JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 12, pp. 2505-2514, Dec. 2019

전자차트데이터 기반 반려동물 생애주기 분류 기법

유 기 진[1] · 이 영 석[2] · 이 헌 규[1*]
[1]가이온 빅데이터연구소
[2]건국대학교 수의과대학, 수의외과학

Prediction of Life Cycle based on EMR data of Companion Animals

Kijin Yu[1] · YoungSeock Lee[2] · Heon Gyu Lee[1*]
[1]BigData Research Center, GAION, Seoul 06167, Korea
[2]Department of Veterinary Surgery, College of Veterinary Medical, Konkuk University, Seoul 05029, Korea

[요 약]

반려동물 양육 가구의 급증에 따라, 반려견의 건강상태를 모니터링 하고 질환 여부를 사전 진단함으로써 반려견의 의료비 지출을 줄일 수 있는 제도적 방안이 필요하다. 본 논문에서는 동물병원의 전자의무기록 (EMR; electronic medical record) 데이터를 이용하여 반려견의 종, 성별 및 생애주기별 주요 질병을 분석하는 인공지능 시스템을 제안한다. 인공지능 분석 시스템은 반려견의 종과 성별, 나이, 질병, 처방 정보로부터 상관성이 높은 의료정보 추출이 가능한 빈발패턴 분석모델과 반려견의 질환 진단에 영향을 주는 생애주기를 자동으로 예측할 수 있는 Deep Belief Network (DBN) 모델을 포함한다. 동물병원에서 수집된 EMR 데이터는 고차원 대용량 데이터이므로, 병렬처리를 통해 인공지능 알고리즘의 수행 속도 개선이 가능한 Map-Reduce 기술을 적용하였다. 국내 4개 동물병원에서 수집된 반려견의 16,139건 EMR 데이터를 이용하여 반려견의 생애주기를 예측한 결과, DBN 모델은 약 88%의 정확도를 보였다.

[Abstract]

As companion animals increase, there is a need for an institutional strategy to reduce health care costs for animals through frequent monitoring of health status and preliminary determination of disease status. In this paper, we propose an artificial intelligence (AI) system that analyzes the major diseases of dogs by breed, gender and life cycle using electronic medical record (EMR) data of animal hospitals. The system includes a frequent pattern analysis model that enables the extraction of highly correlated medical information from breed, gender, life cycle, disease and prescription information, and a Deep Belief Network (DBN) model that can automatically predict the life cycle that affects the diagnosis of a animal's disease. EMR data, a high-dimension bulk data, is managed by applying Map-Reduce technology, which performs parallel processing to improve the performance speed of AI algorithms. Using 16,139 EMR data of dogs collected from four domestic animal hospitals, the DBN model showed 88% accuracy for life cycle prediction in dogs.

색인어 : 반려견, 전자의무기록데이터, 생애주기, 인공지능, 분류모델
Key word : Companion dogs, Electronic medical record, Life cycle, Artificial intelligence, Classification model

http://dx.doi.org/10.9728/dcs.2019.20.12.2505

Received 11 October 2019; **Revised** 30 November 2019
Accepted 15 December 2019

***Corresponding Author; Heon Gyu Lee**

Tel: +82-2-2051-9595
E-mail: hglee@gaion.kr

Ⅰ. 서 론

반려동물에 대한 사회적 인식의 발달과 가구형태의 변화로 인해 우리나라의 반려동물 수가 증가하고 있다[1]. 한국농촌경제연구원의 2017년 조사에 따르면, 우리나라 전체 가구의 29.4%인 574만 가구가 반려동물 874만 마리를 기르고 있으며, 2027년에는 1,320만 마리에 이를 것으로 전망하고 있다. 그에 따라 반려동물을 위한 고품질 양육방법과 헬스케어에 대한 관심이 많아지면서 반려동물의 사료와 의약품, 생활용품, 의료서비스 등을 포함한 연관산업도 급속도로 성장하고 있다[2]. 2017년 2조 3,322억원 규모의 연관산업은 2014년 1조 5,684억원의 규모가 연평균 14.5%씩 성장한 결과이고, 2020년에는 약 6조원 규모로 성장할 것이라는 농림축산식품부의 연구결과가 발표되었다. 그러나 현재 우리나라 반려동물 연관산업은 반려동물에 대한 성숙한 인식을 기반으로 하는 제도나 정책이 필요한 상황이다. 또한, 국내에서는 동물병원의 폐쇄적인 의료데이터 관리와 체계적인 의료시스템의 부재로 인해[3], 의료데이터를 이용한 반려동물 의료산업이 활발하게 이루어지지 않고 있다.

반려동물 양육가구 비율이 높은 미국, 영국, 독일, 프랑스, 일본 등의 국가는 반려동물 건강에 대한 다양한 연구와 데이터 관리를 진행해왔고, 의료상품 및 서비스 산업도 발달하고 있다[2]. 최근에는 인간의 개인맞춤의료서비스와 같이, 반려동물의 건강상태를 수시로 파악하고 검진자료를 바탕으로 솔루션을 제공하는 시스템 및 모바일앱 등이 영국과 일본에서 개발되었다. 그리고 반려동물의 놀이와 운동량 등을 측정하고 동일 품종과 비교하여 정보를 제공하는 헬스케어서비스 제품이 미국과 오스트리아에서 출시되었다.

해외의 반려동물 의료데이터 분석 결과, 반려견의 질환은 품종과 연령대에 따라 다르게 발병하고 수명과 삶의 질을 결정한다고 보고되었다[4], [5]. 순종견은 품종 특이적 질환[8]과 선천적 질환의 발병율과 불임율[6], [7]이 높다는 연구결과가 미국과 영국에서 발표되었다. 피부질환은 유전적 요인보다 환경적 요인에 의해 말티즈, 시추, 요크셔테리어 등 장모종에게 많이 발병한다. 유방암이나 고환암, 전립선암과 같은 생식기 질환이 작은 체형의 잡종견에게 잘 발병하기 때문에, 중성화 수술을 통해 질환을 예방한다는 분석결과가 보고되었다[9].

