홀스타인 착유우 사료에 대사 단백질 공급량에 대한 머신러닝 기반 예측 모델 개발

김동현, 임동현, 박성민, 박지후, 최태정, 엄준식
국립축산과학원 낙농과
e-mail:kimdh3465@korea.kr

Development of Machine Learning-Based Prediction of Metabolizable Protein Supply from Feed in Dairy Cows during Lactation

Dong-Hyeon Kim, Dong-Hyun Lim, Seong-Min Park, Ji-Hoo Park,
Tae-Jeong, Choi, Jun-Sik Eom
Dairy Science Division, National Insititute of Animal Science

요 약

지속 가능한 가축 생산을 위해 영양 및 우유 생산량을 최적화하려면 젖소의 단백질 이용률을 정확하게 예측하는 것이 필수적이다. 본 연구는 식이 단백질 섭취량을 기반으로 반추위 미분해 단백질(RUP)과 미생물체단백질(MicN)을 예측하는 새로운 머신러닝 모델을 개발하는 것을 목표로 했다. 436개의 과학 출판물에서 1,779개의 관측치로 구성된 데이터 세트를 사용하여 지원 벡터 회귀(SVR) 및 랜덤 포레스트 회귀(RFR) 모델을 학습했다. 각 모델에 대해 우유 섭취 일수(DIM), 건물 섭취량(DMI), 식이 섬유 함량, 조단백질 분율과 같은 요인을 포함하여 다양한 예측 변수 세트를 식별했다. 모델 성능은 결정계수(R^2), 예측 평균 제곱근 오차(RMSEP), 일치 상관 계수(CCC)를 포함한 통계적 지표를 사용하여 평가되었으며, 결과는 기존 NASEM(2021) 모델과 비교되었다. RFR 모델은 RUP에 대해 가장 정확하고 편향되지 않은 예측을 제공했으며(R^2 = 0.60, RMSEP = 0.326 kg/d, CCC = 0.71), SVR 모델은 MicN에 대해 가장 효과적이었다(R^2 = 0.76, RMSEP = 42.4 g/d, CCC = 0.86). 두 모델 모두 기존 방법보다 우수한 성능을 보였으며, 이는 기계 학습이 단백질 이용 예측을 개선하는 데 잠재력이 있음을 보여줬다. 향후 연구에서는 기존 모델과 AI 기반 모델을 통합하여 예측 정확도를 높이는 하이브리드 접근 방식을 모색할 수 있다.

1. 서론

착유우의 영양관리는 유제품 생산의 중요한 측면으로 유량, 동물 건강, 낙농 농가의 수익성에 직접적인 영향을 미친다.[1, 2]. 다양한 영양소 중 사료 단백질은 우유 단백질 합성의 중요한 전구체이며, 필수적인 대사 및 생리 기능을 지원한다[3]. 특히 소장에서 흡수 가능한 질소원으로 정의되는 대사 단백질(MP)는 착유우의 단백질 이용률을 결정하는 핵심 요소이다[4, 5]. MP 공급은 주로 반추취 미분해 단백질(RUP)와 미생물체단백질(MicN)에서 유래한다. 다시 말해서, MP 공급의 정확한 예측은 이러한 구성요소의 정확한 추정에 달려있다. 전통적인 기계론적 모델은 사료 성분의 화학적 조성과 총 소화 영양소를 가반으로 MP 공급을 예측하는데 사용되었다[6, 7]. 이러한 모델은 사전의 정의된 방정식과 정적 매개변수에 의존하기 때문에 생물학적 시스템의 본질인 비선형적이고 다인자적 특성을 설명하는 능력이 제한된다. 머신러닝은 이러한 기존 모델링 접근 방식에 대한 잠재적 대안으로 부상했다. 머신러닝 학습 회귀 알고리즘은 큰 데이터 세트를 분석하고 복잡한 패턴을 식별하는데 효과적으로, MP 공급과 같은 생물학적 결과 예측이 필요한 응용 분야에 적합할 수 있음을 시사합니다[8, 9]. 따라서, 본 연구는 젖소 착유우의 단백질 공급량을 예측하기 위해 핵심요소인 RUP와 MicN을 머신러닝 기반 하위 모델을 개발하고 평가하고자 한다.

2. 재료 및 방법

2.1 데이터베이스 구축 및 데이터셋 추출

본 시험에 활용된 데이터 세트는 436개 논문에서 발췌한 1,779개의 관측치로 구성되었다.

