기계학습을 이용한 가축 질병 조기 발견 방안

이웅섭*

*국립경상대학교

Fast Detection of Disease in Livestock based on Machine Learning

Woongsup Lee* · Sewoon Hwang* · Jonghyun Kim*

*Gyeongsang National University

E-mail : wslee@gnu.ac.kr

요 약

최근 기계학습에 기반을 둔 빅데이터 분석이 큰 관심을 받으면서 다양한 학문 분야에 기계 학습 방안들이 접목되고 있다. 그 대표적인 분야 중 하나로 농축산 분야를 들 수 있고 실제 다양한 기계 학습 방안들이 농축산분야에 적용되고 있다. 하지만 농축산에서 활용되는 기계학습의 경우 대부분 농업분야의 기후예측 및 축산분야의 유전자 분석 쪽으로 연구가 집중되어있고, 가축의 생체 데이터를 활용한 기계학습 방안은 많은 연구가 이루어지지 않았다. 본 연구에서는 가축의 실시간 생체 데이터를 이용하여 문제가 발생한 개체를 조기에 발견하는 방안을 제안하였다. 제안 방안에서는 기댓값 최대화 알고리즘을 이용하여 단일 가축 개체들의 실시간 생체 데이터를 2개의 클러스터로 나누고 이 두 클러스터 사이즈의 변화를 통해서 이상 개체를 조기에 판단한다. 특히 단일 개체의 문제와 전염성 질병 여부를 나누어 판단하므로 구제역과 같은 전염성 질병의 경우 빠른 대응을 가능케 하여 국가적 손실을 줄일 수 있게 한다. 더불어 제안 방안은 측정 생체 데이터에 대한 통계적 정보 없이도 적응적으로 클러스터를 형성할 수 있으므로 축사 외부의 환경 요소에 의해서 생체 데이터의 통계적 특성이 변화는 상황에서도 적응적으로 동작할 수 있다.

ABSTRACT

Recently, big data analysis which is based on machine learning has been gained a lot of attentions in various fields. Especially, agriculture is considered as one promising field that machine learning algorithm can be efficiently utilized and accordingly, lots of works have been done so far. However, most of the researches are focusing on the forecast of weather or analysis of genome, and machine learning algorithm for livestock management, especially which uses individual data of livestocks, e.g., temperature and movement, are not properly investigated yet. In this work, we propose fast abnormal livestock detection algorithm based on machine learning, more specifically expectation maximization, such that livestock which has problem can be efficiently and promptly found. In our proposed scheme, livestocks are divided into two clusters using expectation maximization based on their bionic data and the abnormal livestock can be detected by comparing the size of two clusters. Especially, we divide the case in which single livestock has problem and the case in which livestocks have epidemic such that fast response is enabled when epidemic case. Moreover, our algorithm does not need statistical information.

키워드

기계학습, 축산, 빅데이터, 질병 조기 발견, 전염성 질병

Ⅰ. 서 론

현재 다양한 학문 분야에 기계학습에 기반을 둔 빅데이터 분석이 큰 관심을 받고 있다. 기계학습이 적용되는 대표적인 분야 중 하나로 농축산 분야를 들 수 있는데 실제 다양한 기계 학습 방안들이 농축산분야에 적용되고 있다 [1-3]. 농축산 분야에서 많은 기계학습 방안들이 연구되고 있지만 대부분 농업분야의 기후예측 및 축산분야의 유전자 분석 쪽으로 연구가 집중되어있고, 가축의 생체 데이터를 활용한 기계학습 방안은 많은 연구가 이루어지지 않았다.

축산분야에서는 여러 환경 데이터들을 종합하여 각 개체들의 성장을 예측하는 기계학습 방안들이 연구되어 왔다. [1]의 연구에서는 기계학습을 이용하여 젖소의 reproduction rate를 분석하였다. 뉴질랜드의 와이키토 대학 연구팀은 젖소의 reproduction rate를 분석하였고 특히 기계 학습 분석 툴인 WEKA를 개발하였다 [2]. [3]에서는 머신 러닝을 이용하여 모돈의 생산성을 예측하였다. 축산분야에서 여러 기계학습에 기반을 둔 방안들이 연구되었지만 실시간 생체 데이터를 활용한 연구는 이루어지지 않았다.

