Journal of the Korean Data Analysis Society (October 2019)
Vol. 21, No. 5, pp. 2365-2378

지방자치단체 동물보호소의 유기·유실 반려견에 대한 입양확률예측모형*

최성은[1], 유현선[2], 정희운[3], 정희원[4], 박유미[5], 이관제[6]

요 약

농림축산검역본부가 운영하는 동물보호관리시스템 홈페이지에 유기나 유실된 반려동물에 대한 공고들 중에서 2014년부터 2018년까지 반려견의 자료 32만여 건을 크롤링(crawling)하여 전처리하였다. 이 과정에서 분석에 사용할 속성변수들을 정의하였다. 유기나 유실되어 동물보호소에 보호되고 있는 반려견의 입양성공에 영향을 미치는 요인을 알아내고, 입양 확률을 예측하기 위해서 입양확률예측모형을 구축하였다. 전처리된 자료들은 먼저 k-프로토타입(k-prototype) 군집법을 이용하여 두 개의 군집으로 나누어 분석하였다. 두 군집에서 속성별 입양 비율을 분석하였다. 각 모형의 오즈비(odds ratio)와 오즈비의 95% 신뢰구간을 이용하여 속성변수의 수준(level)간의 입양성공에 대한 영향력을 비교하였다. 두 집단의 입양확률예측모형에 대한 정확도, 민감도, 특이도와 이들의 95% 신뢰구간을 구하였고, 모형의 ROC 곡선을 구하였다. 입양확률예측모형과 대체확률(threshold value)을 이용한 모의실험을 통하여 동물보호관리시스템 홈페이지이나 다른 반려견 입양관련 단체에서 입양을 활성화하는 방안을 제시하였다. 국외 반려견 관련 연구의 영향력 속성변수와 비교·논의하였다.

주요용어 : 반려동물, 크롤링, 로지스틱, k-프로토타입, 군집화.

1. 서론

우리나라에서 매년 유기나 유실되는 반려견의 수가 증가하고 있다. 농림축산검역본부(APQA, 2019)에 따르면 유기나 유실되는 반려동물은 2017년 10만 마리를 넘어서더니 2018년 12만여 마리로 전년 대비 18% 증가했다. 유기나 유실 문제의 예방과 효율적인 해결을 위한 방안으로써, 반려동물의 식별번호가 담긴 인식표를 반려견의 목에 걸거나 체내에 삽입하는 '동물등록제'를 2008년부터 2012년까지 지방자치단체에서 자율적으로 시행되다가 2013년 이후 전국적으로 의무화됐다. 농림축산검역본부는 2008년 동물등록제 시행 후 2018년까지 전국에서 총 130만여 마리의 반려견이 등록됐다고 밝혔다. 전국 반려견을 680만 마리로 추정하면 2018년까지 등록된 반려견 비율은 19.1% 수준에 그쳤다. 이 정책은 권장사항이 아닌 필수사항임에도 불구하고 필요성을 느끼지 않거나 해당 제도를 알지 못한다는 이유로 등록제를 활용하지 않고 있다.

---

*본 논문은 농촌진흥청 반려동물 연구 사업단(세부과제번호:PJ0139862019) 연구사업의 지원에 의해 이루어진 것임.
[123456]04620 서울특별시 중구 필동로 1길 30, 동국대학교 통계학과.
[1]동국대학교 강사, 동국대학교 빅데이터연구센터 전문연구원. E-mail : c6300@hanmail.net
[2]석사. E-mail : yyhhss777@naver.com
[3]석사과정. E-mail : jhw941217@gmail.com
[4]석사과정. E-mail : drave11@naver.com
[5]석사과정. E-mail : byumm315@naver.com
[6](교신저자) 교수. E-mail : kwanlee@dongguk.edu
[접수 2019년 9월 20일; 수정 2019년 10월 17일; 게재확정 2019년 10월 20일]

반려동물이 유기나 유실되어 발견되었을 경우에 전국 시, 군, 그리고 구청 등에서 운영하는 동물보호소(animal shelter)에서 보호된다. 농림축산검역본부가 운영하는 동물보호관리시스템(animal protection management system)에서는 위 지방자치단체가 직영하는 동물보호소에서 보호 중인 개와 고양이들에 대한 공고를 하고 있다. 이 시스템의 홈페이지에는 연 8만여 마리의 반려동물이 공고되고, 그 중에서도 개는 60~70%이상을 차지한다. 이 시스템은 홈페이지에 10일 동안 주인을 찾는 공고를 올려 둔다. 공고기한이 지나면 국가법에 따라 해당 관할지역의 소유가 되어 새로운 주인의 입양을 기다리거나 자연사 또는 안락사 된다. 동물보호소에서 안락사 되지 않더라도 보호소의 한정된 관리비용 때문에 보호환경이 열악하여 동물들이 건강한 상태로 유지되기 어렵다는 연구가 있다(Protopopova, Wynne, 2014).

불필요한 안락사나 열악한 환경으로 인한 병사 또는 자연사를 방지하기 위해서는 공고기한 내에 주인에게 돌아가지 못한 반려동물의 입양을 활성화하는 방법을 찾아야 한다. 국외에서 유기동물과 유실동물의 입양을 활성화시키기 위해서 수행한 연구에 로지스틱 회귀모형을 적용하였다(Lepper et al., 2002; Hill, Murphy, 2016; Yoo, 2019). Lepper et al.(2002)의 연구는 로지스틱 회귀모형을 사용하였으나 속성변수들의 전처리방법이 본 연구와 상이하다. 로지스틱 회귀모형은 범주형 자료가 종속변수일 때 사용되는 모수적 통계법으로써 독립변수들이 연속형(Kang et al., 2014)이거나 범주형 자료일 때 사용할 수 있으며, 연속형과 범주형이 혼재해 있는 경우의 최적화(In et al., 2009; Ryu, 2017)에도 사용할 수 있다(Lee et al., 2017). 로지스틱 회귀모형은 지도학습(supervised learning)의 분류방법으로도 사용된다(Kweon, 2010).

