Journal of the Korean Data & Information Science Society
2017, **28**(3), 515–524

http://dx.doi.org/10.7465/jkdi.2017.28.3.515
한국데이터정보과학회지

랜덤포레스트의 크기 결정에 유용한 승리표차에 기반한 불일치 측도[†]

박철용[1]

[1]계명대학교 통계학전공

접수 2017년 4월 17일, 수정 2017년 5월 15일, 게재확정 2017년 5월 16일

요 약

이 연구에서는 분류를 위한 RF (random forest)의 크기 결정에 유용한 승리표차 MV (margin of victory)에 기반한 불일치 측도를 제안하고자 한다. 여기서 MV는 현재의 RF에서 1등과 2등을 차지하는 집단이 무한 RF에서 차지하는 승리표차이다. 구체적으로 -MV가 양수이면 현재와 무한 RF 사이에 1등과 2등인 집단에서 불일치가 생긴다는 점에 착안하여, max(-MV,0)을 하나의 불일치 측도로 제안한다. 이 불일치 측도에 근거하여 RF의 크기 결정에 적절한 진단통계량을 제안하며, 또한 이 통계량의 이론적인 점근분포를 유도한다. 마지막으로 이 통계량을 최근에 제안된 진단통계량들과 소표본 하에서 성능을 비교하는 모의실험을 실행한다.

주요용어: 랜덤포레스트의 크기 결정, 불일치 측도, 승리표차, 진단통계량.

1. 서론

RF (random forest)는 분류 혹은 회귀를 위한 나무 (tree)를 분류기로 사용하는 앙상블 방법 중의 하나이다. 이 연구는 의사결정나무 (decision tree)를 분류기로 사용하는 연구에 한정하고자 한다. 분류를 위한 RF는 훈련용 붓스트랩 표본을 이용하여 서로 독립적인 의사결정나무를 생성하여 사용하는 앙상블 방법인 배깅 (bagging; Breiman, 1996)에 추가적인 무작위성이 추가된 방법이다. 구체적으로 RF에서는 의사결정나무의 각 노드에서 모든 변수를 사용하지 않고 일부 입력변수를 무작위로 선택하여 그 선택된 입력변수들 중에서 최적의 분리를 탐색하게 된다. 이런 무작위적 일부 입력변수 선택 방법이 목적함수를 최적화시키는 분리 (split) 탐색이라는 관점에서는 최적에서 당연히 벗어나지만 오히려 과적합에 강건 (robust)하고 (Breiman, 2001) 잡음 (noise)에 강건한 것으로 알려져 있다 (Hamza와 Larocque, 2005). 또한 대표적인 세 가지 앙상블 방법인 RF, 배깅 및 부스팅 (boosting; Shapire 등, 1998)에 대한 모의실험 비교에서 RF가 전체적으로 제일 성능이 좋은 것으로 나타났다 (Hamza와 Larocque, 2005; Caruana 등, 2008). RF는 입력변수가 많은 경우에서도 성능이 좋은 것으로 알려져 있다 (Dudoit 등, 2002; Caruana 등, 2008).

RF에서 의사결정나무를 몇 개까지 생성시켜야 할지 결정하는 문제는 RF의 실제 응용에서 상당히 중요한 주제임에도 불구하고 지금까지 연구가 그리 활발하게 전개되지 못했다. 기존 연구를 간략히 소개하면 다음과 같다. 새로운 하나의 개체에 대한 연구로는 Hernandez-Lobato 등 (2011), Park (2010)

[†] 이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2015R1D1A1A01056871).

[1] (42601) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학전공, 교수. E-mail: cypark1@kmu.ac.kr

등의 연구가 있는데 이 방법을 여러 개의 개체에 동시에 적용할 경우 생성시켜야 하는 의사결정나무의 숫자가 너무 커지게 될 가능성이 높은 문제점이 있다. 이와 더불어 OOB (out of bag) 오차율의 변동률을 이용한 Banfield 등 (2007)의 연구가 있는데, 이것에 의해 결정되는 RF의 크기는 변동성이 크고 과소하게 선택될 가능성이 많다는 문제점이 있다. Hermandez-Lobato 등 (2013)은 Hernandez-Lobato 등 (2011)의 새로운 하나의 개체에 대한 연구를 여러 개의 개체에 대한 연구로 확장하였다. 그런데 Hermandez-Lobato 등 (2013)은 목표변수가 두 개의 집단 (class)인 경우에만 적용되는 단점이 있다. 또한 이와 관련된 연구로 앙상블의 의사결정나무 깊이에 대한 연구가 진행되었다 (Choi와 Kim, 2016).

