

다중 센서 데이터와 다층 퍼셉트론을 활용한 젖소의 유방염 진단 예측

송혜원*, 박기철**, 박재화***

*중앙대학교 사회복지학부

**중앙대학교 일반대학원 컴퓨터공학과

***중앙대학교 소프트웨어대학 소프트웨어학부

hye2016@cau.ac.kr, rlarla423@cau.ac.kr, jaehwa@cau.ac.kr

Prediction of dairy cow mastitis with multi-sensor data using Multi-Layer Perceptron(MLP)

Hye-Won Song*, Gi-Cheol Park**, JaeHwa Park***

*Department of Social Welfare, Chung-Ang University

**Department of Computer Science and Engineering, Chung-Ang University

***College of Software, Chung-Ang University

요 약

낙농업에서 경제적 손실을 불러일으키고 관찰 시간과 비용이 필요한 젖소의 유방염 관리는 중요하다. 그러나 지금까지의 연구는 유방염 진단에 초점을 맞추고 있고, 예측하려는 시도는 전무하다. 유방염에 걸린 개체는 며칠 동안 우유를 생산할 수 없기 때문에 낙농가에 막대한 피해를 준다. 따라서 젖소가 유방염에 걸려 증상이 나타나기 전에 미리 파악해 조치를 할 수 있도록 하는 것이 중요하다. 이에 본 연구는 유방염 예측을 위해 생체 데이터를 포함한 다중 센싱 데이터를 사용해 유방염 예측 모델을 개발하였다. 모델에 사용된 데이터는 충청남도의 농가에 설치된 로봇 착유기로부터 수집하였으며, 일정 기간 동안의 다중 센싱 데이터를 바탕으로 다음 날의 유방염 여부를 예측한다. 많은 양의 비선형 데이터를 효과적으로 처리하기 위해 다층 퍼셉트론을 사용해 모델을 학습하였다. 그 결과, 81.6%의 예측 정확도를 보였으며 교차 검증을 통해 정확도뿐만 아니라 재현율까지 우수함을 확인할 수 있었다.

1. 서론

유방염은 낙농업에서 경제적인 손실과 젖소의 스트레스 상승, 우유의 질 저하 등의 결과를 불러온다. 경제적인 손실을 줄이고 우유의 품질 향상에 도움이 될 수 있도록 유방염을 미리 진단하고 예방해야 한다. 하지만 사람이 직접 젖소를 관찰하고 진단하는 경우 시간이 걸리고, 그랬을 때는 이미 치료하는 시기를 놓쳐 우유 생산에 차질이 생길 수도 있다. 이러한 시간적, 비용적 한계를 극복하기 위해 낙농업에서는 여러 선행연구들을 통해 유방염 진단 및 예측 모델을 발전시켜왔다.

지금까지의 선행연구들은 생체데이터들을 바탕으로 유방염을 판단했다. 유방염과 관련된 지표로서 젖소의 체온과 우유의 전기전도도, 체세포수, 착유량, 유지방, 유단백 등을 사용하였다. 체세포수 측정 비용의 부담을 줄이고자 전기전도도만을 활용한 선행연구도

비슷한 맥락에서 발전했다.[1] 또한 체온과 관련한 선행연구에서는 생체 온도 인식 센서를 개발해 젖소의 제 1 위 속에 넣어 온도 상승에 따른 유방염 진단을 연구했다.[2] 이와 같은 연구들을 통해 각각의 지표들이 유방염 진단에 미치는 영향 및 상관관계를 확인해 볼 수 있다.

생체데이터들을 활용한 연구들은 예측보다는 진단에 초점이 맞춰져 있다. 또한 전기 전도도, 체온 등 단일 지표만을 사용해 유방염을 진단하고자 하는 한계점을 가지고 있다. 이에 본 연구에서는 다양한 센서들로부터 얻은 데이터들을 유방염 예측에 효과적으로 활용할 수 있음을 확인해보고자 한다.

자동화된 착유 시스템으로 사람이 직접 관찰하여 얻을 수 있는 데이터보다 확보할 수 있는 착유 데이터의 양이 증가했다. 많은 양의 데이터를 효과적으로 분석하기 위해 머신러닝 예측 모델을 적용했다. 그

중에서도 다층 퍼셉트론(Multi-Layer Perceptron)을 활용해 비선형적으로 분리되는 데이터의 예측 정확도를 높이고자 했다.