본 연구에서는 동물보호관리시스템 홈페이지에 연 8만여 건 올라오는 반려동물에 대한 공고들 중에서 반려견의 자료를 크롤링하여 입양여부에 영향을 미치는 요인을 알아내고, 입양확률을 예측하기 위해서 입양확률예측모형(adoption probability prediction model)을 구축한다. 그리고 입양성공요인 분석과 입양확률예측모형을 이용한 모의실험을 통하여 입양을 활성화 할 수 있는 몇 가지 제언을 한다.

반려견의 특성에 따른 입양 여부를 예측하는 분류모형을 학습함으로써 입양확률이 높은 반려견들의 공고문을 홈페이지에서 더 효과적이게 노출시킬 수 있게 한다. 이로 인하여 입양 가능성이 높은 유기 또는 유실된 반려견들이 보호소에서 머무는 시간을 단축시킬 수 있다. 그래서 얻어지는 경제적 효과를 입양되지 못한 나머지 반려견들에게 더 나은 환경을 제공하는 예산으로 확보하거나 또는 안락사를 늦추어 입양기회를 더 줄 수 있다.

본 논문의 구성은 다음과 같다. 1장 서론에서는 유기나 유실된 반려견의 현황 및 동물등록제 그리고 농림축산검역본부의 동물보호시스템에 대해서 소개하고, 2장에서는 자료와 전처리에 대하여 설명한다. 3장에서는 k-프로토타입을 이용한 군집분석 방법을 소개한다. 군집분석에서 얻은 두 개의 군집에 대하여 로지스틱 회귀모형을 이용하여 입양확률예측모형을 구축하고, 입양확률예측모형과 대체확률을 이용한 모의실험을 한다. 4장에서는 입양확률을 향상시키기 위한 홈페이지 운영방법을 논의한다.

2. 자료와 전처리

2014년부터 2018년까지 농림축산검역본부에서 운영하는 동물보호관리시스템의 홈페이지(http://www.animal.go.kr/portal_rnl/index.jsp)에 게시된 공고문들 중에서 반려견에 관한 자료 342,574개를 크롤링하였다. 크롤링에 파이썬(Python) 프로그램의 Beautiful Soup 라이브러리를 이용하였다. 이 동물보호시스템 홈페이지에 올라와 있는 유기 또는 유실된 반려동물의 공고에서는 공고번호, 품종, 색

상, 성별, 중성화 여부, 나이, 체중, 접수 일시, 발생장소, 특징, 공고기한, 보호센터이름, 보호장소, 관할기관, 현재 상태 등의 정보가 있다. 이들 중에서 도시명, 품종, 색상, 성별, 중성화 여부, 나이, 체중, 그리고 상태라는 8개 항목을 선택하였다. 상태라는 항목은 동물보호소에서 현재 반려견의 상태 - 보호중, 반환, 입양, 안락사, 방사, 자연사 - 를 말한다. 8개의 자료 중에서 2개 - 나이, 체중 - 이 연속형 자료이고 나머지 6개가 범주형 자료이다.

범주형 자료 중에서 도시명, 품종, 그리고 색상은 각각의 범주가 아주 많아서 직접 분석에 사용하기가 적합하지 않아 전처리가 필요하다. 도시명은 시군구 기준으로 세분화되어 있기 때문에 국가통계포털(KOSIS)에서 제공하는 2017년 인구조사결과를 참고하여 인구수가 100만 이상인 경우엔 거대도시, 50만 이상 100만 미만인 경우엔 대도시, 10만 이상 50만 미만인 경우엔 중소도시, 그리고 10만 미만인 경우엔 소도시로 재범주화 하였다. 수십 개의 범주를 갖고 있는 품종은 미국의 동물협회인 아메리칸 켄넬클럽(American Kennel Club, AKA : http://www.akc.org)에서 정의한 범주에 따라 8개 범주로 나누었고, '믹스견'이라고 입력되어 있는 반려견들은 'mix'라는 범주로 추가하여 최종 9개의 범주를 갖는 속성변수로 재범주화 하였다. 색상은 색이 섞여 있는 점을 고려하여 갈색, 흰색, 검정색, 검갈색, 흰갈색, 검흰색, 그리고 검흰갈색 등의 7개의 범주와, 표범무늬와 같은 무늬가 포함된 색상을 '기타'로 하여 총 8개의 범주를 갖는 색상변수를 만들었다. 성별은 암컷과 수컷이고 중성화 여부는 예, 아니오, 미상 등이 존재하는데 미상을 결측치로 처리하였다. 그리고 상태변수는 보호중, 반환, 입양, 안락사, 방사, 자연사 6개의 범주로 되어 있다. 이 상태변수의 6개 범주 중에서 반환과 방사된 반려견은 동물보호소에서 보호되고 있지 않으므로 본 연구목적에 맞지 않는다. 그러므로 반환이나 방사된 반려견들을 제외한 상태항목은 입양(adopted)과 나머지 보호중, 안락사, 자연사를 미입양(unadopted)으로 재분류하여 분석에 사용하였다. 전처리 된 8개 속성변수의 이름과 수준의 수는 adoption(입양여부, 2), breed(반려견의 품종, 9), citysize(도시의 규모, 4), color(반려견의 색, 8), gender(성별, 2), neutralization(중성화 여부, 2), age(연속형 자료), weight(연속형 자료)이

