Journal of the Korean Data & Information Science Society
2014, **25**(6), 1371–1383

http://dx.doi.org/10.7465/jkdi.2014.25.6.1371
한국데이터정보과학회지

분류 앙상블 모형에서 Lasso-bagging과 WAVE-bagging 가지치기 방법의 성능비교[†]

곽승우[1] · 김현중[2]

[12]연세대학교 응용통계학과

접수 2014년 8월 15일, 수정 2014년 9월 23일, 게재확정 2014년 10월 17일

요 약

분류 앙상블 모형이란 여러 분류기들의 예측 결과를 통합하여 더욱 정교한 예측성능을 가진 분류기를 만들기 위한 융합방법론이라 할 수 있다. 분류 앙상블을 구성하는 분류기들이 높은 예측 정확도를 가지고 있으면서 서로 상이한 모형으로 이루어져 있을 때 분류 앙상블 모형의 정확도가 높다고 알려져 있다. 하지만, 실제 분류 앙상블 모형에는 예측 정확도가 그다지 높지 않으며 서로 유사한 분류기도 포함되어 있기 마련이다. 따라서 분류 앙상블 모형을 구성하고 있는 여러 분류기들 중에서 서로 상이하면서도 정확도가 높은 것만을 선택하여 앙상블 모형을 구성해 보는 가지치기 방법을 생각할 수 있다. 본 연구에서는 Lasso 회귀분석 방법을 이용하여 분류기 중에 일부를 선택하여 모형을 만드는 방법과 가중 투표 앙상블 방법론의 하나인 WAVE-bagging을 이용하여 분류기 중 일부를 선택하는 앙상블 가지치기 방법을 비교하였다. 26개 자료에 대해 실험을 한 결과 WAVE-bagging 방법을 이용한 분류 앙상블 가지치기 방법이 Lasso-bagging을 이용한 방법보다 더 우수함을 보였다.

주요용어: 가지치기, 데이터마이닝, 배깅, 분류, 앙상블.

1. 서론

모든 분류 모형들은 크게 불안정적인 분류기 (unstable classifiers)와 안정적인 분류기 (stable classifiers)로 구분해 볼 수 있다. 전자의 특징은 분류 결과의 분산 (variance)이 크지만 편의 (bias)가 작다는 점이다. 분산이 크기 때문에 훈련 자료의 작은 변화에도 분류 결과는 많은 변화가 있을 수 있다. 대신 편의가 적기 때문에 평균적으로 더 정확한 예측 결과를 가지게 된다. 반면에 후자의 경우에는 분류 결과의 분산은 작으나 상대적으로 편의가 크다. 따라서 훈련 자료의 작은 변화에는 큰 영향을 받지 않지만, 적합모형이 적절치 않을 때 예측 결과가 편의를 많이 포함할 가능성이 있다는 단점이 있다 (Breiman, 1996).

여러개의 분류기로 부터 얻은 분류결과를 통합하여 더 좋은 분류 예측 모형을 만들려는 시도가 분류 앙상블 방법론이라 할 수 있다. 분류 앙상블 모형의 장점은 하나의 분류기를 사용한 모형보다 예측 정확도가 향상된다는 점이다 (Dietterich, 2000). 앙상블을 구성하는 분류기중에서 하나의 분류기가 분류해 내지 못한 것을 다른 분류기들이 분류해 낼 수 있기 때문에 예측능력을 더욱 향상 시킬 수 있는 원리이다 (Kuncheva, 2005).

[†] 이 논문은 2012년도 연세대학교 학술연구비의 지원과 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2012R1A1A2042177).
[1] (120-749) 서울특별시 서대문구 연세로 50, 연세대학교 상경대학 응용통계학과, 석사과정.
[2] 교신저자: (120-749) 서울특별시 서대문구 연세로 50, 연세대학교 상경대학 응용통계학과, 교수.
E-mail: hkim@yonsei.ac.kr

불안정적인 분류기와 안정적인 분류기 중에서는 불안정적인 분류기가 분류 앙상블 모형에 더 적합하다고 볼 수 있다. 불안정적인 분류기의 결과들을 앙상블 모형으로 통합하면, 분류결과의 분산을 낮출수 있기 때문이다. 이는 마치 평균들의 분산은 원 자료의 분산보다 $1/n$ 만큼 작아진다는 원리와 같은 것이다. 불안정적인 분류기들은 대체로 편의는 작기 때문에, 그것들을 이용한 분류 앙상블 모형은 낮은 편의와 함께 개별 분류기 보다는 더 작은 분산을 달성할 수 있게 된다. 이는 궁극적으로 분류 예측 정확도의 향상으로 귀결된다. 이 점이 분류 앙상블 모형의 가장 큰 장점이라고 할 수 있다.

분류 앙상블 모형을 사용할 때의 단점은 예측된 결과가 정확하더라도 결과를 해석하기가 어렵다는 것이다. 하나의 함수로 대표되는 모형이 아니라 여러개의 개별 분류모형의 결과를 투표 (voting)과정을 통해서 통합 (aggregation)하여 하나의 분류 앙상블 모형을 만들기 때문에 분류의 규칙을 해석하는 것이 쉽지 않다.

분류 앙상블 모형이 성공적이기 위해서는 다음과 같은 두 가지 조건을 만족해야 한다고 알려져 있다. 첫째, 앙상블을 이루는 분류기들의 적합된 함수들이 서로 상이하여 다양성이 높아야 한다. 한 분류기의 예측력이 다소 약하여 실패하더라도, 서로 상이하고 다양한 나머지 분류기들로 인해 올바른 예측을 할 수도 있기 때문이다 (Kuncheva, 2005). 둘째, 앙상블을 이루는 분류기들의 예측력이 좋아서 편의가 적어야 한다. 예측도가 높은 분류기들로 앙상블 모형을 구성한다면 결과적으로 더 예측력이 강한 모형을 만들기 용이하기 때문이다. 앙상블 모형을 대표하는 모형으로 배깅 (bagging; Breiman, 1996), 부스팅 (boosting; Freund와 Schapire, 1997; Kim 등, 2012), 랜덤포레스트 (random forest; Breiman, 2001)와 같은 방법론들이 있다. 이러한 방법들은 위의 첫째 조건인 앙상블을 이루는 분류기들을 서로 상이하고 다양하게 만드는 부분에 중점을 둔 방법이라 할 수 있다.