현재까지 생성된 RF와 무한으로 생성된 RF 사이에 존재하는 괴리를 나타내는 하나의 지표가 MV (margin of victory)이다. 여기서 MV는 현재까지 생성된 RF에서 1등과 2등인 집단이 무한으로 생성된 RF에서 차지하는 승리표차이다. (구체적인 정의는 2.1절을 참조하기 바란다.) 구체적으로 MV가 양이면 현재의 1등과 2등인 집단이 무한으로 생성된 RF에서 그대로 유지되는 것을 의미하며, MV가 음이면 현재의 1등과 2등인 집단이 무한 RF에서 역전되는 것을 의미하게 된다. 따라서 -MV가 커지게 되면 현재와 무한 RF 사이에 불일치가 커지는 것을 의미하게 된다. 이렇게 MV에 근거한 방법을 사용하게 되면 두 개의 집단에만 적용 가능한 Hermandez-Lobato 등 (2013)과는 달리 여러 개의 집단에 적용될 수 있는 장점이 생기게 된다.

Park (2016)은 MV에 근거한 RF의 크기 결정 방법 한 가지를 제안하였다. 이 연구에서 사용한 현재와 무한 RF의 불일치 측도는 $I(-MV > \epsilon)$으로, 여기서 $I(A)$는 A가 사실이면 1, 아니면 0인 지시함수 (indicator variable)이며 ϵ은 아주 작은 고정된 양수값 (예를 들면, 0.03)이다. 구체적으로 -MV가 ϵ보다 큰 값을 가지는 개체의 비율이 0.01보다 작아지는 최초의 값으로 RF의 크기를 결정하는 하나의 방법을 제안하였다.

이 연구에서 제안하고자 하는 현재와 무한 RF의 불일치 측도는 max(-MV,0)이다. $I(-MV > \epsilon)$을 불일치 측도로 사용하면 $-MV > \epsilon$ 여부만 따지지만, max(-MV,0)를 사용하면 -MV의 크기를 고려하기 때문에 보다 효율적인 불일치 측도가 되리라는 기대에서 이 불일치 측도를 고려한 것이다. 이 연구에서 제안하고자 하는 방법은 모든 개체의 max(-MV,0) 평균값이 아주 작은 고정된 양수값 (예를 들면 0.001)보다 작아지는 최초의 값으로 RF의 크기를 결정하는 방법이다.

앞에서도 언급되었듯이 MV에 근거한 방법의 가장 큰 장점은 3개 이상의 집단에서도 적용 가능하다는 점이다. 유일한 대안 연구인 Hermandez-Lobato 등 (2013)은 2개의 집단에서만 적용 가능하며, 아직까지 3개 이상의 집단으로 확장하는 방법이 개발되어 있지 않다. 따라서 3개 이상의 집단이 있는 분류용 RF에서는 MV에 근거한 방법이 현재까지 독보적인 위치에 있다고 할 수 있다.

이 논문의 구성은 다음과 같다. 2절에서는 RF의 크기 결정에 도움이 될 수 있는 새로운 진단통계량을 구체적으로 제안하고, 이 통계량의 이론적 점근분포를 유도한다. 이 때 Park (2016)이 제안한 방법을 동시에 소개하며, 더불어 Hermandez-Lobato 등 (2013)의 RF의 크기 결정 방법도 간략히 소개한다. 3절에서는 이 연구에서 제안하는 방법과 기존의 Hermandez-Lobato 등 (2013), Park (2016)의 방법을 11개 실제 자료집합에서 비교하는 소표본 하에서의 성능비교 모의실험을 실행한다. 4절에서는 이 연구에서 제안하는 방법을 요약하고 소표본 모의실험 결과를 정리한다.

2. RF의 크기 결정 방법들

2.1. 승리표차 MV에 근거한 방법: 제안하는 방법과 Park (2016)의 방법

이 절에서는 현재의 RF와 무한 RF의 불일치도로 승리표차 MV (margin of victory)를 이용하는 방법을 소개한다. 구체적으로 이 연구에서 제안하는 불일치 측도 max(-MV,0)를 중심으로 불일치도를 계

산하는 진단통계량들을 유도하고, 이를 통해 RF의 크기를 결정하는 알고리즘을 제안하고자 한다. 이 과정에서 역시 MV에 근거하여 불일치 측도를 사용하는 Park (2016)의 방법도 간략히 설명하도록 한다.

먼저 구체적인 진단통계량을 정확하게 나타내기 위해 간단한 표기법을 정의하도록 하자. 훈련용 자료는 $(x_1, y_1), \cdots, (x_J, y_J)$로서 J개의 개체로 구성되어 있다고 하자. 여기서 x는 입력변수이며, y는 $y^{(1)}, \cdots, y^{(r)}$의 r개 값을 취할 수 있는 범주형 목표변수이다. 매번 의사결정나무가 생성될 때마다 모든 개체에 대한 OOB (out of bag) 투표의 결과를 정리하여 다음의 통계량을 계산한다. 의사결정나무가 n개 생성되었을 때, i번째 개체에 대한 지금까지의 누적 투표도수 (cumulative frequency of votes)가 1등과 2등인 집단을 각각 c_{in}^1과 c_{in}^2이라고 나타내고, 이 집단 c_{in}^1과 c_{in}^2의 누적 투표도수를 각각 v_{in}^1과 v_{in}^2이라고 나타내자.