Table 1. Attribute names, levels and descriptions

Attributes	# of Levels	Levels(descriptions)	
Adoption	2	0(unadopted)	1(adopted)
Breed	9	Herding(breed for raising livestock)	Mix(mixed breed)
		Hound(breed for hunting)	Miscellaneous(unlicensed breed)
		Non-sporting(non-hunting breed)	Sporting(active breed helping to hunt)
		Terrier(breed to catch rodents)	Toy(small breed)
		Working(breed to help a person's work)	
Citysize	4	Metropolis(over a million)	Big city(500thousand and less than 100million)
		Medium city(more than 100thousand, less than 500thousand)	
		Small city(less than 100thousand)	
Color	8	Brown	Black&Brown
		Black	Etc
		White	White&Brown
		Black&White&Brown	White&Black
Gender	2	Male	Female
Neutralization	2	No	Yes
Age		1~20years old	(continuous variable)
Weight		0~100kg	(continuous variable)

다. 전처리 과정을 통하여 분석에 투입된 속성변수들의 이름(attributes name), 변수들의 수준의 수(number of levels)는 Table 1에 정리하였다.

반환이나 방사된 반려견의 자료, 결측치가 있거나 홈페이지 관리자에 의하여 입력오류로 판단된 자료, 그리고 크롤링할 시기에 보호소에 들어 온지 10일이 되지 않은 반려견(원주인을 찾을 수 있기 때문에)의 자료들은 제거하였다. 위와 같은 전처리 작업을 거친 후 사용하게 되는 자료의 수는 180,646개의 반려견의 자료이다.

3. 통계적 분석

전체 자료는 먼저 k-프로토타입 군집법을 이용하여 두 개의 군집으로 나눈다. 두 집단의 입양비율을 비교검정하기 위하여, 연속형 자료인 경우에는 독립 표본 t 검정(independent two sample t-test)

Table 2. Frequencies, adoption proportions, means and CI of each attribute level

Attributes	Levels of attributes	Adoption		Proportion of adoption	Total
		Unadopted	Adopted		
Breed*	Herding	265	1034	0.80	1299
	Mix	67495	36359	0.35	103854
	Hound	872	1883	0.68	2755
	Miscellaneous	5115	5835	0.53	10950
	Non-sporting	6978	9733	0.58	16711
	Sporting	1865	3108	0.62	4973
	Terrier	2719	2330	0.46	5049
	Toy	14337	17675	0.55	32012
	Working	1244	1799	0.59	3043
Citysize*	Metropolis	3068	6601	0.68	9669
	Big city	20155	21186	0.51	41341
	Medium city	60412	41759	0.41	102171
	Small city	17255	10210	0.37	27465
Color*	Brown	24386	19205	0.44	43591
	Black&Brown	8039	6100	0.43	14139
	Black	7669	6424	0.46	14093
	Etc	1991	1489	0.43	3480
	White	34779	28278	0.45	63057
	White&Brown	14453	10604	0.42	25057
	Black&White&Brown	3544	2630	0.43	6174
	White&Black	6029	5026	0.45	11055
Gender*	Male	59865	48037	0.45	107902
	Female	41025	31719	0.44	72744
Neutralization*	No	93731	71359	0.43	165090
	Yes	7159	8397	0.66	15556
Age*		3.46	2.71		
CI		(3.44, 3.47)	(2.69, 2.72)		
Weight+		6.23	6.2		
CI		(6.2, 6.27)	(6.15, 6.25)		
	Total	100890	79756	0.44	180646

Note. * indicates that the variable is significant at p <0.001
+ p-value for Independent two sample t-test'=0.2834

을, 범주형 자료인 경우에는 카이제곱검정(chi-square test) 또는 피셔의 정확검정(Fisher's exact test)을 이용하여 속성변수들의 입양비율을 비교검정 하였다. 또한 입양확률예측을 위한 통계모형을 구축하기 위하여 후진소거법(backward elimination method)을 이용한 다중 로지스틱 회귀모형(multiple logistic regression model)을 이용하였다. 먼저 전체 데이터의 70%를 학습용 데이터(train data)로, 30%는 평가용 데이터(test data)로 분류한 후 입양예측을 위한 통계모형을 구축하고 평가하였다. 통계분석을 위해 R version 3.6.1(R foundation for statistical computing, Vienna, Austria), SAS version 9.4(SAS Institute Inc., Cary, NC), 파이썬 version 3.6.0을 이용하였다.

3.1. 자료의 군집화

전처리된 반려견 자료들에 대한 입양확률예측모형을 구축하는 첫 단계로 각각의 속성변수의 특징에 따라 군집화한다. 분류분석이나 모형을 구축하기 전에 군집분석을 수행하면 분류나 모형이 최적화되어 분류나 모형의 성능이 또한 증가한다는 연구가 있다 (Rahman, Verma, 2013; Alapati, Sindhu, 2016; Szepannek, 2018). 다만 군집분석을 했을 때 각 군집의 특성이 직관적으로 이해가 쉽거나 군집 간의 차이가 유의해야 군집분석을 하는 의미가 있다. 그러므로 직관적이고 뚜렷한 특성을 갖도록 군집화를 하는 것이 중요하다. 자료를 구성하고 있는 속성변수들이 범주형 자료와 연속형 자료가 혼재해 있으므로 k-프로토타입 군집화(k-prototype clustering) 방법을 이용하여 군집화한다. 군집화된 반려견 집단은 Cluster1과 Cluster2로 정의한다.