반려견의 질환은 품종과 상관없이 연령대에 따라 다르게 발병하는 것을 미국의 연구결과를 통해 알 수 있다[9-11]. 진행성 퇴행성질환인 심장질환은 8세 이상의 고령견에서 발병율이 높고 사망을 유발하는 고위험 인자이다. 유방암은 대표적인 악성종양으로 또 다른 사망의 고위험 인자이고, 50% 이상은 중성화 수술을 하지 않은 고령 암컷견에서 발견된다. 3세 이하 반려견의 사망을 초래하는 가장 큰 원인은 소화기 질환이고, 나이와 무관하게 모든 반려견의 공통 질환은 귀와 피부질환이다. 피부질환 중 아토피 증상의 95% 환자는 5세 이하 반려견이라는 연구[11]가 있고, 아토피의 55~60% 환자는 귀 질환도 동반하는 것으로 관찰되었다[12].

미국동물병원협회 (AAHA; The American Animal Hospital Association)와 미국수의사회 (AVMA; The American Veterinary Medical Association)는 반려견 예방 의료 가이드라인 (Canine Preventive Healthcare Guideline)과 치아관리 (Dental Care), 당뇨병 관리 (Diabetes Management), 행동 (Behavior Management), 영양평가 (Nutritional Assessment), 생애주기 (Life Cycle) 등의 가이드를 제시하여 국가차원에서 반려동물의 연령대와 질환 별로 건강을 관리하고 있다. 이와 같은 정기건강검진 등의 헬스케어제도를 통해 축적된 의료데이터는 미국의 반려동물 질환에 대한 다양한 관점의 통계적 분석 및 연구를 가능하게 하였다.

반려동물의 품종 및 연령대와 관계없이, 영국에서는 피부와 소화기, 근골격계 질환이 가장 많이 발병하고[13], 이탈리아에서는 반려동물 진료항목으로 소화기, 피부, 심혈관 질환이 가장 많았다[14]. 그러나 국내에서는 피부와 심장, 귀 질환의 진료비중이 가장 높았다. 이와 같은 결과는 해외와 다른 한국의 주거형태 등의 환경적 요인과 사람들이 선호하는 견종비율에 의한 것으로 분석되었다[6], [8].

이처럼 반려동물의 질환 특성은 각 국가마다 다르기 때문에, 국내 반려동물의 의료데이터를 국가차원에서 수집 및 통합 관리할 수 있는 시스템의 필요성이 제기되고 있다. 이를 통해 확보된 데이터로부터 국내 호발성 질환 및 전염병 발생 패턴 등의 다양한 분석이 가능하고, 그 결과 미국과 같이 국내 반려동물 건강관리 가이드가 완성될 수 있을 것이다.

또한, 국가차원에서 수집하여 통합한 대용량의 반려동물 의료데이터를 분석할 수 있는 고차원적 방법이 요구된다. 최근 통신과 인공지능 기술 발달에 의해, 기하급수적으로 생산되는 빅데이터를 분석하려는 시도가 의료분야에서도 증가하고 있다. 수의학분야에서는 돼지열병 등의 전염병 발생과 관련된 요인을 식별하거나 양의 신체 사이즈와 유전자형(genotype)간의 연관분석, 양의 출생 체중과 수동면역 (passive immunity)의 관계분석, 종에 따른 소의 형태학적 특징 (morphological feature) 분석 등을 위해 머신러닝 (machine learning) 알고리즘을 적용하였다. 그리고 AI 기법을 이용하여 방사선 사진으로부터 개의 고관절 영역을 식별하거나, 사람의 지문식별과 같이 Kangal 종 개의 코 무늬를 식별, 양 근육 이미지 데이터로부터 지방량을 추정하는 등의 의료영상데이터 연구가 진행되었다[15]. 해외에서 모바일을 통해 반려동물의 건강정보를 수집하고 AI 모델을 이용하여 보호자에게 반려동물의 질환을 예측 및 진단결과를 제공하는 시스템들이 개발되었지만, 대용량의 전자차트데이터 기반 시스템 개발은 이루어지지 않았다.

따라서 본 논문에서는 국내 반려동물의 대용량 전자차트데이터를 수집 및 관리하고, 수의학자가 전문적인 진단처방정보를 검색하고 분석할 수 있는 AI 기반 시스템을 제시한다. 반려동물의 생애주기에 따른 발병질환의 유의한 차이는 반려동물의 진단과 처방에 유의한 정보이기 때문에, 전자차트데이터를 이용하여 반려동물 생애주기를 예측하는 모델을 구축하였다. 데이터 표준화 및 관리가 가능한 분석시스템을 통해, 고품질 및

고신뢰성의 의료데이터를 생성하고 국내 반려동물 특이적 질환연구를 고차원적으로 수행할 수 있을 것이다.

국내 반려동물 의료서비스의 현실화를 위해서 이 논문에서는 다음과 같은 세 가지 핵심 연구를 수행한다.

- 견종, 질환, 처방약 등의 반려동물 전자차트데이터를 표준화하고, 견종별, 성별, 생애주기별 질환의 발병율을 비교하여 유의성을 검증한다.
- 견종, 성, 생애주기, 질환, 처방약 등으로부터 빈발하는 패턴을 분석하고, 신뢰성 보장이 가능한 Deep Belief Network (DBN)[16] 기반 반려동물 생애주기 예측모델을 제안한다.
- 대용량, 고차원 데이터에 대한 분석 알고리즘의 고속처리가 가능하도록 in-memory 기반의 병렬처리 기술인 Map-Reduce 기법을 DBN에 적용한다.
- 사용자가 데이터를 직접 입력하여 분석하고, 입력데이터를 이용하여 DBN 모델을 설계 및 학습할 수 있는 웹기반 반려동물 진료데이터 분석시스템을 제시한다.

논문의 효율적인 이해를 위해, 2장 본론의 내용 구성은 다음과 같다.