본 논문은 분류 앙상블 모형이 성공적이기 위한 두번째 조건인 예측력이 좋은 분류기들로 앙상블을 이루는 방안에 대한 연구이다.

본 논문의 2절에서는 분류앙상블의 대표적 방법인 배깅 방법을 소개하고, WAVE 방법을 이용한 WAVE-bagging 방법을 함께 소개한다. 3절에서는 Lasso 회귀모형과 그를 이용한 분류앙상블 가지치기 방법인 Lasso-bagging 방법을 소개한다. 4절에서는 본 논문에서 사용된 자료와 실험 방법에 대해 소개를 하고, 5절에서는 4절에서 설명한 자료를 각각의 방법론에 적용하여 예측한 결과를 논의한다. 마지막으로 6절에서는 결론과 향후 연구방향에 대한 논의를 할 것이다.

2. 분류 앙상블 방법론

2.1. 배깅 방법

배깅 (bagging)은 Bootstrapping Aggregating의 약자로 훈련 자료 (L)를 대표적 재추출법인 붓스트랩을 통해 k개의 훈련 자료 ($L_b, b = 1, \ldots, k$) 로 만들어 낸 후, L_b들을 바탕으로 여러 개의 분류기 ($C_b, b = 1, \ldots, k$)를 만들어 내는 분류앙상블의 대표적 방법이다. 그리고 예측하고자 하는 관찰치의 변수 (x)값에 따라 분류기($C_b(x), b = 1, \ldots, k$)마다 계급에 대한 예측을 수행하고 그 예측 결과를 단순 투표 알고리즘으로 통합하여 하나의 예측 결과값을 반환하는 모형 ($C(x)$)이다. 여러개의 적합함수를 만들어 하나의 예측 모형을 만드는 과정이 있으므로, 단 하나의 분류기를 사용하는 모형보다 오차가 작다는 장점을 가지고 있다 (Breiman, 1996).

배깅 앙상블 방법에서 사용하는 단순 투표 알고리즘이란 종속변수 y가 연속형 값이라면 붓스트랩을 통해서 얻은 자료에 적합된 모형의 예측 결과들의 평균값을 의미한다. 만약 종속변수 y가 계급이라면, 분류기마다 예측된 계급들을 동일한 가중치로 투표를 하여 가장 많은 표를 얻은 계급이 배깅 모형이 예측한 최종계급이 된다.

붓스트랩을 이용하면 같은 훈련자료로 부터 서로 다른 자료들이 생성되므로 다양한 분류기를 얻기에 용이하다. 하지만, 이 분류기들이 반드시 예측정확도가 높다는 것을 의미하지는 않는다. 더욱이 붓스트랩 자료를 사용하여 만든 분류기들이 다양하지 않고 서로 비슷한 형태를 가질 수도 있다. 이러한 경우에는 각각의 분류기에 동일한 가중치를 주게 되면 유사하게 중복되는 형태의 분류기가 더 높은 가중치를 갖게 되는 효과가 발생한다. 따라서, 더 나은 예측력을 가진 분류 앙상블 모형을 만들기 위해서는 유사한 형태의 분류기는 중복되지 않도록 제외시키고, 예측력이 높은 분류기들을 선택해서 분류 앙상블이 구성되도록 하는 방법을 생각할 수 있다. 특히 분류기의 모형이 의사결정나무 모형일 경우에는 붓스트랩 자료를 사용하여 만든 분류기 전부를 사용하는 것보다 일부를 사용하는 것이 더 나을 수 있다고 알려져 있다 (Zhou와 Tang, 2003). 따라서 배깅의 장점을 이용하면서 여러개의 분류기중 일부를 선택하여 앙상블을 만드는 앙상블 가지치기 방법 (ensemble pruning)에 대한 연구는 의미가 있다.

2.2. WAVE-bagging

훈련자료내 관찰치들은 계급 예측이 용이한 관찰치와 계급 예측이 어려운 관찰치로 구분해 볼 수 있다. 계급 예측이 용이한 자료들은 앙상블을 구성하는 대부분의 분류기에서 비슷한 예측결과를 보일 것이며, 정확하게 예측되고 있을 가능성이 높다. 반면 계급 예측이 난해한 관찰치들에 대해서는 앙상블내 분류기의 예측 결과들이 일관적이지 못할 것이기 때문에 예측정확도도 낮을 것이다. WAVE (weight-adjusted voting ensemble algorithm; Kim 등, 2011)의 의도는 훈련 자료를 기반으로 만든 여러개의 분류기들 중에 더 예측 정확도가 높은 분류기에 가중치를 높게 부여하여 앙상블 모형으로 만드는 가중투표 앙상블 방법론이다. 구체적으로 설명하면, 계급예측이 어려운 관찰치에 대해서 예측정확도가 높은 분류기에 가중치를 높게 부여하고, 평이한 관찰치에 예측 정확도가 높은 분류기에는 가중치를 증가시키지 않은 선택적 앙상블 방법론이라고 할 수 있다 (Kim 등, 2011). 특히 배깅 앙상블 방법에 WAVE 가중 투표방법을 적용하는 경우를 WAVE-bagging 방법이라 칭한다.

