Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 4 pp. 175-182, 2024

https://doi.org/10.5762/KAIS.2024.25.4.175
ISSN 1975-4701 / eISSN 2288-4688

유의성 기반 로지스틱 회귀모형 변수중요도: 워게임 데이터 분석

신웅섭[1], 차영호[2*]
[1]육군 전투지휘훈련단, [2]국방과학연구소

Significance-based Variable Importance in Logistic Regression Model: War-game Data Analysis

WoongSub Shin[1], YoungHo Cha[2*]
[1]Army Battle Command Training Program
[2]Agency for Defense Development

요 약 앙상블 기법은 다중 학습 알고리즘을 이용한 방법으로 높은 예측 성능을 보이지만 블랙박스(black-box)적인 특성으로 인해 모델에 대한 해석이 필요한 경우에는 선호되지 않는다. 부족한 설명력을 보완하기 위해 앙상블 기법에서는 설명변수 간의 상대적인 수치인 변수중요도 척도를 활용하고 있으나, 변수유의성까지 보장해 주지 못하는 한계가 있다. 반면, 회귀모형의 경우 p-value 혹은 벌점화 회귀함수 등을 활용하여 변수유의성 검정은 가능하나 변수중요도는 주어지지 않는다. 본 연구에서는 로지스틱 회귀모형에서의 변수유의성을 기반으로 한 로지스틱 회귀모형 변수중요도를 제안한다. 모의실험 결과, 로지스틱 회귀모형 변수중요도는 설명변수의 종류와 상관없이 앙상블 변수중요도와 유사한 결과가 도출되었고 해당 변수들의 유의성 또한 확인하였다. 본 연구 결과를 육군 전투지휘훈련 워게임 데이터에 적용한 결과, 선정된 10개 설명변수 간 상대적 중요도와 유의성을 판단할 수 있어 본 연구가 군 데이터 분석 시에도 활용도가 높다는 것을 확인할 수 있다.

Abstract The ensemble technique uses a multi-learning algorithm to show high predictive performance but is not preferred when the model needs to be interpreted because of its black-box nature. The ensemble technique uses a variable importance scale, a relative number between explanatory variables, to compensate for the lack of explanatory power, but it cannot guarantee variable significance. In the case of the regression model, however, it is possible to test variable significance using p-value or demerit regression functions, but the variable significance is not given. This paper reports the importance of a logistic regression model variable based on variable significance in a logistic regression model. A simulation showed that the importance of the logistic regression model variable was similar to that of the ensemble variable regardless of the type of explanatory variable, and the significance of the variables was also confirmed. After applying the results of this study to the Army BCTP data, the relative importance and significance of the 10 selected explanatory variables could be found, indicating that this study is suitable for analyzing military data.

Keywords : Logistic Regression Model, Variable Importance, Significance, War-game Data, KAARS

*Corresponding Author : YoungHo Cha(Agency for Defense Development)
email: infcha@empas.com
Received January 24, 2024 Revised March 5, 2024
Accepted April 5, 2024 Published April 30, 2024

1. 서론

빅데이터의 등장은 우리의 삶과 산업 전반에 큰 영향을 주고 있다. 단순히 데이터를 수집하고 저장하는 수준에서 벗어나, 데이터를 분석하고 이를 이용하여 새로운 가치를 창출하는 기술의 발전을 가져오고 있다. 이러한 변화의 흐름은 군이 보유하고 있는 데이터에 대한 인식과 태도에도 영향을 미치고 있다. 군 데이터의 특성상 보안이 강조되다 보니 생성된 데이터를 안전하게 저장, 관리하는 수준에서 크게 벗어나지 못했던 과거와 달리, 최근에는 군에서 축적한 다양한 데이터를 통계 기법들을 활용하여 분석하려는 노력들이 활발하게 이루어지고 있다. 특히, 과학화 전투훈련단(Korea Combat Training Center, KCTC)이 보유한 전투원들에 대한 훈련 데이터와 전투지휘훈련단(Battle Command Training Program, BCTP)이 보유한 워게임 데이터는 전장상황을 간접적으로 체험한 결과물로서 미래 전장상황 예측을 위한 핵심적인 데이터로 활용되고 있다. 이러한 군 데이터 분석 시 군 내부적으로 활용도를 높이기 위해 모델링이 간결하면서도 성능은 뛰어난 기법들이 주로 사용되는데 그중에서 로지스틱 회귀모형과 앙상블 기법이 연구에 많이 사용되고 있다. 예를 들면, [1]에서는 여단급 KCTC 데이터의 곡사화기 전투결과를 분석 시 랜덤포레스트 기법을 활용하여 22개 변수 중 우수부대에 영향을 미치는 중요 변수를 확인하였으며, [2]에서는 KCTC 데이터를 Cox 비례위험모형과 로지스틱 회귀모형을 활용하여 전투원 생존분석을 하였다. [3]에서는 BCTP 데이터를 활용하여 군사작전 성공률 예측모델 연구를 위하여 로지스틱 회귀모형과 랜덤포레스트 기법을 활용하였다. 이렇듯 본 연구에서는 군 데이터 분석 시 활용도가 높은 로지스틱 회귀모형에 대해 설명력 강화 및 성능 향상 방안을 제안한다.