연속형 자료의 군집에 사용되는 k-평균(k-means) 군집법(Hartigan, Wong, 1979)은 유클리드 거리 계산이 가능해야 한다. 범주형 자료에는 빈도수를 이용하는 k-최빈값(k-modes) 군집법을 적용한다. 자료가 연속형 자료와 범주형 자료가 혼재된 경우는 k-프로토타입 군집화(k-prototype clustering)를 이용한다. k-프로토타입 군집화(Huang, 1998)는 거리기반과 빈도수기반으로 군집화가 이루어진다. k-프로토타입 군집법에서 거리를 측정하는 식은 아래 식(3.1)과 같다.

$$d(x_i,\mu_j)=\sum_{m=1}^{q}(x_i^m-\mu_j^m)^2+\lambda\sum_{m=q+1}^{p}\delta(x_i^m,\mu_j^m), \tag{3.1}$$

p개의 속성변수들 중에서 q개의 연속형 변수와 $(p-q)$개의 범주형 변수가 있을 때, 관측치 x_i와 중심점 μ_j사이의 거리 $d(x_i,\mu_j)$는 연속형 변수의 유클리드 거리의 제곱 합과 범주형 변수의 불일치 거리의 가중합(weighted sum)으로 계산되어진다. 불일치 거리는 $\delta(\,)$로 계산되는데, 만일 $a=b$이면 $\delta(a,b)=0$ 그리고 만일 $a\neq b$이면 $\delta(a,b)=1$이다. λ값은 범주형 변수들의 거리가 미치는 영향의 정도를 조절해주는 모수이다. $\lambda=0$이면 범주형 자료의 영향은 없어지고 연속형 자료만 고려하는 경우가 되어 k-평균 군집법이 된다. λ값을 적절히 조절하여 연속형 변수의 영향을 크게 할지 범주형 변수의 영향을 크게 할지 지정해줄 수 있다. λ값은 연구자에 의해 주관적으로 지정할 수 있는데, 일반적으로는 표준화된 수치형 변수들의 분산의 평균값을 이용하여 계산된다(Szepannek, 2018).

k-프로토타입 군집분석의 결과로 만들어진 두 군집의 중심점은 Table 3과 같다. 7개의 속성변수 중에서 5개의 범주형 변수가 있고, 이들 5개의 범주형 속성변수 중에서 3개의 속성 - citysize, gender, neutralization - 에서 중심점이 같다. Table 4와 Table 5에 5개의 범주형 속성들에 대한 입양비율을 정리하였다. Cluster1에서는 herding의 입양비율이 0.81로 가장 높다. 그 다음이 hound와 metropolis 순으로 높게 나타난다. 또 mix와 small city에서 가장 낮은 입양비율이 관측된다(Table 4). Cluster2에서는 metropolis의 입양비율이 0.75로 가장 높다. 그 다음이 herding과 hound 순으로 높게 나타났다(Table 5).

Table 3. Centers of two k-prototype clustered groups

Cluster	Citysize	Breed	Color	Gender	Neutralization	Age	Weight
Cluster1	Medium city	Mix	Brown	Male	No	2.74	7.99
Cluster2	Medium city	Toy	White	Male	No	3.74	3.43

Table 4. Adoption proportions of k-prototype clustered groups - Cluster1

Cluster1		Adoption		Proportion of adoption	Total
		Unadopted	Adopted		
Breed	Mix	53382	28389	0.35	81771
	Herding	228	952	0.81	1180
	Hound	478	1074	0.69	1552
	Miscellaneous	3364	3111	0.48	6475
	Non-sporting	5042	6909	0.58	11951
	Sporting	1562	2670	0.63	4232
	Terrier	422	482	0.53	904
	Toy	86	142	0.62	228
	Working	800	1391	0.63	2191
Citysize	Metropolis	2130	3767	0.64	5897
	Big city	12227	11105	0.48	23332
	Medium city	39017	23501	0.38	62518
	Small city	11990	6747	0.36	18737
Color	Brown	23281	17650	0.43	40931
	Black&Brown	5237	3328	0.39	8565
	Black	5743	3895	0.40	9638
	Etc	1532	1130	0.42	2662
	White	10433	6022	0.37	16455
	White&Brown	11292	7229	0.39	18521
	Black&White&Brown	2767	1887	0.41	4654
	White&Black	5079	3979	0.44	9058
Gender	Female	27021	18895	0.41	45916
	Male	38343	26225	0.41	64568
Neutralization	No	62756	42109	0.40	104865
	Yes	2608	3011	0.54	5619
	Total	65364	45120	0.41	110484

3.2. 로지스틱 분류모형

k-프로토타입 군집화된 Cluster1과 Cluster2에 대하여 각각 다변량 로지스틱 회귀분석을 이용하여 반려견의 입양확률예측모형(adoption probability prediction model)을 구한다. 먼저 단계적 모형선택 방법인 후진소거법(backward elimination method)을 이용하여 최적모형을 선택한다. 모형선택은 아카이케 정보 기준(Akaike information criterion; AIC)을 선택기준으로 한다. Cluster1과 Cluster2에서의 단계적 후진소거법의 결과는 아래 Table 6과 Table 7과 같다. AIC에 근거하여 두 군집에서 모든 속성변수들이 포함된 모형을 최적의 다변량 로지스틱 회귀모형으로 결정되었다(Cluster1 model 1 AIC = 96447.91, Cluster2 model 1 AIC = 60816.59).

Cluster1과 Cluster2의 각 속성별 오즈비(odds ratio, OR)는 Table 8과 Table 9와 같다. 속성변수 citysize의 모든 오즈비는 두 군집에서 유의적($p<0.001$)이며, metropolis를 기준으로 한 다른 도시의

오즈비는 모두 1보다 작다. 그러므로 인구가 100만 이상 도시인 metropolis에서 입양성공률이 가장 높다.