Rokach (2009)와 Kuncheva (2004)에 따르면 분류 앙상블 방법은 결합 단계 (combination level), 분류기 단계 (classifier level), 변수선택 단계 (feature level), 그리고 자료 단계 (data level)의 4가지 형태로 나누어서 구분할 수 있다. 결합단계의 앙상블 방법은 앙상블을 구성하는 분류기들의 결합 (combining)방법을 다양하게 하는 형태이며, 분류기 단계의 앙상블 방법은 앙상블을 구성하는 기본 분류기를 다양하게 하는 형태의 앙상블 방법을 의미한다. 변수선택 단계는 분류기들을 적합시키기 위해서 다양한 변수 부분집합 조합을 사용하는 것을 의미한다. 그리고 마지막 자료 단계는 훈련자료로 부터 재추출한 다양한 자료를 생성하여 분류기를 적합하는 분류 앙상블 방법이다. WAVE 방법론은 위와 같은 분류 앙상블 모형을 만드는 방안 중에서 결합의 과정에서 가중투표 방법을 사용하여 더 나은 예측력을 가진 분류 앙상블 모형을 얻으려는 것으로 설명할 수 있다.

Algorithm 1을 살펴보면 붓스트랩을 통하여 기존의 훈련자료 (L)를 여러 개의 훈련자료 ($L_b, b = 1, \ldots, k$)로 만들고 그것을 바탕으로 분류기 ($C_b, b = 1, \ldots, k$)를 만든다. 그리고 훈련자료 L에 대해서 각 분류기별 계급값의 예측이 옳고 그름을 각기 1 과 0으로 표현하여 행렬 $\boldsymbol{X}$를 만든다. 이 행렬 $\boldsymbol{X}$를 이용하여 훈련자료 L의 예측 난이도에 대한 가중치를 반영하여 난이도가 높은 관찰값에 분류예측을 잘 수행하는 분류기에 대한 가중치 벡터 $\boldsymbol{P}^*$를 생성한다.

예측 난이도가 높은 관찰값을 정확히 예측해 내는 분류기에 대해서 가중치를 더 높게 설정하면 기존의 분류 앙상블 모형에 비해서 예측 정확도를 높일 수 있다. 배깅을 기준으로 설명하면, 계급예측이 용이한 관찰값에 대한 최종 계급 예측값은 분류기마다 거의 같은 계급으로 예측될 것이기 때문에 가중치의 존재 여부에 의존하지 않는다. 하지만, 계급예측이 난해한 관찰값은 분류기마다 서로 다른 계급으로 예측될 가능성이 높으므로, 더 높은 가중치를 갖는 분류기의 결과에 영향을 많이 받게 된다. 이는 곧 기존의 분

Algorithm 1 WAVE-bagging algorithm (Kim 등, 2011)

Input:

- L : training data set composed of n instances
- L_y : class member of L
- k : number of classifiers in an ensemble

Output:

- $C^*(\cdot)$: Weighted combination of outputs of classifiers

Step (1) Generate k bootstrap samples L_b $(b = 1, \ldots, k)$ from L with n instances.

Step (2) Train classifiers C_b using L_b.

Step (3) Acquire an $n \times k$ performance matrix $\boldsymbol{X} = [X_1, \ldots, X_k]$, consisting of 0's (wrong) and 1's (correct).

$X_b = I\{C_b(L_b) = L_y\}$, $n \times 1$ vector

Step (4) $\boldsymbol{T} = \boldsymbol{X}'(\boldsymbol{J}_{nk} - \boldsymbol{X})(\boldsymbol{J}_{kk} - \boldsymbol{I}_k)$

$\boldsymbol{J}_{nk} = n \times k$ matrix of 1
$\boldsymbol{I}_k = k \times k$ identity matrix.

Step (5) Calculate $\boldsymbol{P}^* = \frac{\left(\sum_{i=1}^{r} \boldsymbol{u}_i \boldsymbol{u}_i'\right) \mathbf{1}_k}{\mathbf{1}_k' \left(\sum_{i=1}^{r} \boldsymbol{u}_i \boldsymbol{u}_i'\right) \mathbf{1}_k} = [\boldsymbol{P}_1^*, \ldots, \boldsymbol{P}_k^*]'$.

λ_i = eigenvalues of $\boldsymbol{T}$, $i = 1, \ldots, k$
$\boldsymbol{u}_i$ = eigenvector corresponding to λ_i
r = number of dominating eigenvalues such that
$\lambda_1 = \lambda_2 = \ldots = \lambda_r > \lambda_{r+1}$, $1 \leq r \leq k$

Step (6) Aggregate the k classifiers using weighted combination of classifiers consisting of $\boldsymbol{P}^*$ and $\boldsymbol{X}$.

$$C^*(x) = argmax_y \sum_{b=1}^{k} \boldsymbol{P}_b^* \times I(C_b(x) = L_y)$$

류 앙상블 모형에서 예측 난이도가 높은 관찰값에 대해서 정확성을 보정하였으므로 전체적인 예측 정확도가 더 높아지는 효과로 귀결된다. 결론적으로, WAVE-bagging 모형은 기존의 분류 앙상블 모형의 결합 과정에 가중치를 부여하는 방법을 적용하여 더 높은 예측 정확도를 가진 분류 앙상블 모형을 만드는 방법론이다.

2.3. WAVE-bagging을 이용한 분류 앙상블 가지치기

WAVE-bagging 방법은 분류기의 중요도에 따라 가중치를 계산해 준다는 사실을 이용하여 분류 앙상블의 가지치기에 활용해 볼 수 있다. 예를 들어 가중치의 문턱기준 (threshold)를 설정하고, 문턱기준을 넘지 못하는 분류기에 대해서는 분류 앙상블에서 제외하는 방식으로 가지치기를 하는 방안이다.

문턱기준의 값을 변동시킴에 따라, 분류 앙상블에 포함되는 분류기의 갯수가 변동될 수도 있다. 만약 계급 예측이 어려운 데이터라면 문턱기준을 낮추어 더 많은 분류기가 앙상블을 구성하도록 할 수도 있으며, 그 반대로 계급예측이 용이한 데이터라면 문턱기준을 높이어 소수의 분류기로만 앙상블을 구성하도록 할 수도 있을 것이다.