로지스틱 회귀모형은 종속변수가 이항 분포를 따를 때, 독립변수와의 관계를 모델링하는 통계 기법으로, 의학, 금융, 마케팅, 기후분석, 국방 등 다양한 분야의 분류 문제에서 활용도가 높은 방법론이다. 앙상블 기법은 다중 학습 알고리즘을 이용한 방법으로, 여러 개의 개별 모델들의 조합을 통해 최적의 모델로 일반화함으로써 분류 문제에서 높은 예측 성능을 보인다. 앙상블 기법의 모델들은 높은 성능을 보여주지만 블랙박스(black-box)적인 특성으로 인해 직관적인 해석이 필요한 경우에는 로지스틱 회귀모형이 많이 사용되고 있다. 부족한 설명력을 보완하기 위해 앙상블 기법에서는 변수중요도라는 척도를 통해 설명변수들이 종속변수에 영향을 주는 정도를 상대적인 수치로 제시한다. 그러나 변수중요도를 활용하여 종속변수에 영향을 많이 주는 설명변수를 확인할 수 있으나 이는 다른 설명변수들과의 상대적인 수치일 뿐, 해당 변수의 유의성까지 보장해 주지 못한다는 한계가 있다. 즉, 변수중요도 값이 가장 큰 설명변수는 종속변수에 영향을 가장 많이 준 변수이긴 하나 유의한 설명변수는 아닐 수 있는 것이다. 이와는 반대로 로지스틱 회귀모형의 경우 p-value 혹은 벌점화 회귀함수 등을 활용하여 유의한 설명변수를 찾아주는 다양한 방법들이 제시되어 있는 반면, 앙상블의 변수중요도와 같이 변수 간의 상대적인 중요도를 나타내는 척도는 주어지지 않는다. 본 논문에서는 회귀모형에서의 변수 선택 방법들을 활용하여 변수유의성을 기반으로 한 로지스틱 회귀모형 최적 변수중요도를 제안한다.

본 논문의 2장에서는 앙상블 기법에서의 변수중요도를 소개하고, 유의성 기반 로지스틱 회귀모형 변수중요도를 설명한다. 3장에서는 모의실험을 통해 기존의 앙상블 변수중요도와 유의성 기반 로지스틱 회귀모형 변수중요도를 비교하여 유사성을 확인한다. 4장에서는 BCTP 워게임 데이터에 유의성 기반 로지스틱 회귀모형 변수중요도를 적용하여 어떠한 유용성을 가지는지 논의한다. 마지막 5장에서는 모의실험과 워게임 데이터 분석을 통해 알 수 있는 제안된 방법의 의의 및 추후 연구에 대해 논의한다.

2. 변수중요도

본 연구에서는 사용하는 변수중요도는 앙상블의 대표적인 기법인 랜덤포레스트와 그래디언트 부스팅의 변수중요도이다.

2.1 랜덤포레스트 변수중요도

랜덤포레스트[4]의 대표적인 변수중요도는 mean decrease impurity(MDI)이다[5]. MDI 변수중요도는 의사결정나무에서 각 설명변수가 마디를 분리할 때 발생하는 지니(Gini) 불순도(impurity)의 감소량 평균을 구한 값이다. $m(m=1,2,\cdots,M)$번째 의사결정나무에서 특정 마디 t에서의 설명변수 $X_j(j=1,2,\cdots,J)$에 대한 지니 불순도를 구하는 수식은 Eq. (1)과 같다.

$$G(X_j,t)=\sum_{i=1}^{K}p_{i\mid t}(1-p_{i\mid t})=1-\sum_{i=1}^{K}p_{i\mid t}^2 \quad (1)$$

이때, 종속변수 Y는 K개의 범주 중에 하나의 값을 가지며, n개의 데이터가 특정 범주에 속할 확률을 각각 $p_{i \mid t}(i=1,2,\cdots,K)$라 한다. MDI 변수중요도를 계산하는 알고리즘은 다음과 같이 진행한다.

