머신러닝을 활용한 모돈의 생산성 예측모델

Forecasting Sow's Productivity using the Machine Learning Models

이민수* · 최영찬**
Min Soo Lee · Young Chan Choe

Abstract

The Machine Learning has been identified as a promising approach to knowledge-based system development. This study aims to examine the ability of machine learning techniques for farmer's decision making and to develop the reference model for using pig farm data. We compared five machine learning techniques: logistic regression, decision tree, artificial neural network, k-nearest neighbor, and ensemble. All models are well performed to predict the sow's productivity in all parity, showing over 87.6% predictability. The model predictability of total litter size are highest at 91.3% in third parity and decreasing as parity increases. The ensemble is well performed to predict the sow's productivity. The neural network and logistic regression is excellent classifier for all parity. The decision tree and the k-nearest neighbor was not good classifier for all parity. Performance of models varies over models used, showing up to 104% difference in lift values. Artificial Neural network and ensemble models have resulted in highest lift values implying best performance among models.

주요어(Key words) : 머신러닝(Machine Learning), 모돈도태결정(Sow culling decision), 의사결정나무(Decision Tree), 인공신경망(Neural Network)

* 전북발전연구원 부연구위원. e-mail: minsooo.lee@gmail.com
** 서울대학교 지역정보전공 교수. e-mail: aggi@snu.ac.kr

1. 서 론

지난 10년간 농업경영 효율성 향상을 위해 농가와 농업지원기관들은 다양한 정보시스템을 구축하여 왔다. 이에 따라 다량의 데이터가 정보시스템을 통해 DB형태로 수집되고 분석되고 있다. Gartner Group(1998)은 데이터베이스 기반에서 필요한 정보의 80%는 단순 Query를 통해 추출 가능하지만, 더 큰 가치를 지닌 20%의 정보추출을 위해서는 데이터에 기반한 머신러닝 기법이 필요하다고 지적하고 있다. 농업분야에서도 머신러닝 기법을 예측(prediction), 분류(classification), 모델링문제(modeling problem)에 적용하는 사례가 증가하고 있다(McQueen et al., 1995; Jayas et al., 2000; Schultz et., 2000; Pietersma, 2003). 1980년에 식물의 질병을 판별하기 위해 머신러닝 기법을 적용하였으며, 기존 전문가에 의한 판단이나 전문가시스템에 비해 우수하다는 것이 밝혀졌다(McQueen et al., 1995).

이에 따라 농업경영 컨설팅이나 의사결정지원을 위한 생산량예측, 축산분야의 도태결정, 생산이상 개체 발견 등에 머신러닝 기법이 적용되었다. 최근에는 전자상거래의 활성화로 인해 경영과 판매, 고객관리 등에서 의사결정을 지원하기 위해 머신러닝 기법이 사용되고 있다. 머신러닝 방법론을 농업경영분야에 활용하여 경영 및 마케팅의사결정 지원을 위한 기법과 솔루션을 분석, 개발하여 현장 농기업의 효율적인 경영관리와 고객 및 판매관리 능력을 향상시키는 것은 고급인력이 부족한 현장사정에서 필수적이다. 지난 20여년간 빠르게 생산의 규모가 커진 양돈산업의 경우 47.3%의 농가들이 전산관리를(양돈협회, 2005) 통해 생산 및 경영에 대한 자료를 축적하고 있고 이를 활용한 효율적인 의사결정의 요구는 더욱 높아지고 있어, 머신러닝기법을 적용한 의사결정지원시스템의 도움이 절실하다.

본 연구의 목표는 머신러닝 방법론을 국내에서 가장 대표적으로 사용되는 양돈관리프로그램인 Pigplan을 통해 축적된 자료에 적용하여 농업경영의사결정지원에 활용할 수 있는 방안을 제시하는 데 있다. 이를 위해 우선 Pigplan 사용농가와 관련종사자를 대상으로 요구사항을 수집하였다. 이 후 농업분야의 데이터에 대한 머신러닝 기법 적용에 대한 현실성을 평가하기 위하여 실제 Pigplan 데이터에 머신러닝 방법론을 적용하였다. 이 결과를 토대로 향후 농업분야에 머신러닝기법 적용을 위한 참조모형을 도출하였다.

Ⅱ장에서는 머신러닝에 대한 개념, 대표적 머신러닝 기법, 머신러닝 기법의 특징, 머신러닝 방법론, 농업경영에서의 머신러닝 적용 사례를 논의 하였다. Ⅲ장에서는 본 연구과제를 수행하기위한 연구방법에 관해 논의하였다. 본 연구에서 사용한 사용자 요구분석방법과 머신러닝 기법 적용 방법이 나타나 있다. Ⅳ장에서는 연구분석결과 머신러닝 방법론에 따른 임신사고 예측결과를 제시하였다. Ⅴ장에서는 연구결과를 요약하고 연구의 활용방안과 향후 연구과제를 살펴보았다.

2. 머신러닝의 농업경영 의사결정 적용

머신러닝은 데이터로부터 지식획득의 과정을 컴퓨터를 통하여 자동화하는 방법에 관한 연구이며(Langley and Simon, 1995), 인공지능(artificial intelligence) 분야에서 가장 핵심적인 역할을 하고 있다(Mitchell, 1997). 오늘날 머신러닝은 데이터분석을 위한 필수적인 툴이다. 디지털 혁명을 통해 정보시스템의 데이터는 수집되고, 모니터링되고, 공유되고 있으며, 머신러닝 기법은 수집된 대량의 데이터를 분석하는 데 가장 적합한 방법론으로 (Kononenko, 2001), 세 개의 주요 머

신러닝 분파로 나뉘어진다.

첫 번째는 Hunt et al.(1966)에 의해 제시된 기호적 학습(symbolic learning)이며, 주요 알고리즘은 의사결정나무(induction of decision tree), 의사결정규칙(decision rules), 논리프로그램(induction of logic programs) 등이 있다. 두 번째는 Nilsson(1965)에 의해 제시된 통계적 방법론(statistical methods)이며, 통계 혹은 패턴인식 기법으로 불리며, 주요 알고리즘으로 k-NN(k-nearest neighbors), 판별분석(discriminant analysis), 베이지안 분류기(Bayesian classifiers) 등이 있다. 마지막 세 번째는 Hunt et al.(1962)에 제시된 인공신경망(neural networks) 방법론이며, 알고리즘으로는 역전파학습(backpropagation learning), Kohonen SOM(Kohonen's self-organizing network), Hofield 연상메모리(Hofield's associative memory) 등이 있다.