본 연구에서는 가축의 실시간 생체 데이터를 기반으로 기계 학습을 활용하여 문제가 발생한 개체를 실시간으로 발견하는 방안을 제안하였다. 제안 방안에서는 기댓값 최대화(Expectation Maximization)을 이용하여 개체들의 실시간 생체 데이터를 2개의 클러스터로 나누고 이 두 클러스터 사이즈의 변화를 통해서 이상 개체를 조기에 판단한다. 실시간으로 개체의 생체 정보를 모니터링 하므로 이상 개체를 실시간으로 찾아낼 수 있다. 더불어 제안 방안은 측정 생체 데이터에 대한 통계적 정보가 없는 환경에서 적응적으로 클러스터를 구성하므로 어떠한 환경에서도 적응적으로 동작할 수 있다.

제안 방안에서는 개체의 문제를 단일 개체 이상과 전염성 질병 발병으로 나누어 판단한다. 따라서 구제역과 같은 전염성 질병의 발병을 조기에 진단할 수 있고 전염성 질병에 대한 빠른 대응을 가능케 하여 국가적 손실을 줄일 수 있게 한다.

Ⅱ. 본 론

본 연구에서는 가축의 생체 데이터가 실시간으로 수집되는 환경을 고려하였다. 또한 다른 종류의 생체 데이터들(예를 들어서 각 개체들의 체온, 운동량, 체중 등)이 수집될 수 있다고 가정하였고 수집되는 생체 데이터 들 간 연관성 (Co relation)이 존재할 수 있다고 가정하였다. 수집되는 생체 데이터의 통계적 특성을 모르기 때문에 생체 데이터들의 값이 가우시안 분포를 따른다고 가정하였다. 또한 가축들은 정상 상태와 이상 상태를 가질 수 있고 어느 상태에 있느냐에 따라서 다른 가우시안 분포 파라미터를 가진다고 가정하였다.

본 연구에서는 실시간 생체 데이터의 클러스터링을 위해서 기댓값 최대화 알고리즘(Expectation Maximization)을 사용하였다. 기댓값 최대화에서는 다른 특성을 지닌 두 가우시안 분포가 있을 때 이를 나눠준다. 분포의 통계적 특성이 없을 때 Maximum Likelihood를 최대화 하는 방향으로 가우시안 분포의 파라미터를 예측해서 업데이트 한다.

기댓값 최대화 알고리즘에서는 Bayes Theorem을 이용한 conditional probability를 예측하는 E-step과 이를 이용하여 두 가우시안 분포의 파라미터를 예측하는 M-step으로 이루어진다. 알고리즘에서는 E-step과 M-step을 번갈아 가면서 업데이트 하면서 최적의 값을 찾는다. 그림 1에서 볼 수 있듯이 가우시안 분포의 초기 값이 잘못된 값으로 설정되어 있다면 원본 클러스터와 다른 클러스터로 나눌 수 있지만 iterative한 업데이트를 통해서 최종적으로 올바른 cluster로 나누는 것을 확인할 수 있다.

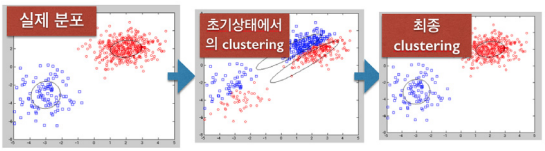

그림 1. 제안 기댓값 최대화를 통한 클러스터링 예제

기존의 기댓값 최대화 알고리즘을 이용하여 정상 개체와 이상 개체를 올바르게 나눌 수 있는 것을 확인할 수 있었다. 하지만 기존의 기댓값 최대화 알고리즘을 사용하면 모든 개체들이 정상 상태에 있는 경우에도 두 개의 클러스터로 나누게 되므로 클러스터로 나눌 수 있다는 것만으로는 특정 개체들이 이상 상태에 있다고 결론 내리기 어렵다. 또한 제안한 기댓값 최대화 방안을 활용하였을 때 단일 개체에 문제가 발생한 것과 전염성 개체가 있어서 시간이 변함에 따라 이상 상태의 수가 증가하는 경우를 구별할 수 없다.