Step1. 특정 마디를 분리할 때 지니 불순도 감소량 $\triangle G(X_j,t)$을 최대로 하는 설명변수 X_j를 Eq. (2)와 같이 선택한다.

$$\arg\max_j \triangle G(X_j,t) = \arg\max_j \left[\left(1-\sum_{i=1}^{K} p_{i \mid t}^2\right) - \frac{N_{t_L}}{N_t}\left(1-\sum_{i=1}^{K} p_{i \mid t_L}^2\right) - \frac{N_{t_R}}{N_t}\left(1-\sum_{i=1}^{K} p_{i \mid t_R}^2\right) \right] \quad (2)$$

이때, t_L, t_R은 각각 특정 마디 t의 왼쪽 마디, 오른쪽 마디를 의미하며, N_t, N_{t_L}, N_{t_R}은 각각 t, t_L, t_R마디에 속하는 표본의 수를 의미한다.

Step2. Step1에서 발생하는 지니 불순도 감소량을 Eq. (3)과 같이 의사결정나무 T_m마다 각 설명변수 X_j별로 누적한다.

$$I_{jm} = \sum_{t \in T_m} \triangle G(X_j,t) \quad (3)$$

Step3. Step2에서 계산된 누적 값 I_{jm}을 M으로 나눠주어 설명변수 X_j에 대한 변수중요도 I_j를 Eq. (4)와 같이 구한다.

$$I_j = \frac{1}{M}\sum_{m=1}^{M} I_{jm} \quad (4)$$

Step4. Steps 2~3을 반복하여 모든 설명변수의 변수중요도를 계산한다.

2.2 그래디언트 부스팅 변수중요도

그래디언트 부스팅[6]에서의 변수중요도는 의사결정나무에서 특정 설명변수를 사용했을 때 얻어지는 정보획득량(Information Gain)의 합을 구한 값이다. 여기서 정보 획득량은 특정 설명변수에 의해 종속변수의 엔트로피 $H(Y)$가 얼마나 감소하였는가를 나타내는 값으로, 엔트로피와 정보 획득량은 Eq. (5)와 같이 정의된다.

$$H(Y) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i)), \quad (5)$$
$$IG(Y,X_j) = H(Y) - H(Y \mid X_j)$$

여기서 $H(Y \mid X_j)$는 설명변수 X_j에 대한 종속변수 Y의 조건부 엔트로피를 의미하며, $p(x_i)$는 각 마디에 속한 데이터의 클래스 비율이다. 변수중요도를 계산하는 알고리즘은 다음과 같이 진행한다.

Step1. 특정 의사결정나무 T_m에서 설명변수 X_j를 사용했을 때 얻어지는 정보 획득량의 합을 Eq. (5)과 같이 계산한다.

$$I_{jm} = \sum_{j=1}^{L-1} IG(Y,X_j)1(X_j=j) \quad (6)$$

이때, L은 해당 트리의 리프 노드(leaf node)의 개수이며, $1(X_j=j)$는 지시 함수(indicator function)이다.

Step2. Step1에서 계산된 I_{jm}을 M으로 나눠주어 설명변수 X_j에 대한 변수중요도 I_j를 Eq. (7)과 같이 구한다.

$$I_j = \frac{1}{M}\sum_{m=1}^{M} I_{jm} \quad (7)$$

Step3. Steps 1~2를 반복하여 모든 설명변수의 변수중요도를 계산한다.

2.3 유의성 기반 로지스틱 회귀모형 변수중요도

회귀모형에서 유의한 설명변수를 찾을 때 주로 사용되는 척도는 회귀계수이다. 특정 설명변수의 계수가 유의한 지를 p-value 등 기준 통계치를 선정하여 판단하거나, 벌점함수를 활용하여 유의하지 않은 계수를 0으로 만들어 제거하는 방식으로 유의한 설명변수를 식별한다. 즉, 종속변수에 영향을 주는 설명변수들의 계수 값은 0이 아닌 특정 값을 가지게 되며, 계수 값이 클수록 종속변수의 변동이 커지기 때문에 계수 값이 큰 설명변수가 상대적으로 종속변수에 더 많은 영향을 준다고 볼 수 있다. 이러한 회귀계수 특성을 활용하여 회귀모형에서의 변수중요도를 정의할 수 있다. 이 논문에서 제시하는 유의성 기반 로지스틱 회귀모형 변수중요도의 계산 방법은 아래와 같다.