MV의 보다 엄밀한 정의를 위해서 추가적으로 다음의 표기법을 정의한다. 처음 k개의 의사결정나무가 생성되었을 때 개체 i에 대한 OOB 투표 중 집단 c에 투표된 것의 비율을 $RF(i, c, k)$라고 나타내자. 그러면 처음 n개의 의사결정나무에서 계산되는 MV는 다음과 같이 정의할 수 있다.

$$\widehat{d}_{in} = \frac{v_{in}^1 - v_{in}^2}{v_{in}^1 + v_{in}^2} = \frac{RF(i, c_{in}^1, n) - RF(i, c_{in}^2, n)}{RF(i, c_{in}^1, n) + RF(i, c_{in}^2, n)}.$$

여기서 $\widehat{d}_{in}$의 hat은 c_{in}^1, c_{in}^2 집단에 대해 무한 RF에서 계산된 MV인 d_{in}의 추정량이라는 의미로서 사용되었다. 따라서 d_{in}은 다음과 같이 계산될 수 있다.

$$d_{in} = \lim_{k\to\infty} \frac{RF(i, c_{in}^1, k) - RF(i, c_{in}^2, k)}{RF(i, c_{in}^1, k) + RF(i, c_{in}^2, k)}.$$

만약 $d_{in} > 0$이면 개체 i에 대한 크기 n인 RF에서의 1등과 2등 집단 c_{in}^1, c_n^2의 누적투표 순서가 무한 RF에서도 그대로 유지되게 된다. 반대로 $d_{in} < 0$이면 무한 RF에서 집단 c_{in}^2에 대한 투표가 집단 c_n^1에 대한 투표보다 커지게 되어, 무한 RF와 크기 n인 RF의 투표의 결과가 일치하지 않게 된다. 이 연구에서는 $d_{in} < 0$일 때 $-d_{in}$의 크기를 반영하는 다음의 불일치 측도를 사용하고자 한다.

$$D_n = \frac{1}{J}\sum_{i=1}^{J} \max(-\delta_{in}, 0).$$

이 불일치 측도에 대한 추정량으로 다음을 사용할 수 있다.

$$\widehat{D}_n = \frac{1}{J}\sum_{i=1}^{J} \xi\left(-\widehat{d}_{in}, (1 - \widehat{d}_{in}^2)/t_{in}\right). \tag{2.1}$$

여기서

$$\xi(a, b) = \sqrt{\frac{b}{2\pi}} \exp\left(-\frac{a^2}{2b}\right) + a\Phi\left(\frac{a}{\sqrt{b}}\right),$$

Φ는 표준정규분포의 누적분포함수이며, $t_{in} = v_{in}^1 + v_{in}^2$는 처음 n개의 의사결정나무에서 개체 i에 대한 투표가 1등과 2등인 집단의 누적 투표도수의 합계이다.

참고로 Park (2016)에서는 $\max(-d_{in}, 0)$의 크기가 아닌 $-d_{in} > 0.03$ 여부를 반영하는 다음과 같은 불일치 측도와 그에 대한 추정량을 사용하였다.

$$D_{n,.03}^* = \frac{1}{J}\sum_{i=1}^{J} I(-d_{in} > .03), \quad \widehat{D}_{n,.03}^* = \frac{1}{J}\sum_{i=1}^{J} \Phi\left(-(.03 + \widehat{d}_{in})\sqrt{\frac{t_{in}}{1 - \widehat{d}_{in}}}\right). \tag{2.2}$$

여기서 $I(A)$는 A가 사실이면 1, 아니면 0의 값을 취하는 지시함수 (indicator function)이다.

다음으로 $\max(-d_{in}, 0)$의 추정량으로

$$\xi\left(-\widehat{d}_{in}, (1-\widehat{d}_{in}^2)/t_{in}\right)$$

을 사용할 수 있는 통계적 이론의 근거를 간략하게 설명하도록 하자.