머신러닝 기법은 그동안 비즈니스와 공학 영역에서 광범위하게 활용되어 왔으나 최근 농업분야에서도 머신러닝과 데이터마이닝을 이용한 연구들이 늘어나고 있다(이용범, 2004). 농업생명공학의 성장과 더불어 바이오 인포메틱스(Bio-informatics) 분야에서 많이 활용되고 있으며, 기상정보분석, 농산물 품질판정, 정밀농업 등에서도 머신러닝기법이 활용되고 있다. 그러나 농업생산 및 경영 의사결정에 머신러닝을 활용한 연구는 많지 않다. 축산분야의 생산데이터와 경영데이터에 머신러닝기법을 적용한 연구로는 McQueen et al.(1995), Scott Mitchell et al.(1996), Pietersma et al.(2003), Kirchner et al.(2004a,b), Kirchner et al.(2006) 등의 연구 등이 있다.

McQueen et al.(1995)는 젖소의 도태에 대한 의사결정에 의사결정나무기법을 이용하여 건강상태와 난산기록연령보다는, 송아지 생산능력(breeding index), 우유생산능력(production index)이 중요한 변인임을 밝혔으며, 도태의사결정에 있어서 머신러닝의 유용성과 향후 연구

과제를 제시하였다. Scott Mitchell et al.(1996)은 의사결정나무기법인 C4.5(Quinlan, 1992)와 First Order Inductive Learner(FOIL)(Quinlan, 1990)을 적용하여 젖소의 발정을 예측하였다. 연구의 결과 개별 개체의 평균 산유량 보다는 산유량의 변화량 패턴이 발정 예측의 주요변인으로 파악되었으며, C4.5가 FOIL보다 우수한 것으로 나타났다.

Pietersma et al.(2003)은 의사결정나무기법을 사용하여 각 산차별(1, 2~3, 4산차이상) 젖소의 비유곡선(lactation curve), 초기 우유생산, 최대생산시기 등을 예측하는 모형을 설정하였다. 이들은 체세포수(Somatic Cell Count), 단백질지방비율(Protein to Fat Ratio), CAR 유산지수(CAR Code abortion) 등을 주요 변인으로 사용하였으며, 의사결정나무에서 설정조건을 변경하거나 k-fold cross validation 기법 등의 사용이 정확도를 향상시키는 것으로 파악하였다.

양돈분야의 생산 및 경영의사 결정에 머신러닝을 사용한 경우는 Kirchner et al.(2004a, b, 2006)의 연구가 있다. 이들의 처음 모형에서는 모돈 도태에 대한 의사결정을 지원하기 위해서 의사결정나무 알고리즘의 하나인 C4.5를 이용하였으며, 2개 농가의 데이터를 이용해 농가에 따른 도태결정의 차이를 규명하기 위해 산차, 평균분만두수, 평균분만사고두수, 평균이유두수, 교배회수, 이유사고두수, 이유후 발정기간, 재발률 등을 사용하였으며 실측지의 적용결과 정확도는 85% 정도로 나타났다. 또한, 질병이나 사고로 인해 생성된 데이터들은 제외한 후, 의사결정나무를 적용한 결과는 92~95%의 정확도를 나타냈다. 또 두 농가의 도태전략도 차이가 있는 것으로 나타났다. 이런 결과를 토대로 연구자들은 농가에 따른 의사결정 구조의 차이를 밝혀내고, 의사결정나무를 통한 의사결정 지원이 가능함을 보여주었다.

Kirchner et al.(2006)는 후속연구에서 동일한 연구모형을 사용하였으나 모돈 도태 결정에서 C4.5 알고리즘의 우수성을 파악하기위해 시뮬

레이션 기법을 통해 3개수준(생산성 높음, 중간, 낮음)의 농가 데이터를 생산하여 4개의 실험 비교를 통하여 C4.5 알고리즘의 가능성과 한계를 파악하였다. 실험비교로 사용된 4개의 방법은 모돈 생산성(이유두수)만 고려하여 도태, 무작위 도태, C4.5 알고리즘, 가지치기 알고리즘을 포함한 C4.5 알고리즘이였다. 이 연구에서 연구자들은 C4.5 알고리즘을 이용한 의사결정이 매우 우수함을 보여주면서, 아울러 대량의 데이터를 통해서 도태의사결정을 자동화할 수 있음을 보여주었다. 하지만 이들의 연구는 머신러닝의 다양한 방법중에서 의사결정나무만을 이용하였고, 최근에는 머신러닝의 다양한 기법들이 연구에 활용되고 있다. 본 연구에서는 로짓모형(logistic regression), 의사결정나무(decision tree), 인공신경망(artificial neural network), kNN(k-nearest neighbor), 앙상블 모형(ensemble) 등 머신러닝기법을 이용하여 이들의 효율성을 검증하고 모돈의 임신사고 발생에 대한 예측력을 비교분석하여 양돈농장 모돈 관리에 대한 의사결정을 자동화 할 수 있을지 여부를 판단하고자 한다.

3. 연구의 방법

머신러닝 분야에서 각 알고리즘을 비교한 선행 연구들을 보면 알고리즘에 따라 상당한 정도의 예측력 차이가 존재한다(Michie et al. 1994). 그러나 특정 알고리즘이 모든 문제영역(business domains)에서 예측력이 우수한 경우는 거의 없다. 각각의 문제영역, 즉 자료집합(data set)의 형태에 따라 알고리즘의 예측력은 달라진다(Wolpert, 1996). Michie et al.(1994)은 20개의 자료집합(datasets)을 대상으로 20개의 알고리즘을 적용하여 각 알고리즘의 성과(performance)를 비교분석하므로써, 자료집합의 성격, 독립변인의 형태와 개수, 종속변인의 형태

에 따라 우수한 알고리즘을 일반화하려 하였으나, 자료집합의 성격에 따라 각 결과가 상대적인 것으로 나타났으며, 제한된 범위내에서의 일반화만 찾아내었다.

첫째, 많은 머신러닝 기법보다 로짓모형이나 판별분석이 우수한 경우가 많았으며, 변인들에 대한 정규분포의 가정이 충족되지 않을 경우 머신러닝 기법들이 효과적이라는 것을 밝혔다. 둘째, k-Nearest Neighbor (kNN) 기법이 머신러닝의 여타 모형에 비해 일반적으로 분류 및 예측 능력이 우수한 것으로 나타났으며, 셋째, kNN의 분류 및 · 예측 능력이 떨어지는 경우, 대부분 의사결정나무가 상대적으로 좋은 결과를 나타내었다.