본 논문에서는 WAVE-bagging을 이용한 분류 앙상블 가지치기 방법을 26개 실제 데이터에 적용하였고, 여러가지 문턱기준에 대하여 분류 예측 정확도의 추이를 살펴보았다.

3. Lasso 회귀분석과 분류앙상블 가지치기에의 적용

3.1. Lasso 회귀분석

만약 $(\mathbf{x}_i, y_i), i = 1, \ldots, N$와 같은 자료가 있다할 때, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$은 개수가 p 개인 설명변수이고, y_i는 반응변수라 하자. 일반적인 회귀분석에서와 마찬가지로 각각의 관찰치는 독립이거나 반응변수들이 주어진 설명변수에 대해서 조건부 독립이라고 하자. 여기서 x_{ij}는 $\sum_i x_{ij}/N = 0, \sum_i x_{ij}^2/N = 1$을 만족한다. Lasso (least absolute shrinkage and selection operator) 회귀분석 (Tibshirani, 1996)은 회귀모형 계수의 절대값의 합이 어떤 양의 상수보다 작아야 한다는 제약 조건하에서 식 (3.1)과 같이 회귀모형의 계수를 추정하는 것이다. $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)'$라 하면 Lasso의 추정치 $\left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right)$은 식 (3.1)과 같이 정의할 수 있다. 식 (3.1)에서 $t \geq 0$는 제약조건으로서 계수를 0에 가깝게 만들거나 0으로 만드는 역할을 하게 된다.

$$\left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right) = \text{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\}, \text{ subject to} \sum_j |\beta_j| \leq t \tag{3.1}$$

Lasso 회귀분석의 주요 특징은, 제약 조건에 따라서 계수의 크기가 0을 향해서 줄어들게 하고, 일부 변수의 계수를 0으로 만드는 것이다. 이러한 특징은 계수의 제곱의 합을 어떤 숫자의 이하로 만드는 제약조건을 가진 능선회귀분석 (ridge regression)과는 차별되는 것이다.

Lasso 회귀모형은 선형회귀모형에 비해서 편의 (bias)는 조금 높아지지만, 분산 (variance)을 줄여서 전체적인 오차 제곱합이 줄어들게 된다. 더불어 모형의 독립변수들 사이에 상관관계가 높을수록 계수가 0이 되는 경향이 나타나게 되어 모형에 유효한 변수들이 자동적으로 선택된다. 유사한 특징을 가진 모형으로 부분집합 선택 (subset selection) 회귀모형과 능선회귀모형이 있으나, 해석이 쉬운 부분집합 선택 회귀모형은 자료의 변화에 민감하다는 점에서 그리고 상대적으로 안정적인 능선회귀모형은 변수선택이 용이하지 않다는 점에서 단점이 존재한다. 두 모형의 장점을 취할 수 있는 모형이 Lasso 회귀모형이라 할 수 있다 (Tibshirani, 1996).

3.2. Lasso-bagging

Lasso 회귀모형은 원래 종속변수가 연속형인 경우에 사용하는 방법이다. 하지만 Chen과 Jin (2010)은 Lasso 회귀모형을 분류 앙상블에 적용하고자 하였다. 계급의 갯수가 2개인 경우에 한하여, 분류 앙상블을 구성하는 분류기의 계급 예측 결과를 독립변수로 하고 원래의 계급을 종속변수로 취급한다면, 이는 회귀분석이 가능한 상황이 된다. 다만, 계급의 갯수가 두 개로 한정되었으므로, 독립변수와 종속변수는 모두 0 과 1로 구성되었을 것이다.

분류 앙상블을 구성하는 분류기 중에서 실제 계급을 잘 예측하는 분류기는 Lasso 회귀분석의 결과 0이 아닌 계수값을 가질 것이지만, 실제 계급에 대한 예측 정확도가 낮은 분류기는 Lasso 회귀분석에서는 0의 계수값을 보일 것이다. 0이 아닌 계수값을 가지는 분류기로만 분류 앙상블에서 구성하면, 앙상블에 대한 가지치기의 효과를 가져올 수 있다.

또한 Lasso 회귀모형은 독립변수들 사이에 상관관계가 높은 경우에, 대표되는 하나의 독립변수를 제외한 나머지에 대해서 회귀계수가 0이 되는 성질을 가지고 있다. 이것을 배깅에 적용할 경우에 계급 예

Algorithm 2 Lasso-bagging (Chen과 Jin, 2010)

Input:

- L : training data set composed of n stances
- L_y : class member of L
- k : number of classifiers in an ensemble

Input:

- $C(\cdot)$: Combination of outputs of classifiers

Step (1) Generate k bootstrap samples L_b $(b = 1, \ldots, k)$ from L with n instances.

Step (2) Train classifiers C_b using L_b.

Step (3) Set $\{C_1(L_1), \ldots, C_k(L_k)\} = \left\{\widehat{Y_1}, \ldots, \widehat{Y_k}\right\}$ as input variable and $L_y = Y$ as response variable.

Step (4) Filter out classifiers $\left\{\widehat{Y_1}, \ldots, \widehat{Y_k}\right\}$ using Lasso regression.

Step (5) Aggregate selected classifiers $C_j, j = 1, \ldots, s$ and select the class having the plurality in them as the predicted class of x.

$$C(x) = argmax_y \sum_{j=1}^{s} I(C_j(x) = L_y)$$

측값 사이의 상관관계가 높은 분류기들 중에서 하나의 분류기만을 선택하게 될 것이다. 따라서 선택된 분류기들은 서로 상이한 분류기들로 이루어진다.