RF를 구성하는 의사결정나무는 독립적으로 생성되기 때문에 n개의 의사결정나무가 생성되고 나면 개체 i에 대한 OOB 투표 $(W_1, \ldots, W_r)$은 다항분포 $MN(n^*, (p_1, \ldots, p_r))$를 따르게 된다. 여기서 $n^* \approx n/e$로서 근사적인 OOB 투표의 개수이다. 두 개의 고정된 집단 j, k를 고려하자. 그러면 다음이 성립하는 것을 쉽게 알 수 있다.

$$W_j | (W_j + W_k = t) \sim B\left(t, p_j/(p_j+p_k)\right).$$

여기서 $B(n,p)$는 시행횟수 n, 성공의 확률 p인 이항분포이다. $d = (p_j - p_k)/(p_j + p_k)$, $\widehat{d} = (W_j - W_k)/(W_j + W_k)$라고 정의하자. 그러면 중심극한정리에 의해 $t = W_j + W_k$가 클 때 다음이 근사적으로 성립한다.

$$(\widehat{d} - d) | (W_j + W_k = t) \approx N\left(0, (1-d^2)/t\right).$$

그러므로 d에 무정보 사전분포를 주고 $\widehat{d}$가 주어진 d의 사후분포를 계산하면 $t = W_j + W_k$가 클 때 다음이 근사적으로 성립한다.

$$d | (W_j + W_k = t) \approx N\left(\widehat{d}, (1-\widehat{d}^2)/t\right).$$

따라서 $t = W_j + W_k$가 클 때 다음이 근사적으로 성립한다.

$$E\left(\max(-d >, 0)\right) \approx \xi\left(-\widehat{d}_{in}, (1-\widehat{d}_{in}^2)/t_{in}\right).$$

이를 통해 (2.1)에 제시된 진단통계량을 사용할 수 있는 통계적 이론의 근거가 마련되었다. (2.1)의 진단통계량을 사용하여 RF 크기를 결정하는 구체적인 방법을 다음과 같이 제안한다.

$$\widehat{D}_n \le \epsilon \text{이 만족되는 최소의 } n\text{으로 } RF\text{의 크기를 결정한다.} \tag{2.3}$$

여기서 ϵ은 아주 작은 고정된 양수값 (예를 들면, 0.001)이다. 참고로 Park (2016)에서 (2.2)의 진단통계량을 이용하여 RF 크기를 결정하는 방법은 다음과 같다.

$$\widehat{D}^*_{n,.03} \le 0.01 \text{이 만족되는 최소의 } n\text{으로 } RF\text{의 크기를 결정한다.} \tag{2.4}$$

2.2. Hermandez-Lobato 등 (2013)에 의한 방법

Hermandez-Lobato 등 (2013)이 제안한 방법은 Park (2016)에서도 이미 설명되어 있지만 이 논문의 자체 완성도를 위해 간단한 설명을 제공한다. 이 방법은 목표변수 y가 두 개의 값 $y^{(1)}, y^{(2)}$만 가지는 경우에 적용된다. 크기 n인 RF의 의사결정나무 분류기들의 앙상블을 $\{h_j(\cdot)\}_{j=1}^n$이라고 나타내자. 그러면 입력변수 x에 대한 다수결 (majority vote)에 의한 크기 n인 RF의 앙상블 예측은 다음과 같다.

$$\widehat{y}_n = \text{argmax}_y \sum_{j=1}^{n} I(h_j(x) = y),\ y \in \left\{y^{(1)}, y^{(2)}\right\}.$$

여기서 $I(A)$는 A가 사실이면 1, 아니면 0을 값을 취하는 지시함수이다. 그러면 크기 n인 RF의 앙상블 예측과 무한 RF의 앙상블 예측이 같아지는 확률은 다음과 같이 계산된다 (Hermandez-Lobato 등, 2013).

$$P(\widehat{y}_n = \widehat{y}_\infty | p_1) = I_{\max(p_1,1-p_1)}\left(\left[\frac{n}{2}\right]+1, n-\left[\frac{n}{2}\right]\right).$$

여기서 $p_1 = P(y = y^{(1)})$이며 $I_x(a,b) = P(B \le x)$는 모수가 a,b인 베타분포의 누적분포함수이다. $I_{\max(p_1,1-p_1)}([n/2]+1, n-[n/2])$이 n의 증가함수라는 사실을 이용하면 1에 가까운 값 0.99에 대해 다음을 만족하는 최소의 n이 개체가 하나일 때 RF 크기를 결정하는 하나의 방법이 될 수 있다 (Hermandez-Lobato 등, 2013).

$$.99 \le I_{\max(p_1,1-p_1)}\left(\left[\frac{n}{2}\right]+1, n-\left[\frac{n}{2}\right]\right). \tag{2.5}$$

하나의 개체에 대한 (2.5)의 규칙을 여러 개의 개체인 경우에도 적용할 수 있게 평균 신뢰도 개념을 사용하여 다음을 만족시키는 최소의 n으로 RF의 크기를 결정할 수 있다.

$$.99 \le P(\widehat{y}_n = \widehat{y}_\infty) = \int_0^1 I_{\max(p_1,1-p_1)}\left(\left[\frac{n}{2}\right]+1, n-\left[\frac{n}{2}\right]\right) f(p_1)dp_1.$$

