방법이다. 고정 반지름 γ는 0.1로 설정하였으며 총 56개의 프로토타입이 선택되었다. (b)는 가변 반지름의 초월 구를 이용한 프로토타입 선택방법이다. 프로토타입 영역 내 포함되는 데이터의 수는 1개 이상이며 총 110개의 프로토타입이 선택되었다. 각 프로토타입의 영역 내 서로 다른 클래스의 데이터를 고려하여 반지름을 설정하기 때문에 반지름의 사전 정의가 필요하지 않다. (c)는 제안하는 방법으로 대각선은 최대 길이 θ는 0.4이며, 총 47개의 프로토타입이 선택되었다.

Fig. 4는 프로토타입 선택 알고리즘별 프로토타입 선택에 소요된 시간을 비교한다. 사용된 데이터는 무작위로 생성한 3개 클래스의 2차원 데이터이며, 300~3000개까지 데이터의 수를 다양화했다. IPS는 프로토타입 영역 내 서로 다른 클래스의 데이터를 고려하지 않기 때문에 프로토타입 선택시간이 적다. PBL의 경우, 프로토타입 영역 내 다른 클래스의 데이터 포함 여부와 곱셈 연산식을 사용하기 때문에 데이터의 수가 증가함에 따라 프로토타입 선택시간이 높게 분석된다. HRPS는 PBL과 같이 프로토타입 영역 내 다른 클래스의 데이터 포함 여부를 계산하지만, 제안하는 거리 계산식을 통해 IPS와 유사한 프로토타입 선택시간을 갖는다.

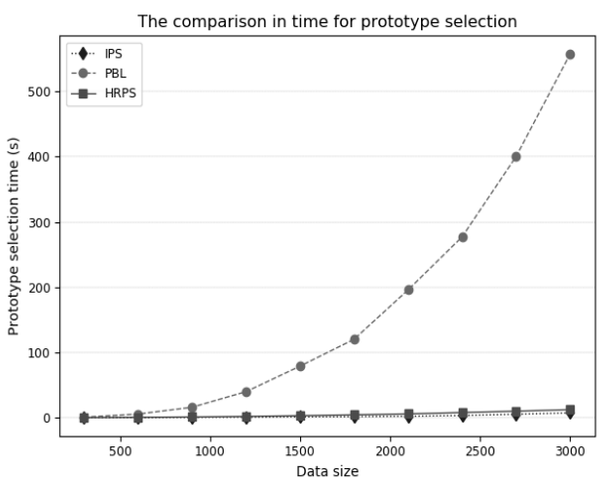


Fig. 4. The Comparison in Time of a Toy Problem

4. 실 험

제안하는 방법과 고정 반지름 및 가변 반지름 초월 구의 성능 비교를 위해 13개의 UCI Machine Learning Repository [18] 데이터를 사용한다. 13개의 데이터 중 2개는 nominal 데이터이며 나머지는 numeric 데이터이다. Table 1은 주어진 각 벤치마크 분류 문제에서 문제의 이름(Problem), 데이터 수(No.), 속성의 수(Attr), 데이터 형식(Numerical[Num], Nominal [Nom]), 클래스 수(Cls)를 정리하였다. 최근접 이웃 규칙과 의사 결정 나무 학습 알고리즘을 사용하여 순수 분류

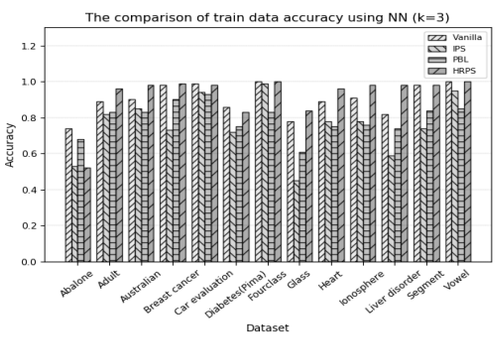
Table 1. UCI Benchmark Problems

Problem	No.	Attr	Num	Nom	Cls
Abalone	4,177	8	8	0	3
Adult	48,842	14	14	0	2
Autralian	690	7	7	0	2
Breast Camcer	683	9	0	9	2
Car Evaluation	1,728	6	0	6	4
Diabetes(Pima)	768	8	8	0	2
Fourclass	862	2	2	0	2
Glass	214	9	9	0	7
Heart	270	13	13	0	2
Ionospere	351	34	34	0	2
Liver Disorder	345	6	6	0	2
Segment	2,310	19	19	0	7
Vowel	990	13	13	0	11