2-1절에서는 전반적인 반려동물 전자차트데이터의 구성과 통계분석 방법 및 결과를 기술하고, 2-2절에서는 전자차트데이터로부터 빈발하는 패턴을 분석한다. 2-3절에서는 분산/병렬처리를 실행하는 반려동물 생애주기 분류모델의 성능평가 결과를 제시하고, 끝으로 3장에서 결론을 맺는다.

Ⅱ. 본 론

2-1 표준 의료데이터와 통계 분석

본 논문은 동물병원에서 진료를 받은 3,071마리 반려견의 16,139개 전자차트로부터 추출한 진료데이터를 분석하였다. 4개 동물병원의 전체 273,268개 전자차트데이터로부터 반려견의 호발성 질환을 기준으로 데이터를 추출하였고, 주요 10가지 질환은 다음과 같다; 관절염 (ART; arthritis), 심장판막의 이상 (MVI; mitral valve insufficiency), 기관 허탈 (TRC; trachea collapse), 귀 질환 (OTT; otitis), 당뇨병 (DIB; diabetes), 벼룩 및 진드기 감염 (IFT; infection), 슬개골 탈구 (PLX; patella luxation), 신부전 (RNF; renal failure), 췌장염 (PCT; pancreatitis), 피부 질환 (DMT; dermatitis). 진료데이터는 4개 동물병원의 통합 데이터이기 때문에, 반려동물의 품종 및 성별, 질환명, 처방 약품명 등의 용어 표준화 작업을 수행하였다. 또한, 반려동물의 출생정보와 전자차트데이터를 확보한 시점을 기준으로 나이를 계산하였다. 나이는 6가지 범주 (A: 0~2세, B: 3~5세, C: 6~8세, D: 9~11세, E: 12 ~ 14세, F: 15세 이상)로 나누어 생애주기로 표준화하였다.

표 1. 주요 4가지 품종과 질환 비율
Table 1. Disease rate by breed

	Maltese	Toy Poodle	Shih Tzu	Yorkshire Terrier	F-value
ART	215	124	49	24	9.51
MVI	1,041	213	416	218	16.97
TRC	18	1	7	13	1.33
OTT	925	432	300	96	11.92
DIB	2	13	5	1	1.93
IFT	66	39	5	10	6.24
PLX	325	129	33	62	9.13
RNF	568	147	306	314	13.89
PCT	96	91	36	54	4.23
DMT	1,136	761	791	349	17.65
Total no.	4,392	1,950	1,948	1,141	

진료데이터는 잡종을 포함한 68 종 반려견의 전자차트이며 그 중 말티즈의 데이터가 27% (4,392개)로 가장 많았다 (Table 1). 토이 푸들과 시추가 각 12% (1,950개와 1,948개), 잡종견이 8% (1,319개), 요크셔 테리어가 7% (1,141개), 포메라니안이 6% (1,058개) 순으로 많이 차지하였다. 그리고 아토피를 포함한 피부 질환 (DMT)때문에 병원을 방문한 횟수가 5,420번으로 10가지 질환 중 가장 많았다. 다음으로 심장판막의 이상 (MVI)으로 2,991번, 귀 질환 (OTT)으로 2,895번, 신부전 (RNF)으로 2,163번 진료를 많이 받았다. 토이 푸들은 다른 견종에 비해 심장판막의 이상과 신부전이 더 적게 발병하지만, 피부질환은 더 많은 것을 알 수 있다.

표 2. 성별과 질환 비율
Table 2. Disease rate by gender

	Castrated Male	Male	Spayed Female	Female	F-value
ART	361	100	225	130	20.28
MVI	894	301	1,305	485	68.22
TRC	33	6	13	7	1.73
OTT	1,287	249	916	436	7.72
DIB	6	0	30	4	8.53
IFT	120	15	28	75	15.76
PLX	420	70	306	201	1.91
RNF	949	141	724	348	7.73
PCT	256	52	133	60	7.95
DMT	2,245	450	1,540	1,164	23.91
Total no.	6,571	1,384	5,220	2,910	

Table 2는 차트의 성별 비율과 성별에 따른 질환의 발병율을

나타낸다. 중성화 암컷은 중성화하지 않은 암컷보다 1.8배가 많고 중성화 수컷은 일반 수컷에 비해 4.7배 많은 것을 확인할 수 있다. 그리고 심장판막의 이상 (MVI)은 일반 수컷과 중성화 암컷에서 많이 발병하고, 췌장염 (PCT)은 암컷보다 중성화 수컷과 일반 수컷에서 1.5배~2배 많이 나타난다.

질환의 발병 빈도는 품종이나 성별 차이보다 생애주기 차이에서 더 뚜렷하게 달라지는 것을 확인하였다 (Table 3). 9~11세 반려견의 전자차트 21%와 12~14세의 42%, 15세 이상의 40%는 심장판막이상 (MVI)으로 동물병원에서 진단을 받은 기록이다. 전체 심장판막이상 (MVI)으로 진단받은 전자차트 중 12세 이상의 반려견이 79%를 차지하는 것과 유사하게, 전체 신부전 (RNF) 진단기록 중 12세 이상의 반려견이 66%를 차지한다.

표 3. 생애주기와 질환 비율
Table 3. Disease rate by life cycle

	0 ~ 2	3 ~ 5	6 ~ 8	9 ~ 11	12 ~ 14	15 ~	F-value
ART	174	141	191	119	141	65	12.20
MVI	17	42	72	502	1,325	1,033	637.84
TRC	0	8	11	17	8	15	3.37
OTT	530	745	644	473	287	216	79.32
DIB	0	0	9	4	16	11	4.07
IFT	103	55	29	29	12	9	47.10
PLX	241	295	208	185	47	23	59.34
RNF	135	183	182	242	730	691	158.18
PCT	88	76	95	83	93	67	1.50
DMT	1,078	1,386	1,269	756	464	467	218.82
Total no.	2,366	2,931	2,710	2,410	3,123	2,597	

그림 1. 분석시스템의 반려동물 진료데이터 통계적 정보
Fig. 1. Statistical information of EMR in web-based analysis system

이와 반대로 피부질환 (DMT) 진단기록 69%는 8세 이하의 반려견의 진료이고, 귀질환 (OTT) 진단기록의 66%가 8세 이하 반려견의 진료결과이다. 따라서 심장판막이상 (MVI)과 신부전 (RNF)은 고령견에서 쉽게 발병하지만, 피부 (DMT)와 귀질환 (OTT)은 나이가 들수록 발병율이 낮아지는 것을 볼 수 있다. 반려견 진료데이터의 품종별, 성별, 생애주기별 통계적 분석결과는 논문에서 제시하는 웹 분석시스템에 Fig. 1과 같이 시각화하여 제공된다.