여기서 $f(\cdot)$는 p_1의 사전분포이다. 모집단에 근거한 이 방법을 표본에서 사용하기 위해서 p_1에 대한 추정량으로 i $(i = 1,\ldots,J)$번째 훈련용 자료에 대해서는 OOB 투표에서 $y^{(1)}$로 예측된 비율인 $\widehat{p}_{1i}$가 선택된 다음과 같은 방법을 사용할 수 있다.

$$.99 \le \frac{1}{J}\sum_{i=1}^{J} I_{\max(\widehat{p}_{1i},1-\widehat{p}_{1i})}\left(\left[\frac{n}{2}\right]+1, n-\left[\frac{n}{2}\right]\right)$$
를 만족하는 최소의 n으로 RF의 크기를 결정한다. (2.6)

3. 세 가지 RF의 크기 결정 방법들의 성능 비교

3.1. 분석 대상 자료집합

이 분석에서 사용된 자료집합 11개는 Hermandez-Lobato 등 (2013) 및 Park (2016)의 모의실험에서도 사용되었던 자료집합이다. 이 자료들은 Hermandez-Lobato 등 (2013)에서 사용되었기 때문에 당연히 목표변수값이 2개이며, 더 많은 자료집합 중에서 결측값이 거의 없는 것들만 선정되어 사용되었다. 왜냐하면 의사결정나무가 테스트용 자료에서 훈련용 자료에는 없던 새로운 결측값이 등장하면 이를 분석할 수 있는 명확한 방법이 없기 때문이다. 구체적으로 분석 대상 자료는 Table 3.1에 설명되어 있다.

3.2. 모의실험 방법 및 결과

이 모의실험에서는 11개의 각 자료집합에서 2/3의 개체를 훈련용 자료, 나머지 1/3의 개체를 테스트용 자료로 랜덤하게 분할하여 사용하였다. 이 훈련용 자료를 이용하여 이 연구에서 제안한 $\widehat{D}_n$에 의한 (2.3) 방법, Park (2016)에 의해 제안된 $\widehat{D}^*_{n,.03}$에 의한 (2.4) 방법과 Hermandez-Lobato 등 (2013)에 의해 제안된 (2.6) 방법을 만족하는 RF의 크기를 찾기 위해 순차적인 방법을 사용하였다. 구체적으로 n을 1부터 하나씩 증가시키면서 (2.3), (2,4) 및 (2.6)을 만족시키는 값으로 각각의 RF의 크기를 결정하였다. 훈련용 자료에 대해 이렇게 결정된 세 가지 방법의 RF 크기를 테스트용 자료에 적용하여 테스트 오차율 (test error rate)을 계산하였다. 이 과정을 500번 반복하여 세 가지 방법에 의한 RF 크기와 그 때의 테스트 오차율을 계산하였다.

Table 3.1 Data sets used for our analysis: Missing cases are omitted.

Name of data sets	No. of cases	No. of input variables
abalone	4177	8
australian	690	14
banana	5299	2
german	1000	20
heart	270	13
liver	345	6
magic	19020	10
phoneme	5403	5
spam	4601	57
tictactoe	958	9
whitewine	4898	11

먼저 이 연구에서 제안한 방법 (2.3)에 대한 적절한 ϵ을 선정하도록 하자. 구체적으로 (2.3)에서 $\epsilon = .002, .001, .0005$에 대한 RF의 크기를 (2.4), (2.6)에 의한 RF의 크기들과 (이상값에 상대적으로 덜 민감한) 중앙값 관점에서 비교한 결과가 Table 3.2에 정리되어 있다.

Table 3.2 The median values of RF sizes determined by (2.3) for $\epsilon = .002, .001, .0005$ compared with those determined by (2.4) (denoted by $\widehat{D}^*_{n,.03}$) and (2.6) (denoted by hms)

Date set	$\epsilon = 0.002$	$\epsilon = 0.001$	$\epsilon = 0.0005$	$\widehat{D}^*_{n,.03}$	hms
abalone	176.0	**352.5**	696.5	416.0	412.0
australian	149.0	**289.0**	561.5	325.0	246.5
banana	*91.0*	*182.0*	361.5	165.5	112.5
german	373.0	*743.5*	*1492.0*	993.5	1684.0
heart	258.0	**503.0**	936.5	633.0	619.0
liver	470.0	*925.0*	*1822.0*	1250.0	2179.0
magic	139.0	**274.0**	544.0	300.0	256.0
phoneme	139.0	**278.0**	556.0	304.5	257.0
spam	*72.0*	*136.0*	263.0	107.5	68.0
tictactoe	*137.0*	*240.0*	424.0	249.0	187.0
whitewine	228.0	*459.0*	*913.5*	579.0	695.5