Michie et al.(1994)은 이러한 연구결과를 토대로 비교연구에서는 최소한 로짓모형이나 판별모형, k-Nearest Neighbor, 의사결정나무(decision tree)기법 등이 포함되어야 한다고 하였다. 최근에는 Lim et al.(2000)이 22개의 의사결정나무 알고리즘, 9개의 통계기법, 2개의 인공신경망 알고리즘을 32개의 데이터셋에 적용한 결과 데이터셋의 형태에 따라 각 알고리즘의 성능이 상대적인 것으로 파악하였다. 따라서, 모돈의 생산정 예측을 위한 분석모형으로 Michie et al.(1994)이 제시한 로짓 모형, kNN, 의사결정나무 알고리즘의 하나인 CHAID를 선택하였으며, 이와 함께 기존 머신러닝기법에서 인공신경망이 매우 우수한 결과를 나타낸 결과가 많았음을 고려하여(Bound and Ross ,1997; Mouninho et al., 1994;, Bentz and Merunka, 2000), 인공신경망을 추가로 선택하였다. 이와 함께 이들 4개의 모형을 복합적으로 사용하는 앙상블 기법도 추가하여 그 결과를 비교하였다.

3.1. 모돈의 생산성 예측을 위한 머신러닝 모형의 설정

로짓모형의 경우 모돈의 생산성을 예측하기 위해 모돈유지를 위한 최

소 산자수의 기준이 되는 7두를 넘는지 여부를 묻는 선택을(0=8두이상, 1=7두이하) 독립변인으로 도태의사에 영향을 주는 종속변인들을 단계별로 투입하는 방법(stepwise)를 사용하여 로짓모형을 생성하였다. 최우도측정법(Maximum Likihood Estimation)이 사용되었으며, 투입(entry)과 유지(stay)에 적용된 유의수준은 0.05로 설정하였다. 의사결정나무모형의 경우 최적분리기준으로는 CHAID 알고리즘을 사용하였다. CHAID 알고리즘은 카이스퀘어 검정(x^2 test)을 사용하여, 부모마디로부터 분리되는 자식마디들이 최대한 서로 다르도록 만드는 것이다. 최적분리는 x^2 검정을 통해 계산된 가장 적은 p값(p-value)을 갖는 분리(split)를 선택한다. 하나의 마디에 너무 많은 관측치가 있거나 너무 적은 관측치가 있을 경우에는 오류가 증가한다. 만약 마디가 너무 적은 관측치를 가지고 있을 경우에는 통계적인 유의성의 부족으로 오류가 발생할 가능성이 많아진다.

또 너무 많은 관측치를 가질 경우에는 더 좋은 마디가 제거되었을 가능성과(1종 오류), 더 나쁜 마디가 생성되었을 가능성(2종 오류)이 존재한다(Levin and Zahavi, 2001). 따라서 마디에 포함될 관측치의 개수에 대한 설정이 필요하며, 이와 함께 설정에 따라 최종 나무의 크기가 작거나 크지 않도록 종료구칙을 두어야 한다. 만약 너무 작은 나무가 생성되면 유용성이 떨어지며, 너무 많은 나무는 해석이 어려운 문제가 있다(Levin and Zahavi, 2001).

본 연구에서 설정한 종료규칙은, 첫째, 분리된 마디는 최소한 30개 이상의 관측치를 가져야 하며, 둘째, 마디의 관측치가 100개이하일 경우에는 더 이상 분리하지 않으며, 셋째, 나무의 깊이(depth of tree)의 최대값은 6으로 설정하였고, 넷째, x^2 검정시 p값이 0.2 이상일 경우에는 더 이상 분리하지 않는다.

종료규칙을 토대로 생성된 의사결정나무의 경우도 훈련에 사용되지 않

은 새로운 데이터는 잘 판별하지 못하는 과도적합(overfitting)의 문제가 존재할 수 있다. 따라서 유효성검정데이터(validation data)를 이용하여 가지치기(Pruning)을 실시하며, 가지치기 방법은 첫째, 종료규칙에 의해 최종 의사결정나무를 생성하며, 둘째, 최종 의사결정나무로부터 하위의사결정나무들(subtrees)을 생성한다. 셋째, 하위의사결정나무는 최종의사결정나무의 잎(leaf) 수를 순차적으로 하나씩 줄여가면서 만든다. 이 경우 동일한 잎 수를 가진 여러 하위의사결정나무가 나타날 수 있어. 유효성검정데이터를 적용하여 가장 잘 분류하는 하나의 나무만 선택한다. 넷째, 잎 수가 하나인 하위의사결정나무로부터 시작하여 최종의사결정나무까지 유효성검정데이터를 적용하여 분류정확도를 계산하며, 다섯째, 유효성검정데이터(validation set)의 분류정확도가 더 이상 증가하지 않거나 감소하는 시점에서 가지치기를 한다.

인공신경망 모형의 경우 다층 퍼셉트론(multi-layer perceptron)과 역전파학습(back-propagation) 알고리즘으로 입력계층과 출력계층, 그리고 하나의 은닉계층을 가지는 3층 퍼셉트론(three layer perceptron)을 사용하였다. 은닉노드의 수를 결정하기 위해서 은닉노드 수를 1에서 증가시켜본 결과 3개 이상에서 더 이상 validation error가 낮아지지 않았다. 따라서 은닉노드수는 3개로 결정하였다. 학습중 validation set의 에러가 증가할 경우 학습을 멈추는 validation error를 통한 early stopping을 사용하였다.

kNN의 경우 비모수분석(Non-Parametric Analysis)으로 모형을 생성하지 않는다. 새로운 사례가 투입되면 가장 가까운 사례들을 추출하여 새로운 사례가 어디에 속하지를 결정한다. 이때 결정할 파라미터는 k의 개수이다. k를 선택하는 대략적인 기준(rule of thumb)은 training set의 사례개수의 제곱근을 사용하는 것이다(Dasarathy, 1991). 그러나 k가 클 경우에는 너무 많은 계산시간이 소요된다. 따라서 본 연구에

서는 SAS E-miner에서 제시하는 16-NN을 사용하였다. 앙상블모형은 네트웍의 학습 알고리즘을 다르게 하는 방법을 사용하였다. 도태결정 문제는 분류 문제이므로, 여러 네트웍의 결과에서 다수결인 경우를 해답으로 선택하였다. 선택된 네트웍은 위에서 생성한 로짓회귀, 의사결정나무, 인공신경망, kNN 이었다.