2. 배경 연구

유방염을 진단하기 위해 여러 지표들을 활용한 선행연구를 살펴보면 각각의 지표가 유방염 진단에 가지는 의미를 알 수 있다. 젖소의 체온, 우유의 체세포수, 전기전도도, 유지방, 유단백 등의 지표들을 바탕으로 진행된 연구들을 그 예로 들 수 있다.

무증상의 유방염을 진단하는 연구에서는 전기전도도가 유의미한 지표임을 보인다. 전기전도도만을 활용해 유방염을 진단해 소규모의 농가에서 체세포수 측정 지표를 제외하고 진단할 수 있도록 하였다. 하지만 전기전도도만을 사용했을 때 비교적 낮은 정확도와 민감도를 보여준다는 한계가 있다.[1]

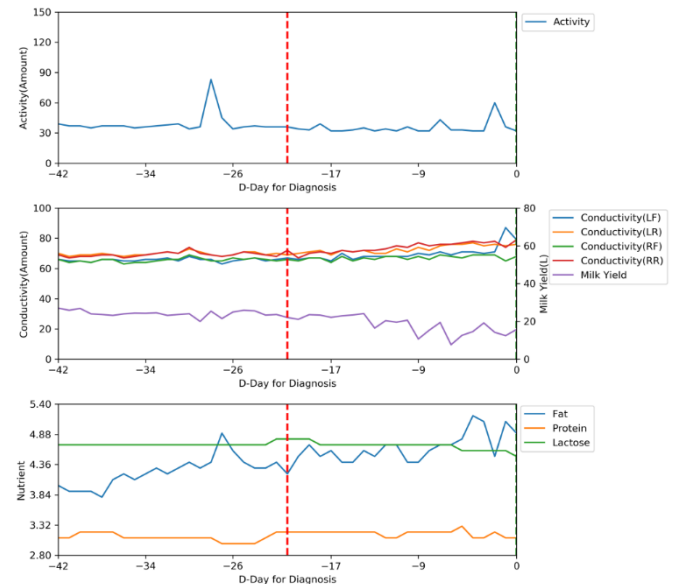
또 다른 연구에서는 젖소의 제 1 위 온도를 측정하여 유방염을 진단했다. 젖소가 생체 온도 인식 센서를 삼켜 제 1 위에 센서가 안착하면 일정한 시간간격으로 온도변화 데이터를 전송하는 방식으로 측정했다. [2] 정상시의 제 1 위의 온도는 평균 39.16°C이고 유방염 진단을 받았을 때와의 편차의 평균은 2.42°C이다. 따라서 온도 상승에 따른 신호를 농장 관리인이 받고 유방염을 빠르게 진단할 수 있다. 하지만 이 때 사용된 생체인식온도 측정기는 수명이 짧아 상용 가능성에 대한 한계가 존재한다.

다층 퍼셉트론을 예측 모델로 사용했을 때 비교적 높은 정확도를 가지는 것을 선행 연구에서 확인했다. [3] 유방염 진단에 있어 중요한 지표인 체세포 수를 예측하고자 한 연구에서 사용된 4 가지 모델 중 하나인 다층 퍼셉트론은 평균보다 낮은 에러율을 가지는 것을 볼 수 있다. 해당 연구에서는 얼마나 많은 데이터가 사용 가능한지가 다층 퍼셉트론 모델이 체세포 수를 얼마나 정확히 예측하는지에 영향을 끼칠 것이라 말한다. 체세포 수 데이터를 포함하고 있는 본 연구에서는 이러한 모델을 적용하는 것이 효과적이다.

3. 데이터 및 연구 방법

본 논문에서는 로봇 착유기를 사용하는 2 개 목장의 착유 데이터를 사용한다. 총 240 마리의 젖소로부터 92,680 개의 데이터를 수집하였다. 착유 데이터는 활동량, 4 개의 유방별 전기 전도도, 유지방, 유단백 등 다중 센서로부터 수집한 데이터를 포함한다. 유방염 최초 진단일을 기준으로 6 주 전까지의 데이터를 합쳐 하나의 분석 단위로 활용한다. 6 주간의 데이터를 3 주씩 나눠 양성과 음성 데이터로 사용한다. 그림 1은 한 분석 개체의 센싱 데이터를 일별 시계열

그래프로 나타낸 것이다. 첫번째 그래프는 자이로스 코프로부터 얻은 젖소의 활동량, 두번째 그래프는 유량과 분방별 전기전도도, 세번째 그래프는 우유의 지방, 단백질, 유당 함량을 나타낸다. 최초 진단일 이전 데이터를 반으로 나눠 붉은색 점선으로 표시하고 양성 및 음성 데이터를 구분한다.



