한국컴퓨터정보학회 동계학술대회 논문집 제27권 제1호 (2019. 1)

기계학습 모델을 이용한 신용 승인 데이터 분석

김동현[O], 김세준[*], 이병준[*], 김경태[**], 윤희용[*]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {kdh7263, ksj105, byungjun, youn7147}@skku.edu[O*], kyungtaekim76@gamil.com[**]

Analysis of Credit Approval Data using Machine Learning Model

Dong-Hyun Kim[*], Se-Jun Kim[*], Byung-Jun Lee[*], Kyung-Tae Kim[**], Hee-Yong Youn[*]
[O*]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[**]Dept. of Software, Sungkyunkwan University

● 요 약 ●

본 논문에서는 다양한 기계학습 모델을 이용한 신용 데이터 분석 기법에 대해 서술한다. 기계학습 모델은 크게 Canonical models, Committee machines, 그리고 Deep learning models로 분류된다. 이러한 다양한 기계학습 모델 중 일부 학습 모델을 기반으로 Benchmark dataset인 Credit Approval 데이터를 분석하고 성능을 평가한다. 성능 평가에는 k-fold evaluation method를 사용하며, k-fold evaluation 결과에 대한 평균 성능을 측정하기 위해 Accuracy, Precision, Recall, 그리고 F1-score가 사용되었다.

키워드: 기계학습(machine learning), 신용승인(credit approval), k-fold 교차검증(k-fold evaluation)

I. Introduction

벤치마크(Benchmark; 비교평가)란 본래 컴퓨터의 부품과 같은 성능을 평가하여 점수를 내거나, 경쟁 기업 대비 자사의 생상 능력을 평가하기 위한 용도로 사용되었다 [1]. 이와 비슷한 개념으로 빅데이터 기반의 시스템에 대한 객관적인 성능을 평가 하기 위해 벤치마크 데이터를 활용할 수 있다. 본 논문에서는 다양한 기계학습 모델의 분류 및 예측 정확도 성능을 측정하기 위해 UCI Benchmark dataset를 활용하였다. UCI는 UC Irvine에서 지원하는 기계학습 저장소(UCI Machine learning repository)로, 다양한 종류 및 형식의 데이터 세트를 지원하고 있다. 본 연구에서는 이러한 UCI 데이터 중에서 신용 승인 데이터 [2]를 활용하여 각 기계학습 모델별 성능을 측정하고 비교한다.

II. Preliminaries

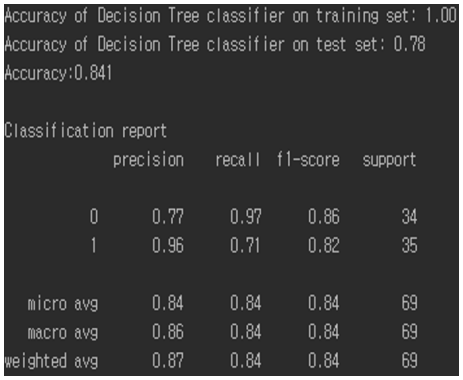
1. Related works

1.1 Credit Approval Data (CAD)

CAD는 신용카드의 승인 여부에 해당하는 데이터 [2] 로 전체 690줄로 구성되어 있으며, 각 줄은 16개의 특징 값을 갖는다. 첫 0~14에 해당하는 특징은 신용카드 사용자의 성별, 나이, 고용기간 등을 의미하며, 마지막 15에 해당하는 특징은 신용카드의 승인 여부이다. 여기서 승인여부는 + 또는 -로 구분되는데 +는 Positive, -는 Negative를 의미한다. 또한 전체 데이터 세트에는 일부 Missing value가 존재한다. 일반적으로 Missing value는 최대 빈도수 또는 평균값으로 대체된다. 이러한 신용카드 승인여부는 사용자의 개인적 또는 금융 환경에 따라 승인되거나 거절된다.

III. The Proposed Scheme

본 연구에서는 UCI 데이터 세트에 대해 Decision tree, Support Vector Machine(SVM), Ada-boost, 그리고 Convolution Neural Network(CNN)를 이용하여 데이터의 분류 정확도 및 Confusion matrix를 평가하였다. 아래 그림은 각 기계학습 모델 별 성능 평가 결과이다.