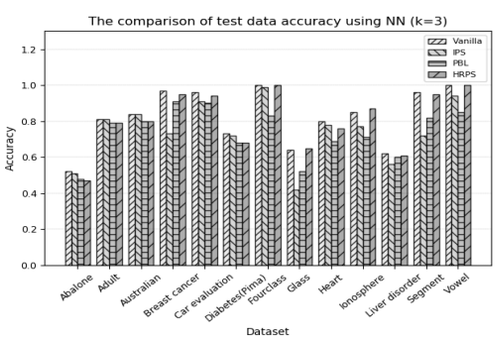
성능 및 3종류의 프로토타입 선택 알고리즘을 적용한 분류 성능을 실험한다. 최근접 이웃 규칙 알고리즘은 클래스 경계면에 위치한 데이터들이 분류 성능에 영향이 미비한 반면에 의사 결정 나무 알고리즘은 클래스 경계면에 위치한 데이터들이 분류 성능에 큰 영향을 준다. 따라서, 프로토타입 선택 알고리즘별 데이터 분포에 따른 분류 성능의 비교를 위해 두 학습 알고리즘을 사용하였고, 3-겹 교차 검증(3-way cross-validation)을 수행했다.

Table 2는 주어진 벤치마크 문제에서 프로토타입을 적용하지 않았을 경우(Vanilla), IPS, PBL, 그리고 HRPS의 훈련 데이터와 테스트 데이터의 정확도(Accuracy)와 F1 score를 나타내며, 사전 실험을 통해 프로토타입 선택 알고리즘별 성능을 최대로 갖는 파라미터를 각각 설정했다.

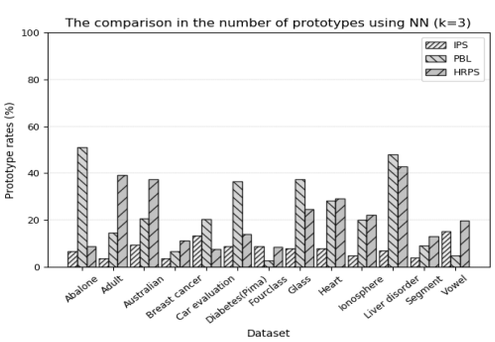
Fig. 5는 벤치마크 분류 문제에 대한 순수 최근접 이웃 규칙 및 각 프로토타입 선택 알고리즘을 적용한 분류 성능을 비



(a) Accuracy of Train Data

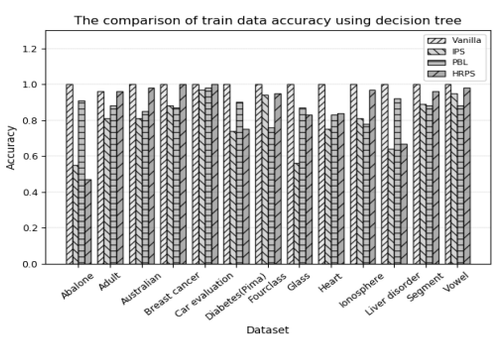


(b) Accuracy of Test Data

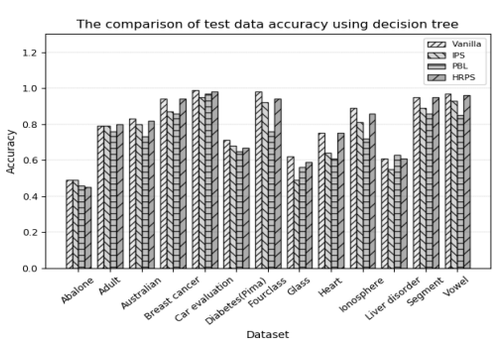


(c) The number of Prototypes

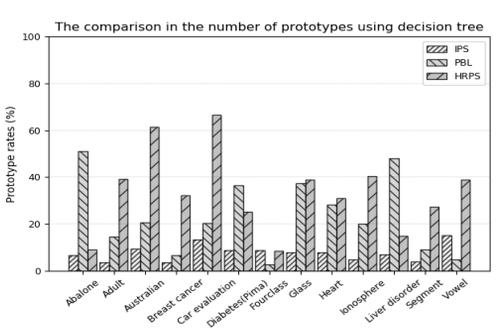
Fig. 5. Performance Comparison using Nearest Neighbor



(a) Accuracy of Train Data



(b) Accuracy of test data



(c) The Number of Prototypes

Fig. 6. Performance Comparison using Decision Tree

교한다. (a)는 훈련 데이터의 분류 성능, (b)는 테스트 데이터의 분류 성능, (c)는 IPS, PBL, HRPS 알고리즘을 통해 선택된 프로토타입의 수를 나타내며 프로토타입 선택 수는 평균적으로 PBL, HRPS, IPS 순으로 높다.