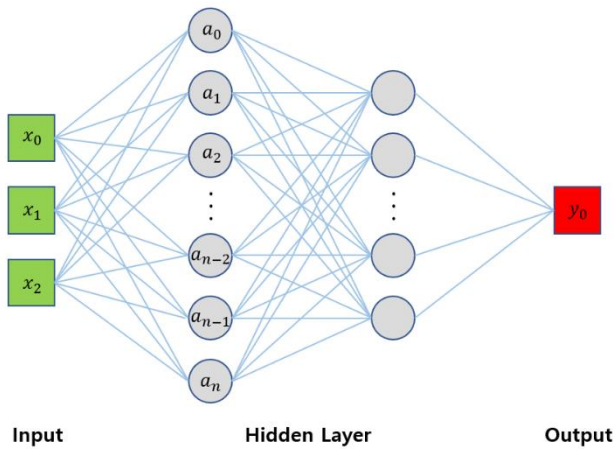
(그림 1) 유방염 진단 데이터

수집된 데이터에는 유방염 진단 이력이 없는 개체, 진단일 전후로 데이터가 비어 있는 개체, 유방염 진단일이 단 하루 밖에 되지 않아 거짓 양성으로 의심되는 개체 등이 포함되어 있다. 효과적인 예측 모델을 위해서는 부정확하거나 불완전한 데이터를 제거하는 등의 적절한 데이터 전처리 과정이 필요하다.

본 연구에서는 수집한 데이터를 3 단계에 걸쳐 정제한다. 전처리 1 단계는 유방염 진단일이 존재하지 않아 분석에 불필요한 데이터들을 제거한다. 2 단계는 각 분석 단위별로 6 주간의 데이터가 모두 채워져 있지 않은 불완전 데이터를 제거한다. 마지막 3 단계는 거짓 양성으로 판단되는 데이터를 제거한다. 거짓 양성으로 판단되는 데이터는 최초 진단일로부터 7 일 이내에 새로운 유방염 진단 데이터가 존재하지 않는 경우이다. 질병 특성상 감염된 젖소의 경우 진단 데이터가 연속적으로 나타나야 하기 때문이다. 필터링 된 데이터는 각각의 특성들 간의 크기 차이가 있어 Z-정규화를 사용하여 정규화 과정을 거친다.

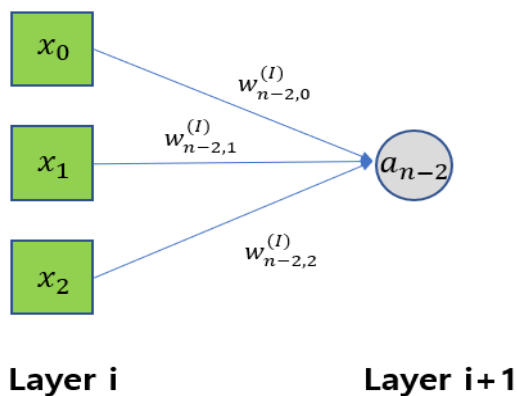
유방염 예측을 위한 모델로는 다층 퍼셉트론(Multi-Layer Perceptron)을 활용한다. 다층 퍼셉트론 모델은 입력층과 출력층 사이에 하나 이상의 은닉층(Hidden Layer)을 두어 비선형적으로 분리되는 데이터에 대해서도 학습할 수 있도록 고안되었다. 본 연구에서는 분석 단위별 3 주간의 착유 데이터의 모든 특성을 입

력 데이터로, 다음날의 유방염 진단 여부를 출력 값으로 설정한다. 다층 퍼셉트론 모델은 퍼셉트론으로 이루어진 층 여러 개를 순차적으로 연결하여 구성한다.



(그림 2) 다층 퍼셉트론 모델 구조

위의 그림은 다층 퍼셉트론 구조의 한 예를 보인 것이다. 입력층에서 전달되는 값은 은닉층의 모든 노드로 전달되며 은닉층의 모든 노드의 출력값 역시 출력층의 모든 노드로 전달된다. 이러한 방식으로 값이 전달되는 것을 순전파(Feed-Forward) 방식이라고 한다. 한 층에서 다음 층으로 값을 전달할 때 각 노드의 값과 가중치를 벡터 연산 후 다음 층으로 전달한다. 이때 벡터 연산을 통해 계산된 값을 순입력 함수값이라고 하고, 이를 활성화함수를 통해 출력값을 판단하게 된다.