이상의 아이디어를 정리해보면 Lasso-bagging 방법은 붓스트랩을 통해서 만든 분류기 $(C_b, b = 1, \ldots, k)$의 계급 예측값들을 독립변수로, 본래 자료의 반응값 (L_y)을 종속변수로 하여 Lasso 회귀분석을 수행한다. Lasso 회귀분석의 특성에 따라 일부 분류기에 대한 계수는 0이 된다. 회귀계수가 0인 분류기는 제외하고, 0이 아닌 계수를 가지는 총 s개의 분류기가 선택되었다면 선택된 분류기 $(C_j, j = 1, \ldots, s)$를 단순 투표 방법론에 의해서 하나의 앙상블 모형으로 구성한다. 이 방법을 알고리즘의 형태로 정리해 보면 Algorithm 2와 같다.

4. 실험방법

비교를 위해서 사용한 분류 앙상블 모형은 Lasso-bagging과 WAVE-bagging, 그리고 가지치기가 없는 배깅의 세 가지 모형이다. 분류 앙상블 모형들의 공정한 비교를 위해서 가지치기의 결과로 선택된 분류기들은 동일한 가중치를 부여하여 계급예측값을 투표하는 방식을 사용하였다. 4.1절에서는 본 논문에서 사용한 데이터에 대해서 설명을 하고, 공정한 비교를 위해서 사용된 실험방법에 대해서 4.2절과 4.3절에서 자세히 설명할 것이다.

4.1. 자료 설명

Table 4.1은 실험에 사용된 실제 데이터에 대한 간단한 정보이다. 사용된 자료들은 22개의 실제 자료와 4개의 합성 자료 (Cir, Rng, Trn, Twn)로 이루어졌고 두 개의 계급으로 분류되는 반응변수 (binary response variable)를 가지고 있다. 대부분의 자료는 University of California at Irvine에서 제공하는

Table 4.1 Data description

Dataset	Description	# instances	# variables	Source
Aba	Abalone	4,177	8	UCI
Ail	Aileron	13,750	12	Loh (2009)
Aus	Australian credit approval	690	14	UCI
Bcw	Breast cancer Wisconsin	699	10	UCI
Bld	Liver Disorder	345	6	UCI
Bod	Body dimension	507	24	Heinz 등 (2003)
Cir	Circle in a square	10,000	10	mlbench
Cre	Credit approval	690	15	UCI
Cyl	Cylinder bands	540	35	UCI
Dia	Diabetes	532	7	Loh (2009)
Ech	Echocardiogram	132	12	UCI
Ger	German credit	1,000	20	UCI
Hea	Statlog (Heart)	270	13	UCI
Hep	Hepatitis	155	20	Loh (2009)
Int	Chessboard	1,000	10	Kim 등 (2011)
Ion	Ionosphere	351	34	UCI
Mam	Mammographic mass	961	6	UCI
Pid	Pima indians diabetes	768	8	UCI
Pks	Parkinsons	197	23	UCI
Rng	Ringnorm	1,000	10	mlbench
Snr	Sonar	208	61	mlbench
Spe	SPECTF heart	267	44	UCI
Tel	MAGIC Gamma Telescope	19,020	10	UCI
Trn	Threenorm	1,000	10	mlbench
Twn	Twonorm	1,000	10	mlbench
Vot	Congressional voting records	435	16	UCI

Machine Learning Repository (Frank와 Asuncion, 2010)에서 구했고, 일부자료는 R 라이브러리 중의 하나인 'mlbench' 에서 구했다. 이렇게 만든 자료의 반응변수 (y)는 0, 1의 두 개의 값으로 변환시켰다.

4.2. Lasso-bagging의 실험방법

공정한 예측 정확도를 얻기 위해서 26개 자료에 10겹 교차검증 (10-fold cross validation)을 이용하여 예측 정확도를 얻는다. 전체의 자료중 임의로 9/10의 자료를 훈련자료로 선택하고, 200개의 자료 $(L_1, \ldots, L_{200})$를 붓스트랩을 통해서 만든 후, 이것을 바탕으로 CART (Breiman 등, 1996) 방법을 이용하여 분류기 $C_1, \ldots, C_{200}$을 생성한다. 예측결과 $(C_b(x))$를 독립변수로 사용하고 원래 자료의 계급값을 종속변수로 사용하여 Lasso 회귀분석을 수행한다.

Lasso 회귀분석에서 계수값이 0으로 나온 분류기를 제외한 나머지 분류기로 앙상블을 구성한다. Lasso 회귀분석의 계수값은 앙상블 가지치기 기준으로만 활용되며 일단 선택된 분류기들은 계수값과 관계없이 같은 가중치를 갖는 투표방식을 따른다. 1/10의 검증자료에 선택된 분류기의 계급예측값을 구하고, 동일 가중치에 의한 투표방식을 적용하여 계급 예측값을 구한 후 실제 계급과 비교하여 예측 정확도를 측정한다. 이 과정을 10회 반복하여 10겹 교차검증을 완성한다. 이 과정을 100번에 걸쳐서 반복하여 평균적인 분류 예측 정확도를 계산한다.

4.3. WAVE-bagging의 실험방법

Wave-bagging은 Lasso-bagging의 실험방법과 동일한 방식으로 수행한다. Lasso-bagging과 차이점은 CART 방법을 이용하여 분류기 $C_1, \ldots, C_{200}$을 만들고 훈련자료의 y값인 L_y와의 비교를 통해서 $\boldsymbol{X}$를 만든 후, 이 예측치를 이용하여 가중치 벡터 ($\boldsymbol{P}^*$)를 만드는 부분이다. 또한, 가중치의 문턱기준에 의하여, 문턱기준을 넘은 분류기를 선택하고 이들로 구성된 분류 앙상블을 생성한다. 1/10의 검증자료를 대상으로 선택된 분류기의 계급예측값을 구하고, 동일 가중치에 의한 투표방식을 적용하여 계급 예측값을 구한 후 실제 계급과 비교하여 예측 정확도를 측정한다. 이 과정을 10회 반복하여 10겹 교차검증을 완성한다. 이 과정을 100 번에 걸쳐서 반복하여 평균적인 분류 예측 정확도를 계산한다.