Tale 3.2에는 각 자료집합에서 $\widehat{D}^*_{n,.03}$라고 표시된 (2.4) 방법과 hms라고 표시된 (2.6) 방법 모두에서 가장 가까운 (2.6) 방법의 ϵ의 중앙값을 진하게 표시하고 있다. 또한 (2.4)와 (2.6) 방법에 의해 가까운 결과가 다른 경우에는 해당되는 두 개의 ϵ의 중앙값을 이탤릭으로 표시하고 있다. 이 결과를 요약하면 다음과 같다. $\epsilon = .001$를 사용하는 (2.3) 방법이 가장 적절한 것으로 나타났다. 구체적으로 전체 11개 자료집합 중 5개에서는 (2.4)와 (2.6) 방법 모두에서 가장 가까운 결과를 얻었으며 (진하게 표시), 나머지 6개의 자료집합에서도 (2.4) 방법에서 가장 가까운 결과를 얻은 것을 알 수 있다 (이탤릭으로 표시). 따라서 $\epsilon = .001$을 사용하는 (2.3)의 방법이 추후분석에서 사용될 최종적인 방법으로 선택되었다. 이런 방식으로 (2.3) 방법에서 사용될 ϵ을 선택한 이유는 (2.4)와 (2.6)는 이미 성능이 좋은 것으로 알려져 있어 그에 가까운 ϵ을 사용하는 것이 적절하다고 판단하였기 때문이다.

이렇게 선정된 $\epsilon = .001$를 사용하는 (2.3)의 방법을 (2.4)의 방법 및 (2.6)의 방법과 11개 자료집합에서 비교하는 모의실험 결과를 Figure 3.1부터 Figure 3.4에서 제시하였다.

Figure 3.1에서 Figure 3.4에는 각각 abalone, australian, banana 자료집합들, german, heart, liver 자료집합들, magic, phoneme, spam 자료집합들, 그리고 tictactoe, whitewine 자료집합들에 대한 상자그림 비교가 나타나 있다. 구체적으로 세 개의 상자그림이 들어 있는 왼쪽 박스는 RF의 크기, 그리고 세 개의 상자그림이 들어 있는 오른쪽 박스는 테스트 오류율을 비교하고 있다. 이 그림들을 살펴보면 자료집합에 상관 없이 테스트 오차율은 세 가지 방법이 거의 비슷한 값을 나타내고 있으나, RF의 크기

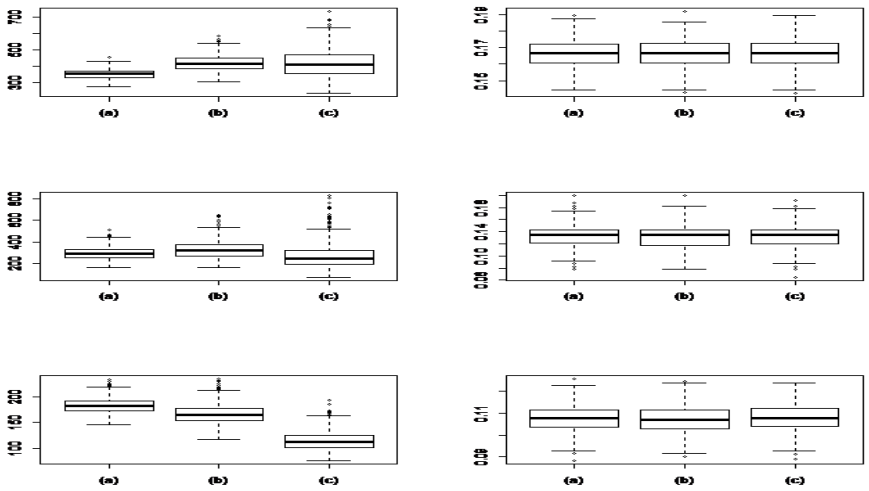

Figure 3.1 Boxplots for the number of trees (left box) and test error rate (right box) of abalone, australian, and banana data sets (three rows): The results by (2.3) with $\epsilon = .001$, (2.4) and (2.6) are denoted by (a), (b), (c), respectively.