Step1. 설명변수가 연속형 변수일 경우 모형 적합 전 정규화를 Eq. (8)을 이용하여 실시한다.

$$X_j^* = \frac{X_j - \overline{X_j}}{S.D(X_j)} \quad (8)$$

Step2. Step1이 완료되면 유의한 설명변수를 찾는 다양한 방법론들을 회귀모형에 N번 적용하여 그 결과로 얻어지는 설명변수 X_j^*의 회귀계수 추정량 $\hat{\beta}_{ij}$의 절대값의 평균을 Eq. (9)와 같이 구한다.

$$\overline{|\hat{\beta}_{ij}|} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\beta}_{ij}| \quad (9)$$

Step3. Step2에서 계산된 $\overline{|\hat{\beta}_{ij}|}$을 $S.E(|\hat{\beta}_{ij}|)$으로 나눠주어 설명변수 X_j에 대한 변수중요도 I_j를 Eq. (10)과 같이 구한다.

$$I_j = \frac{\overline{|\hat{\beta}_{ij}|}}{S.E(|\hat{\beta}_{ij}|)} \quad (10)$$

Step4. Steps 2~3을 반복하여 모든 설명변수의 변수중요도를 계산한다.

3. 모의실험

3장에서는 기존의 앙상블 변수중요도와 2장에서 제시한 로지스틱 벌점화 회귀모형 변수중요도를 모의실험으로 서로 비교한다. 설명변수의 종류와 선형관계 여부 등을 고려하여 총 8가지 모형으로 모의실험을 진행한다. 변수는 유의한 변수 5개, 잡음변수 5개로 표본 500개를 생성하여 사용한다. 앙상블 모형의 변수중요도는 랜덤포레스트와 그래디언트 부스팅을 활용하여 각각 100번씩 계산한 변수중요도들의 평균값을 사용하며, 로지스틱 벌점화 회귀모형 변수중요도는 벌점화 함수로 Lasso[7], SACD[8], MCP[9], 단계선택법(Stepwise method)을 적용하여 변수중요도를 계산한다.

3.1 모형 1 : 표준정규분포, 선형 모형

모형 1은 10개의 설명변수 $X_1, X_2, X_3, \cdots, X_{10}$가 모두 표준정규분포를 따르며, 반응변수 Y는 X_1, X_2, X_3, X_4, X_5를 이용하여 Eq. (11)과 같이 생성한다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5 + X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5},\ Y_i \sim B(1,\ \pi_i) \quad (11)$$

이렇게 생성한 Y에 대해서 설명변수 $X_1, X_2, X_3, \cdots, X_{10}$를 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

Fig. 1. Comparison of variable importance for Model 1

Fig. 1에서 회귀모형 변수중요도를 보면 유의한 설명변수 $X_1, \ldots, X_5$와 잡음변수 $X_6, \ldots, X_{10}$의 크기가 구분되어 나타난다. 전체적인 그래프 형태가 앙상블 모형에서의 변수중요도와 유사한 형태를 보이며, 잡음변수의 경우에는 회귀모형 변수중요도에서 0 또는 0에 가까운 값으로 측정된다는 것을 확인할 수 있다. 이는 변수 간의 상대적 중요도뿐만 아니라 해당 수치값을 통해 유의성 여부도 확인 가능하다는 것을 알 수 있다.

3.2 모형 2 : 표준정규분포, 잡음변수 없는 선형 모형

모형 2는 5개의 설명변수 X_1, X_2, X_3, X_4, X_5가 모두 표준정규분포를 따르며, 반응변수 Y는 X_1, X_2, X_3, X_4, X_5를 이용하여 Eq. (12)와 같이 생성한다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5 + X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5},\ Y_i \sim B(1,\ \pi_i) \quad (12)$$

이렇게 생성한 Y에 대해서 잡음변수를 제외한 설명변수 X_1, X_2, X_3, X_4, X_5만을 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

Fig. 2에서의 5개 변수는 잡음변수 없이 모두 유의한 설명변수이다. 5개 변수 모두 모형 적합에 사용되었기 때문에 변수중요도 상에서 변수 간 큰 차이 없이 나타난다는 것을 알 수 있다. 또한 6개 방법론 모두 X_1가 가장 큰 변수중요도 값을 가지는 것으로 나타난다. 설명변수에 잡음변수가 없고 모두 표준정규분포일 경우 회귀모형 변수중요도가 앙상블 변수중요도와 유사하게 나타난다고 볼 수 있다.