MP는 NASEM[6]에서 정의한 RUP와 MciN의 합으로 정의되었다. RUP(kg/d)는 십이지장 비암모니아, 비미생물질소 흐름에서 추정된 내인성 질소 흐름을 뺀 값으로 계산되었으며, 질소 값에 6.25를 곱하여 조단백질로 변환했다. RUP(kg/d)와 MciN(g/d)의 후보 설명 변수는 다음과 같이 설정되었다.

동물정보(비유일수, 산차, 체중, 건물섭취량, 유량) 사료성분(유기물 함량, 조단백질, 중성세제불용성섬유, 산성

세제불용성섬유, 비전분탄수화물, 조지방, 조회분, 전분). NASEM 사료 라이브러리를 사용하여 각 단백질 분획(kg/d)에 대해 계산된 조단백질 섭취량도 후보변수로 포함했다. 우유 성분, 총 소화관 소화율, 반추위 특성 데이터 그리고 서료 섭취 후 소화 결과데이터는 후보변수에서 제외했다.

2.2 모델 개발 및 평가

RUP와 MicN 예측을 위해 제한된 양의 변수를 데이터 집합에서 제외했다. 정제된 데이터 집합은 훈련(80%)과 테스트(20%) 세트로 나뉘었다. 훈련 집합은 입력 변수와 하이퍼파라미터의 최적 조합을 결정하기 위해 하위 훈련(80%)과 검증(20%) 세트로 다시 나뉘었다. 최적의 입력 변수와 하이퍼파라미터는 검증 집합에서 평가했을 때 가장 높은 정밀도와 정확도를 달성한 모델을 기반으로 선택되었다. 최종 최적 모델은 전체 훈련 집합을 사용하여 재훈련하고 테스트 집합에서 평가했다. 이후, 최종 모델에 선택된 입력 변수만 포함하는 새로운 데이터 집합을 원본 데이터 집합에서 추출했다. 이 데이터 집합을 사용하여 5-겹 교차 검증을 수행하여 모델의 강건성을 평가했다. 동일한 평가 조건에서 NASEM[6] 모델식과도 직접적인 성능 비교를 위해 5-겹 교차 검증 방법을 사용하여 평가했다.

Support vector regression (SVR)와 random forest regression (RFR) 기법 모두 RUP와 MicN을 예측하는 데 사용되었습니다. 지원 벡터 머신 알고리즘을 적용한 SVR은 가우시안 방사형 기저 함수(RBF) 커널을 활용했다. 최적의 감마와 비용은 R의 e1071 패키지에 있는 tune 함수를 사용하여 10겹 교차 검증 프레임워크 내에서 그리드 탐색을 통해 결정되었다. RFR은 트리 수(ntree)와 각 분할에서 고려되는 변수 수(mtry)라는 두 가지 하이퍼파라미터를 포함했다. 이러한 하이퍼파라미터는 그리드 탐색을 통해 최적화되었으며, 모델링은 R의 randomForest 패키지를 사용하여 수행되었습니다.

모델 적절성은 Tedeschi[10]가 설명한 여러 통계적 측정을 기반으로 평가되었습니다. 모든 모델의 정밀도와 정확도는 결정 계수(R2), 예측 오차의 평균 제곱근(RMSEP), 일치 상관 계수(CCC)를 사용하여 평가되었습니다. CCC 값은 Hinkle et al.(1988)에 따라 무시할 수 있음(0.00-0.30), 낮음(0.30-0.50), 보통(0.50-0.70), 높음(0.70-0.90), 매우 높음(0.90-1.00)으로 분류되었습니다. 잔차 분석을 수행하여 평균 및 기울기 편향을 평가했습니다. 모든 통계 분석 및 모델링은 R 소프트웨어(버전 4.3.1)를 사용하여 수행되었습니다. 통계적 유의성은 $p < 0.05$에서 결정되었고 추세는 $0.05 \leq p < 0.1$에서 확인되었습니다.

3. 결과 및 고찰

RUP 예측을 위한 모델 개발의 최적의 입력 변수 조합은 비유일수, 건물섭취량, 사료 건물함량, 조단백질 B와 C 분획 섭취량으로 확인되었다. 모델 적합성은 5겹 교차 검증을 사용하여 평가되었다. RFR과 SVR 모델에 대해 관찰된 RUP 예측은 각각 60%와 53%를 설명했으며, 이는 NASEM 모델식의 설명력의 약 2배였다. 평가된 모델에서 통계적으로 유의미한 편향은 관찰되지 않았다. 가장 좋은 성능을 보인 모델은 RFR 모델로 평균오차(RMSEP)와 일치도(CCC)가 각각 0.33kg/d 및 0.71이었다.

