머신러닝을 이용한 유기견 안락사 예측

이예슬[O], 이세훈[**], 존 킨[*]
[O]맨체스터대학교 컴퓨터과학과,
[*]맨체스터대학교 컴퓨터과학과,
[**]인하공업전문대학 컴퓨터시스템공학과
e-mail: ashleylee9404@gmail.com[O], john.keane@manchester.ac.uk[*], seihoon@inhatc.ac.kr[**]

Prediction of the Shelter Dog Outcome using Machine Learning Models

Ye-Seol Lee[O], Se-Hoon Lee[**], John Keane[*]
[O]Dept. of Computer Science, Manchester University, UK,
[*]Dept. of Computer Science, Manchester University, UK,
[**]Dept. of Computer Systems & Engineering, Inha Technical College

● 요 약 ●

The number of abandoned dogs were increasing every year in South Korea. However, many dogs are euthanized in the shelter because of the lack of budget. This project predicts euthanasia of abandoned dogs using machine learning algorithm. It collects data from the public data portal where Korea government provides a public dataset as a form of open API. This project uses recent three-year data 2017 to 2019 and 263371 cases were founded. This project implements random forest and logistic regression models. This project attained an average 72% of prediction accuracy.

키워드: 유기견(Shelter dog), 안락사(euthanasia),
머신러닝(Machine learning), 예측 모델(Prediction Model)

I. Introduction

Every year, 120 thousand of dogs were abandoned and about 30000 dogs which are 20% of shelter dogs are euthanized in South Korea. Moreover, 26% of shelter dogs were died from natural causes because of the bad condition and hygiene of the shelter. Besides, the number of abandoned dogs increases every year. This became a serious social problem regarding animal welfare and social cost. But, South Korea still have a lack and weak law for animal welfare that many abandoned dogs have died in the shelter. This project aims to raise awareness of South Korea's shelter dog problem and give an idea to reduce the euthanasia rate by predicting and analyzing the factors.

II. Experiment

In this project, data was collected by using open API provided by Korea government. The selected feature was, 'happDt', 'OrgNm', 'age', 'kindCd', 'sexCd', 'neuterYn', 'weight', 'colorCd', 'processState', 'specialMark'. This project uses 2017 to 2019 data and 263371 cases were collected. The size of the dataset for 2017 was 72612, 2018 was 89884, 2019 was 100875.

Data preprocessing was operated by three data types; Categorical, numerical, text. This project use one-hot-encoding for categorical data and label numerical data as divided specific standards. 'specialMark' variable indicates a condition of the dog, and it is consisting of short sentences in the Korean language. In this project, it extracts the condition of the dog, labeled by three conditions: good signal, bad health, and cloth word. It extracted noun words for every sentence and stem it by using 'KONLPY' python library. Then trained the word2vec for both continuous bag of words and skip-gram methods by window size 5. After the dictionary is ready, then start labeling for each of the conditions.

In this project, 2 prediction models are implemented, random forest, logistic regression. For logistic regression model, it uses grid search k-fold to find the best C. The best C was 0.001. Randomly split the dataset into 70% of training data, 30% of testing data. To evaluate these models, it uses accuracy, precision, recall and f1-score.

III. Results and Analysis

Random forest model shows better prediction accuracy, 72.2% than the logistic regression model, 71.8%. From these models, we could get the feature importance score of each feature. It gives a score for each feature and a higher score means it is more relevant to the target from the predictive models. Therefore, comparing each model's feature importance scores will help to analyze what affects the outcome of the dog.

Table 1. Result of model

Model	Accuracy	Precision	Recall	F1-score
Random forest	0.727	0.78	0.86	0.82
Logistic regression	0.72	0.72	1	0.83

There are top two feature importance, month and places. For the month feature, it is clear that from June to July and August, abandon dog rapidly increases. According to the research, there are two reasons. The first reason was that July and August are the summer vacation season. Many Koreans go on vacation and abandon their dog in the holiday spot. The second reason was that in summer, many houses open their door because of the heat that many dogs run away from the open door, lose their way. For the place feature, 'Gyeonggi-do' province shows the largest number of abandoned dogs. However, the euthanasia rate is 31.23% which is the 5th all through other provinces. This is because 'Gyenggi-do' province picks out the shelter dogs who have good sociality then provide the treatment, training and disease vaccination. Additionally, they provide the free matching system with the new owners to encourage the adoption of the shelter dogs. On the other side, 'Jeju' province shows the highest euthanasia rate of 58.98% and it has 4th largest number of shelter dogs. Ironically, 'Jeju' province has the biggest animal shelter in Korea, but it has exceeded the capacity. Most of the residents in 'Jeju' province is elderly and they raise the mutt dog by letting it loose. This cause a low adoption rate and an increase in euthanasia rate.

IV. Conclusions

This project implements the random forest and logistic regression model to predict the euthanasia of the shelter dogs. Random forest shows better accuracy than the logistic regression model. This both model shows two common highest feature which affects euthanasia; month and place. As the number of abandoned dogs increases than the rate of euthanasia also increases. So in summer, when the number of abandoned dogs rapidly increases, the shelter has to try to expand the capacity of the shelter. But by looking at 'Jeju' province case, the shelter has to not only increasing capacity but also try as like 'Gyenggi-do' province, build strategy to encourage adoption rate by providing matching, training system. For further studies, it will try to use other models for the prediction.