10가지 질환이 품종, 성별, 생애주기에 따라 유의하게 발생하는지 테스트하기 위해 분산분석 (ANOVA; analysis of variance)[17]을 실행하였다. 검정통계량은 다음의 식 (1)을 따른다.

$$F = \frac{MSgroup}{MSerror} \tag{1}$$

$$= \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 / (k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 / (n-k)}$$

식에서 $\bar{y}$는 전체 질환 발생빈도의 평균이고 $\bar{y}_j$는 j그룹의 질환 발생빈도 평균, y_{ij}는 j그룹의 i번째 샘플, k는 전체 그룹의 수, n_j는 j그룹의 샘플 수이다. 각 질환의 그룹 내 발병빈도와 그룹 간 발병빈도의 차이를 비교하는 방법으로, F-value가 클수록 그룹 간 차이가 뚜렷한 질병을 의미한다.

Table 1,2,3의 각 10가지 질환 F-value를 비교한 결과, 질환의 발병율 차이가 품종과 성별 간에서 뚜렷하지 않았다 (Table 1,2). 하지만 생애주기에 따라 심장판막의 이상 (MVI)과 귀 질환 (OTT), 신부전 (RNF), 피부 질환 (DMT)의 발병율 차이가 유의하다 (Table 3). 따라서 본 논문에서는 10가지 질환 발생빈도를 기반으로 반려동물의 생애주기를 분류하는 방법을 연구한다.

2-2 빈발패턴 분석

앞에서는 반려동물 품종과 질환, 성별과 질환, 그리고 생애주기와 질환의 관계를 분석하였다. 이번 장에서는 품종과 성별, 생애주기, 질환, 처방약 등 통합된 데이터에서 빈발하는 패턴을 검색하여 그들의 관계를 분석하였다. 빈발패턴 분석은 트랜잭션 데이터에서 빈번히 발생하는 패턴을 검색하는 방법이다[18]. 대표적으로 기업의 구매 상품 집합이나 접속 웹페이지 집합을 분석하기 위해 사용되고, 생물학에서는 질병의 원인 유전자 집합을 발견하기 위해 사용된다. 모든 항목 (item)의 집합은 다음과 같다; $I = \{i_1, i_2, \cdots, i_k\}$. 그리고 여러

개의 항목으로 구성된 트랜잭션 (transaction) T는 집합 I의 부분 집합이고 ($T \subseteq I$), 패턴 P를 포함할 수 있다 ($P \subseteq T$). 최소 지지도 (minimum support)보다 빈발하게 발생하는 패턴 P를 빈발 패턴으로 정의한다.

Table 4는 최소 지지도 $0.01\,(16{,}139 \times 0.01 = 161)$보다 빈번하게 발생하고, 견종, 성별, 생애주기, 질환, 처방약 중 3종류 이상의 항목을 포함하는 전자차트 (트랜잭션)의 패턴을 검색한 결과이다.

빈발패턴의 주요 내용을 해석하면 다음과 같다.

- 12~14세 중성화 암컷에서 심장판막의 이상이 가장 많이 발병함
- 3~8세 중성화 수컷은 피부질환을 가장 많이 겪은 것을 알 수 있음
- 12~14세 말티즈 (maltese)와 15세 이상의 중성화 암컷, 12~14세 중성화 수컷이 심장판막이상으로 병원을 많이 내원함
- 12세 이상의 중성화 수컷과 15세 이상의 중성화 암컷이 신부전 진단을 많이 받음

빈발패턴 분석을 통해 견종과 성별, 생애주기 등에 따라 다르게 질환이 발병하는 것을 확인할 수 있었다. 이처럼 반려동물의 신체정보와 의료정보들 간의 상관관계를 나타내는 빈발패턴을 이용하여, 다음 절에서 반려동물의 생애주기를 분류하는 모델을 학습한다. 그러나 모델의 과적합 (overfitting) 문제를 피하지만 정확한 모델을 구축하기 위해, 나이 정보를 포함하지 않은 빈발패턴을 검색하여 학습데이터에 추가하였다.

표 4. 품종과 성별, 연령대, 질환의 빈발패턴
Table 4. Frequent patterns between breed, gender, age, and disease

Frequent Patterns	Count
Spayed Female, Age 12~14, MVI	639
Castrated Male, Age 6~8, DMT	591
Castrated Male, Age 3~5, DMT	567
Castrated Male, Age 3~5, OTT	449
Maltese, Age 12~14, MVI	443
Spayed Female, Age 15~, MVI	398
Castrated Male, Age 12~14, RNF	328
Castrated Male, Age 15~, RNF	289
Spayed Female, Age 15~, RNF	284
Castrated Male, Age 12~14, MVI	271

2-3 반려동물 생애주기 분류

표준화된 진료데이터의 통계적 분석과 빈발패턴 분석을 통해 반려동물의 품종과 성별보다 생애주기별 질환의 차이가 뚜렷한 것을 알 수 있었다. 이 장에서는 반려동물의 생애주기를 고성능으로 예측할 수 있는 머신러닝 모델을 검색하기 위해, 진료데이터로부터 특징을 추출하여 학습데이터를 생성하고 최적화된 DBN 모델을 구축하였다.