Table 4에서 보면 Cluster1에서의 metropolis의 입양비율은 0.64이다. Table 5에서는 Cluster2의 metropolis의 입양비율은 0.75로 Cluster1에서 보다 높다. 품종변수인 breed는 Cluster1에서는 모든 오

Table 5. Adoption proportions of k-prototype clustered groups - Cluster2

Cluster2		Adoption		Proportion of adoption	Total
		Unadopted	Adopted		
Breed	Mix	14113	7970	0.36	22083
	Herding	37	82	0.69	119
	Hound	394	809	0.67	1203
	Miscellaneous	1751	2724	0.61	4475
	Non-sporting	1936	2824	0.59	4760
	Sporting	303	438	0.59	741
	Terrier	2297	1848	0.45	4145
	Toy	14251	17533	0.55	31784
	Working	444	408	0.48	852
Citysize	Metropolis	938	2834	0.75	3772
	Big city	7928	10081	0.56	18009
	Medium city	5265	3463	0.40	8728
	Small city	21395	18258	0.46	39653
Color	Brown	1105	1555	0.58	2660
	Black&Brown	2802	2772	0.50	5574
	Black	1926	2529	0.57	4455
	Etc	459	359	0.44	818
	White	24346	22256	0.48	46602
	White&Brown	3161	3375	0.52	6536
	Black&White&Brown	777	743	0.49	1520
	White&Black	950	1047	0.52	1997
Gender	Female	21522	21812	0.50	43334
	Male	14004	12824	0.48	26828
Neutralization	No	30975	29250	0.49	60225
	Yes	4551	5386	0.54	9937
	Total	35526	34636	0.49	70162

Table 6. Summary of variable significance and Akaike information criterion (AIC) values for multivariate logistic regression model for adoption success - Cluster1

	Age	Breed	Citysize	Color	Gender	Neutralization	Weight	AIC
Model1	√	√	√	√	√	√	√	96447.91
Model2	√	√	√	√	√		√	96490.1
Model3	√	√	√		√	√	√	96523.58

Table 7. Summary of variable significance and Akaike information criterion (AIC) values for multivariate logistic regression model for adoption success - Cluster2

	Age	Breed	Citysize	Color	Gender	Neutralization	Weight	AIC
Model1	√	√	√	√	√	√	√	60816.59
Model2	√	√	√	√	√		√	60862.21
Model3	√	√	√		√	√	√	60883.55

Table 8. List of OR and 95% CI - Cluster1

Attributes	Levels of attributes	Odds ratio	95% CI for odds ratio	p-value
Breed	Herding	1.00		
	Mix	0.11	[0.09, 0.14]	<0.001
	Hound	0.63	[0.5 , 0.78]	<0.001
	Miscellaneous	0.23	[0.19, 0.27]	<0.001
	Non-sporting	0.41	[0.34, 0.49]	<0.001
	Sporting	0.48	[0.4 , 0.59]	<0.001
	Terrier	0.37	[0.29, 0.48]	<0.001
	Toy	0.48	[0.33, 0.71]	<0.001
	Working	0.42	[0.34, 0.52]	<0.001
Citysize	Metropolis	1.00		
	Big city	0.47	[0.44, 0.51]	<0.001
	Medium city	0.33	[0.31, 0.35]	<0.001
	Small city	0.33	[0.3 , 0.35]	<0.001
Color	Brown	1.00		
	Black&Brown	0.97	[0.91, 1.04]	0.497
	Black	1.11	[1.04, 1.17]	<0.001
	Etc	1.15	[1.03, 1.27]	<0.001
	White	0.92	[0.87, 0.98]	<0.001
	White&Brown	1.13	[1.08, 1.18]	<0.001
	Black&White&Brown	1.03	[0.95, 1.11]	0.464
	White&Black	1.18	[1.11, 1.25]	<0.001
Gender	Male	1.00		
	Female	1.2	[1.17, 1.24]	<0.001
Neutralization	No	1.00		
	Yes	1.28	[1.19, 1.38]	<0.001
Age		0.62	[0.61, 0.64]	<0.001
Weight		1.09	[1.06, 1.11]	<0.001

즈비가 매우 유의적(p <0.001)이나, Cluster2에서는 한 가지 품종 mix(OR=0.15, CI=(0.09, 0.25))에서만 p<0.001이다. breed에서 herding을 기준으로 모든 오즈비가 1보다 같거나 작다. Herding의 입양비율이 0.81과 0.69로 제일 높고, hound가 두 군집에서 0.69와 0.67로 그 다음으로 높다(Table 4와 Table 5). mix의 오즈비가 0.11(CI=(0.09, 0.14))과 0.15(CI=(0.09, 0.25))로 가장 입양이 안 되는 것으로 나타났고, 이들의 입양비율도 0.35와 0.36으로 낮다. 두 군집의 중성화에 대한 오즈비가 1.28(CI=(1.19, 1.38))과 1.23(CI=(1.16, 1.31))으로 중성화 시술을 받은 반려견이 입양에 유리하다. 암컷인 반려견의 오즈비는 약 1.2로 입양비율이 높다. 연속형 속성변수 age와 weight의 오즈비는 1과 비슷하거나 1보다 작다. 나이가 많거나 몸무게가 나가는 반려견들은 입양성공에 유리하지 않다. 두 연속변수 age와 weight의 평균의 95% 신뢰구간이 겹치지 않으므로 두 군집의 모평균 간에 유의적인 차이가 있다(유의수준 $\alpha=0.05$).

