한국컴퓨터정보학회 동계학술대회 논문집 제27권 제1호 (2019. 1)

효율적인 기계학습을 위한 데이터 전처리

김동현O, 유승언*, 이병준*, 김경태**, 윤희용*
O*성균관대학교 정보통신대학 전자전기컴퓨터공학과
**성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {kdh7263, seyoo90, byungjun, youn7147}@skku.eduO*, kyungtaekim76@gamil.com**

Data preprocessing for efficient machine learning

Dong-Hyun Kim*, Seung-Eon Yoo*, 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 ●

데이터를 기반으로 한 기계학습은 데이터의 양, 학습 모델, 그리고 데이터의 특징 등 다양한 환경에 민감한 특징을 지니고 있어, 보다 효율적인 기계학습을 위해 데이터의 전처리 과정을 필요로 한다. 데이터의 전처리 과정이란 특징 선택(Feature selection), 노이즈 데이터의 제거, 차원 감소(Demension reduction), 클러스터링(Clustering) 등 보다 효율적인 기계학습을 위한 방법이다. 따라서 본 논문에서는 다양한 환경에서 보다 효율적인 기계학습을 위한 데이터 전처리 기술의 종류 및 간단한 특징에 대해 서술한다.

키워드: 기계학습(machine learning), 데이터(data), 전처리(preprocessing)

I. Introduction

데이터를 기반으로 한 기계학습은 학습 데이터의 양, 학습 모델, 그리고 데이터의 특징의 수 등 학습 환경에 따라 모델의 성능이 크게 좌우된다. 또한 데이터의 처리 방법에 따라서도 상이한 결과를 보인다. 따라서 보다 정확하고 효율적인 기계학습을 위해서는 데이터의 전처리 과정이 요구된다. 여기서 데이터의 전처리란 학습 이전에 학습에 불필요한 데이터와, 학습에 반드시 필요한 데이터를 구분하여 제거 및 분류하는 과정이다. 이러한 데이터의 전처리 과정은 기계학습 모델의 처리 속도 및 정확도 등 다양한 성능을 향상시킬 수 있다. 따라서 본 논문에서는 기계학습 모델의 전처리 과정의 종류와 각 전처리 방법에 대한 간단한 특징에 대해 서술한다.

II. Preliminaries

1. Related works

1.1 특징 선택(Feature selection)

기계학습을 위한 데이터는 다양한 특징을 가지고 있다. 여기서 데이터의 특징이란 사람으로 예를 들면, 키, 체중, 성별, 머리 길이, 시력, IQ 등 다양한 요소를 포함한다. 하지만 실제 사람과 동물을 구분하기 위한 필요한 특징은 위에서 언급한 모든 요소를 포함하지 않는다. 이렇듯 학습에 반드시 필요한 특징만을 선택하는 방법이다. 이러한 특징 선택 기법에는 Filter, Wrapper, Embedded 방법 등이 있으며, 기계학습 모델의 학습 정확도를 향상 시킬 수 있다.

1.2 노이즈 데이터(Noise data) 제거

노이즈 데이터란 손상되거나(Corrupted) 왜곡된 (Distorted) 혹은 중복된 데이터를 의미한다 [2]. 이러한 노이즈 데이터는 기계학습 모델의 부정확하고 잘못된 결과를 야기할 수 있기 때문에 학습 이전에 제거되어야 한다. 이러한 노이즈 데이터의 제거에는 Sorted Neighborhood Methods(SNM), Filter-and-refine paradigm [2] 등이 있으며, 불필요한 데이터를 제거함으로써 학습 정확도 및 학습 속도를 향상시킬 수 있다.

1.3 차원 감소(Demension reduction)

데이터가 많은 특징을 포함할 경우 데이터의 분석에 어려움이 있고, 3개 이상의 차원이 있을 경우 시각화가 어려워 진다. 이는 특징 선택 기법과 비슷한 개념이지만, 차원 감소에는 특징 선택(Feature selection) 뿐만 아니라 특징 추출(Feature extraction)을 포함한다. 특징 추출이란 학습에 필요한 원본 데이터와 상이한 형태의 데이터를 추출하는 방법이다. 이러한 차원 감소 방법에는 대표적으로 주성분 분석(Principal Component Analysis; PCA)가 있으며, 학습 정확도 및 학습 속도를 향상시킬 수 있다.

1.4 클러스터링(Clustering)

데이터의 클러스터링이란 비지도 학습으로 유사한 성격을 지닌 데이터를 묶어 그룹으로 구성하는 것으로, 학습에 필요한 데이터를 기반으로 유사성을 이용하여 보다 효율적인 분석을 가능케 한다. 클러스터링의 대표적인 방법으로 *k*-means 클러스터링 기법이 있다. *k*-means 클러스터링은 *k*개의 클러스터로 데이터를 분류하는데, 이는 패턴인식, 음성인식 등 불분명한 데이터의 분류를 가능케 하며, 학습 정확도를 향상시킨다.

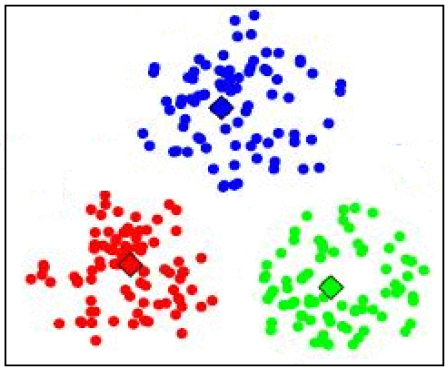

Fig. 1. *k*-means clustering example with *k*=3

IV. Conclusions

본 연구에서는 데이터를 기반으로 한 기계학습 모델의 학습 효율성을 높이기 위한 데이터 전처리 기법에 대해 서술하였다. 여기서 기계학습을 위한 데이터는 방대하거나 다양한 종류로 구성된 혹은 불필요한 데이터를 포함할 경우 학습 과정에 많은 시간을 소요하며, 잘못된 결과를 초래할 수 있다. 이러한 문제점을 해결하기 위한 방법으로 다양한 데이터의 전처리 방법이 있으며, 이에 대해 간략히 설명하고 각 전처리 방법별로 가장 기본적인 학습 알고리즘에 대해 서술하였다.

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://en.wikipedia.org/wiki/Noisy_data

[2] H. Xiong et al. "Enhancing data analysis with noise removal," pp.304-319, 2006