

지역 전문가의 앙상블 학습

이병우^o 양지훈

서강대학교 컴퓨터공학과

elva1212@sogang.ac.kr, yangjh@sogang.ac.kr

Ensemble Learning of Region Experts

Byungwoo Lee^o Jihoon Yang

Department of Computer Engineering, Sogang University

1. 서론

앙상블 학습(Ensemble Learning)이란 여러 개의 단일 분류기(Classifier)를 학습하고 그것들의 예측을 결합함으로써 새로운 가설(Hypothesis)을 학습하는 방법이다. 다양한 분류기의 결과를 결합함으로써 단일 분류기보다 신뢰성이 높은 예측값을 얻는 것이 앙상블 학습의 목표이다. 대표적인 앙상블 학습 방법에는 배깅(Bagging)과 부스팅(Boosting)이 있다.

본 논문에서 제시하는 지역 전문가의 앙상블 학습은 결정 트리와 비슷하게 데이터를 분할하면서 학습을 한다. 그러나 결정 트리와는 달리 말단 노드(Leaf Node)뿐만 아니라 모든 노드에 분류기가 존재하며 데이터를 분할하기 위한 속성(Attribute)을 선택하는 방법도 결정 트리와 다르다. 또한 배깅이나 부스팅이 기본 학습기(Base Learner)를 생성할 때 모두 같은 학습 알고리즘을 사용하는 것과는 달리 기본 학습 알고리즘은 기본 학습기마다 다르다. 앞으로 이 앙상블 학습을 ELRE(Ensemble Learning of Region Experts)라 지칭한다.

2. 본론

2.1. 학습 데이터 분할 과정

각 속성에 대하여 학습 데이터를 이분할 하는 값을 구한다. 정확히 이등분이 되지 않을 경우 최대한 이등분이 되는 값을 구한다. 이를 위해 속성값이 연속적인 값(Continuous Value)인 데이터만을 고려한다. 식 (1)[1], (2)를 이용하여 각 속성의 Score를 구하고 그 값이 가장 큰 속성을 선택한다.

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (1)$$

$$Score(S, A) = Entropy(S) - \frac{||S_1| - |S_2||}{|S|} \sum_{v=1}^2 \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (2)$$

, S 는 학습 데이터, c 는 클래스 개수, p_i 는 S 중에서 클래스 i 에 속하는 데이터의 비율 A 는 데이터의 속성 S_1 과 S_2 는 A 의 분할 값에 의하여 분할된 데이터이다. 이러한 방법으로 속성을 선택하고 데이터를 계속해서 분할해 나가며 학습 데이터의 개수가 10개 미만이거나 Entropy가 0일 경우 분할을 종료한다.

2.2. 기본 학습기 학습 과정

ELRE는 분할된 데이터를 이용하여 다양한 기본 학습기를 생성한다. 분류기를 생성할 때 주어진 학습 알고리즘 중 그 지역에서 가장 좋은 성능을 보이는 학습 알고리즘을 이용한다. 이를 위해 주어진 학습 알고리즘 중 10겹 교차검증(10-fold Cross-Validation) 정확도가 가장 높은 알고리즘을 선택한다. 10겹 교차검증 정확도가 같을 경우 학습 정확도(Training Accuracy)가 가장 정확한 알고리즘을 선택한다. 두 정확도가 모두 동일한 경우 만약 부모 노드(Parent Node)가 있다면 뿌리 노드(Root Node)부터 부모 노드까지의 경로(Path)에 존재하는 분류기에서 가장 많이 선택된 학습 알고리즘을 선택한다. 만약 부모 노드가 없다면 즉 뿌리 노드라면 미리 정해 놓은 학습 알고리즘의 우선순위에 따라 학습 알고리즘을 선택한다. 이렇게 생성된 기본 분류기를 그 지역의 전문가(Expert)라 하자.

2.3. 분류 과정

새로운 패턴을 분류할 때에는 <그림 1>과 같이 그 패턴이 포함된 지역을 담당하는 전문가들만이 투표를 한다. 각 전문가마다 학습 정확도가 다르기 때문에 서로 다른 가중치를 주어 투표를 한다. 식 (3)은 전문가의 가중치를 구하는 식이다.

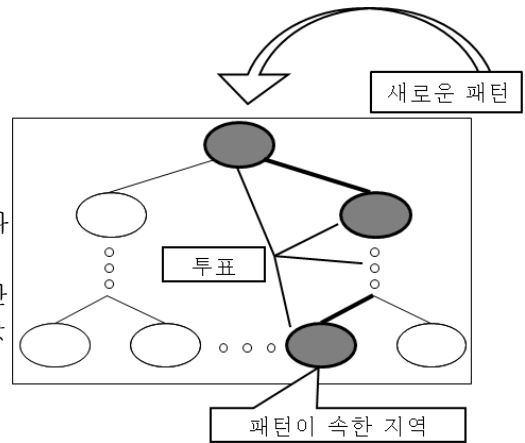
$$Classifier_Weight(N) = \frac{|D|}{|PD|} \cdot PW \quad (3)$$

, N 은 가중치를 구하고자 하는 전문가 D 는 N 에 할당된 학습 데이터 중 N 이 정확하게 예측한 데이터 PD 는 N 에 할당된 학습 데이터 중 부모 노드의 전문가가 정확하게 예측한 데이터 PW 는 N 의 부모 노드에 있는 전문가의 가중치이다