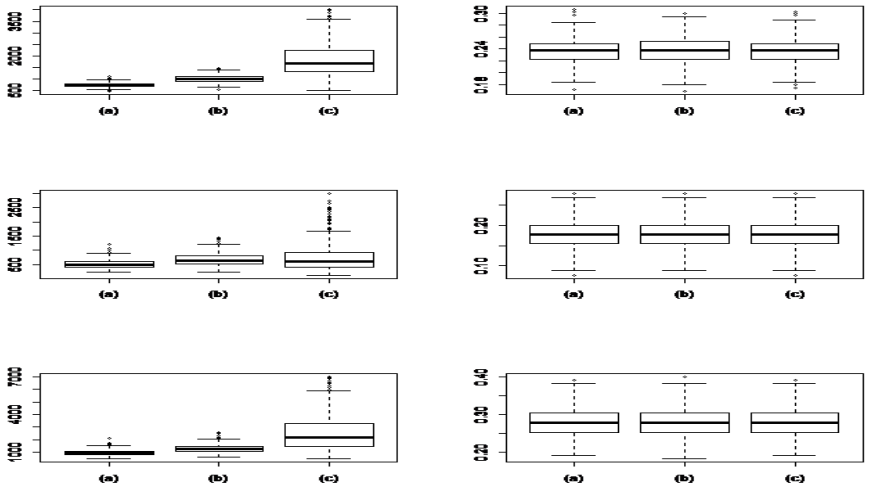

Figure 3.2 Boxplots for the number of trees (left box) and test error rate (right box) of german, heart, and liver data sets (three rows): The results by (2.3) with $\epsilon = .001$, (2.4) and (2.6) are denoted by (a), (b), (c), respectively.

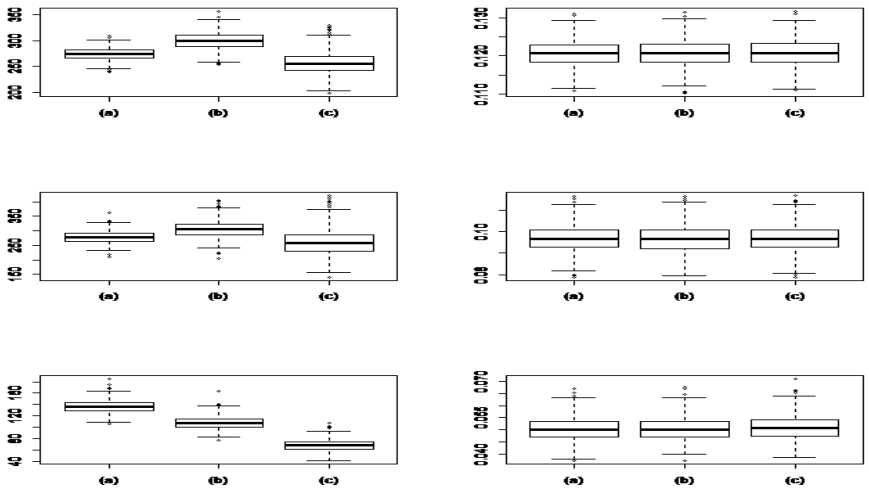

Figure 3.3 Boxplots for the number of trees (left box) and test error rate (right box) of magic, phoneme, and spam data sets (three rows): The results by (2.3) with $\epsilon = .001$, (2.4) and (2.6) are denoted by (a), (b), (c), respectively.

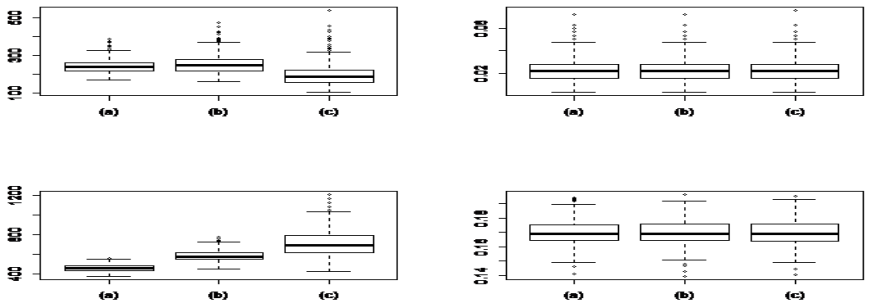

Figure 3.4 Boxplots for the number of trees (left box) and test error rate (right box) of tictactoe and whitewine data sets (two rows): The results by (2.3) with $\epsilon = .001$, (2.4) and (2.6) are denoted by (a), (b), (c), respectively.

는 세 가지 방법의 차이가 뚜렷하게 나타나고 있다. Figure 3.1부터 Figure 3.4에서 (c)로 표시된 hms 방법은 대부분 400보다 작게 RF 크기가 결정되는 자료집합인 australian, banana와 magic, phonem 및 spam, tictactoe에서 평균적으로 가장 작은 RF의 크기를 나타냈다. 그에 반해 대부분 400보다 크게 RF 크기가 결정되는 나머지 자료집합에서는 (a)로 표시되는 $\epsilon = .001$을 사용하는 (2.3) 방법, (b)로 표시되는 (2.4) 방법이 변동폭이 작으면서도 평균적으로 작은 RF의 크기를 제공하며, 특히 $\epsilon = .001$을 사

용하는 (2.3) 방법이 변동폭과 평균이 동시에 가장 작은 RF의 크기를 제공하는 것으로 나타났다.