한가지 유의할 점은 결합의 단계에서는 Lasso-bagging 모형이 가중치를 적용하지 않으므로, 공정한 비교를 위해서 WAVE-bagging에서도 마찬가지로 결합의 과정에서는 가중치를 사용하지 않기로 한 점이다. 마지막으로, 가지치기가 없는 배깅에 대한 실험은 Lasso-bagging과 WAVE-bagging의 실험에서 가지치기를 하지 않은 단계에서의 결과를 산출하면 된다.

5. 실험 결과

이 절에서는 26개 자료를 이용한 실험에서 얻은 결과를 바탕으로 상대적 개선도, 대응표본 t-통계량, 정확도 그리고 dominance rank를 이용하여 각 분류 앙상블 방법의 성능을 비교하고자 한다.

5.1. 상대적 개선도와 t-통계량

WAVE-bagging과 Lasso-bagging 모형을 통해서 앙상블 가지치기를 수행하고 얻은 예측 정확도를 상대적 개선도 (relative improvement)와 대응표본 t-검정 통계량 값을 이용하여 비교할 것이다. 상대적 개선도는 모형 B의 오분류율 (e_B)과 모형 A의 분류율 (e_A)을 비교하는 것으로 다음과 같이 정의되었다.

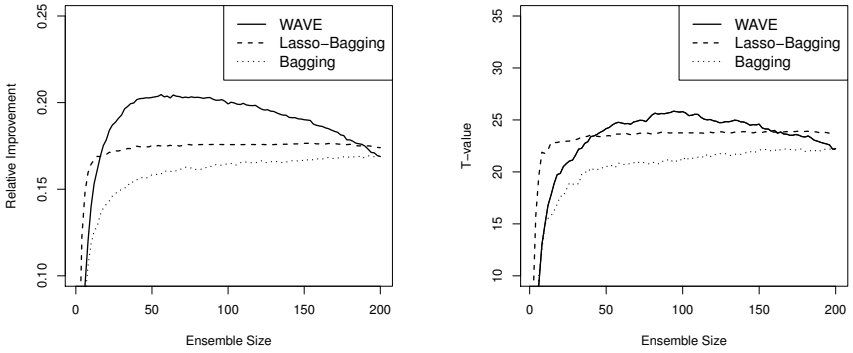
$$B\text{에 대한 } A\text{의 상대적 개선도} \equiv \frac{e_B - e_A}{e_B}. \tag{5.1}$$

식 (5.1)은 A의 모형이 B에 비해서 상대적으로 우수한 정도를 측정할 수 있는 측도이다. 그리고 t-통계량은 A의 오분류율이 B의 오분류율보다 낮다는 가설검정 ($H_0 : e_A = e_B$ vs $H_1 : e_A < e_B$)에 대한 대응표본 t 검정의 통계량을 의미한다.

Lasso-bagging과 WAVE-bagging, 배깅 등의 분류 앙상블 방법과 CART 모형을 비교한 상대적 개선도의 값을 그래프로 그리면 Figure 5.1(a)와 같이 나타낼 수 있다. 여기에서 배깅 방법의 경우에 앙상블을 이루는 분류기의 개수가 많아지면 모형의 상대적 개선도가 증가하는 것을 알 수 있다. Lasso-bagging의 경우에 계수값이 높은 소수의 분류기 만으로도 매우 높은 수준의 개선도를 보인다. 이는 Lasso-bagging 방법이 상당히 효율적인 앙상블 가지치기 방법임을 보인 결과이다.

앙상블을 이루는 분류기의 개수에 따라서 WAVE-bagging이 Lasso-bagging에 비해서 더 좋은 개선도 즉 예측 정확도를 나타내는 경우가 있다는 것을 확인할 수 있다. 이는 WAVE-bagging에 의한 가중치 방법이 Lasso-bagging보다 더 정교한 방법일 수 있음을 제시하고 있다. 즉, 문턱기준이 적절히 선택된다면 WAVE-bagging에 의한 가지치기 방법이 Lasso-bagging에 의한 가지치기 방법보다 우수할 수 있다. 배깅 방법은 분류기를 모두 사용했을 때 최대의 예측 정확도를 갖는 것에 비해서, Lasso-bagging이나 WAVE-bagging 방법은 앙상블 가지치기의 결과로서 비교적 적은 수의 분류기로 구성되었을 때에도 예측 정확도가 배깅을 상회한다. 다만, WAVE-bagging에서 문턱기준을 매우 높게 설정하면 앙상블 가지치기가 수행되지 않기 때문에 배깅의 결과와 일치하게 된다는 점도 확인된다.

Bagging이나 WAVE-bagging과는 다르게 Lasso-bagging을 이용한 분류 앙상블 모형은 분류기의 개수가 일정수를 넘어서면 예측 정확도에 큰 변화가 없는 것을 알 수 있다. 이는 앙상블 모형을 이루는 분류기의 개수가 증가하여도 Lasso 회귀분석에서 0이 아닌 가중치를 가진 분류기의 갯수가 일정하기 때문에 나타난 현상이다.



(a) Relative improvement (b) Matched pair t-statistics

Figure 5.1 Relative improvement over single tree and matched pair t-statistics

Figure 5.1(b)에서는 Figure 5.1(a)와 비슷한 형태의 대응표본 t-통계량 그래프가 나타난다. 여기서, 대응표본 t-통계량은 CART 모형에 대해서 WAVE-bagging, Lasso-bagging, 그리고 배깅이 더 우수한 예측 정확도를 가진다는 대립가설에 대한 검정통계량이다. 배깅은 초기에는 급격하게 증가하다가 완만하게 증가하여 가장 많은 분류기를 포함시켰을 때 최대의 검정통계량에 도달하는 형태를 보인다. Lasso-bagging 은 앙상블 모형의 크기가 어느 정도 이상이 되면 대응표본 t-통계량이 정체하는 현상이 보이고, WAVE-bagging의 경우 증가하다가 감소하는 형태를 보인다. 세가지 분류 앙상블 모형은 모두 CART 모형에 비해서 좋은 예측 정확도를 갖는다는 것을 보여주고 있다. WAVE-bagging과 Lasso-bagging을 이용한 앙상블 가지치기 방법은 서로 유사한 예측 성능을 보인다고 할 수 있으며, 두 방법은 모두 배깅 보다는 더 정확한 방법이라 할 수 있다.