3.2. 변인의 설정

모돈의 생산성 예측 모형의 설계를 위해 먼저 총산자수를 결정하는 변인들을 설정하였다. 모돈의 총산자수는 주로 모돈의 유전적 변인과 농가의 사양 및 환경관리에 의해 좌우된다. 본 연구에서는 투입변인으로는 이전 3산차까지의 성적을 투입변인으로 설정하였다. 과거 산차의 성적에 사용된 투입변인은 임신사고회수, 총사고 일령, 분만일령, 이유일령, 이유 후 교배일령, 총산, 실산, 생시체중, 이유두수, 이유체중이다. 산차에 관계없는 변인으로는 초교배일령을 추가하였다. 각 투입변인에 대한 자세한 설명은 〈표 1〉에 나타나 있다. 모돈 임신사고 예측의 목표변인은 총산자수가 실제로 7두이하인 경우와 그렇지 않은 경우로 설정되며, 해당 산차에서 총산자수가 7두 이하인 경우 값이 1이되고, 그렇지, 그렇지 않으면 0이 된다. 우리나라 모돈의 산차당 평균 총산자수가 2006년 편재 10.47에 이르고 있으며(대한양돈협회. 2006), 총산자수 7두 이하의 경우 모돈 도태의 대상으로 고려되고 있다.

〈표 1〉 도태예측 변인 설명

변인명		변인 설명(단위)	비고
투입변인	FmateDay	초교배일령(일)	초교배일－출생일
	m0c_Abrt1	임신사고회수1(회)	교배후 46일까지의 불임사고 (조기재발, 1차재발, 불규칙재발, 2차재발)
	m0c_Abrt2	임신사고회수2(회)	47일 이후의 불임사고 (지연재발, 공태, 분만사불임)
	m0c_Abrt3	임신사고회수3(회)	불임이외의 임신사고 (유산, 분만돈사고)
	m0d_Abrt	임신사고일령(일)	임신사고일－교배일
	m0d_mf	분만일령(일)	분만일－교배일
	m0d_fw	이유일령(일)	이유일 - 분만일
	m0c_tpigs	총산(두)	-
	m0c_apigs	실산(두)	-
	m0w_apigs	생시체중(kg)	총생시자돈체중/실산
	m0c_wpigs	이유두수(두)	-
	m0w_wpigs	이유체중(kg)	총이유자돈체중/이유두수
	m0c_mate	교배회수(회)	-
목표변인	tPigs	총산	1=7두이하, 0=8두이상

3.3. 자료의 전처리

모형을 구축하기 전에 우선 데이터를 훈련데이터 셋(training set) 40%, 유효성데이터 셋(validation set) 30%, 검정데이터 셋(test set) 30%로 나누었다. 훈련데이터 셋(training set)은 모형을 생성하는 데 사용되며, test set은 모형을 비교·평가하는 데 사용된다. 유효성데이터 셋(validation set)은 인공신경망 모형과 의사결정나무에서 사용된다. 유효성데이터 셋(validation set)은 인공신경망에서는 과도적합(overfitting)을 막기위한 방법으로 유효성데이터 오류(validation error)를 통한 조

기 멈춤(early stopping)에 사용되었으며, 의사결정나무에서도 과도적합을 막기위한 가지치기(pruning)에 사용되었다.

의사결정나무의 경우는 결측치를 하나의 정보로 받아들인다. 그러나 나머지 모형의 경우는 결측치가 있을 경우 분석에서 제외된다. 따라서 로짓모형, kNN, 인공신경망 모형에는 결측치를 등간척도와 서열척도 모두 의사결정나무 기법을 이용해 대체(replacement)한 자료가 분석에 사용되었다. kNN의 경우에는 수치가 큰 변인에 따라 거리 값이 큰 영향을 받는다. 따라서 등간척도의 경우에는 변인을 표준화한 후 kNN에 적용하였다.

3.4. 모형평가 방법

Calder and Malthous(2003)은 모형을 평가하는 기준으로 'fit'과 'performance'를 제시하였다. 'fit'은 실제 값과 모형이 예측한 값 사이의 유사정도를 의미하며, 'performance'는 모형의 예측 값을 바탕으로 마케팅을 실시할 경우 실제 응답한 사람이 얼마나 되는지를 측정하는 것이다. 모형의 'fit'은 여러 방법으로 측정될 수 있는 데 가장 많이 사용되는 것은 정확도(accuracy)이다. 정확도는 모형이 예측한 값이 실제값과 얼마나 동일한지를 측정하는 것으로, 바르게 예측한 빈도를 전체 빈도로 나누어서 계산한다. 정확도는 아래와 같다.

$$\text{정확도}(accuracy) = \frac{TP+TN}{TP+TN+FP+FN}$$

TP(true positive)는 모돈의 생산성이 낮은 모돈, 즉 총산자수가 7두 이하인 모돈을 정확하게 예측한 빈도, FP(false positive)는 생산성

이 높은 낮은 모돈을 생산성이 높은 모돈, 즉 총산자수가 7두 이상인 모돈으로 예측한 빈도, TN(true negative)는 생산성이 높은 모돈을 정확히 예측한 빈도, FN(false negative) 생산성이 높은 모돈을 생산성이 낮은 모돈으로 예측한 빈도이다.

Performance를 측정하는 대표적인 방법은 응답률(response rate)이다. 모돈의 생산성을 예측하여 모돈의 관리에 대한 의사결정에 사용하는 경우, 생산성이 낮은 모돈을 파악하여 도태여부를 결정해야 한다. 조사대상 모돈의 9.36%가 생산성이 낮은 현실에서, 전체 자료를 사용하여 측정된 정확도(accuracy)를 토대로 모형을 선택하는 것은 의미가 없다. 따라서 이 경우에는 모형을 토대로 생산성이 낮을 확률이 높은 10%를 추출했을 경우, 어느 모형이 더 정확히 응답자를 추출해내는지를 파악해야 한다. 응답률(RR: response rate)은 이를 파악하기위한 지표이다.

응답률(RR: response rate)은 선택된 집단이 실제 응답자일 가능성을 나타내며, 각 분위(quantile)에서의 응답자 비율로 나타낸다. 리프트(lift)는 모형을 통해 응답자를 선택할 경우, 무작위로 선택하는 것에 비해 응답자일 확률이 얼마나 증가하는지를 파악하기 위한 지표이다.

$$\text{응답률}(RR_j) = \frac{A_j}{A_j + B_j}$$

$$\text{리프트}(LIFT_j) = \frac{RR_j}{RR_T}$$

여기서 A_j는 j번째 분위(quantile)에서의 응답자(responder)의 수이며, B_j는 j번째 분위(quantile)에서의 미응답자(non-responder)의 수이다. RR_T는 전체 데이터에서의 응답자 비율이다. 리프크 값이 클수

록 모형의 예측력이 높다는 것을 의미한다.