IPS는 선택된 프로토타입 영역 내 서로 다른 클래스를 포함하고 있어 프로토타입 선택 비율이 낮으며, PBL은 클래스 경계면에 소규모 영역을 갖는 프로토타입이 빈번하게 선택되기 때문에 프로토타입 선택 비율이 높은 것으로 분석된다. 제안하는 방법은 기존 훈련 데이터에서 약 7.5%~42.9%를 사용하여 순수 최근접 이웃 규칙과 유사한 분류 성능을 얻었으며 높은 일반화 성능을 보였다.

Fig. 6은 벤치마크 분류 문제에 대한 의사 결정 나무 알고리즘의 성능을 보여준다. 프로토타입 선택 수는 IPS, PBL. HRPS 순으로 적게 선택되었다. 의사 결정 나무 알고리즘은 분류 경계면의 데이터를 기반으로 나무를 구성한다. HRPS에서 초월 사각형의 크기가 큰 경우, 선택된 프로토타입은 클래스 경계면과 떨어진 곳에 위치할 가능성이 높다.

해당 학습 알고리즘은 클래스 경계면에 위치한 데이터가 분류 성능에 큰 영향을 주기 때문에 초월 사각형의 최대 크기를 작게 설정하여 클래스 경계면에 프로토타입을 위치시켰다. 그러나 초월 사각형의 영역이 축소되어 최근접 이웃 규칙보다 프로토타입 선택 비율이 높다. 제안하는 프로토타입 선택방법은 기존 훈련 데이터에서 약 8.5%~66.6%를 사용하여 순수 의사 결정 나무 알고리즘과 유사한 분류 성능을 보였다.