Figure 5.1을 통해서 알 수 있는 것은 상대적 개선도와 대응표본 t-통계량에 따른 차이는 있지만 앙상블의 크기에 따라서 WAVE-bagging 방법이 Lasso-bagging 방법에 비해서 더 예측 정확도가 높은 구간이 존재한다는 것이다.

5.2. 분류 정확도

26개 자료에 대해서 WAVE-bagging과 Lasso-bagging, 그리고 배깅 방법들의 분류 예측 정확도를 Table 5.1에 제시하였다. 각 데이터 별로 가장 분류 예측 정확도가 높은 방법은 굵은 인쇄체로 표현되었다. 여기서 분류 예측 정확도는 10겹 교차검증 결과의 100회 평균값을 사용하였다. WAVE-bagging 방법의 분류 예측 정확도가 다른 방법에 비해 더 우수한 횟수가 많다는 것을 관찰 할 수 있다. 여기서 WAVE-bagging의 문턱기준은 총 200개의 분류기 중에서 가중치가 높은 50개의 분류기가 선택되도록 선정되었다. 분류 예측 정확도에 대한 유의적 차이 비교는 5.3절에서 진행하도록 한다.

Table 5.1 Accuracy of ensemble methods on 26 datasets

Data	WAVE-bagging	Lasso-bagging	bagging
Aba	**.7795** (.0004)	.7767 (.0003)	.7751 (.0005)
Ail	.8521 (.0004)	**.8598** (.0001)	.8401 (.0005)
Aus	.8609 (.0005)	.8568 (.0006)	**.8639** (.0004)
Bcw	.9611 (.0006)	**.9655** (.0003)	.9580 (.0007)
Bld	**.6940** (.0023)	.6788 (.0020)	.6911 (.0025)
Bod	.9330 (.0009)	**.9459** (.0005)	.9229 (.0011)
Cir	**.8216** (.0005)	.7535 (.0010)	.8087 (.0005)
Cre	.8563 (.0009)	.8580 (.0006)	**.8583** (.0007)
Cyl	.7348 (.0016)	**.7584** (.0009)	.7184 (.0015)
Dia	**.7573** (.0010)	.7506 (.0010)	.7561 (.0009)
Ech	.6818 (.0016)	.6512 (.0029)	**.6887** (.0016)
Ger	**.7459** (.0010)	.7452 (.0009)	.7438 (.0011)
Hea	.8072 (.0017)	**.8128** (.0013)	.8043 (.0019)
Hep	.8184 (.0013)	.7987 (.0019)	**.8201** (.0014)
Int	.9006 (.0007)	**.9015** (.0007)	.8145 (.0010)
Ion	.9121 (.0014)	**.9151** (.0009)	.9057 (.0015)
Mam	**.8325** (.0005)	.8315 (.0004)	.8302 (.0004)
Pid	.7720 (.0012)	.7644 (.0010)	**.7731** (.0014)
Pks	**.9138** (.0008)	.8899 (.0016)	.9035 (.0010)
Rng	**.8962** (.0006)	.8726 (.0007)	.8843 (.0006)
Snr	**.8085** (.0017)	.7902 (.0021)	.7945 (.0014)
Spe	.8001 (.0017)	.7917 (.0015)	**.8040** (.0014)
Tel	.8319 (.00020)	**.8381** (.0001)	.8258 (.0003)
Trn	**.8442** (.0006)	.8101 (.0010)	.8365 (.0006)
Twn	**.9521** (.0004)	.9131 (.0008)	.9459 (.0004)
Vot	.9489 (.0008)	.9538 (.0005)	**.9552** (.0005)

5.3. 승패를 통한 Dominance 비교

Dominance는 A모형의 예측 정확도와 B모형의 예측 정확도를 비교하는 대립가설 ($H_1 : acc_A > acc_B$)에 대한 대응표본 t검정을 수행한 결과를 요약한 것으로서, 유의수준 5%하에서 유의한 경우에는 A가 B를 '이겼다 (win)'로, ($H_1 : acc_A < acc_B$)일 때 유의한 경우에는 '졌다 (lose)'로 나타내어, 26개 데이터에 대해서 A 방법이 B 방법을 이긴 횟수에서 진 횟수를 뺀 값을 의미한다. 만약 세 모형 A, B, C가 있다면, 총 26개 데이터 각각에 대하여 (A 방법이 B를 이긴 회수 + A 방법이 C를 이긴 횟수) − (A 방법이 B에게 진 횟수 + A 방법이 C에게 진 횟수)로 정의된다.

26개 자료에 대해서 Lasso-bagging과 WAVE-bagging, 배깅의 모형을 비교한 Dominance의 결과인 Table 5.2를 바탕으로 설명을 하면, 각각의 모형을 비교하므로 52개의 비교 쌍이 존재한다. 모형에 따라서 유의한 차이가 난 경우는 총 47개의 경우이고. WAVE-bagging의 경우에 배깅과 Lasso-bagging에 대해서 총 33번 통계적으로 유의하게 예측 정확도가 높았고, 14번 통계적으로 예측 정확도가 낮았다. 따라서 Dominance는 이긴 경우에서 진 경우를 뺀 19가 된다. 결국 WAVE-bagging 방법이 가장 예측 정확도가 높은 방법인 것으로 보인다.

