데이터가 지속적으로 쌓이는 상황에서는 모델의 구축에 사용한 학습 데이터와 추가되는 데이터의 분포가 다른 수가 있기 때문에 모델의 성능이 필요하다. 이러한 상황에서 쉽게 적용할 수 있는 해결책 중 하나는 앵상블 기법을 이용한 점진적인 학습 방법이다. 그러나 이러한 경우에 발생할 수 있는 문제점 중 하나는 지속적으로 추가되는 앵본클러가 중 어느 것을 선택하는 것이 가장 핵심적인 가치 판단하기에 매우 어렵다는 것이다. 본 논문에서는 이러한 문제의 해결을 위하여 베이지안 모형 결합법을 이용한 효율적인 앵본클러의 선택 방법을 제시하였다. 제안한 방법론은 추가되는 데이터를 이용하여 만들어진 다수의 앵본클러들을 이용하여 베이즈적해에 가장 가까운 모델을 구할 수 있도록 앵본클러들의 가중치를 결정하고 이에 기반하여 모델을 선정하게 된다.

1. 서론

우선홍성 기술의 발달으로 인하여 다양한 종류의 정보기기가 나타났다. 이는 기술과 결합하여 사용되며 이와 함께 기존의 사용자 단말에서 처리되던 많은 작업들이 온라인 상에서 이루어지고 있다. 특히 음성인식이나 밀기체인식 등의 기술이 확장되어 발생한 큰 변화 중 하나는 사용자 데이터의 지속적인 수집과 이로 인한 데이터의 급격한 증가이다. 이러한 상황에서 대용량 데이터에 대한 처리기술과 함께 지속적으로 증가하는 데이터에 대해 유연하게 대처할 수 있는 점진적인 학습법이나 모델의 공학 기술에 대한 관심이 높아지고 있다[1, 2].

이러한 문제를 해결하기 위한 접근법 중 하나는 앵상블 모델을 이용한 방법[3, 4, 5, 6]들이 연구되고 있으나 앵상블 모델의 접근법이 가지고 있는 어려움 중 하나가 데이터가 지속적으로 증가함에 따라서 계속해서 추가되는 앵본클러 중 어느 것을 선택하는 것이 가장 효율적인지를 판단하기에 어렵다는 것이다. 이러한 문제점을 해결하기 위해서 본 논문에서는 베이지안 모형 결합법을 이용한 효율적인 앵본클러의 선택 방법을 제시하였다. 베이지안 모형 결합법은 앵상들 내에 존재하는 각각의 앵본클러들이 전체 데이터를 얼마나 잘 설명하고 있는지를 판단하는 것이 아니라, 전체 데이터를 가장 잘 설명하는 앵본클러의 조합을 찾는 것이 목적이다. 이로써 뿐만 아니라 지속적으로 추가되는 데이터의 양을 수용하므로 데이터의 다양성이나 특성을 갖는다.

2. 베이지안 모형 결합법

베이지안 모형 결합법(Bayesian Model Averaging)은 개별 모형이 얼마나 중요한 예측력을 낼 것인지에 대한 확률을 계산하고 추정된 개별 모형의 예측력을 이용하여 각각의 가중치를 결정한 후에 이를 이용하여 모델을 생성하는 방법이다. 크기가 n인 데이터 집합 D가 있을 때, 각각의 데이터 d는 특징 벡터(feature vector) x와 범주(class) y로 구성되고 가설 공간이 가설 h의 집합인 H로 표현된다. 다중 데이터 인스턴스가 범주 y에 속할 확률은 다음과 같이 계산된다.

\[ p(y|x, D, H) = \sum_{h \in H} p(y|h) p(h|D) \]  \hspace{1cm} (1)
베이지스 정리에 의해 데이터 집합 $D$에 대한 $h$의 사후확률(posterior probability)는 이하와 같이 계산된다. 이 때, $p(h|D)$는 $h$의 사전확률(prior probability)을 의미하고 우도함수(likelihood)는 $p(d_i|h)$의 곱으로 표현된다.

$$p(h|D) = \frac{p(h \mid D)}{p(D)} \prod_{i=1}^{n} p(d_i|h)$$

(2)

베이지안 모형 결합법은 베이지안 모형 평균법의 표본화 방법을 수정한 것으로, 식 1을 아래와 같이 수정하여 정의할 수 있다[7]. 이 때, $e$는 가설들의 조합으로 만들어진 사전확률 $E$의 원소이다.

$$p(y|x, D, H, E) = \sum_{e \in E} p(y|x, e)p(e|D)$$

(3)

식 3에서 알 수 있듯이 베이지안 모형 평균법이 각각의 가설을 개별적으로 표본화 하는 것과는 달리 베이지안 모형 결합법은 모든 가능한 열망을 공간에서 표본화을 함으로써 베이지안 모형 평균법에서 발생할 수 있는 특정 모델로의 가중치 산정 현상을 극복할 수 있다. 그림 1에 베이지안 모형 결합법의 작동과정이 알고리즘으로 기술되어 있다.

