기계학습의 문제점 및 해결방안

임환희[O], 김세준[*], 이병준[*], 김경태[*], 윤희용[**]
[O*]성균관대학교 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {lhh423,ksj105,byungjun}@skku.edu[O*], kyungtaekim76@gmail.com[*], youn7147@skku.edu[**]

Problems and Solutions for Machine Learning

Hwan-Hee Lim[O], 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 ●

기계학습이란 인공지능의 한 분야이다. 컴퓨터에 명시적인 프로그램 없이 배울 수 있는 능력을 부여하는 연구 분야이며, 사람이 학습하듯이 컴퓨터에도 데이터들을 줘서 학습하게 함으로써 새로운 지식을 얻어내게 하는 분야이다. 기계학습 종류에는 크게 Supervised Learning, Unsupervised Learning, Reinforcement Learning이 있다. 본 논문에서는 기계학습 종류 및 컴퓨터가 데이터들을 학습하면서 생기는 문제점을 알아보고, 문제점의 종류 및 해결방안을 제시한다.

키워드: 기계학습(Machine Learning), 데이터, 과적합(Overfitting), Learning Rate

I. Introduction

기계학습이란, 경험적 데이터를 기반으로 학습을 진행하고 예측을 수행하며, 스스로 성능을 향상시키는 시스템과 이를 위한 알고리즘을 연구하고 구축하는 기술이다[1].

기계학습의 종류에는 크게 세 가지로 분류된다. 감독 학습(Supervised Learning), 비 감독 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning)이 있으며, 감독 학습은 입력과 이에 대응하는 미리 알려준 출력을 mapping하는 함수를 학습하는 것이다. 비 감독 학습은 출력 없이 입력만으로 모델을 구축하고 학습하는 형태이며, 강화 학습은 학습자가 행동을 선택하여 환경에 영향을 미치고 피드백을 받아 학습하는 형태이다.

이러한 기계학습 기법들은 최근 들어 실생활에서 자주 쓰인다. 예를 들어, 소비자가 평소에 구매한 물건의 패턴을 분석하여 관심을 가질만한 상품을 예측해 정보를 제공하거나, 스마트폰으로 사용자가 찍은 사진을 특정 콘텐츠 별로 분류를 하거나, 개인의 취향에 맞는 음악이나 영화 등을 추천해주는 시스템 등 이 있다.

이러한 기계학습을 수행하면서 주의해야 할 부분도 있다. 적절하지 않은 학습 계수를 유지하거나, 크기가 너무 큰 학습 데이터를 가지고 학습을 진행하거나, 학습이 너무 잘되어 실제 데이터를 대상으로 예측을 못하는 문제점 등이 있다.

본 논문에서는 컴퓨터가 데이터를 학습하면서 생기는 문제점에 대해 자세히 알아보고, 문제점의 종류 및 해결방안을 제시한다. 2장에서는 기계학습의 종류에 대해서 제시하며 3장에는 기계학습을 이용해 학습 시 문제점과 해결방안을 제시한다.

II. Preliminaries

1. Related works

1.1 기계학습의 종류

감독 학습은 감독한 내용으로 학습하는 것이다. 일반적으로 입력 데이터와 출력 데이터를 나타내는 하나의 Data Set을 가지고 새로운 입력 데이터를 받았을 때 출력 데이터를 예측하는 방법이다. 주로 분류와 회귀에 쓰이며, Support Vector Machine (SVM), Naive Bayes, Decision Tree, Logistic Regression, Linear Regression 등이 있다. 비 감독 학습은 사람 없이 컴퓨터가 스스로 레이블 되어 있지 않은 데이터에 대해 학습하는 것이다. 주로 군집화와 분포 추정에 쓰이며, K-means, Clustering, Density estimation 등이 있다. 강화 학습은 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 이로부터 보상을 얻으면서 학습을 진행하며, 보상을 최대화 하도록 학습이 진행된다. 주로 게임이나, 로봇틱스 환경에 주로 쓰이며, Q-learning, Markov Decision Process 등이 있다.

III. 기계학습의 문제점 및 해결방안

다음으론 기계학습을 이용해 학습 진행 시 문제점에 대해 제시한다.

기계학습 기법을 이용해 학습할 시 주의해야할 것은 크게 3가지가 있다.

첫 번째로, 학습 계수(Learning Rate)는 적절한 값을 유지해야 하는 것이다.

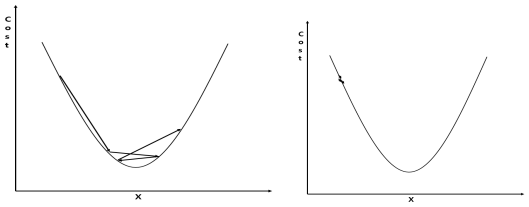

Fig. 1. 학습 계수 비교 그래프

이와 같이 학습 계수가 너무 크면 최솟값을 제대로 찾지 못하고, 너무 작으면 최솟값을 찾는데 오랜 시간이 걸릴 수 있다.

두 번째로, Overfitting[2] 문제가 있다. Overfitting 이란, 과도하게 데이터에 대해 모델을 학습한 경우를 의미한다.

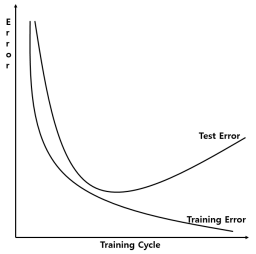

Fig. 2. Overfitting

과도하게 학습 데이터에 대해 학습을 진행한 경우, 학습 데이터에 대해선 정확도가 높아지지만 테스트 데이터에 대해선 정확도가 떨어지는 문제이다.

이러한 문제점들을 해결하기 위한 방안은 대표적으로 2가지가 있다. 사용하는 데이터의 Feature 수를 줄이는 방법과, Regularization을 이용하는 방법이 있으며, 향후, 기계학습의 해결방안을 좀 더 구체적으로 연구할 예정이다.

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), 삼성전자, BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] David M. Dutton, Gerard V. Conroy, “A review of machine learning”, The Knowledge Engineering Review, Vol. 12, No. 4, pp. 341-367, Dec. 1997

[2] T. Dietterich, “Overfitting and undercomputing in machine learning”, ACM Computing Surveys, Vol. 27, No. 3, pp. 326-327, Sept. 1995