Table 5.2 Dominance 1: WAVE-bagging, Lasso-bagging and bagging

Methods	Dominance	Wins	Losses
WAVE-bagging	19	33	14
Lasso-bagging	-9	19	28
Bagging	-10	18	28

Lasso-bagging과 WAVE-bagging 두 방법을 직접적으로 비교하기 위해서 Table 5.3을 살펴보면, 전체 26개 자료 중에서 24개 자료에서 유의한 차이가 있었고, WAVE-bagging 방법을 이용한 예측 정확

도가 Lasso-bagging 방법론을 이용한 예측 정확도에 비해서 통계적으로 유의하게 좋았던 자료의 개수가 15개라는 것을 알 수 있다. Table 5.2와 마찬가지로 WAVE-bagging 방법이 더 높은 예측 정확도를 지니고 있다는 것을 보여준다.

Table 5.3 Dominance 2: WAVE-bagging and Lasso-bagging

Methods	Dominance	Wins	Loses
WAVE-bagging	6	15	9
Lasso-bagging	-6	9	15

6. 결론 및 향후 과제

분류앙상블 가지치기 방법론의 하나인 Lasso-bagging과 가중치를 사용한 앙상블 방법론의 하나인 WAVE를 적용한 WAVE-bagging 의 비교를 통해서 앙상블 가지치기 방법을 비교해 보았다. 실제 자료와 가상의 자료 26개에 대해서 각 방법을 적용한 결과를 보았을 때, WAVE-bagging의 예측 정확도가 더 높은 구간과 Lasso-bagging의 예측 정확도가 더 높은 구간이 존재하였다. 결과적으로 Dominance를 기준으로 보았을 때, WAVE-bagging을 사용한 경우에 더 많은 데이터에서 통계적으로 유의하다는 사실을 알았다.

WAVE-bagging 방법의 문턱기준을 너무 높게 선정하게 되면 예측정확도가 감소하여 배깅과 같은 수준으로 수렴한다는 점을 알 수 있다. 반면에 Lasso-bagging의 경우에는 앙상블을 이루는 분류기의 갯수가 일정 정도 이상 되면 예측 정확도를 유지할 수 있다는 장점이 있었다.

예측 정확도가 일관되게 높은 모형을 좋은 모형이라고 생각한다면, Lasso-bagging의 예측 정확도를 WAVE-bagging의 예측 정확도가 가장 좋았던 구간 만큼 높일 수 있거나, WAVE의 가장 높은 구간을 계속 유지할 수 있는 앙상블 모형이 좋은 모형일 것이다. 이러한 관점에서 두 가지 방법론의 장점을 결합하여 Lasso 회귀분석으로 선택된 분류기를 바탕으로 가중치를 부여하되, 그 기준이 되는 방법론은 WAVE를 적용하는 것도 하나의 방법이 될 수 있을 것이다. Lasso-bagging의 계수를 가중치로 사용하는 것도 좋은 방법이 될 수 있다. 하지만 계수가 0보다 작은 수가 될 수 있기 때문에 이 계수를 그대로 가중치로 사용할 수 없다. 따라서 앙상블을 이루는 분류기의 가중치가 0보다 같거나 크다는 제약조건을 추가하여 모형을 만드는 것도 고려해 볼 수 있을 것이다.

References

Asuncion, A. and Newman, D. J. (2007). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, `http://archive.ics.uci.edu/ml`.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and regression trees*, Chapman and Hall, New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5-32.

Chen, K. and Jin, Y. (2010). An ensemble learning algorithm based on lasso selection. *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, **1**, 617-620.

Dietterich, T. G. (2000). *Ensemble methods in machine learning*, Springer, Berlin.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119-139.

Heinz, G., Peterson, L. J., Johnson, R. W. and Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, **11**, `http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html`.

Kim, A., Kim, J. and Kim, H. (2012). The guideline for choosing the right-size of tree for boosting algorithm. *Journal of the Korean Data & Information Science Society*, **23**, 949-959.

Kim, H., Kim, H., Moon, H. and Ahn, H. (2011). A weight-adjusted voting algorithm for ensemble of classifiers. *Journal of the Korean Statistical Society*, **40**, 437-449.

Kuncheva, L. (2004). *Combining pattern classifiers: Methods and algorithms*, Wiley, New Jersey.

Kuncheva, L. (2005). Diversity in multiple classifier systems. *Information Fusion*, **6**, 3-4.

Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, **3**, 1710-1737.

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, **53**, 4046-4072.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Zhou, Z. H. and Tang, W. (2003). Selective ensemble of decision trees. *Lecture Notes in Computer Science*, **2639**, 476-483.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1371–1383

http://dx.doi.org/10.7465/jkdi.2014.25.6.1371
한국데이터정보과학회지

Comparison of ensemble pruning methods using Lasso-bagging and WAVE-bagging[†]

Seungwoo Kwak[1] · Hyunjoong Kim[2]

[12]Department of Applied Statistics, Yonsei University

Received 15 August 2014, revised 23 September 2014, accepted 17 October 2014

Abstract

Classification ensemble technique is a method to combine diverse classifiers to enhance the accuracy of the classification. It is known that an ensemble method is successful when the classifiers that participate in the ensemble are accurate and diverse. However, it is common that an ensemble includes less accurate and similar classifiers as well as accurate and diverse ones. Ensemble pruning method is developed to construct an ensemble of classifiers by choosing accurate and diverse classifiers only. In this article, we proposed an ensemble pruning method called WAVE-bagging. We also compared the results of WAVE-bagging with that of the existing pruning method called Lasso-bagging. We showed that WAVE-bagging method performed better than Lasso-bagging by the extensive empirical comparison using 26 real dataset.

Keywords: Bagging, classification, data mining, ensemble, pruning.

[†] This work was supported (in part) by the Yonsei University Research Grant of 2012 and Basic Science Research program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (No. 2012R1A1A2042177).
[1] Master candidate, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea.
[2] Corresponding author: Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea. E-mail: hkim@yonsei.ac.kr