이를 해결하기 위해서 제안 방안에서는 클러스터의 사이즈 변화를 관찰하여 이상개체를 파악하는 방안을 제안하였다. 우선 모든 개체들이 정상 상태에 있을 경우에는 나눠지는 두 클러스터의 사이즈가 비슷할 것이다. 다음으로 단일 개체에 문제가 발생 했을 경우에는 나눠지는 클러스터 중 한 클러스터의 크기가 매우 작을 것이다. 마지막으로 전염성 질병이 발생한 경우에는 클러스터의 사이즈가 시간에 따라서 변화할 것이다. 이를 기반으로 다음과 같은 실시간 생체 데이터 기반 이상개체 파악 알고리즘을 제안하였다.

Algorithm 1 Algorithm to detect abnormal activity

1: Set $\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2$ to initial values
2: Initial $F_1 = F_2 = 0$
3: **while** $F_1 \leq \theta_1$ AND $F_2 \leq \theta_1$ **do**
4: Monitor biological data X
5: Store $\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2$ to $\mu_1^{prev}, \mu_2^{prev}, \sigma_1^{prev}, \sigma_2^{prev}, \tau_1^{prev}, \tau_2^{prev}$
6: Update $\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2$ using EM algorithm based on X
7: **if** $\eta_d \leq |\tau_1 - \tau_1^{prev}| \leq \eta_u$ **then** $F_1 \leftarrow F_1 + 1$
8: **if** $\min(\tau_1, \tau_2) \leq \eta_u$ **then** $F_2 \leftarrow F_2 + 1$

그림 2. 제안 알고리즘

제안 알고리즘에서 μ, σ, τ는 각 클러스터의 평균, 분산 및 클러스터에 속할 확률이고 Θ는 카운터 값의 threshold를 의미한다. 또한 X는 모니터링 되고 있는 생체 데이터의 크기를 나타내고 F는 특정 이벤트가 발생한 횟수를 저장하는 카운터 값이다.

제안 알고리즘에서는 각 클러스터의 크기를 나타내는 τ를 모니터링 하면서 τ의 변화량이 일정 이상으로 계속 클 경우에는 전염성 질병이 있다고 판단하고 τ의 절대 값이 일정시간이상 정해진 thredhold 이하일 경우에는 단일 개체에 문제가 있다고 판단한다. 제안 방안의 사용을 통해서 정상상태의 개체들, 단일 개체 문제, 전염성 질병의 발병 등을 실시간으로 파악할 수 있다. 제안 알고리즘에서는 카운터 F의 값이 계속 증가하므로 결국 알고리즘이 halting하게 될 것이다.

제안 방안의 성능을 확인하기 위해서 컴퓨터 기반 시뮬레이션을 수행하였다. 시뮬레이션에서는 단일 개체에 문제가 있을 경우와 전염성 질병 문제가 있는 상황을 나누어서 고려하였다. 또한 4종류의 생체 데이터를 사용한다고 가정하였고, 정상개체의 mean = [1;2;3;4], covariance = [1 0.2 0.2 0.2;0.2 1 0.2 0.2;0.2 0.2 1 0.2;0.2 0.2 0.2 1]라고 가정하였고 비정상개체의 mean = [-3;-2;-1;0], covariance = [1 0.2 0.2 0.2;0.2 1 0.2 0.2;0.2 0.2 1 0.2;0.2 0.2 0.2 1]라고 가정하였다. 즉 각 생체 데이터 사이의 연관성을 고려하였다.

우선 단일 개체 문제가 발생한 환경에서의 성능 분석을 하였다. 본 성능 분석에서는 100개의 개체 중 2개의 개체에 문제가 발생한 환경을 고려하였다. 이러한 환경에서 제안 알고리즘이 이상 개체를 판단할 때 까지 걸리는 시간의 histogram을 그림 3에 나타냈다. 결과에서 볼 수 있듯이 평균 33번의 iteration 이후 이상 개체를 발견할 수 있는 것을 확인할 수 있고 전염성 질병으로 잘못 예측하는 경우는 없는 것을 확인할 수 있다. 즉 제안 방안을 통해 정확하고 신속하게 단일 개체의 문제를 발견할 수 있는 것을 확인할 수 있다.