데이터 분포는 선택된 프로토타입의 수와 분류 성능에 큰

Table 2. Classification Performance for Benchmark Problems

Problem		Nearest Neighbor (k=3)				Decision Tree			
		Vanilla	IPS	PBL	HRPS	Vanilla	IPS	PBL	HRPS
Abalone	Accuracy	0.74/ 0.52	0.53/ 0.51	0.68/ 0.48	0.52/ 0.47	1.00/ 0.49	0.55/ 0.49	0.91/ 0.46	0.47/ 0.45
	F1-score	0.73/ 0.51	0.52/ 0.50	0.68/ 0.48	0.52/ 0.47	1.00/ 0.49	0.55/ 0.49	0.91/ 0.46	0.47/ 0.45
Adult	Accuracy	0.89/ 0.81	0.82/ 0.81	0.83/ 0.79	0.96/ 0.79	0.96/ 0.79	0.81/ 0.79	0.88/ 0.76	0.96/ 0.80
	F1-score	0.84/ 0.74	0.75/ 0.74	0.83/ 0.78	0.96/ 0.79	0.95/ 0.71	0.74/ 0.71	0.88/ 0.76	0.96/ 0.80
Australian	Accuracy	0.90/ 0.84	0.85/ 0.84	0.83/ 0.80	0.98/ 0.80	1.00/ 0.83	0.81/ 0.80	0.85/ 0.73	0.98/ 0.82
	F1-score	0.90/ 0.83	0.84/ 0.84	0.83/ 0.80	0.98/ 0.80	1.00/ 0.83	0.81/ 0.80	0.85/ 0.73	0.98/ 0.82
Breast Cancer	Accuracy	0.98/ 0.97	0.73/ 0.73	0.90/ 0.91	0.99/ 0.95	1.00/ 0.94	0.88/ 0.87	0.87/ 0.86	1.00/ 0.94
	F1-score	0.98/ 0.97	0.55/ 0.55	0.90/ 0.91	0.99/ 0.95	1.00/ 0.94	0.85/ 0.84	0.87/ 0.86	1.00/ 0.94
Car Evaluation	Accuracy	0.99/ 0.96	0.94/ 0.91	0.93/ 0.90	0.98/ 0.94	1.00/ 0.99	0.97/ 0.95	0.98/ 0.97	1.00/ 0.98
	F1-score	0.99/ 0.96	0.93/ 0.90	0.93/ 0.90	0.98/ 0.94	1.00/ 0.99	0.97/ 0.95	0.98/ 0.97	1.00/ 0.98
Diabetes(Pima)	Accuracy	0.86/ 0.73	0.72/ 0.72	0.75/ 0.68	0.83/ 0.68	1.00/ 0.71	0.74/ 0.68	0.90/ 0.65	0.75/ 0.67
	F1-score	0.84/ 0.69	0.68/ 0.66	0.75/ 0.68	0.83/ 0.68	1.00/ 0.67	0.71/ 0.65	0.90/ 0.65	0.75/ 0.67
Fourclass	Accuracy	1.00/ 1.00	0.99/ 0.99	0.83/ 0.83	1.00/ 1.00	1.00/ 0.98	0.94/ 0.92	0.76/ 0.76	0.95/ 0.94
	F1-score	1.00/ 1.00	0.99/ 0.99	0.83/ 0.83	1.00/ 1.00	1.00/ 0.98	0.93/ 0.92	0.76/ 0.76	0.95/ 0.94
Glass	Accuracy	0.78/ 0.64	0.45/ 0.42	0.61/ 0.52	0.84/ 0.65	1.00/ 0.62	0.56/ 0.49	0.87/ 0.56	0.83/ 0.59
	F1-score	0.72/ 0.51	0.25/ 0.23	0.61/ 0.52	0.84/ 0.65	1.00/ 0.60	0.46/ 0.41	0.87/ 0.56	0.83/ 0.59
Heart	Accuracy	0.89/ 0.80	0.78/ 0.78	0.75/ 0.69	0.96/ 0.76	1.00/ 0.75	0.75/ 0.64	0.83/ 0.61	0.84/ 0.75
	F1-score	0.89/ 0.79	0.78/ 0.78	0.75/ 0.69	0.96/ 0.76	1.00/ 0.75	0.74/ 0.63	0.83/ 0.61	0.84/ 0.75
Ionoshpere	Accuracy	0.91/ 0.85	0.78/ 0.77	0.76/ 0.71	0.98/ 0.87	1.00/ 0.89	0.81/ 0.81	0.78/ 0.72	0.97/ 0.86
	F1-score	0.90/ 0.82	0.67/ 0.67	0.76/ 0.71	0.98/ 0.87	1.00/ 0.88	0.77/ 0.76	0.78/ 0.72	0.97/ 0.86
Liver Disorder	Accuracy	0.82/ 0.62	0.59/ 0.56	0.74/ 0.60	0.98/ 0.61	1.00/ 0.61	0.64/ 0.55	0.92/ 0.63	0.67/ 0.61
	F1-score	0.82/ 0.61	0.58/ 0.55	0.74/ 0.60	0.98/ 0.61	1.00/ 0.60	0.61/ 0.52	0.92/ 0.63	0.67/ 0.61
Segment	Accuracy	0.98/ 0.96	0.74/ 0.72	0.84/ 0.82	0.98/ 0.95	1.00/ 0.95	0.89/ 0.89	0.88/ 0.86	0.96/ 0.95
	F1-score	0.98/ 0.96	0.74/ 0.72	0.84/ 0.82	0.98/ 0.95	1.00/ 0.95	0.89/ 0.89	0.88/ 0.86	0.96/ 0.95
Vowel	Accuracy	1.00/ 1.00	0.95/ 0.94	0.85/ 0.85	1.00/ 1.00	1.00/ 0.97	0.95/ 0.93	0.88/ 0.85	0.98/ 0.96
	F1-score	1.00/ 0.98	0.83/ 0.79	0.85/ 0.85	1.00/ 1.00	1.00/ 0.89	0.84/ 0.77	0.88/ 0.85	0.98/ 0.96

영향을 미친다. PBL은 클래스 경계면의 정보를 유지하기 때문에 Abalone, Diabetes(Pima), Glass, Liver disorder등과 같이 클래스 경계면에 다수의 데이터가 위치한 분류 문제의 경우, 의사 결정 나무 알고리즘에서 좋은 분류 성능을 보였으나, 높은 프로토타입 선택율을 갖는다는 단점이 있다. HRPS는 데이터의 분포와 분류 알고리즘의 특성을 고려하고, 초월 사각형의 최대 크기를 미리 설정하여 낮은 프로토타입 선택율과 높은 분류 성능을 기대한다.

5. 결 론

본 논문에서는 초월 사각형을 이용한 프로토타입 선택방법을 제안했다. 주어진 데이터에서 각 클래스를 대표할 수 있는 프로토타입을 선택하였고, 선택된 프로토타입이 새로운 훈련 데이터로 사용된다. 새로운 훈련 데이터의 수는 기존 훈련 데이터의 수보다 적어 저장 공간 및 학습 시간을 단축할 수 있다. 이러한 전략은 모든 분류 알고리즘에 적용되며 높은 일반화 성능을 얻을 수 있다.