1) 학습 데이터

모든 전자차트데이터는 아파치 루씬(Apache Lucene) 기반의 오픈소스 분산 검색 엔진인 Elasticsearch에 의해 웹시스템에 저장된다. Elasticsearch는 데이터베이스와 달리 인덱싱을 통해 빠르게 데이터를 저장하고 검색할 수 있다. 또한, 대용량 데이터의 분산처리 기능을 통해, 신속하게 통계적 분석을 실행하고 머신러닝 모델을 학습할 수 있다.

본 논문은 반려동물 전자차트데이터를 빠르고 정확하게 분석할 수 있지만 복잡한 전처리작업을 필요로 하지 않는 고성능의 분류모델을 구축하려고 한다. 따라서 전체 68가지 견종과 4가지 성별, 10가지 질환, 201가지 처방약 정보를 이용하여 모델을 학습하고 모델의 성능을 평가하였다.

대부분의 처방약은 질환을 치료하려는 목적의 치료약과, 소화제와 같이 질환과 무관하게 처방되는 일반약으로 구성되어 있다. 이 연구는 처방약 데이터가 반려동물의 생애주기를 예측하는 데에 중요한 정보인지 확인하기 위해 생애주기별 처방약의 분산분석을 실행하였다. 그 결과 고령 반려동물에게 발병율이 높은 심장과 신장 질환을 타겟으로 하는 치료약이 반려동물의 생애주기 예측에 중요한 특징으로 확인되었다 (Table 5).

빈발패턴 분석을 통해 생애주기와 관련된 견종, 성별, 질환, 처방약 사이의 패턴이 다양하게 검색되었다. 반려동물의 생애주기를 더 정확하게 예측하기 위해, 반려동물의 나이를 제외한 데이터에서 최소 지지도 0.01을 만족하는 179개 빈발 패턴을 검색하여 학습데이터에 추가하였다. 또한 주성분 분석 (PCA; principal components analysis)[19]을 통해 모든 데이터로부터 10개 주성분을 추출하여 모델 학습에 이용하였다.

따라서 반려동물의 품종과 성별, 질환데이터 뿐만 아니라 생애주기와 관련된 처방약과 빈발패턴을 추가함으로써, 고성능의 반려동물 생애주기 분류모델을 구축할 수 있을 것이다.

표 5. 반려동물 연령대 분류에 중요한 처방약
Table 5. Significant drugs for life cycle classification of companion animals

Prescription drugs	F-value
Furosemide	1,688.34
Pimobendan	1,493.28
Enalapril	1,076.04
Torsemide	700.65
Itraconazole	299.67

2) 분산처리 가능한 Deep Belief Network 기반 분류모델

DBN은 입력층 (input layer)과 은닉층 (hidden layer)으로 구성된 restricted Boltzmann machine (RBM)을 multi-layer perceptron (MLP) 구조로 쌓은 딥러닝 모델이다. 오류역전파 (error back-propagation) 알고리즘을 적용한 심층 신경망은 출력층보다 입력층의 가중치가 잘 교정되지 않는 오차소멸 (vanishing gradient) 문제를 가진다. 이 문제를 해결하기 위해, 층을 쌓으면서 가중치 (weight)를 계산하는 층별 사전훈련 (layerwise pre-training) 방법을 적용한 DBN이 제안되었다. DBN은 입력층 데이터만으로 사전훈련하는 RBM의 은닉층 값을 다음 단계 RBM의 입력층에 전달하는 비감독학습 (unsupervised learning) 구조이다. 하지만, 분류 문제를 풀기 위해 MLP의 가중치를 역으로 학습 및 조정해 나가는 감독학습을 이용하였다.

Fig. 2과 같이 RBM1은 입력 벡터 (V)로부터 사전훈련하여 은닉층 (h_1)을 학습하고 가중치 (w_1)를 산출한다. 그리고 RBM1의 은닉층 (h_1)을 RBM2의 입력벡터 (h_1)로 취급하여 다음 은닉층 (h_2)을 학습한다. 결국 이 과정을 n번 반복하면 n개 RBM의 은닉층 (h_1, h_2, ... ,h_{n-1}, h_n)과 가중치 (w_1, w_2, ..., w_{n-1}, w_n)로 구성된 DBN을 생성한다. 마지막으로 반려동물의 생애주기를 분류하기 위해, 오류 역전파 알고리즘을 통해 출력층부터 입력층까지 생성되어있는 가중치들을 조정하여 DBN을 학습한다. 따라서 감독학습 방법을 추가하여 가중치를 미세조정 (fine-tuning)함으로써, 분류모델의 성능을 향상시킨다.

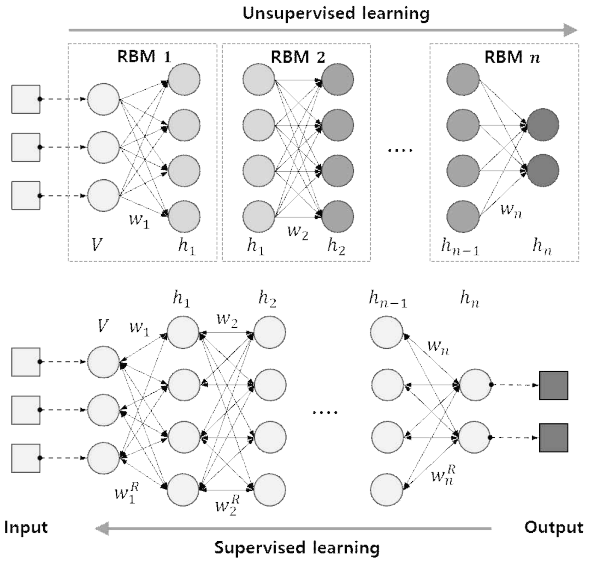

그림 2. DBN 구조와 학습 과정
Fig. 2. DBN structure and learning process

그러나 DBN의 미세조정 과정은 학습 데이터의 사이즈가 클수록 더 오랜 시간이 필요하다. 또한, 대용량의 전국단위 동물병원 전자차트 데이터를 분석하기 위해 효율적으로 처리할 수 있는 기법이 필요하다. 따라서 본 논문에서는 기존의 DBN 알고리즘의 확장성 및 통합 진료데이터의 분산병렬처리를 위해 MapReduce[20] 기법을 적용한다.