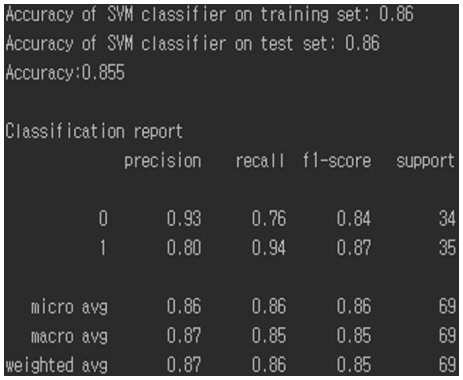


Accuracy of Decision Tree classifier on training set: 1.00
Accuracy of Decision Tree classifier on test set: 0.78
Accuracy:0.841

Classification report

	precision	recall	f1-score	support
0	0.77	0.97	0.86	34
1	0.96	0.71	0.82	35
micro avg	0.84	0.84	0.84	69
macro avg	0.86	0.84	0.84	69
weighted avg	0.87	0.84	0.84	69

Fig. 1. Decision tree

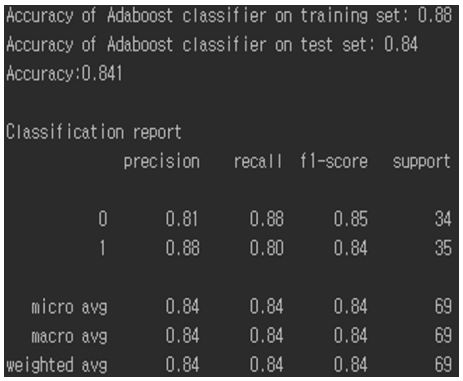


Accuracy of SVM classifier on training set: 0.86
Accuracy of SVM classifier on test set: 0.86
Accuracy:0.855

Classification report

	precision	recall	f1-score	support
0	0.93	0.76	0.84	34
1	0.80	0.94	0.87	35
micro avg	0.86	0.86	0.86	69
macro avg	0.87	0.85	0.85	69
weighted avg	0.87	0.86	0.85	69

Fig. 2. SVM

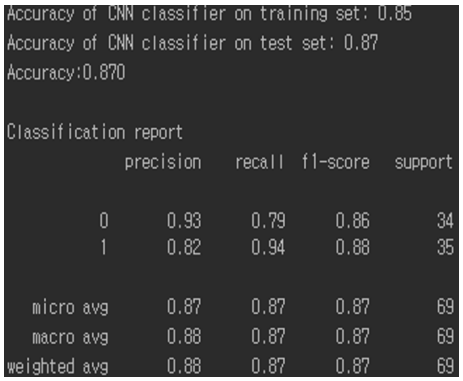


Accuracy of Adaboost classifier on training set: 0.88
Accuracy of Adaboost classifier on test set: 0.84
Accuracy:0.841

Classification report

	precision	recall	f1-score	support
0	0.81	0.88	0.85	34
1	0.88	0.80	0.84	35
micro avg	0.84	0.84	0.84	69
macro avg	0.84	0.84	0.84	69
weighted avg	0.84	0.84	0.84	69

Fig. 3. Ada-boost



Accuracy of CNN classifier on training set: 0.85
Accuracy of CNN classifier on test set: 0.87
Accuracy:0.870

Classification report

	precision	recall	f1-score	support
0	0.93	0.79	0.86	34
1	0.82	0.94	0.88	35
micro avg	0.87	0.87	0.87	69
macro avg	0.88	0.87	0.87	69
weighted avg	0.88	0.87	0.87	69

Fig. 4. CNN

시뮬레이션 결과 Ada-boost 모델의 학습 정확도가 가장 높았는데, 이는 오답에 대해 높은 가중치를 부여하고 정답에 대해 낮은 가중치를 부여하여 오답에 더욱 학습을 집중시켰기 때문이다.

IV. Conclusions

본 연구에서는 다양한 기계학습 모델을 기반으로 Benchmark 데이터에 대한 정확도 및 Confusion matrix에 대한 성능을 평가하였으며, 시뮬레이션 결과 Ada-boost 모델의 학습 정확도가 가장 높은 결과를 보였다. 이러한 이유는 오답에 대한 높은 가중치 할당으로 오차를 줄여 학습을 하였기 때문이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016 -0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업(No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구), BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] https://ko.wikipedia.org/wiki/벤치마크
[2] http://archive.ics.uci.edu/ml/datasets/Credit+Approval