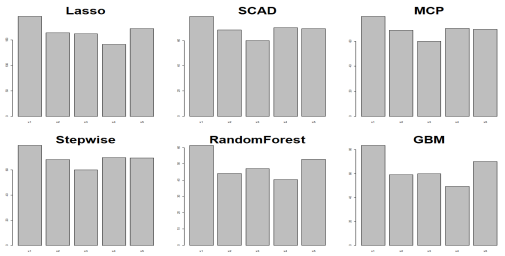

Fig. 2. Comparison of variable importance for Model 2

3.3 모형 3 : 표준정규분포, 상수항 모형

모형 3은 5개의 잡음변수 $X_6, X_7, X_8, X_9, X_{10}$가 모두 표준정규분포를 따르며, 반응변수 Y는 상수항에 의해 Eq. (13)과 같이 생성한다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5,\ Y_i \sim B(1,\ \pi_i) \tag{13}$$

이렇게 생성한 Y에 대해서 잡음변수 $X_6, X_7, X_8, X_9, X_{10}$를 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

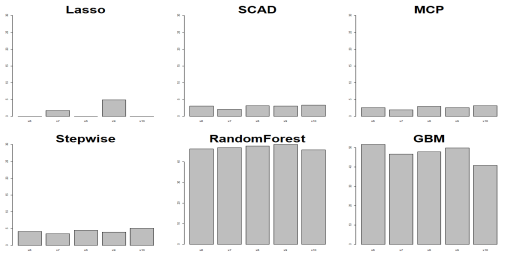

Fig. 3. Comparison of variable importance for Model 3

모형 적합 시 잡음변수 $X_6, X_7, X_8, X_9, X_{10}$가 모두 사용되지 않았기 때문에 변수중요도는 설명변수가 모두 유사한 값을 가지거나 0 또는 0에 가까운 값을 가지는 것이 적합하다. Fig. 3에서 랜덤포레스트와 SCAD, MCP, Stepwise 변수중요도가 이에 부합하는 결과라고 볼 수 있으며, 나머지 방법론들도 어느 정도 유사한 수치의 변수중요도 결과를 보여준다고 할 수 있다. Lasso의 경우, 100번의 모형 적합에서 X_6, X_8, X_{10}를 모두 제거하였으며, X_7, X_9 또한 0에 가까운 값으로 변수중요도 값이 나타남을 확인할 수 있다. 따라서 상수항 모형에서도 회귀모형 변수중요도가 변수 간 상대적 중요도를 적절하게 표현할 뿐만 아니라 유의성 또한 값으로 잘 표현함을 알 수 있다.

3.4 모형 4 : 혼합 분포, 선형 모형

모형 4는 설명변수의 분포가 모두 다른 경우이며, 유의한 설명변수 5개와 잡음변수 5개 각각 1개씩 짝을 지어 Eq. (14)와 같이 동일한 분포를 따르도록 생성한다.

$$\begin{aligned} &X_1, X_6 \sim \chi^2_{(2)},\ X_2, X_7 \sim U(-2, 2), \\ &X_3, X_8 \sim Exp(0.5), X_4, X_9 \sim I(0.5, 1), \\ &X_5, X_{10} \sim N(0, 1) \end{aligned} \tag{14}$$

반응변수 Y는 X_1, X_2, X_3, X_4, X_5를 이용하여 Eq. (15)와 같이 생성한다.

$$\begin{aligned} &\log\left(\frac{\pi_i}{1-\pi_i}\right) = \\ &0.5 + X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5},\ Y_i \sim B(1,\ \pi_i) \end{aligned} \tag{15}$$

이렇게 생성한 Y에 대해서 설명변수 $X_1, X_2, X_3, \cdots, X_{10}$를 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

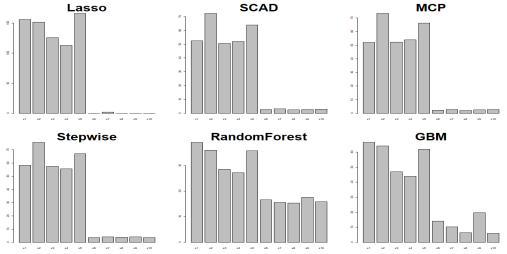

Fig. 4. Comparison of variable importance for Model 4

Fig. 4를 보면 Fig. 1에서와 유사하게 유의한 설명변수 $X_1, \ldots, X_5$와 잡음변수 $X_6, \ldots, X_{10}$가 변수중요도 크기 차이로 잘 구분되어 나타난다. 설명변수의 분포와 상관없이 선형 로지스틱 모형에서는 유의한 설명변수와 잡음변수의 변수중요도 값의 차이가 그래프 상으로 확인 가능한 것이다. 또한 잡음변수의 경우 0 또는 0에 가까운 값을 가지고 있으므로 상대적 중요도뿐만 아니라 유의성 또한 유의성 기반 로지스틱 회귀모형 변수중요도를 통해 확인할 수 있다.