MapReduce는 두 함수 Map과 Reduce로 구성되고, 정수, 실수, 문자열, 바이트열 또는 임의 구조 형태로 (key, value) 쌍의 자료구조를 가진다. 대용량 데이터를 고정 크기의 블록으로 분할하여 여러 노드에 분산-입력하고, 각 노드 내에서 동일한 Map 함수가 병렬적으로 실행한다. Map 함수는 데이터를 변형 (transformation)하여 (key, value) 결과를 출력하고, key를 기준으로 결과를 정렬 및 병합하여 동일한 key를 가지는 (key, value) 쌍 그룹들을 반환한다. 각 그룹에 동일한 Reduce 함수를 적용하여 병렬적으로 데이터를 집계연산 (aggregation)하고, 산출된 결과를 분산 파일시스템에 기록한다.

DBN 학습 과정 중 미세조정이 가장 많은 시간을 소비하기 때문에[21], 미세조정 과정을 MapReduce를 이용하여 병렬적으로 실행하였다. Fig. 3과 같이 모든 진료데이터에 동일한 Map함수를 적용하여 DBN의 새로운 가중치를 산출하고, 사전훈련 시 생성된 기존 가중치와의 오차를 계산한다. 동일한 Reduce 함수가 진료데이터로부터 산출된 각 가중치의 오차 합을 계산하고, 합계된 가중치 값을 이용하여 DBN을 학습한다.

이처럼 오류역전파 알고리즘의 분산처리를 통해 효율적으로 반려동물 생애주기 분류기를 구축하는 방법을 제시한다. 그리고 사용자가 학습데이터와 DBN 구조를 입력하여 분류기를 구축할 수 있는 기능을 웹분석시스템에서 이용할 수 있다.

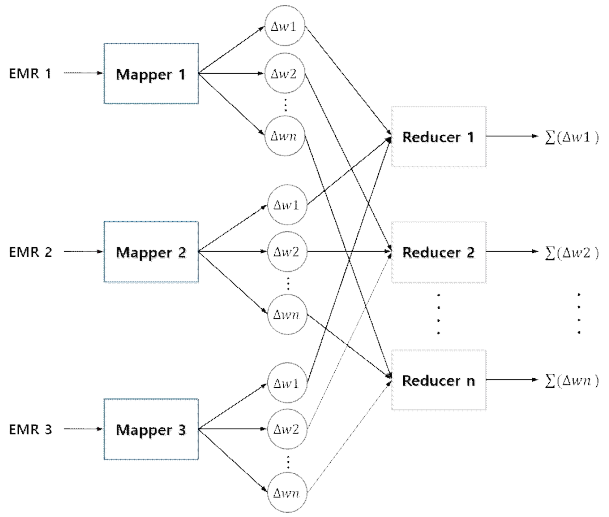

그림 3. MapReduce를 이용한 오류역전파 알고리즘 과정
Fig. 3. The process of error back-propagation using MapReduce

3) DBN 모델 성능평가

본 논문에서 제안한 MapReduce 기반 DBN 알고리즘의 생애주기 분류모델의 실행시간 평가를 위해서, 아마존 웹서비스(AWS)의 Elastic MapReduce 분산/병렬처리 환경에서 실험하였다.

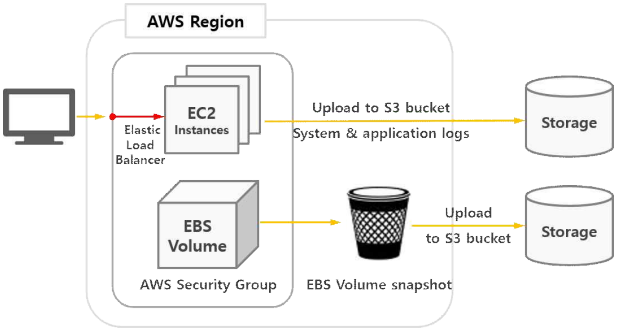

그림 4. AWS 분산처리 환경
Fig. 4. Distributed computing architectures with AWS

Elastic MapReduce에 DBN 알고리즘과 반려동물 학습데이터를 업로드한 후, 가상 인스턴스(EC2)를 1에서부터 최대 16가지 생성하였다. 단일머신에서 모델을 학습할 때의 시간과 2^1, 2^2, 2^3, 2^4대의 머신에서 동시에 분산/병렬 학습할 때의 실행시간을 비교하였다. AWS의 EC2 인스턴스를 생성하여 분산 및 병렬처리하는 과정은 Fig. 4와 같다.

3개의 은닉층 (각 은닉층의 300개, 150개, 30개 노드)으로 구성된 DBN이 다수의 머신을 이용하여 분산/병렬처리 하였을 때 소요된 시간을 비교한 결과는 Fig. 5와 같다. 1개의 노드에서 16,139개 전자차트데이터의 75%를 이용하여 DBN을 10번 반복 (iteration=10)하여 학습하고 25% 데이터를 테스트하는 방법을 4번 실행 (4-fold cross validation)한 경우, 실행시간 126초가 경과하였다. 노드의 수가 증가할수록 실행 시간이 감소하는 것을 확인할 수 있고, 16개의 노드를 이용하여 분산처리할 경우 가장 빠른 시간 (27초)이 소요되었다. Fig. 6은 각 프로세스가 노드 4개와 16개에서 분산/병렬처리 되는 과정과 소요되는 시간을 보여준다.