4. 요약 및 결론

이 연구에서는 분류를 위한 RF의 크기 결정에 유용한 승리표차 MV (margin of victory)에 기반한 불일치 측도를 제안하였다. 구체적으로 -MV가 양수이면 현재와 무한 RF 사이에 1등과 2등인 집단에서 불일치가 생긴다는 점에 착안하여, max(-MV,0)를 하나의 불일치 측도로 제안하였다. 이 불일치 측도에 근거하여 RF의 크기 결정에 적절한 진단통계량을 제안하였으며, 이 통계량의 이론적인 점근분포를 유도하였다. 마지막으로 이 연구에서 제안한 방법을 최근에 제안된 방법들과 UCI 자료저장소의 11개 자료집합에서 소표본 성능을 비교하는 모의실험을 실행하였다.

11개 자료집합에 대한 모의실험에 근거하여 $\epsilon = 0.001$을 사용하는 (2.3)의 방법이 추후분석에서 사용될 우리의 방법으로 선택되었다. 또한 $\epsilon = 0.001$을 사용하는 (2.3) 방법을 최근에 제안된 $\widehat{D}^*_{n,.03}$에 근거한 (2.4) 방법 및 Hermandez-Laboto 등 (2013)에 의한 (2.6) 방법과 RF의 크기와 테스트 오차율 두 가지 관점에서 모의실험에 의해 비교하였다. 이 모의실험 결과에 의하면 테스트 오차율 관점에서는 거의 차이가 나타나지 않았지만, RF의 크기 관점에서는 뚜렷한 차이가 나타났다. 구체적으로 이 모의실험을 살펴보면 Hermandez-Laboto 등 (2013)에 의한 방법은 대부분 RF의 크기가 작게 결정되는 자료집합에서 좋은 성능을 보였고, MV에 근거하는 (2.3), (2.4) 방법은 대부분 RF의 크기가 크게 결정되는 자료집합에서 우수한 성능을 보였다. 특히 (2.3) 방법은 대부분 RF의 크기가 크게 결정되는 경우에 평균뿐만 아니라 변동폭도 세 가지 방법 중 가장 작게 결정되는 우수한 성능을 보였다. MV에 근거하는 (2.3), (2.4) 방법은 다항 목표변수에도 적용 가능하기 때문에 이진 목표변수에만 적용 가능한 Hermandez-Laboto 등 (2013)에 의한 방법보다 뛰어난 확장성을 가진다고 할 수 있다.

References

Banfield, R. E., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P. (2007). A comparison of decision tree creation techniques. *IEEE Transactions on Pattern Recognition and Machine Learning*, **29**, 173-180.

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

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

Caruana, R., Karampatziakis, N. and Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, 96-103.

Choi, S. H. and Kim, H. (2016). Tree size determination for classification ensemble. *Journal of the Korean Data & Information Science Society*, **27**, 255-264.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, **75**, 629-643.

Hernandez-Lobato, D., Martinez-Munoz, G. and Suarez, A. (2011). Inference on prediction of ensembles of infinite size. *Pattern Recognition*, **44**, 1426-1434.

Hernandez-Lobato, D., Martinez-Munoz, G. and Suarez, A. (2013). How large should ensembles of classifiers be? *Pattern Recognition*, **46**, 1323-1336.

Park, C. (2016). A simple diagnostic statistic for determining the size of random forest. *Journal of the Korean Data & Information Science Society*, **27**, 855-863.

Shapire, R., Freund, Y., Bartlett, P. and Lee, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, **26**, 1651-1686.

Journal of the Korean Data & Information Science Society
2017, **28**(3), 515–524

http://dx.doi.org/10.7465/jkdi.2017.28.3.515
한국데이터정보과학회지

A measure of discrepancy based on margin of victory useful for the determination of random forest size[†]

Cheolyong Park[1]

[1]Major in Statistics, Keimyung University

Received 17 April 2017, revised 15 May 2017, accepted 16 May 2017

Abstract

In this study, a measure of discrepancy based on MV (margin of victory) has been suggested that might be useful in determining the size of random forest for classification. Here MV is a scaled difference in the votes, at infinite random forest, of two most popular classes of current random forest. More specifically, max(-MV,0) is proposed as a reasonable measure of discrepancy by noting that negative MV values mean a discrepancy in two most popular classes between the current and infinite random forests. We propose an appropriate diagnostic statistic based on this measure that might be useful for the determination of random forest size, and then we derive its asymptotic distribution. Finally, a simulation study has been conducted to compare the performances, in finite samples, between this proposed statistic and other recently proposed diagnostic statistics.

Keywords: Determination of random forest size, diagnostic statistic, margin of victory, measure of discrepancy.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01056871).

[1] Professor, Major in Statistics, Keimyung University, Daegu 42601, Korea. E-mail: cypark1@kmu.ac.kr