전문가로 새로운 패턴을 분류할 때 예측값을 구하는 것이 아니라 각 클래스별로 클래스 확률(Class Probability)을 구하고 이 클래스 확률에 그 전문가의 가중치를 곱하여 각 클래스별로 투표값을 구한다. 각 전문가에서 구한 투표값을 클래스별로 합하고 그 중 투표값이 가장 큰 클래스를 ELRE의 예측값으로 정한다 식 (4)는 클래스의 투표값을 구하는 식이다

$$Voting(C, P) = \sum_{v=1}^L Classifier_Weight(N_v) \cdot Class_Prob(C, P) \quad (4)$$

, C 는 클래스, P 는 분류하고자 하는 패턴 L 은 투표에 참여하는 전문가의 개수, N_v 는 투표에 참여하는 전문가 $Class_Prob(C, P)$ 는 패턴 P 를 클래스 C 라 예측할 클래스 확률이다



<그림 1> ELRE의 분류 과정

3. 실험 및 결론

각 노드의 전문가를 생성하기 위한 기본 학습 알고리즘으로 나이브 베이즈(Naive Bayes), SVM을 구현한 SMO, C4.5를 사용하였으며 우선순위는 SMO, 나이브 베이즈, C4.5 순으로 하였다. 다른 앙상블 학습과의 성능 비교를 위해 부스팅의 한 종류인 Adaboost, 배깅과 정확도 비교를 하였다 ELRE는 레벨 3까지 확장하였으며 공정한 비교를 위해 Adaboost와 배깅도 21개의 기본 분류기를 생성하였다 10개의 데이터 세트를 이용하여 실험하였으며 10겹 교차검증으로 정확도를 측정하였다 <표 1>은 정확도 비교를 보여주며 정확도가 가장 높은 것을 굵게 표시하였다

<표 1> ELRE와 각 방법들과의 정확도 비교

	SMO			Naive Bayes			C4.5			ELRE
	단일	Ada	Bagging	단일	Ada	Bagging	단일	Ada	Bagging	
Vehicle	74.55	74.43	74.90	46.28	46.28	46.52	75.02	76.92	74.32	77.27
Glass	53.94	58.14	55.22	47.45	47.92	50.76	69.39	77.42	76.56	74.18
Abalone	25.31	25.31	25.58	23.80	23.80	23.64	20.90	22.15	24.47	27.10
Hypothyroid	89.85	95.81	89.77	96.75	95.81	97.21	94.91	94.42	93.49	97.23
Ionosphere	88.03	87.78	88.04	82.90	91.76	82.62	88.05	93.18	91.20	92.33
Ecoli	84.22	86.30	83.63	86.29	86.23	86.27	70.30	82.43	84.50	86.62
SPECTF	79.42	76.40	79.40	67.86	68.54	70.41	74.54	80.15	79.78	81.30
Segment	92.90	93.20	92.90	79.65	79.65	79.96	96.80	98.27	97.45	97.45
Page-Blocks	92.91	92.91	93.42	90.15	90.15	90.02	97.06	97.19	97.37	97.17
Spambase	90.39	90.78	90.81	79.50	79.50	79.70	93.04	95.35	94.24	93.74

ELRE는 모든 학습 알고리즘의 단일 분류기보다 정확도가 높았으며 10개 중 5개의 데이터 세트에서는 기존의 Adaboost와 배깅보다도 정확도가 높았다 나머지 5개의 데이터 세트에 대해서는 SMO, 나이브 베이즈를 사용한 Adaboost와 배깅보다는 정확도가 높았으나 C4.5를 사용한 Adaboost 보다 정확도가 낮고 배깅과는 비슷한 성능을 보였다. 이는 C4.5를 사용한 Adaboost가 ELRE보다 정확도를 향상시킬 수도 있기 때문이다 실제 과거 연구들을 보면 부스팅과 배깅이 C4.5와 같이 불안정한 알고리즘(Unstable Algorithm)에서 큰 정확도 상승을 보인다고 하였다[2].

ELRE는 패턴 분류에 알맞은 학습 알고리즘을 선택해주기 때문에 분류기를 학습시킬 때 어떤 학습 알고리즘을 사용할지 고민할 필요가 없이 단지 데이터에 적합한 학습 알고리즘의 종류만을 선택해주면 된다)로 인해 하나의 학습 알고리즘만으로는 표현할 수 없는 가설 공간을 다양한 학습 알고리즘으로 표현할 수 있게 되어 정확도를 향상시킬 수 있다. 또한 기본 학습기가 특정 지역만을 학습하며 분류하고자 하는 패턴을 학습한 기본 학습기만을 이용하여 투표하므로 보다 정확한 예측값을 얻을 수 있다 그리고 데이터를 분할하면서 학습할 경우 말단 노드로 갈수록 과잉적합(Overfitting)의 문제점이 발생할 수 있으나 조상 노드를 투표에 참여시킴으로써 완화시킬 수 있다

4. 참고문헌

- [1] Quinlan, J., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
- [2] Breiman, L., Bagging Predictors, Machine Learning, 24(2), 123-140, 1996.