4. 분석 및 결과

4.1. 자료와 기술적 통계

모돈의 생산성을 예측하는 머신러닝 모델들의 적용을 위해 국내 양돈농가에서 가장 많이 사용하고 있는 양돈 생산 경영 관리프로그램인 Pigplan에서 수집된 데이터를 대상으로 분석하였다. 모돈의 산차별 총산자수 예측의 효율성을 높이기 위해 2006년 현재 Pigplan의 사용기간이 가장 오래된 도드람양돈농협 조합원 농가들의 자료를 대상으로 하였으며. 총 109농가중 2년이상의 데이터를 보유하고 있는 79농가의 모든 자료를 머신러닝 모델들의 분석에 이용하였다. 분석에 사용된 모돈수는 52,066두였다. 모형의 비교는 test set을 통해 이루어졌다. 훈련(train)데이터를 통해 생성된 5개의 모형에 test set을 적용하여 모형을 평가하였다. 각 산차별 모돈수와 해당 산차 train set, validation set, test set의 총산자수가 7두 이하인 비율은 〈표 2〉에서 제시된 바와 같다.

총산자수가 7두 이하인 모돈들은 산차가 커질수록 증가하는 경향을 보이고 있는데, 3산의 경우 8,89%의 총산자수를 보이고 있으며, 4산과 5산에서는 큰 차이를 보이지 않았으나 6산에서는 7두이하를 출산하는 모돈의 수가 증가하여 10.49%를 차지하고 있으며, 7산에 이르러서는 그 비율이 더욱 늘어나 11.64%의 모돈이 7두이하의 총사자수를 보이고 있는 것으로 나타났다. 이는 모돈의 생산성이 산차가 커질수록 감소하는 일반적인 경향을 반영하고 있는 것으로 보여진다. 3산과 5산은 validation set의 총산 7두 이하의 비율이 전체 모돈의 평균치 보다 높았고, 4산과

6산의 경우는 train set의 총산 7두 이하의 비율이 평균치에 비해 높았다. 모든 산차에서 분할된 데이터 셋의 총산 7두 이하의 비율은 크게 차이가 없었으며, 5산의 경우는 validation set의 모돈에서 총산자수 7두 이하의 비율이 1% 정도 높은 것으로 나타났으나, 측정과 평가에 영향을 줄만큼 유의미한 차이는 없는 것으로 보인다. 따라서, 각 데이터셋의 분할이 적절히 이루어진 것으로 판단된다.

〈표 2〉 산차별 총산 7두 이하 비율

산차	모돈수	총산(7두 이하비율)			
		전체	train (40%)	validation (30%)	Test (30%)
3산	19,135	8.89	8.88	9.09	8.71
4산	16,285	8.99	9.21	8.95	8.74
5산	13,482	8.4	9.09	9.59	9.35
6산	10,321	10.49	10.64	10.27	10.53
7산	7,096	11.64	11.59	11.41	11.94

4.2. 모형별 임신사고예측 정확도

로짓(Logit)모형, 인공신경망(NN)모형, 의사결정나무(DT)모형, 앙상블(Ensemble)모형, k-Nearest Neighbor(kNN)모형의 정확도를 구하면 〈표 3〉과 같다. 3산, 4산, 5산의 경우의 경우는 모든 모형이 90.5%-91.3%의 정확도를 보였으며, 6산의 경우는 89.2%~89.5%의 정확도를 나타내었다. 7산의 경우는 87.6%~88.1%의 정확도를 나타내었다. k-Nearest Neighbor(kNN)모형만 다른 모형에 비해 상대적으로 낮은 정확도를 보였고, 로짓(Logit)모형, 인공신경망(NN)모형, 의사결정나무(DT)모형, 앙상블(Ensemble)모형들은 모든 산차에서 모두 유

사한 정확도를 보였다. Kirchner et al. (2006)의 C4.5 의사결정나무 모형이 85%의 정확도를 보인것과 비교하여 본 연구에서 사용한 다섯가지 모형이 모두 모든 산차에서 87.5%~91.3%의 높은 정확도로, 5~12% 정도의 오차를 보이고 있어, 효율적으로 모돈 임신사고를 예측하고 있음을 보여주고 있다. 따라서, 머신러닝의 모형들 모두 효과적인 의사결정지원에 사용될 수 있음을 보여준다.

전체적으로 보면, 산차가 증가할수록 모형의 정확도는 미미하게 감소하였으나, 모형들간의 예측력은 비슷한 것으로 나타났다. 이는 산차가 증가할수록 모돈의 생산능력이 감소하여, 7두이하의 낮은 총산자수를 나타내는 모돈의 수가 늘어나서 총산자수의 예측력에 조금씩 떨어지는 것에서 기인한 것으로 사료된다.

〈표 3〉 모형별 총산 7두 이하 예측 정확도

	Logit	NN	DT	Ensemble	kNN
3산	91.3%	91.3%	91.3%	91.3%	91.2%
4산	91.3%	91.3%	91.3%	91.3%	91.1%
5산	90.7%	90.7%	90.7%	90.7%	90.5%
6산	89.5%	89.5%	89.5%	89.5%	89.2%
7산	88.0%	88.1%	88.1%	88.1%	87.6%

4.3. 모돈 임신사고 예측의 머신러닝기법 활용 효율성

3산차 모돈의 리프트(lift) 값을 통해 모형의 성능(performance)를 비교해 보면, 인공신경망(NN)모형이 가장 우수한 성능을 나타내었으며, 그 다음으로는 앙상블(Ensemble)모형, 로짓(Logit)모형, 의사결정나무(DT)모형 순이었다. k-Nearest Neighbor(kNN)은 다른 모형에 비해

현저히 성능이 떨어지는 것으로 나타났다(〈표 4〉). 리프트(lift) 값은 모형사용에 따른 효과를 직접적으로 보여주는 지표이다. 리프트 값을 통해 무작위로 추출했을 경우에 비해 해당 모형이 몇 배나 더 높은 예측 확률을 가지고 있는지 파악할 수 있다. 조사대상 농가들의 모돈의 산차별 총산자수가 7두이하인 비율이 산차별로 8-11%에 머무르고 있는 점을 고려하여 생산성이 낮은 모돈순으로 상위 10%의 모돈을 추출했을 경우의 모형의 리프트값을 모형별로 비교하였으며, 생산성이 낮은 하위 20%의 모돈까지 확대하여 결과를 비교하였다.