3.5 모형 5: 혼합 분포, 잡음변수 없는 선형 모형

모형 5는 유의한 설명변수 5개의 분포가 서로 다르며, 반응변수 Y는 X_1, X_2, X_3, X_4, X_5를 이용하여 다음과 같이 생성한다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)= 0.5+X_{i1}+X_{i2}+X_{i3}+X_{i4}+X_{i5},\quad Y_i\sim B(1,\pi_i) \tag{16}$$

이렇게 생성한 Y에 대해 설명변수 X_1, X_2, X_3, X_4, X_5만을 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

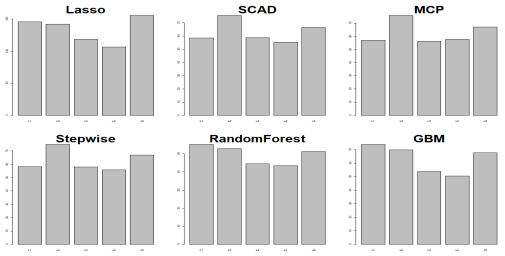

Fig. 5. Comparison of variable importance for Model 5

Fig. 5에서 6개 방법론 모두 유사한 변수중요도 그래프를 보여준다. 즉, 잡음변수가 없는 선형 모형에서는 설명변수의 분포와 상관없이 앙상블 변수중요도와 유의성 기반 로지스틱 회귀모형 변수중요도 모두 유사한 성능을 보여준다고 할 수 있다.

3.6 모형 6 : 혼합 분포, 상수항 모형

모형 6은 잡음변수 $X_6, X_7, X_8, X_9, X_{10}$의 분포가 서로 다르며, 반응변수 Y는 상수항에 의해 다음과 같이 생성한다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5,\ Y_i\ \sim\ B(1,\ \pi_i) \tag{17}$$

이렇게 생성한 Y에 대해서 잡음변수 X_6, X_7, X_8, X_9, X_{10}만을 사용하여 모형 적합을 100번 실행한 후 변수중요도를 확인한다.

모형 적합 시 잡음변수 X_6, X_7, X_8, X_9, X_{10}가 모두 사용되지 않았기 때문에 변수중요도는 설명변수가 모두 유사한 값을 가지거나 0 또는 0에 가까운 값을 가지는 것이 적합하다. Fig. 6을 보면 랜덤포레스트 변수중요도, 유의성 기반 로지스틱 회귀모형 변수중요도가 이에 부합하는 결과를 보여주고 있다. 그래디언트 부스팅 변수중요도에서도 어느 정도 유사한 수치의 변수중요도 결과를 보여준다고 할 수 있다. Fig. 3에서와 마찬가지로 상수항 모형에서 잡음변수에 대한 변수중요도를 잘 표현하고 있음을 알 수 있다.

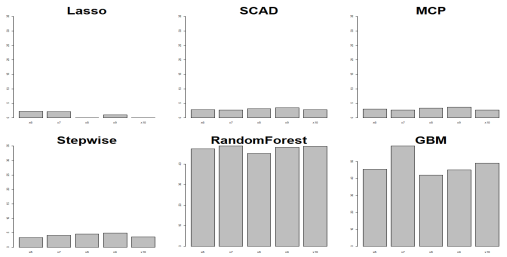

Fig. 6. Comparison of variable importance for Model 6

모의실험 결과, 유의성 기반 로지스틱 회귀모형 변수중요도는 6개 모형에서 모두 기존 앙상블 변수중요도와 그래프 상으로 유사한 형태를 보여주었다. 또한 설명변수와 잡음변수가 함께 있는 경우, 두 그룹의 차이를 명확하게 나타내어 주어 그 자체로 유의성까지도 충분히 확인할 수 있었다. 특히 유의한 변수들 혹은 잡음변수들만으로 이루어진 경우, 앙상블 변수중요도는 Fig. 2, 3, 5, 6에서 볼 수 있듯이 그래프 형태가 모두 유사하여 그 차이가 구분되지 않는 반면, 유의성 기반 로지스틱 회귀모형 변수중요도의 경우, 설명변수가 모두 유의한 변수일 경우에는 Fig. 2, 5와 같이 기존 앙상블 변수중요도와 유사한 그래프를 보여주었으며 모두 잡음변수일 경우에는 Fig. 3, 6에서 볼 수 있듯이 0 또는 0에 가까운 일정한 그래프를 보여줌으로써 상대적 변수중요도 뿐만 아니라 유의성까지도 잘 나타내준다는 것을 확인할 수 있다.

4. 워게임 데이터 분석

4장에서는 본 논문에서 제시한 유의성 기반 로지스틱 회귀모형 변수중요도와 기존 앙상블 변수중요도를 워게임 데이터에 적용하여 어떠한 차이를 보이는지 확인한다. 6가지 방법론에 적용한 결과를 비교하여 본 논문에서 제안한 방법이 어떠한 유용성을 가지는지 논의한다.