로지스틱 회귀분석에서 정확도(accuracy), 민감도(sensitivity), 그리고 특이도(specificity)는 대체확률(threshold value)의 선택에 전적으로 결정된다. 로지스틱 회귀모형으로부터 추정된 입양확률이 대체확률보다 작으면 입양에 실패한 것으로 추정하고, 추정된 입양확률이 대체확률보다 크면 입양에 성공한 것으로 추정하여 계산된다. 어떠한 값을 대체확률로 선택하기 전에 세 통계량 중에서 어느 통계량이 연구목적에 맞는지를 고려하여야 한다. Hill, Murphy(2016)는 반려견의 입양성공에 관한

Table 9. List of OR and 95% CI - Cluster2

Attributes	Levels of attributes	Odds ratio	95% CI for odds ratio	p-value
Breed	Herding	1.00		
	Mix	0.15	[0.09, 0.25]	<0.001
	Hound	1.02	[0.59, 1.7]	0.935
	Miscellaneous	0.57	[0.33, 0.93]	0.031
	Non-sporting	0.73	[0.43, 1.2]	0.232
	Sporting	0.61	[0.35, 1.03]	0.073
	Terrier	0.62	[0.36, 1.02]	0.072
	Toy	0.64	[0.38, 1.04]	0.093
	Working	0.5	[0.29, 0.84]	0.012
Citysize	Metropolis	1.00		
	Big city	0.37	[0.33, 0.41]	<0.001
	Medium city	0.27	[0.25, 0.3]	<0.001
	Small city	0.22	[0.2 , 0.25]	<0.001
Color	Brown	1.00		
	Black&Brown	0.62	[0.54, 0.72]	<0.001
	Black	0.7	[0.62, 0.81]	<0.001
	Etc	0.59	[0.48, 0.72]	<0.001
	White	0.72	[0.65, 0.8]	<0.001
	White&Brown	0.64	[0.56, 0.72]	<0.001
	Black&White&Brown	0.54	[0.46, 0.64]	<0.001
	White&Black	0.72	[0.61, 0.84]	<0.001
Gender	Male	1.00		
	Female	1.19	[1.14, 1.24]	<0.001
Neutralization	No	1.00		
	Yes	1.23	[1.16, 1.31]	<0.001
Age		0.47	[0.46, 0.49]	<0.001
Weight		0.91	[0.89, 0.93]	<0.001

연구에서 세 통계량의 중요도에 대하여 논의하였고, 저자들의 연구에는 특이도를 중요시하겠다는 이유를 설명하였다. 우리나라에서는 입양요인분석이나 입양예측분석 연구결과가 동물보호소나 관련 홈페이지를 운용하는데 선순환하는 사례가 아직은 활발하지 않은 상황이다. 이 점을 고려하여 본 연구에서는 입양 또는 미입양의 한쪽에 가중하는 민감도나 특이도 보다는 입양과 미입양을 같이 고려하는 정확도를 중요시하고 이 정확도를 최대로 하는 대체확률을 선택하였다. 대체확률이 0.5일 때 두 모형의 정확도가 0.66(CI=(0.655, 0.666))과 0.665(CI=(0.659, 0.671))로 가장 큰 값이다. 이때의 {특이도, 민감도}는 각각 {0.868(CI=(0.863, 0.873)), 0.363(CI=(0.355, 0.371))}과 {0.693(CI=(0.684, 0.702)), 0.636(CI=(0.628, 0.645))}이다. Table 10과 Table 11에 여러 개의 대체확률에 따라 정확도, 민감도, 그리고 특이도와 이들의 부츠트랩 95% 신뢰구간을 제시하였다. Figure 1과 Figure 2에 두 입양확률예측모형의 ROC 곡선을 제시하였다.

3.3. 모의실험

위 입양확률예측모형을 이용하면 Table 12에 있는 속성을 갖는 3마리 반려견의 입양예측확률을 구할 수 있고, 이 예측확률과 대체확률 0.5를 비교하여 이 반려견들을 입양가능 또는 입양불가로

잠정 결정할 수 있다. Table 12의 세 반려견은 각각 추정된 입양확률에 따라서 Example1은 입양가능으로 나머지 두 마리는 입양불가로 잠정 평가된다. 입양불가 판정을 받은 Example2와 Example3에 대해서는 속성들의 다른 조합을 제시하여 입양을 독려할 수 있다.

Example3과 같은 속성조합을 갖는 반려견의 일부 속성을 바꾸면 어떻게 입양확률이 다르게 추정되는지를 모의실험을 하여 보자. Table 13에서 보는 바와 같이 Example3 반려견의 속성조합 중에서 나이를 7살에서 1살로 대체한 Example3-1 반려견의 추정입양확률은 0.353으로 증가한다. Example3 반려견을 같은 조건을 갖는 herding인 Example3-2 반려견으로 바꾸면 추정입양확률이 0.618로 증가한다. 다시 Example3-2 반려견을 나이가 1살인 Example3-3 반려견으로 바꾸면 추정입양확률이 0.822가 된다. 이와 같이 입양성공에 영향을 주는 속성들로 대체함으로서 입양확률을 증가시킬 수 있다.