제안하는 방법에서는 데이터 간 유사도 및 초월 사각형의 크기 계산을 차원 별 위치 차이로 이용하기 때문에 일반적인 거리 계산보다 비교적 낮은 시간 복잡도를 갖는다. 최소의 프로토타입을 선택하여 영역 내 모든 훈련 데이터를 포함하는 것은 집합 덮개 최적화 문제의 해가 된다. 초월 사각형은 서로 다른 클래스 데이터를 포함하지 않기 때문에 클래스별 독립적인 집합 덮개 최적화 문제로 변형할 수 있고, 병렬 처리를 통해 시간을 단축할 수 있다.

실험에서는 초월 사각형의 크기에 따라 학습 알고리즘의 분류 성능을 보였다. 초월 사각형의 크기가 큰 경우, 프로토타입 선택 비율은 낮아지나 클래스 경계면의 정보가 축소된다. 추후, 클래스 경계면의 정보를 유지함과 동시에 낮은 프로토타입 선택 비율을 갖는 연구가 필요하다.

References

[1] N. Bhatia, Vandana, "Survey of Nearest Neighbor Techniques," *International Journal of Computer Science and Information Security*, Vol.8, No.2, 2010.

[2] I. Triguero, J. Derrac, S. Garcia, and F. Herrea, "A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification," *IEEE Transactions on Systems, Man, and Cybernetics Part C(Application And Reviews)*, Vol.42, No.1, pp.86-100, 2012.

[3] R. M. Cruz, R. Sabourin, and G. D. Cavalcanti, "Prototype selection for dynamic classifier and ensemble selection," *Neural Computing and Applications*, Vol.29, pp.447-457, 2016.

[4] R. M. Curz, R. Sabourin, and G. D. Cavalcanti, "Analyzing different prototype selection techniques for dynamic classifier and ensemble selection," *International Joint Conference on Neural Networks*, pp.3959-3966, 2017.

[5] E. Pekalska, R. P. W. Duin, and P. Paclik, "Prototype selection for dissimilarity-based classifier," *Pattern Recognition*, 39, pp.189-208, 2006.

[6] J. A. Olvera-Lopez, J. A. Carrasco-Ochoa, J. F. Martinez Trinidad, and J. Kittler, "A review of instance selection methods," *Artif Intell Rev*, Vol.34, No.2, pp.133-143, 2010.

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

[8] 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 Machine Intelligence*, Vol.34, No.3, pp.417-435, 2012.

[9] S. Choi, S. Cha, and C. Tappert, "A Survey of Binary Similarity and Distance Measures," *J. Systemics, Cybernetics and Informatics*, Vol.8, No.1, pp.43-48, 2010.

[10] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, Vol.5, No.4, pp.2403-2424, 2011.

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

[12] R. Younsi and A. Bagnall, "A randomized sphere cover classifier," *International Conference on Intelligent Data Engineering and Automated Learning*, pp.234-241, 2010.

[13] S. Seyong and H. Doosung, "Prototype based Classification by Generating Multidimensional Spheres per Class Area," *Journal of The Korea Society of Computer and Information*, Vol.20, No.2, 2015.

[14] S. Arora, D. Karger, and M. Karpinski, "Polynomial time approximation schemes for dense instances of NP-hard problems," *Journal of Computer and System Sciences*, Vol.58, pp.193-210, 1999.

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

[16] 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, 2007.

[17] A. H. Cannon and L. J. Cowen, "Approximation algorithms for the class cover problem," *Annals of Mathematics and Artificial Intelligence*, Vol.40, No.3-4, pp.215-223, 2004.

[18] UCI Machine Learning Repository [Online]. Available: https://archive.ics.uci.edu/ml/.

백 병 현

https://orcid.org/0000-0002-1818-4181
e-mail : byunghyun8@naver.com
2016년 단국대학교 컴퓨터과학과(학사)
2019년 케이사인 보안기술연구소 연구원
2019년 ~ 현 재 단국대학교 소프트웨어학과 석사과정
관심분야 : Machine Learning, Computer Vision

어 성 율

https://orcid.org/0000-0002-0759-6148
e-mail : startyou@ksign.com
1997년 아주대학교 컴퓨터공학과(학사)
1999년 아주대학교 컴퓨터공학과(석사)
2018년 ~ 현 재 단국대학교 소프트웨어학과 박사과정
관심분야 : Machine Learning, Threat Intelligence

황 두 성

https://orcid.org/0000-0003-1840-9296
e-mail : dshwang@dankook.ac.kr
1986년 충남대학교 계산통계학과(학사)
1990년 단국대학교 전자계산학과(석사)
2003년 Wayne State Univ. 컴퓨터학과 (박사)
2003년 ~ 현 재 단국대학교 소프트웨어학과 교수
관심분야 : Machine Learning, Parallel processing, Computer Vision