5개의 모형을 통해 총산이 7두 이하일 확률이 가장 높은 모돈 10%를 추출했을 경우, 인공신경망(NN)은 무작위추출(baseline)에 비해 2.21배, 앙상블(Ensemble)모형은 2.15배, 로짓(Logit)모형은 2.09배, 의사결정나무(DT)모형은 1.88배, k-Nearest Neighbor(kNN)은 1.38배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다. 총산이 7두 이하일 확률이 높은 모돈의 추출비율을 높여도 인공신경망(NN)모형이 가장 우수한 성능을 보였으며, k-Nearest Neighbor(kNN)모형의 성능이 다른 모형에 비해 가장 떨어지는 것으로 나타났다. 5개의 모형을 통해 총산이 7두 이하일 확률이 가장 높은 모돈 10%를 추출했을 경우, 인공신경망(NN)은 무작위추출(baseline)에 비해 1.90배, 앙상블(Ensemble)모형은 1.85배, 로짓(Logit)모형은 1.76배, 의사결정나무(DT)모형은 1.51배, k-Nearest Neighbor(kNN)은 1.27배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다.

〈표 4〉 3산차 모돈의 총산 7두 이하 예측 리프트값 비교

decile	baseline	DT	NN	Logit	kNN	Ensemble
10	1.00	1.88	2.21	2.09	1.38	2.15
20	1.00	1.51	1.90	1.76	1.27	1.85

4산차 모돈의 경우 리프트(lift) 값을 통해 모형의 성능(performance)를 비교해 보면, 앙상블(Ensemble)모형과 인공신경망(NN)모형이 가장 우수한 성능을 나타내었으며, 그 다음으로는 로짓(Logit)모형, 의사결정나무(DT)모형 순이었다. k-Nearest Neighbor(kNN)은 다른 모형에 비해 현저히 성능이 떨어지는 것으로 나타났다(〈표 5〉). 5개의 모형을 통해 총산이 7두 이하일 확률이 가장 높은 모돈 10%를 추출했을 경우, 앙상블(Ensemble)모형은 무작위추출(baseline)에 비해 2.30배, 인공신경망(NN)모형은 2.23배, 로짓(Logit)모형은 2.13배, 의사결정나무(DT)모형은 1.84배, k-Nearest Neighbor(kNN)은 1.61배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다. 모돈 20%까지 추출했을 경우에는 의사결정나무(DT)은 1.63배, 인공신경망(NN) 모형은 1.87배, 앙상블(Ensemble) 모형은 1.93배, 로짓(Logit) 모형은 1.87배, k-Nearest Neighbor(kNN)은 1.42배를 무작위 추출에 비해 더 정확히 예측하는 것으로 나타나, 추출비율을 달리하여도 4산차 모돈의 생산성 예측은 앙상블(Ensemble)모형이 가장 우수한 것으로 나타났다.

〈표 5〉 4산차 모돈의 총산 7두 이하 예측 리프트값 비교

decile	baseline	DT	NN	Logit	kNN	Ensemble
10	1.00	1.84	2.23	2.13	1.61	2.30
20	1.00	1.63	1.87	1.87	1.42	1.93

5산차 모돈의 경우 리프트(lift) 값을 통해 모형의 성능(performance)을 비교해 보면, 앙상블(Ensemble)모형, 로짓(Logit)모형, 인공신경망(NN)모형, 의사결정나무(DT)모형의 순으로 우수한 성능을 나타내었으며, k-Nearest Neighbor(kNN)은 다른 모형에 비해 현저히 성능이 떨어지는 것으로 나타났다(〈표 6〉). 5개의 모형을 통해 총산이 7두 이하

일 확률이 가장 높은 모돈 10%를 추출했을 경우, 앙상블(Ensemble)모형은 무작위추출(baseline)에 비해 2.35배, 로짓(Logit)모형은 2.33배, 인공신경망(NN)모형은 2.25배, 의사결정나무(DT)모형은 1.87배, k-Nearest Neighbor(kNN)은 1.39배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다. 총산이 7두 이하일 확률이 높은 모돈의 추출비율을 높여도 앙상블(Ensemble)모형이 가장 우수한 성능을 보였으며, k-Nearest Neighbor (kNN)모형의 성능이 다른 모형에 비해 가장 떨어지는 것으로 나타났다. 총산이 7두 이하일 확률이 가장 높은 모돈 20%를 추출했을 경우, 앙상블(Ensemble)모형은 무작위추출(baseline)에 비해 2.00배, 로짓(Logit)모형은 1.98배, 인공신경망(NN)모형은 1.94배, 의사결정나무(DT)모형은 1.52배, k-Nearest Neighbor(kNN)은 1.42배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다.

〈표 6〉 5산차 모돈의 총산 7두 이하 예측 리프트값 비교

decile	baseline	DT	NN	Logit	kNN	Ensemble
10	1.00	1.87	2.25	2.33	1.39	2.35
20	1.00	1.52	1.94	1.98	1.42	2.00

6산차 모돈의 경우 리프트(lift) 값을 통해 모형의 성능(performance)를 비교해 보면, 인공신경망(NN)모형과 앙상블(Ensemble)모형 가장 우수한 성능을 나타내었으며, 의사결정나무(DT)모형과 k-Nearest Neighbor (kNN)은 다른 모형에 비해 다소 성능이 떨어지는 것으로 나타났다(표 7). 5개의 모형을 통해 총산이 7두 이하일 확률이 가장 높은 모돈 10%를 추출했을 경우, 인공신경망(NN)모형은 무작위추출(baseline)에 비해 2.39배, 앙상블(Ensemble)모형은 2.35배, 로짓(Logit)모형은 2.20배, 의사결정나무(DT)모형은 1.62배, k-Nearest Neighbor(kNN)은

1.35배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다. 인공신경망(NN)모형과 앙상블(Ensemble)모형은 총산이 7두 이하일 확률이 높은 모돈의 추출비율을 높여도 가장 우수한 성능을 보였다. 총산이 7두 이하일 확률이 가장 높은 모돈 20%를 추출했을 경우, 인공신경망(NN)모형은 무작위추출(baseline)에 비해 2.00배, 앙상블(Ensemble)모형은 2.02배, 로짓(Logit)모형은 1.82배, 의사결정나무(DT)모형은 1.62배, k-Nearest Neighbor(kNN)은 1.29배 더 정확한 예측 확률을 가지고 있는 것으로 나타났다.