4. 결론 및 논의

Lepper et al.(2002)의 미국 캘리포니아 Sacramento 동물보호소의 반려견들의 입양에 관한 연구결과와 본 연구결과를 비교하여 보면, 입양성공 요인에 대한 분석결과의 통계치에는 차이가 있으나 입양성공 요인들은 방향을 같이 한다. 예를 들면 두 연구에서 성별의 선호도에서 암컷 반려견이 수컷 반려견보다 선호된다(Sacramento의 OR=1.15, CI=(0.96, 1.39), 본 연구의 OR=1.2, CI=(1.17, 1.24)). Sacramento 자료에서도 중성화는 입양성공에 유리하게 나왔다(Sacramento의 OR=1.75 이상, 본 연구의 OR=1.28). 반려견의 색상에 대해서는 범주화 하는 방법과 기준들이 달라서 직접 비교가

Table 10. Accuracy, specificity, sensitivity and 95% bootstrap confidence intervals – Cluster1

	Accuracy	Specificity	Sensitivity
	95% bootstrap CI		
0	0.409	0	1
	0.406, 0.416	-	-
0.1	0.42	0.017	0.997
	0.415, 0.425	0.015, 0.019	0.996,0.998
0.2	0.45	0.084	0.975
	0.445, 0.456	0.08, 0.088	0.973, 0.978
0.3	0.521	0.28	0.866
	0.516, 0.526	0.273, 0.286	0.86, 0.872
0.4	0.636	0.685	0.565
	0.631, 0.641	0.678, 0.692	0.557, 0.574
0.5	0.66	0.868	0.363
	0.655, 0.666	0.863, 0.873	0.355, 0.371
0.6	0.651	0.927	0.255
	0.646, 0.656	0.924, 0.931	0.248, 0.262
0.7	0.62	0.976	0.111
	0.615, 0.626	0.973, 0.978	0.106, 0.117
0.8	0.6	0.995	0.035
	0.595, 0.606	0.994, 0.996	0.032, 0.038
0.9	0.59	0.999	0.002
	0.585, 0.594	-	0.002, 0.003
1	0.588	1	0
	0.584, 0.594	-	-

Table 11. Accuracy, specificity, sensitivity and 95% bootstrap confidence intervals – Cluster2

Threshold	Accuracy	Specificity	Sensitivity
	95% bootstrap CI		
0	0.492	0	1
	0.485, 0.499	-	-
0.1	0.504	0.027	0.997
	0.497, 0.511	0.024, 0.030	0.996, 0.998
0.2	0.536	0.108	0.978
	0.529, 0.543	0.102, 0.114	0.975, 0.981
0.3	0.579	0.237	0.931
	0.572, 0.585	0.229, 0.245	0.926, 0.936
0.4	0.64	0.494	0.791
	0.634, 0.647	0.484, 0.504	0.783, 0.799
0.5	0.665	0.693	0.636
	0.659, 0.671	0.684, 0.702	0.628, 0.645
0.6	0.647	0.807	0.481
	0.640, 0.653	0.799, 0.814	0.471, 0.491
0.7	0.587	0.925	0.239
	0.581, 0.594	0.920, 0.929	0.231, 0.247
0.8	0.531	0.989	0.058
	0.524, 0.538	0.987, 0.991	0.054, 0.062
0.9	0.511	0.999	0.006
	0.504, 0.518	-	0.005, 0.008
1	0.508	1	0
	0.502, 0.514	-	-

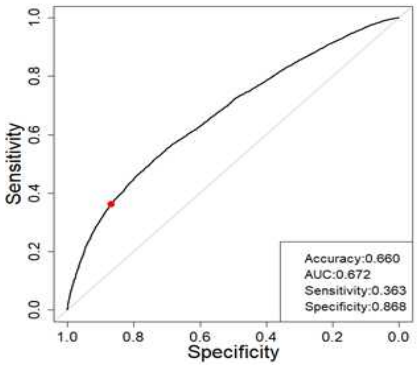

Figure 1. ROC curve of Cluster1

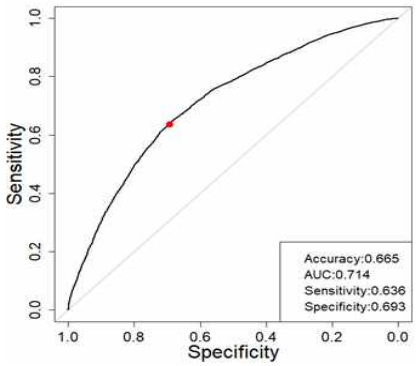

Figure 2. ROC curve of Cluster2

가능하지 않았다.

두 군집에서 citysize의 {metropolis, big city}의 입양성공률이 {medium city, small city}의 입양성공률 보다 높은 이유는 동물보호소의 접근성 때문인 것으로 사료된다. 반려견을 보호하고 있는 동물보호소까지 먼 거리로 인하여 또는 다른 이유로 접근이 어려운 경우에는 입양에 성공할 수 없다. 동물보호소의 위치가 입양에 영향을 미친다.

또 다른 외국의 반려견 입양에 관련한 연구에서는 본 연구에서 다루지 못했던 몇 가지 중요한 속성들이 있다. DeLeeuw(2010)에는 입양을 거절당하는 이유 중에서 40%가 반려견의 행동문제 때

Table 12. Hypothetical evaluation of adoption and attribute combination

Attributes	Example1	Example2	Example3
Breed	Herding	Mix	Mix
Citysize	Medium city	Medium city	Medium city
Color	Brown	White&Brown	Brown
Gender	Male	Female	Male
Neutralization	No	Yes	No
Age	1	1	7
Weight	25	6.2	6.2
Probability of adoption success	0.858	0.49	0.16
Potential adoption	Yes	No	No

Table 13. Hypothetical evaluation of adoption and attribute combination

Attributes	Example3-1	Example3-2	Example3-3
Breed	Mix	Herding	Herding
Citysize	Medium city	Medium city	Medium city
Color	Brown	Brown	Brown
Gender	Male	Male	Male
Neutralization	No	Yes	No
Age	1	7	1
Weight	6.2	6.2	6.2
Probability of adoption success	0.353	0.618	0.822
Potential adoption	No	Yes	Yes

문이고, 그 다음이 반려견의 건강문제이다(26%). 입양성공에 유리하게 작용하는 속성은 반려견이 교육을 받았는지 여부이다. 반려견의 행동, 건강문제, 그리고 교육여부에 대한 자료가 입양확률예측모형을 구축하는데 사용할 수 있다면 보다 정확한 모형을 구할 수 있을 것이다.