### Bayesian Model Combination($T$)

For each hypothesis, $h$, in the ensemble:

weight[$h$] = 0, sum_weight = 0, $z$ = -infinity

Let $n$ be some number of weightings to sample.

for i from 0 to n - 1:

For each hypothesis, $h$, in the ensemble:

$v[h] = -log(random\_uniform(0,1))$

Normalize $v$ to sum to 1

$$\text{log\_likelihood}\_i = |T| * (x * log(x) + (1-x) * log(1-x)),$$

where $|T|$ is the number of training patterns in $T$, $x$ is the predictive accuracy of the entire ensemble.

If log\_likelihood\_i > $z$:

For each hypothesis, $h$, in the ensemble:

weight[$h$] = weight[$h$] * exp($z$ - log\_likelihood\_i[j])

$z$ = log\_likelihood\_i

$w$ = $exp(log\_likelihood\_i[j] - z)$

For each hypothesis, $h$, in the ensemble:

weight[$h$] = weight[$h$] * sum_weight

$$\text{sum\_weight} = sum\_weight + w + v[h]$$

Normalize the model weights to sum to 1.

### 그림 1. Bayesian Model Combination

<table>
<thead>
<tr>
<th>악분류기 개수</th>
<th>정확도</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.725</td>
<td>0.725</td>
</tr>
<tr>
<td>0.765</td>
<td>0.765</td>
</tr>
<tr>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>0.793</td>
<td>0.793</td>
</tr>
<tr>
<td>0.801</td>
<td>0.801</td>
</tr>
<tr>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>0.822</td>
<td>0.822</td>
</tr>
</tbody>
</table>

### 그림 2. Bayesian Model Combination을 이용한 성능 비교: 데이터 집합 A

3. 실험 및 결과

제안한 방법론의 성능을 평가하기 위해서 다수의 사용자로부터 수집된 13만 8천여개의 온라인 평가 데이터를 학습데이터 집합 A로, 그리고 데이터집합 A에서 특수문자 제외하고 올바른 글자를(italic writing) 11만여개가 추가된 데이터집합 B를 사용하였고, 9만여개의 UNIPEN[8] Train-R01/V07 데이터를 테스트데이터로 사용하였다. 데이터집합 A는 대손자의 모양이 동일한 일부 경우와 동일한 형태의 단어들이 존재하는 숫자 0,1을 제외한 숫자와 알파벳 대소문자, 그리고 �棻안 3종류의 특수문자로 이루어진 57개의 클래스로 구성되어 있고, 데이터집합 B는 특수문자와 대소문자의 모양이 동일한 일부 클래스가 제외된 총 50개의 클래스로 구성되어 있다.

본 논문에서는 데이터가 지속적으로 촉발되는 상황을 가장하기 위하여 데이터집합 A는 다섯개, 데이터집합 B는 여섯개로 나누어서 각각에 대하여 악분류기를 학습시킨 후에 이를 이용하여 양상별 모델을 구축하고 그 결과를 확인해 보았다. 양상별 모델을 위한 악분류기는 인공신경망을 사용하였고, 양상별 모델은 배깅(Bagging)을 사용하였다[9].

그림 2에 데이터집합 A를 이용한 실험결과가 보여지고 있다. 전체 데이터 집합에서 복원추출을 사용하는 일반적인 배깅은 커다란 데이터가 순차적으로 들어오는 상황을 가장하기 위하여 전체 데이터 집합에서 복원추출을 이용하지 않고 다섯개로 나누어서 각각에 대한 악분류기를 만든 후에 이를 이용하여 양상별 모델을 구축하였다. 그림에서 확인할 수 있듯이 다수결(majority voting)에 기반한 일반적인 배경의 경우보다 전체적으로 좋은 성능을 보이고 있음을 확인할 수 있다.
그림 3. Bayesian Model Combination을 이용한 성능 비교: 데이터 집합 B

4. 결론

본 논문에서는 다수개의 악분류기를 존재할 경우에 베이지안 모형 결정법을 이용하여 효율적으로 악분류기를 선택할 수 있는 방법론을 제시하였다. 제안한 방법론은 추가되는 데이터를 이용하여 만들어진 다수개의 악분류기를 이용하여 베이지안 최적해에서 가장 가까운 모델을 구할 수 있도록 악분류기의 가중치를 결정하고 이에 기반하여 모델을 선정하게 된다. 제안된 모델은 데이터가 증가함에 따라서 계속적으로 증가하는 악분류기 중 어느 것을 이용하여 양상을 모델을 구축할 것인가에 대한 효과적인 방향을 제시하였을 뿐만 아니라 불균형 데이터인 필기체 데이터의 성능을 극복할 수 있는 부스팅(bootstrapping) 기법과의 결합에서 발생하는 파라메터 학습 문제에도 도움이 되리라 생각한다.

감사의 글

이 논문은 삼성전자 지원으로 수행된 연구이며, 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(NRF-2010-0017734-Videome,)을 일부 받았음.

참고문헌