[표 1] 반추위 미분해 단백질(RUP, kg/d) 예측을 위한 개발된 모델과 기존 모델의 적절성 평가

Model	Performance					
	R^2	RMSEP, kg/d	% RMSEP			
			Mean Bias	Slope Bias	Random Bias	CCC
RFR	**0.60**	**0.326**	**4.7**	**5.2***	**90.1**	**0.71**
SVR	0.53	0.349	3.7	1.2	95.1	0.68
NASEM	0.27	0.437	2.9	3.2	93.9	0.45

RFR, random forest regression; SVR, Support vector regression; NASEM, NRC dairy (2021); RMSEP, root mean square error of prediction; CCC, concordance correlation coefficient
*Statistically significant slope bias ($P < 0.05$) was shown in 1 out of 5 folds.

MicN 예측을 위한 모델개발의 최적의 입력 변수 조합은 비유일수, 건물섭취량, 사료 건물함량과 중성세제불용성섬유, 조단백질 A, B와 C 분획 섭취량으로 확인되었다.

[표 2] 미생물체단백질(MciN, kg/d) 예측을 위한 개발된 모델과 기존 모델의 적절성 평가

Model	Performance					
	R^2	RMSEP, kg/d	% RMSEP			
			Mean Bias	Slope Bias	Random Bias	CCC
RFR	0.69	52.0	4.9	14.1*	81.1	0.73
SVR	**0.76**	**42.4**	**1.8**	**5.6**	**92.7**	**0.86**
NASEM	0.04	90.7	5.5**	6.4	88.2	0.13

RFR, random forest regression; SVR, Support vector regression; NASEM, NRC dairy (2021); RMSEP, root mean square error of prediction; CCC, concordance correlation coefficient
*Statistically significant slope bias ($P < 0.05$) was shown in 2 out of 5 folds.
**Statistically significant mean bias ($P < 0.05$) was shown in 1 out of 5 folds.

모델 적합성은 5겹 교차 검증을 사용하여 평가되었으며, SVR 모델이 가장 높은 정밀도와 정확도(R^2 = 0.76, RMSEP = 42.4g/d)를 보였다. RFR 모델은 특정 폴드에서 상당한 기울기 편향을 보였고, NASEM 모델은 정확도와 정밀도가 매우 낮았다.

두 모델 모두 기존 방법보다 우수한 성능을 보였으며, 이는 기계 학습이 단백질 이용 예측을 개선하는 데 잠재력이 있음을 보여줬다. 향후 연구에서는 기존 모델과 AI 기반 모델을 통합하여 예측 정확도를 높이는 하이브리드 접근 방식을 모색할 수 있다.

참고문헌

[1] Colman, D.R.; Beever, D.E.; Jolly, R.W.; Drackley, J.K. Gaining from technology for improved dairy cow nutrition: Economic, environmental, and animal health benefits. Prof. Anim. Sci. 2011, 27, 505-517.

[2] VandeHaar, M.J.; St-Pierre, N. Major advances in nutrition: Relevance to the sustainability of the dairy industry. J. Dairy Sci. 2006, 89, 1280-1291.

[3] Clark, J.H.; Davis, C.L. Future improvement of milk production: Potential for nutritional improvement. J. Anim. Sci. 1983, 57, 750-764.

[4] Allen, M.S. Do more mechanistic models increase accuracy of prediction of metabolisable protein supply in ruminants? Anim. Prod. Sci. 2019, 59, 1991-1998.

[5] Rius, A.G.; McGilliard, M.L.; Umberger, C.A.; Hanigan, M.D. Interactions of energy and predicted metabolizable protein in determining nitrogen efficiency in the lactating dairy cow. J. Dairy Sci. 2010, 93, 2034-2043.

[6] NASEM. Nutrient Requirements of Dairy Cattle; National Academy Press: Washington, DC, USA, 2021.6

[7] NRC. Nutrient Requirements of Dairy Cattle, 8th rev. ed.; National Academy Press: Washington, DC, USA, 2001.

[8] Tedeschi, L.O. ASN-ASAS Symposium: Future of Data Analytics in Nutrition: Mathematical modeling in ruminant nutrition: Approaches and paradigms, extant models, and thoughts for upcoming predictive analytics. J. Anim. Sci. 2019, 97, 1921-1944.

[9] Tedeschi, L.O. ASAS-NANP Symposium: Mathematical modeling in animal nutrition: The progression of data analytics and artificial intelligence in support of sustainable development in animal science. J. Anim. Sci. 2022, 100, skac111.

[10] Tedeschi, L.O. Assessment of the adequacy of mathematical models. Agric. Syst. 2006, 89, 225-247.