〈표 7〉 6산차 모돈의 총산 7두 이하 예측 리프트값 비교

decile	baseline	DT	NN	Logit	kNN	Ensemble
10	1.00	1.62	2.39	2.20	1.35	2.35
20	1.00	1.62	2.00	1.82	1.29	2.02

7산차 모돈의 경우 리프트(lift) 값을 통해 모형의 성능(performance)를 비교해 보면, 앙상블(Ensemble)모형, 인공신경망(NN)모형, 로짓(Logit)모형이 우수한 성능을 나타내었으며, 의사결정나무(DT)모형과 k-Nearest Neighbor(kNN)은 다른 모형에 비해 다소 성능이 떨어지는 것으로 나타났다(표 8). 5개의 모형을 통해 총산이 7두 이하일 확률이 가장 높은 모돈 10%를 추출했을 경우, 로짓(Logit)모형과 앙상블(Ensemble)모형은 무작위추출(baseline)에 비해 2.60배, 인공신경망(NN)모형은 2.51배, 의사결정나무(DT)모형은 1.75배, k-Nearest Neighbor(kNN)은 1.64배 더 정확한 예측 확률을 가지고 있는 것으로 나타났으나, 추출비율을 20%이상으로 하였을 경우, 앙상블(Ensemble)모형과 인공신경망(NN)모형이 더 우수한 성능을 보였다. k-Nearest Neighbor(kNN)모형은 전반적으로 가장 낮은 성능을 보여, 의사결정나무(DT)은

1.61배, 앙상블(Ensemble)모형은 2.05배, 로짓(Logit)모형은 1.91배, k-Nearest Neighbor(kNN)모형은 1.32배, 인공신경망(NN)모형은 2.11배를 무작위 추출에 비해 더 정확히 예측하는 것으로 나타났다.

〈표 8〉 7산차 모돈의 총산 7두 이하 예측 리프트값 비교

decile	baseline	DT	NN	Logit	kNN	Ensemble
10	1.00	1.75	2.51	2.60	1.64	2.60
20	1.00	1.61	2.11	1.91	1.32	2.05

머신러닝 모형에 의한 산차별 총산 예측 결과를 요약하면 다음과 같다. 산차가 증가할수록 모형의 정확도는 미미하게 감소하였으며, 모형들 간의 정확도 차이는 미미한 것으로 나타났다. 모든 산차에서 앙상블 모형이 가장 우수한 것으로 나타났다. 로짓모형과 인공신경망 모형의 경우도 상대적으로 우수한 성능을 나타내었다. 그러나 의사결정나무 모형과 k-Nearest Neighbor 모형은 성능이 떨어지는 것으로 나타났다. 리프트 값은 산차가 증가할수록 높아지는 것으로 나타났다. 산차에 상관없이 가장 성능이 좋은 모형의 리프트 값과 가장 나쁜 성능의 리프트 값 차이는 1.0배 정도로 나타났다. 이는 성능이 좋은 모형은 나쁜 모형에 비해 2배이상 높은 확률로 임신사고 모돈을 예측할 수 있음을 보여준다.

5. 결론 및 제언

본 연구에서는 머신러닝 방법론을 사용하여 축산농가의 의사결정지원에 활용할 수 있는 방안을 제시하고자 국내에서 가장 대표적으로 사용되는 양돈관리프로그램인 Pigplan을 통해 축적된 자료에 적용하여 양돈농

장의 모돈의 산차별 생산력 예측모형을 도출하고자 하였다. 머신러닝 기법 적용에 대한 현실성을 평가하기 위하여 Pigplan 데이터를 대상으로 사례분석을 실시하기 위해 먼저 선행연구 고찰을 통해 사용할 다섯가지 머신러닝 모형을 선정하였으며, 선정된 모형들을 평가할 방법을 고찰하고, 본 연구에서 사용할 평가방법에 대해 논의하였다.

모형의 측정에 사용된 Pigplan사용농가들의 총산자수가 7두 이하인 모돈들은 산차가 커질수록 증가하는 경향을 보이고 있는데, 3산의 경우 8,89%의 총산자수를 보이고 있으며, 4산과 5산에서는 큰 차이를 보이지 않았으나 6산에서는 7두이하를 출산하는 모돈의 수가 증가하여 10.49%를 차지하고 있으며, 7산에 이르러서는 그 비율이 더욱 늘어나 11.64%의 모돈이 7두이하의 총사자수를 보이고 있는 것으로 나타났다. 이는 모돈의 생산성이 산차가 커질수록 감소하는 일반적인 경향을 반영하고 있는 것으로 보여진다. Kirchner et al. (2006)의 C4.5 의사결정나무 모형이 85%의 정확도를 보인것과 비교하여 본 연구에서 사용한 다섯가지 모형이 모두 모든 산차에서 87.5% 이상의 높은 정확도로, 12.5% 이내의 오차를 보이고 있어, 효율적으로 모돈의 생산성을 예측하고 있음을 보여주고 있다.

3산, 4산, 5산의 경우의 경우는 모든 모형이 90.5%-91.3%의 정확도를 보였으며, 6산의 경우는 89.2%~89.5%의 정확도를 나타내었다. 7산의 경우는 87.6%~88.1%의 정확도를 나타내었다. k-Nearest Neighbor (kNN)모형만 다른 모형에 비해 상대적으로 낮은 정확도를 보였고, 로짓(Logit)모형, 인공신경망(NN)모형, 의사결정나무(DT)모형, 앙상블(Ensemble)모형들은 모든 산차에서 모두 유사한 정확도를 보였다. 전체적으로 보면, 산차가 증가할수록 모형의 정확도는 미미하게 감소하였으나, 모형들간의 예측력은 비슷한 것으로 나타났다. 이는 산차가 증가할수록 모돈의 생산능력이 감소하여, 7두이하의 낮은 총산자수를 나타내

는 모돈의 수가 늘어나서 총산자수의 예측력에 조금씩 떨어지는 것에서 기인한 것으로 사료된다.

또한, 모형의 예측효율성을 나타내는 리프트 값은 산차가 커질수록 증가하였으며, 산차별로 가장 성능이 좋은 모형의 리프트 값과 가장 나쁜 성능의 리프트 값 차이는 3산차에서 .83. 4산차에서 0.62, 5산차에서 0.96, 6산차에서 1.04, 7산차에서 0.96으로 비교적 높게 나타났다. 이는 성능이 좋은 모형은 나쁜 모형에 비해 1.6배-2.04배 이상 높은 확률로 생산성이 떨어지는 모돈을 예측할 수 있음을 보여준다. 인공신경망모형이 대체로 높은 리프트값을 보였으며, 앙상블 모형이 그 다음을 이루고 있으며, k-Nearest Neighbor(kNN)모형이 가장 낮은 리프트 값을 보이고 있어서 인공신경망모형과 앙상블 모형이 가장 효율성이 높은 것으로 나타났다.