반려견 입양성공모형의 결과를 통하여 입양성공에 유리한 반려견의 속성들을 찾았고, 모의실험을 통하여 어떻게 이들 속성들을 이용하여 추정입양성공률을 높일 수 있는 방법을 알아보았다. 여기에 더하여 입양을 원하는 사람들이 고려하는 다른 속성 - 반려견의 행동, 건강문제, 교육여부 - 들을 동물관리보호시스템의 입양공고 사이트에 포함시키고, 모의실험의 제안된 방법을 프로그램화하여 적용시키면 입양성공률을 높일 수 있을 것이다.

References

Alapati, Y. K., Sindhu, K. (2016). Combining clustering with classification: A technique to improve classification accuracy, *International Journal of Computer Science Engineering (IJCSE),* 5(6), 336-338.

Animal and Plant Quarantine Agency (APQA) (2019). *The result of 2018 national awareness survey on animal protection,* Animal and Plant Quarantine Agency. (in Korean).

DeLeeuw, J. L. (2010). *Animal shelter dogs: Factors predicting adoption versus euthanasia,* Unpublished doctoral dissertation, Wichita State University, Wichita, Kansas, US.

Hartigan, J. A., Wong, M. A. (1979). A k-means clustering algorithm, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1), 100-108.

Hill, S. E., Murphy, N. C. (2016). Analysis of dog adoption success and failure using surveys with vignettes, *Journal of Applied Animal Welfare Science,* 19(2), 144-156.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery,* 2(3), 283-304.

In, O. N., Yoon, Y. S., Han, J. S. (2009). An analysis of airline selective factors through logistic regression and optimal scaling, *Journal of the Korean Data Analysis Society,* 11(2), 935-948. (in Korean).

Kang, G. M., Kim, K. K., Kang, C. (2014). A study of cancer-related gene exploration using PCA logistic regression, *Journal of the Korean Data Analysis Society,* 16(3), 1241-1248. (in Korean).

Kweon, Y. R. (2010). The comparative analysis of predictors of suicidal ideation on middle school students using decision tree and logistic regression, *Journal of the Korean Data Analysis Society*, 12(6), 3103-3115. (in Korean).

Lee, B. E., Joo, Y. S., Jung, H. J. (2017). Rare bankruptcy event prediction with missing data, *Journal of the Korean Data Analysis Society,* 19(1), 129-139. (in Korean).

Lepper, M., Kass, P. H., Hart, L. A. (2002). Prediction of adoption versus euthanasia among dogs and cats in a California animal shelter, *Journal of Applied Animal Welfare Science,* 5(1), 29-42.

Protopopova, A., Wynne, C. D. L. (2014). Adopter-dog interactions at the shelter: Behavioral and contextual predictors of adoption, *Applied Animal Behaviour Science,* 157, 109-116.

Rahman, A., Verma, B. (2013). Cluster based ensemble of classifiers, *Expert Systems,* 30(3), 270-282.

Ryu, J. I. (2017). The regional factors of economic activity for persons with disabilities based on propensity scores and multi-level logistic regression model, *Journal of the Korean Data Analysis Society,* 19(4), 1877-1886. (in Korean).

Szepannek, G. (2018). ClustMixType: User-friendly clustering of mixed-type data in R, *The R Journal*, 10(2), 200-208.

Yoo, H. S. (2019). *A study on the adoption prediction model of dogs impounded in APMS,* Master Thesis, Dongguk University. (in Korean).

Adoption Probability Prediction Model of Dogs in Animal Shelters of Local Government*

Sung Eun Choi[1], *Hyunsun Yoo*[2], *Hee Woon Jeong*[3], *Hee Won Jeong*[4], *Yumi Park*[5], *Kwan Jeh Lee*[6]

Abstract

Data of over 320 thousand abandoned dogs impounded in animal shelters of local governments during years of 2014 and 2018 are crawled from the homepage of animal protection management system run by Animal and Plant Quarantine Agency. Data are preprocessed and divided into two groups using k-prototype clustering method for modeling adoption probability prediction model. In order to find the optimal model stepwise logistic regression method of backward elimination is used based on AIC. Factors for adoption success are found using odds ratio. Odds ratios and their 95% confidence intervals are obtained and used for comparison of effects on adoption success of levels of each attribute. Accuracy, sensitivity, specificity, and ROC curves of the optimal models are obtained. And also their 95% bootstrap confidence intervals are calculated. The proper use of the threshold and predicted adoption probabilities is discussed. Some effective method for improving adoption probability is suggested through the simulation.

Keywords : companion animals, crawling, logistic, k-prototype, clustering.

---

*This work was carried out with the support of "Cooperative Research Program of Center for Companion Animal Research (Project No. PJ0139862019)" Rural Development Administration, Republic of Korea.

[123456]Department of Statistics, Dongguk University, 30, Pildong-ro 1-gil, Jung-gu, Seoul, 04620, Republic of Korea.

[1]Instructor, Professional Researcher, Dongguk University Bigdata Research Center.
E-mail : c6300@hanmail.net

[2]Master. E-mail : yyhhss777@naver.com

[3]Master Student. E-mail : jhw941217@gmail.com

[4]Master Student. E-mail : drave11@naver.com

[5]Master Student. E-mail : byumm315@naver.com

[6](Corresponding Author) Professor. E-mail : kwanlee@dongguk.edu

[Received 20 September 2019; Revised 17 October 2019; Accepted 20 October 2019]