(그림 1) 순입력 함수값 전파 과정

아래 그림은 위의 그림의 입력층에서 a_{n-2} 노드로 값을 전달하는 과정을 확대한 것이다. I 층의 K 번째 노드와 $I+1$ 층의 j 번째 노드를 연결하는 가중치를 $w_{j,k}^{(I)}$ 라고 가정하면 a_{n-2} 노드의 순입력 함수 값은 다음과 같다.

$$a_{n-2} = w_{n-2,0}^{(I)} * x_0 + w_{n-2,1}^{(I)} * x_1 + w_{n-2,2}^{(I)} * x_2$$

본 연구에서는 활성화함수로 ReLU 함수를 사용한다.

ReLU 함수는 입력으로 0 보다 작은 값이 들어오면 0을 반환하고, 0 보다 큰 값이 나온 경우 그 값을 그대로 반환하는 함수이다.

다층 퍼셉트론 모델의 경우 은닉층의 개수, 학습률 같은 초 매개변수(Hyper Parameter)의 값에 따라 모델의 성능이 크게 달라진다. 따라서 최적의 결과를 얻기 위해서는 적절한 탐색 방법을 통해 초 매개변수 값을 조정해야 한다.[4] 필터링 알고리즘을 거친 376개의 분석 개체를 6:2:2 비율로 나눠 각각 훈련 데이터, 검증 데이터, 테스트 데이터로 설정한다. 훈련 데이터를 통해 모델을 학습하고 검증 데이터를 예측함으로써 모델의 정확도와 함께 과적합정도를 확인할 수 있다. 검증 데이터를 활용해 모델의 초 매개변수를 조정한다. 훈련 데이터와 검증 데이터를 통해 학습된 모델로 테스트 데이터를 예측하여 정확도를 측정한다.

모델의 성능을 검증하기 위해 교차 검증(Cross Validation)을 진행한다. 교차 검증에는 Stratified K-Fold 검증을 사용하였다. Stratified K-Fold는 기존의 K-Fold 검증과 다르게 각 Fold 안의 데이터 Label 분포가 전체 데이터 Label 분포를 따른다는 장점이 있다.[6]

본 연구에서는 다중 센싱 데이터에 체세포 수를 포함하지 않는다. 체세포 수는 유방염 진단에 있어 매우 중요한 지표지만 다른 데이터들에 비해 측정 시간이 오래 걸린다는 단점이 있기 때문에 빠른 예측에는 사용할 수 없다.[5] 그렇기 때문에 체세포 수를 입력 데이터에 포함한 모델과 그렇지 않은 모델의 정확도를 비교하여 체세포 수를 사용하지 않고도 얼마나 정확하게 유방염을 예측할 수 있는지 확인하고자 한다.

다층 퍼셉트론의 우수함을 평가하기 위해 다른 예측 모델과의 비교를 진행한다. 비교 대상이 되는 모델은 랜덤 포레스트, 그래디언트 부스팅 모델(LightGBM), 순환 신경망(Recurrent Neural Network)이다. 각각의 모델에 체세포 수를 포함한 데이터와 그렇지 않은 데이터를 차례로 학습하고 유방염을 예측한 정확도를 비교한다.

4. 결과

<표 1> 테스트 데이터 예측 결과

	Precision	Recall	F1-Score
Non-Mastitis	0.821	0.788	0.801
Mastitis	0.8	0.83	0.813

학습된 모델은 테스트 데이터를 81.6%의 정확도로 예측했다. 정확도는 전체 데이터에서 올바르게 예측된 데이터의 비율을 나타낸 값이다. 하지만 정확도만

으로는 모델의 성능을 제대로 판단하기 힘들기 때문에 본 연구에서는 모델의 성능을 판단하기 위한 지표로 재현율(Recall)과 정밀도를 추가로 제시한다. 재현율은 실제로 참인 데이터에 대해 모델이 참이라고 판단한 비율이다. 이를 통해 전체 정확도와는 무관하게 제대로 분류해줄 것들에 대해서 모델이 얼마나 잘 분류하는지에 대한 척도가 될 수 있다. 정밀도의 경우 모델이 참이라고 분류한 데이터에 대해 실제 참인 데이터의 비율을 나타낸다. 정밀도가 높다는 것은 얼마나 진짜 사용자가 원하는 결과를 제대로 분류했는지에 대한 척도로 활용할 수 있다.