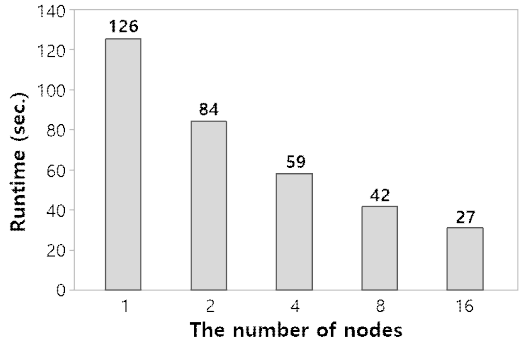

그림 5. 노드 수에 따른 DBN 학습시간
Fig. 5. DBN runtime by the number of nodes

그림 6. 다중 노드를 이용한 분산처리 화면
Fig. 6. A screenshot of distributed processing with multiple nodes

마지막으로 DBN 기반 반려동물 생애주기 분류기의 성능을 4가지 분류성능지표를 이용하여 평가하였다. 생애주기 6가지 범주 (A: 0~2세, B: 3~5세, C: 6~8세, D: 9~11세, E: 12 ~ 14세, F: 15세 이상)의 예측성능에 대한 정확도 (accuracy)와 정밀도 (precision), 민감도 (recall), 조합평가 지표인 F_1–value의 식은 다음과 같다.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = TP/(TP + FP) \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

(※ TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative)

DBN은 오차율 (MAE; mean absolute error) 0.069와 전체 정확도 88%, F_1-value 0.76 성능으로 반려동물 생애주기를 분류하였다. 그리고 Table 6처럼 각 생애주기의 정확도를 확인한 결과, C (6~8세)와 F (15세 이상) 생애주기가 가장 좋은 성능으로 분류되었고 A(0~2세)와 E (12~14세) 생애주기의 예측률이 가장 낮았다. Table 7의 혼잡매트릭스를 통해 실제 생애주기와 예측된 결과의 매칭 정보를 확인할 수 있다. 15세 이상의 전자차트데이터는 모두 정확하게 예측되었고, 12~14세 데이터가 0~2세로 오분류된 비율이 높은 것을 볼 수 있다. 또한 0~5세 데이터가 12~14세로 잘못 예측되는 경향을 보인다.

웹분석시스템은 전자차트데이터를 기반으로 구축된 반려동물 생애주기 분류모델을 이용하여 사용자의 입력데이터를 예측할 수 있는 기능을 제공한다. 그리고 사용자가 입력한 데이터와 파라메터를 기반으로 DBN을 학습하고 결과를 다운로드 받을 수 있다.

표 6. 반려견 생애주기 분류모델의 성능평가
Table 6. Performance for life cycle classifier of companion animal

	Precision	Recall	F_1-value
A (age 0 ~ 2)	0.671	0.873	0.759
B (age 3 ~ 5)	0.745	0.845	0.792
C (age 6 ~ 8)	0.993	0.726	0.839
D (age 9 ~ 11)	0.738	0.878	0.802
E (age 12 ~ 14)	0.693	0.432	0.532
F (age 15 ~)	0.995	1.000	0.998
MAE	0.069		

표 7. 반려견 생애주기 분류행렬
Table 7. Confusion matrix for classification result

Actual class	Predicted class					
	0 ~ 2	3 ~ 5	6 ~ 8	9 ~ 11	12 ~ 14	15 ~
0 ~ 2	87.29%	0%	0.42%	0%	12.29%	0%
3 ~ 5	0%	84.54%	0%	0%	14.98%	0.48%
6 ~ 8	3.85%	12.98%	72.6%	0.48%	4.81%	5.29%
9 ~ 11	0.98%	9.76%	0%	87.80%	0%	1.46%
12 ~ 14	40.36%	5.45%	0.45%	0%	44.18%	9.55%
15 ~	0%	0%	0%	0%	0%	100%

Ⅲ. 결 론

국내 반려가구의 수가 증가함에 따라 반려동물 양육에 대한 경제적 부담이 증가하고 있다. 반려동물의 진료데이터를 통해 품종별, 성별, 생애주기별 건강정보를 분석하여 반려동물의 건강상태를 이해하고 더욱 체계적으로 관리할 필요가 있다. 또한 수의학적 발달에 의한 반려동물의 고령화 사회에서 반려동물의 질환을 예방할 수 있는 사회적 시스템이 요구되고 있다.

본 논문에서는 동물병원에서 진료를 받은 반려동물의 정보와 전자차트데이터의 생애주기와 질환명, 처방약 등을 표준화하고, 반려동물의 품종, 성별, 생애주기 비율과 그에 따라 발병하는 질환을 분석하였다. 그 결과, 신부전과 심장판막의 이상은 고령 반려동물의 대표적인 질환인 것을 확인하였다. 그리고 전체 표준화된 전자차트데이터에서 빈발패턴을 분석하여, 고령 수컷견은 신부전이 많이 발생하고 고령 암컷견은 심장판막의 이상이 많이 발병하는 것을 검색하였다. 진료데이터와 빈발패턴 정보를 통합하여 반려동물의 생애주기를 분류하는 DBN 기반 모델을 구축하였다. Fig. 7과 같이, 전자차트데이터의 분산저장 및 고속 검색이 가능한 데이터 저장 시스템을 개발하고, 통계 및 빈발패턴 분석과 분산/병렬처리가 가능한 분류 모델을 웹시스템에 구현하여 사용자가 반려동물 진료데이터를 분석할 수 있도록 제공하였다.

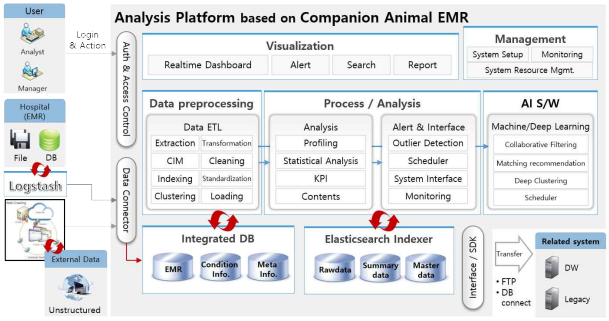

그림 7. 반려동물 진료데이터 분석시스템 아키텍처
Fig. 7. System architecture for analysis of companion animal EMR

본 연구 결과를 바탕으로, 반려동물의 신체정보나 처방약의 용량, 백신접종 유무 등의 정보를 추가하여 분류기의 성능을 높이고, 더 다양한 시각으로 전자차트데이터를 분석할 것이다. 그리고 협업필터링 등의 알고리즘을 추가로 적용하여 의료추천 시스템을 개발할 것이다.