4.1 데이터 소개

본 논문에서 사용한 워게임 데이터는 '22 ~ '23 육군 전투지휘훈련 간 발생한 훈련 데이터를 활용한 것으로, 전투지휘훈련단 사후검토체계(KAARS: Korea After Action Review System)로 추출할 수 있는 변수들의 평균값과 표준편차를 계산하여 정규분포 가정 하 데이터를 생성하여 표본으로 활용한다. 해당 방법은 [3]에서 제안

한 것으로, 전투지휘훈련 간 발생한 훈련 데이터는 군 비밀로 관리되기 때문에 해당 데이터를 직접 사용하는 것은 불가능하여 사후검토체계를 활용하여 위와 같은 방법을 통해 제한적으로 사용하는 방식이다.

사후검토체계에서 추출할 수 있는 독립변수 중 군사전문가의 의견, 군사교범, 사후검토자료 등을 참고하여 역습작전에 의미가 있는 10개의 독립변수를 선정하고, 해당 변수들의 평균값과 표준편차를 활용하여 변수별 500개의 표본을 생성한다. 종속변수는 역습작전 시행 여부를 나타내는 범주형 변수이다. 해당 변수들을 로지스틱 회귀모형으로 적합하여 6가지 방법론에 적용한 변수중요도 결과를 비교한다.

Table 1. War-game data variable description

Variable name	Variable description
Block	Block unit combat power
Counter	Counter unit combat power
Enemy	Enemy combat power
Distance	Distance between enemy units
CAS	Retained quantity of CAS
Arty	Artillery combat power
Depth	Depth of breakthrough
Width	Width of breakthrough
Rain	Precipitation status
Support	Support unit combat power

4.2 변수중요도

워게임 데이터 적합 결과 6개 방법론에서 공통적으로 Block, Counter, Enemy, Arty, Support의 변수중요도가 상대적으로 높게 나타난다. 반면 Distance, CAS, Depth, Width, Rain의 변수중요도는 상당이 낮거나 CAS, Rain의 경우 값이 0인 것을 확인할 수 있다. 유의성 기반 로지스틱 회귀모형 변수중요도와 앙상블 변수중요도가 워게임 데이터 분석에서도 상당히 유사한 결과를 보여준다는 것을 확인할 수 있다. 또한 앙상블 변수중요도와 달리 Lasso, Stepwise 변수중요도의 경우 CAS, Rain의 중요도 값이 0으로 나타났기 때문에 해당 변수들을 불필요한 변수로 판단하여 제거하는 등의 유의성 검정까지도 가능하고 볼 수 있다.

Fig. 7을 바탕으로 역습작전 시행 여부를 판단해 보면, 역습작전이 가능하기 위해서는 가장 큰 변수중요도 값을 보이는 지속지원부대의 전투력(Support)이 무엇보다도 높게 유지되어야 하며, 적의 공격을 막고 있는 저지부대의 전투력(Block)과 역습을 시행하는 공격 부대의 전투력(Counter)이 충분하게 유지되어야 한다고 해석할 수 있다. 또한 역습 시행 전 충분한 포병 화력(Arty)을 통해 적의 전투력(Enemy)을 감소시켜야 한다는 결론을 내릴 수 있다. 유의성 기반 로지스틱 회귀모형 변수중요도를 통해서 실제 교리적, 전술적으로 타당한 결론을 도출해 낼 수 있는 것이다.

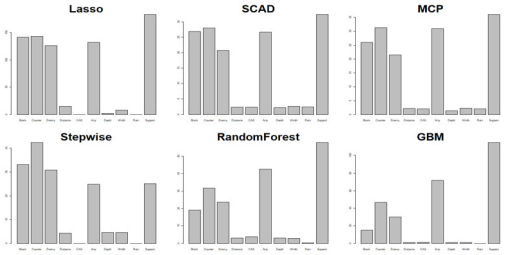

Fig. 7. Comparison of variable importance for War-game data

반면, 현재 돌파구의 종심(Depth)과 너비(Width), 적과의 거리(Distance)는 유의성 기반 변수중요도 값이 상당히 낮고, 특히 CAS와 강수 여부는 값이 0이기 때문에 해당 변수들은 지휘관의 역습작전 계획 시 큰 영향을 미치지 못하므로 우선순위를 낮게 잡을 수 있다. 전장상황에서 여러 가지 변수들을 즉각적으로 판단해야 하는 지휘관 입장에서 불필요한 변수들을 사전에 알 수 있다는 것은 핵심적인 변수들만을 활용하여 더욱 신속하게 상황판단을 할 수 있다는 이점을 가져다준다고 할 수 있다.