표 2는 모델이 테스트 데이터를 예측한 결과를 나타낸 것이다. 정확도뿐만 아니라 재현율과 정밀도 또한 약 0.8로 나온 것으로 보아 모델의 성능이 우수함을 확인할 수 있다.

<표 2> Stratified K-Fold 교차 검증 결과

	Precision	Recall	F1-Score
Non-Mastitis	0.82	0.79	0.81
Mastitis	0.79	0.81	0.8

<표 3> 모델별 예측 정확도

	Random-Forest	Lightgbm	MLP	RNN
SCC Excluded	0.772	0.788	0.816	0.792
SCC Included	0.793	0.788	0.84	0.803

표 3은 다층 퍼셉트론 모델을 포함한 4가지 예측 모델의 정확도를 비교한 것이다. 체세포 수를 포함하거나 그렇지 않은 모델 모두 다층 퍼셉트론 모델의 정확도가 가장 높은 것을 확인할 수 있다. 또한 체세포 수를 포함하지 않은 모델의 정확도가 포함된 모델과 비교해서 2% 이내의 차이를 보이는 것으로 보아, 체세포 수를 포함하지 않고도 유방염을 높은 수준의 정확도로 예측할 수 있음을 보인다.

5. 결론

유방염은 낙농업이 해결해야 할 과제 중 하나로 이를 극복하기 위해 많은 연구들이 진행되어 왔다. 착유 데이터를 사용해 유방염 진단에 집중한 연구들은 유방염과 착유 데이터 간의 관계를 설명해준다. 착유 데이터 중 단일 지표만 사용하는 연구들이 유방염과의 상관관계를 확인했으나 예측의 정확도를 확인하기에는 부족한 경우가 있었다. 따라서 본 연구는 다중 센서 데이터를 사용해 다양한 착유 데이터를 바탕으로 유방염을 예측하고자 했다.

또한 질병 진단에 단서가 되는 착유 데이터들은 진단 시기를 앞당기도록 하는 연구에 활용되었다. 하지만 유방염을 효과적으로 관리하기 위해서는 이러한

데이터들을 바탕으로 시간에 따른 발병 여부를 예측해내는 것이 필요하다.

이러한 배경으로 본 연구는 다중 센서 데이터를 활용한 다층 퍼셉트론 모델을 사용하여 유방염을 정확하게 예측하고자 했다. 약 9 만개의 착유 데이터를 3 단계의 데이터 전처리 과정을 거친 후 모델에 학습시켜 유의미한 결과를 도출하였다.

유방염을 예측하는데 필요한 3 주간의 다중 센서 데이터들은 모델의 훌륭한 데이터 셋으로 사용되었다. 다층 퍼셉트론 모델의 예측 정확도는 81.6%로 선행 연구들과 비교해서 예측 수준이 높다고 볼 수 있다. 또한 교차검증 결과에서 높은 정확도와 재현율을 보여줌으로써 모델의 우수성을 입증하였다.

감사의 글

본 결과물은 농림축산식품부의 재원으로 첨단생산기술개발사업의 지원을 받아 연구되었음(318005-4).

참고문헌

- [1] Cheng-Chang Lien et al, "Online detection of dairy cow subclinical mastitis using electrical conductivity indices of milk", Engineering in Agriculture, Environment and Food, Volume 9, Issue 3, Pages 201-207, July 2016
- [2] Heejin Kim et al, "Real-time temperature monitoring for the early detection of mastitis in dairy cattle: Methods and case researches", Computers and Electronics in Agriculture, Volume 162, Pages 119-125, July 2019
- [3] Dorota Anglart et al, "Comparison of methods for predicting cow composite somatic cell counts", Journal of Dairy Science, Volume 103, Issue 9, Pages 8433-8442, September 2020
- [4] Leo Weissbart, "Performance Analysis of Multilayer Perceptron in Profiling Side-Channel Analysis", Cryptology ePrint Archive, Report 2019/1476, 2019
- [5] Jose Carlos de Figueiredo Pantoja, Pamela Ruegg, "Understanding and using somatic cell counts to improve milk quality", Irish Journal of Agricultural and Food Research Volume 52, Issue 21, Pages 101-117, December 2013
- [6] Ron Kohavi, "A Study of Cross-Validation and BootStrap for Accuracy Estimation and Model Selection", Appears in the International Joint Conference on Artificial Intelligence, Volume 2, Pages 1137-1143, August 1995