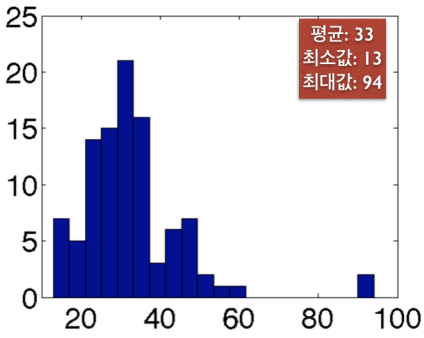

그림 3. 단일 개체에 문제가 있을 경우 알고리즘 stop 까지 걸리는 시간의 histogram

다음으로 전염성 질병이 발생한 경우에서의 성능 분석을 하였다. 본 성능 분석에서는 전염성 질병이 SIR (Susceptible-Infectious-Recover) 모델을 따른다고 가정하였다. 또한 초기 Infection 개체들이 3개체라고 가정하였다. 이 때 전염성 질병을 발견할 때 까지 걸린 시간의 histogram을 그림 4에 표시하였다.

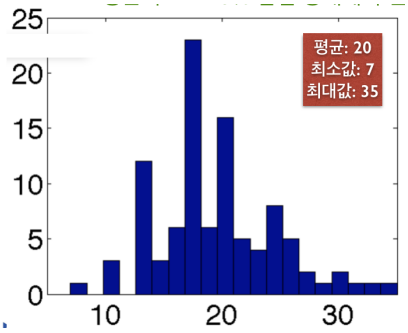

그림 4. 전염성 질병이 있을 경우 알고리즘 stop 까지 걸리는 시간의 histogram

그림 4의 결과에서 볼 수 있듯이 평균 20번의 iteration이후에 전염성 질병을 파악할 수 있는 것을 확인할 수 있다. 또한 전염성 질병을 단일 개체의 문제로 오인할 확률이 0%임을 확인할 수 있었다. 이를 통해서 제안 방안이 전염성 질병 여부를 빠르고 정확하게 파악할 수 있다고 결론지을 수 있다. 또한 평균적으로 전체 개체 중 25%의 개체에 전염성 질병이 퍼졌을 경우에 전염성 질병을 파악할 수 있음을 확인할 수 있다.

III. 결 론

본 연구에서는 기계학습의 기댓값 최대화 알고

리즘을 이용한 가축 생체 데이터의 실시간 분석 방안에 대해서 제안하였다. 제안 방안에서는 가축의 생체 데이터를 모니터링하여 2개의 클러스터로 나누고 각 클러스터의 크기 변화를 관찰하여 가축의 이상을 판단하였다. 제안 방안의 사용을 통해 단일 개체에 이상이 발생한 경우와 전염성 질병이 발생한 경우를 분리하여 예측할 수 있고 전염성 질병의 경우 조기 대응을 통해서 국가적 손실 감소를 가능케 한다.

감사의 글

본 성과물은(논문, 산업재산권, 품종보호권 등)은 농촌진흥청 연구사업(세부과제명: 가축 관리 및 돈사환경 관측 빅데이터 활용에 관한 연구, 세부과제번호: PJ010541022015)의 지원에 의해 이루어진 것임

참고문헌

[1] DZ. Caraviello, KA Weigel, M. Craven, D. Gianola, NB. Cook, KV. Nordlund, PM. Fricke, MC. Wiltbank, “Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms,” *J Dairy Sci*. 89(12) 2006.

[2] R. Scott Mitchell, Robert A. Sherlock, Lloyd A. Smith, “An investigation into the use of machine learning for determining oestrus in cows,” *Computers and Electronics in Agriculture*, 15(3), 1996.

[3] 이민수, 최영찬, “머신러닝을 활용한 모돈의 생산성 예측모델,” *한국농촌지도학회*, 16(4), 2009

[4] P. Flach, “Machine learning: the art and science of algorithms that make sense of data," *Cambridge University Press*, 2012.