이러한 워게임 데이터 분석 결과를 통해 상대적 중요도와 각 변수의 유의성을 동시에 제공하는 유의성 기반 로지스틱 회귀모형 변수중요도는 군 데이터 분석 시 효율성 높여주고 설명력을 보다 더 강화하는 방법론으로 충분히 활용 가능하다는 것을 확인할 수 있다.

5. 결론

본 논문에서는 로지스틱 회귀모형에서의 변수유의성을 활용하여 유의성 기반 로지스틱 회귀모형 변수중요도를 제안하였으며, 모의실험과 워게임 데이터 분석을 통해 앙상블 기법에서의 변수중요도와 비교하였다. 유의성 기반 로지스틱 회귀모형 변수중요도는 3장에서 확인한 것과 같이 설명변수의 종류와 상관없이 앙상블 변수중요도와 유사한 결과를 보였으며 변수중요도를 통해 해당 변수들의 유의성 또한 확인할 수 있었다. 4장에서 워게

임 데이터에 적용한 결과, 유의성 기반 변수중요도를 통해 실제 교리적, 전술적으로 타당한 결론을 도출해 낼 수 있다는 것을 확인하였다. 변수중요도와 함께 유의성까지 확인이 가능하다는 장점을 가지고 있어서 군 데이터 분석 시 로지스틱 회귀모형의 활용도를 높이고 설명력을 강화하는 방법론으로 활용 가능하다는 것을 확인하였다.

이번 논문에서 제안한 유의성 기반 로지스틱 회귀모형 변수중요도는 종속변수가 이진 분류인 경우에만 적용한 것이기 때문에 향후에 다항 로지스틱 회귀모형에 적용하는 연구를 진행한다면 유의성 기반 로지스틱 회귀모형 변수중요도의 활용도를 더욱 높일 수 있을 것이라 예상된다.

References

[1] J. Y. Park, J. H. Kim, C. D. Kim, T. S. Kim, H. S. Moon, "An analysis for Indirect fire results of the brigade-level KCTC training", *Korean Journal of Military Art and Science*, vol.77, no.1, pp.460-481, Feb. 2021.
DOI: http://doi.org/10.31066/kjmas.2021.77.1.017

[2] K. K. Kim, K. C. Won, H. J. Lee, "A survival analysis of combatants using the Cox proportional hazards model and the logistics regression model", *Journal of The Korean Data Analysis Society*, vol.24, no.4, pp.1289-1304, Aug. 2022.
DOI: https://doi.org/10.37727/jkdas.2022.24.4.1289

[3] D. H. Go, "Development of a Military Operation Success Rate Prediction Model Using Machine Learning Techniques", *Journal of Military Innovation*, 2022.

[4] L. Breiman, "Random forests", *Machine learning*, vol.45, no.1, pp.5-32. Oct. 2001.

[5] G. Biau, E. Scornet, "A random forest guided tour", *Test*, vol.25, no.2, pp.197-227 Apr. 2016.
DOI: http://dx.doi.org/10.1007/s11749-016-0481-7

[6] J. Friedman, Stochastic gradient boosting, Department of Statistics in Stanford University, 1999.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso" *Journal of the Royal Statistical Society: Series B (Methodological)*, vol.58, no.1, pp.267-288, 1996.
DOI: http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

[8] J. Fan, R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties" *Journal of the American statistical Association*, vol. 96, pp. 1348-1360, Dec. 2001.
DOI: http://dx.doi.org/10.1198/016214501753382273

[9] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty" *The Annals of Statistics*, vol.38, no.2, pp.894-942, Apr. 2010.
DOI: http://dx.doi.org/10.1214/09-AOS729

신 웅 섭(WoongSub Shin) [정회원]

- 2014년 2월 : 육군사관학교 정보과학과 (정보과학학사)
- 2023년 2월 : 고려대학교 일반대학원 통계학과 (통계학석사)
- 2023년 3월 ~ 현재 : 육군전투지휘훈련단 전투모의처 기동논리장교

〈관심분야〉
머신러닝, M&S, 네트워크, 최적화

차 영 호(YoungHo Cha) [정회원]

- 2005년 3월 : 미국 공군대학원 OR (OR석사)
- 2009년 8월 : KAIST 산업 및 시스템공학과 (산업공학박사)
- 2018년 12월 ~ 2023년 11월 : 육군전투지휘훈련단 전투모의처 전투모의과장
- 2023년 11월 ~ 현재 : ADD 군전력연구센터 현역연구원

〈관심분야〉
최적화, 스케줄링, M&S, AI