연구의 한계로는 산차가 높을수록 모형들의 정확도는 떨어지고 효율성은 높아지고 있어서 선도농가에 대한 모형과 자료의 확충을 통해 보완할 필요가 있다. 산차에 따라 농가의 성격에 따라 의사결정의 효율성이 모형에 따라 달라지는 것으로 보아 산차마다 모돈의 성적에 영향을 주는 요인들이 존재할 수 있음을 보여준다. 좀 더 향상된 생산성 예측을 위해 데이터탐색과정과 현장전문가의 참여를 통한 변인의 선정과 데이터의 분할을 통해 농가의 성격에 따른 생산성 분석이 필요할 것으로 판단된다.

■ 참 고 문 헌 ■

대한양돈협회. (2005). 전업 양돈농가 실태보고서. 대한양돈협회.

대한양돈협회. (2007). 2007년 양돈장 질병보고서. Pig & Pork.

이용범. (2004). 데이타마이닝의 농업적 활용. *Journal of Biosystems Engineering 29*(1): 79-96.

Bentz, Y., and Merunkay, D. (2000). Neural Networks and the Multinomial Logit for Branch Choice Modeling: a Hybrid Approach. *Journal of Forecasting 19*(3): 177-200.

Bitchler, M. and Kiss, C. (2004). *A Comparison of Logistic Regression, k-Nearest Neighbor, and Decision Tree Induction for Campaign Management*. Proceedings of the Tenth Americas Conference on Information Systems. New York. August: 1918-1925.

Bound, D., and Ross, D. (1997). Forecasting Customer Response with Neural Network. *Handbook of Neural Computation*. G6.2. 1-7.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics 24*(6): 2350-2383.

Breiman, L., J. Friedman, Olshen, R., and Stone C. (1984). *Classification and Regression and Regression Trees*. Belmont, CA: Wadsworth.

Cho, S., M. Jang, et al. (1997). Virtural sample generation using a population of networks. *Neural Processing Letters 12*: 88-89.

Chung, H. M. and P. Gray. (1999). Data Mining. *Journal of Management Information Systems 16*(1): 11-17.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control. Signals, and Systems 2*(4): 303-314.

Freund, Y. and R. E. Schapire. (1996). *Game theory, on-line prediction and boosting*. Proceedings of the Annual ACM Conference on Computational Learning Theory.

Freund, Y. and R. E. Schapire. (1999). Large margin classification using the perceptron algorithm. *Machine Learning 37*(3): 277-296.

Gray, P. and H. J. Watson. (1998). Professional Briefings...Present and Future Directions in Data Warehousing. *Database for Advances in Information Systems 29*(3): 83-90.

Gray, P. and H. J. Watson. (1998). *Decision Support in the Data Warehouse*. N.J.: Upper Saddle River.

Han, J. and M. Kamber. (2001). *Data Mining: Concepts and Techniques San Francisco*. Morgan-Kaufmann Academic Press.

Hand, D. J. (1998). Data Mining: Statistics and More?. *The American Statistician 52*(2): 112-118.

Hornik, K., M. Stinchcombe, et al. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks 3*(5): 551-560.

Hunt, E., J. Martin, et al. (1966). *Experiments in induction*. New York: Academic Press.

Iddings, R.K., and Apps, J.W. (1990). What Influences Farmers' Computer Use?. *Journal of Extension 28*(1)(http://www.joe.org/joe/1990spring/a4.html.2004/10/1).

Jayas, D. S., J. Paliwal, et al. (2000). Multi-layer neural networks for image analysis of agricultural products. *Journal of Agricultural and Engineering Research 77*(2): 119-128.

Kass, G. (2001). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics 29*(1980): 119-127.

Kirchner, K., K. H. Tolle, et al. (2004a). Decision tree technique applied to pig farming datasets. *Livestock Production Science 90*(2-3): 191-200.

Kirchner, K., K. H. Tolle, et al. (2004b). The analysis of simulated sow herd datasets using decision tree technique. *Computers and Electronics in Agriculture 42*(2): 111-127.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine 23*(1): 89-109.

Kuhlmann, F., and Brodersen, C. (2001). Information technology and farm management: developments and perspectives. *Computers and Electronics in Agriculture 30*: 71-83.

Langley, P. and H. A. Simon. (1995). Applications of machine learning and rule induction. *Communications of the ACM 38*(11): 54-64.

Levenberg, K. (1994). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly Journal of Applied Mathematics 2*(2):164-168.

Levin, N. and Zahavi, J. (2001). Predictive Modeling Using Segmentation. *Journal of Interactive marketing 15*: 2-22.

Lim, T.S., Loh, W.Y., and Shin, Y.S. (2000). A comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning 40*: 203-228.

McQueen, R. J., S. R. Garner, et al. (1995). Applying machine learning to agricultural data. *Comput. Electron. Agric. 12*(4): 275-293.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Moutinho, L., Curry, B., Davies, F., and Rita, P. (1994). *Neural Network in Marketing*. New York: Routledge.

Murthy, K. S. (1998). Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey. *Data Mining and Knowledge Discovery 2*: 345-389.

Nilsson, N. (1965). *Learning machines*. New York: McGraw-Hill.

Peacock, P. R. (1998). Data mining in marketing: Part 1. *Marketing Management 6*(4): 9.

Peacock, P. R. (1998). Data mining in marketing: Part 2. *Marketing Management 7*(1): 15.

Pietersma, D., R. Lacroix, et al. (2003). Induction and evaluation of decision trees for lactation curve analysis. *Computers and Electronics in Agriculture 38*(1): 19-32.

Quinlan, J. R. (1993). *C4.5: Program of Machine Learning*. CA.: Morgan Kaufman Publishing.

Rumelhart, D. E., B. Widrow, et al. (1994). Basic ideas in neural networks. *Communications of the ACM 37*(3): 87-92.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Learning Internal*

Representation by Error Propagation. in Parallel Distributied Processing: Explorations in the Microstructure of Cognition. D.E. Rumelhart and J.A. McClelland(Eds.). Cambridge. MA: MIT Press.

Schultz, A., R. Wieland, et al. (2000). Neural networks in agroecological modelling-Stylish application or helpful tool?. *Computers and Electronics in Agriculture 29*(1-2): 73-97.

Scott Mitchell, R., L. A. Smith, et al. (1996). An investigation into the use of machine learning for determining oestrus in cows. *Computers and Electronics in Agriculture 15*(3): 195-213.

Sonquist, J., Baker, E., and Morgan, J. N. (1971). *Searching for Structure, Survey Research Center*, Ann Arbor: University of Michigan.

논문투고일: 2009. 10. 30
1차수정일: 2009. 11. 28
게재확정일: 2009. 12. 18