감사의 글

본 연구는 2019년도 농촌진흥원의 지원 (PJ013951)에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] 문화체육관광부와 농촌진흥청. 2018년 반려동물에 대한 인식 및 양육 현황 조사 보고서[Internet]. Available: http://www.korea.kr/archive/expDocView.do?docId=38280&call_from=rsslink.

[2] 농림축산식품부. 반려동물 연관산업 분석 및 발전방향 연구 [Internet]. Available: http://www.prism.go.kr/homepage/entire/retrieveEntireDetail.do?research_id=1543000-201600024.

[3] 박주연. “반려동물 의료체계의 문제점 및 제도개선 방안”, 환경법과 정책, Vol. 19, pp. 99-130, 2017.

[4] B. N. Bonnett and A. Egenvall, “Age patterns of disease and death in insured Swedish dogs, cats and horses,” *Journal of Comparative Pathology,* Vol. 142, pp.33–38, 2010.

[5] A. Egenvall, B. N. Bonnett, A. Hedhammar, and P. Olson, “ Mortality in over 350,000 insured Swedish dogs from 1995–2000: II. Breed-specific age and survival patterns and relative risk for causes of death,” *Acta Veterinaria Scandinavica,* Vol. 46(3), pp. 121–136. 2005.

[6] L. Asher, G. Diesel, J. F. Summers, P. D. McGreevy, and L. M. Collins, “Inherited defects in pedigree dogs. Part 1: disorders related to breed standards,” *Veterinary Journal,* Vol. 182(3), pp. 402–411, 2009.

[7] C. R. Dorn, “Canine Breed-Specific Risks of Frequently Diagnosed Diseases at Veterinary Teaching Hospitals,” *AKC Canine Health Foundation*, 2000.

[8] W. H. Miller, C.E. Griffin, K.L.Campbell, and G.H. Muller, “Muller and Kirk’s Small Animal Dermatology,” *Maryland Heights: Elsevier Health Sciences*, 2013.

[9] M. V. Kustritz, “Determining the optimal age for gonadectomy of dogs and cats,” *Journal of the American Veterinary Medical Association*, Vol. 231(11), pp.1665–1675, 2007.

[10] K. Sorenmo, “Canine mammary gland tumors,” *Veterinary Clinics: Small Animal Practice*, Vol. 33(3), pp. 573–596, 2003.

[11] G. Zur, P. J. Ihrke, S. D. White, and P. H. Kass, “Canine atopic dermatitis: a retrospective study of 266 cases examined at the University of California, Davis, 1992–1998. Part I. clinical features and allergy testing results,” *Veterinary Dermatology*, Vol. 13(2), pp. 89–102, 2002.

[12] G. Zur, B. Lifshitz, and T. Bdolah-Abram, “The association between the signalment, common causes of canine otitis externa and pathogens,” *Journal of Small Animal Practice*, Vol. 52(5), pp. 254–258, 2011.

[13] D. G. O’Neill, D. B. Church, P. D. McGreevy, P. C. Thomson, and D. C. Brodbelt, “Longevity and mortality of owned dogs in England,” *Veterinary Journal*, Vol. 198(3), pp. 638–643, 2013.

[14] M. Martini, R. Busetto, R. Cassini, M. Drigo, C. Guglielmini, I. Masiero, M. L. Menandro, D. Pasotto, and M. Fenati, “A surveillance system of diseases of small companion animals in the Veneto Region (Italy),” *International Journal of Infectious Diseases*, Vol. 53, pp. 117, 2016.

[15] P. Cihan, E. Gökçe, and O. Kalıpsız, “A Review of Machine Learning Applications in Veterinary Field,” *Kafkas Univ Vet Fak Derg*, Vol. 23 (4), pp. 673–680, 2017.

[16] G. Hinton, “Deep belief networks,” *Scholarpedia*, Vol. 4(5), pp.5947, 2009.

[17] G. E. P. Box, “Non-Normality and Tests on Variances,” *Biometrika*, Vol. 40(3/4), pp. 318–335, 1953.

[18] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Record*, Vol. 22(2), pp.207–216, 1993.

[19] K. Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space,” *Philosophical Magazine*, Vol. 2(11), pp. 559–572, 1901.

[20] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Communications of the ACM (CACM)*, Vol. 51(1), 2008.

[21] K. Sun, X. Wei, G. Jia, R. Wang, and R. Li, “ Large-scale Artificial Neural Network: MapReduce-based Deep Learning,” *arXiv*, 2015.

유기진 (Kijin Yu)

2008년 : 충북대학교 대학원 전자계산학과 (공학석사)
2019년 : 충북대학교 대학원 컴퓨터과학과 (공학박사)

2008년～2010년: 한국생명공학연구원
2010년～2013년: 질병관리본부 국립보건연구원
2017년～2019년: 아산생명과학연구원
2019년～현 재: ㈜가이온 빅데이터연구소
※관심분야 : 데이터마이닝, 머신러닝, 데이터베이스, 바이오인포매틱스, 바이오메디컬 등

이영석 (YoungSeok Lee)

2019년 : 건국대학교 대학원 수의학과 (수의학석사 수료)

2018년~현 재: 건국대학교 대학원 수의외과학
※관심분야 : 소동물 임상수의학

이헌규 (Heon Gyu Lee)

2004년 : 충북대학교 대학원 전자계산학과 (이학석사)
2009년 : 충북대학교 대학원 전자계산학과 (공학박사)

2009년~2015년: 한국전자통신연구원
2016년~현 재: ㈜가이온 빅데이터연구소장
※관심분야 : 데이터마이닝, 머신러닝, 인공지능, 바이오인포매틱스, 바이오메디컬, 데이터베이스, 빅데이터 등