최대 엔트로피 이론을 이용한 학습 데이터 분류

김민우[O], 김동현[*], 이병준[*], 김경태[*], 윤희용[**]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과

e-mail:{kimmw95, kdh7263, byungjun}@skku.edu[O*],
kyungtaekim76@gmail.com[*], youn7147@skku.edu[**]

Classification Learning Data using Maximum Entropy Theory

Min-Woo Kim[O], Dong-Hyun 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 ●

빅 데이터 활용의 증가로 인해 효율적으로 데이터를 분류하는 것은 머신러닝의 주요 과제이다. 제한적인 자원을 가지고 이에 맞는 처리능력을 갖기 위해서는 단일 기기의 자원 관리능력을 향상시키는 방향의 연구가 필요하다. 본 논문에서는 머신러닝을 위한 학습 데이터를 최대 엔트로피 이론을 적용시켜 효과적으로 분류하는 방법을 제안한다. 최대 엔트로피에 대한 간단한 설명과 최대 엔트로피 이론을 적용시키기 위한 간단한 사전 작업들의 방향 등에 대한 설명을 토대로 기술하였다. 또한 본 연구를 통해 얻게 된 문제점들과 향후 연구에 필요한 피드백을 갖는다.

키워드: 최대 엔트로피(Maximum entropy), 분류(classification), 머신러닝(machine learning)

I. Introduction

최근의 연구에 따르면 대규모 데이터베이스를 위한 고성능 데이터 마이닝 기술이 등장하면서 데이터 분류의 정확성과 확장성 측면의 경쟁력이 중요시 된다. 본 연구에서 제안하는 최대 엔트로피 모델은 머신러닝을 위한 빅 데이터의 분류를 통해 학습 시 속도 향상을 목표로 하는 초기 연구 단계이다. 본 연구 결과의 목표를 바탕으로 자능화된 IoT 디바이스의 프레임 워크에서 빠른 계산 처리를 가능하게 할 수 있으며 분산 진화 인공지능 시스템에서 효율적인 분산 지능처리를 기대해 볼 수 있다[1]. 2장에서는 기존에 연구되었던 최대 엔트로피 관련된 내용들과 활용되어진 방법들에 대해 서술하였고 3장은 본 논문에서 제안한 최대엔트로피 이론에 대한 간단한 설명[2]과 최대 엔트로피 모델의 연구 방향, 기본 이론 및 문제점 등에 대해 서술하였다. 마지막으로 4장에서는 본 논문에 대한 결론과 향후 연구에 필요한 요소들에 대하여 서술하였다.

II. Preliminaries

1. Related works

최대 엔트로피 모델은 머신러닝 알고리즘 중에서 자연 언어 처리의 언어 분류 문제를 해결하기 위해 사용되었고 최대 엔트로피 분류기는 프로그램이 새 데이터를 발견했을 때 이미 학습된 데이터를 확률 모델로 만들어 파악하는 방법 등으로 사용되었다. 또한 최근 수년간 최대 엔트로피 응용 프로그램을 사용하여 시퀀스 분석, 신경 생물학, 구조 모델링 및 언어 예측과 같은 다양한 분야에서 꾸준한 연구가 이루어졌다[3][4]. 데이터 분류에 관한 앞선 연구들은 특히 IoT 환경이 아닌 생물학이나 로봇과 같은 기계적인 시스템에 초점이 맞추어져 있어 본 연구의 목표와 같이 국내외 모두 초기 단계 수준이다.

III. The Proposed Scheme

1. 최대 엔트로피 원리

1.1 기본 원리 및 문제점

정보 이론(엔트로피) 측면에서 정의된 각 분류에 의해 전달되는 정보의 양은 확률 값이 엔트로피인 확률 변수가 된다. 최대 엔트로피의 원칙은 새로운 데이터에 기반 하여 확률 분포를 구성하는 일반적인

절차로 초기의 모델이 실험과 상충하는 결과를 제공할 때 완벽하게 사용되어 진다. 본 연구에서 정의한 정보 이론의 기본개념은 분류 기준에 대해 더 많이 알수록 새로운 정보가 줄어들어 분류 시 효율성을 높이는 것이다. 측정된 양에 대한 데이터들은 확률 분포로 표현 될 수 있다고 가정하고 평균값에 대한 정보를 얻는다. 이때 최대 엔트로피 모델은 나타나진 모든 데이터를 모델링하지만 알려지지 않은 데이터에 대해서는 확률적인 작업을 하지 않는다. 이로 인해 엔트로피 비율의 추정치가 인위적으로 낮아지게 되면 정확한 확률 분포를 알지 못하며 결과 값을 통해 재구성 하는 것이 불가능해진다. 측정된 데이터들은 직관적으로 보았을 때 모든 값들이 하나의 가치를 가지지만 전체 분포에 대한 평균값을 통해 분류하는 것은 비합리적일 수 있다. 이에 따라 본 연구에서 빅 데이터에 대한 시뮬레이션 실험 결과 정략적이지 않은 결과를 나타내기도 하고 불안정한 정확도를 보였다. 또한 때때로 시뮬레이션의 결과 값이 실험에 사용된 데이터와 양적으로 일치하지 않는 것을 알 수 있었다. 이 문제에 대한 해결책으로는 시뮬레이션 데이터에 포괄적인 분류기준 보다 디테일한 부분을 보완할 수 있는 알고리즘 함수를 추가해야 한다.

IV. Conclusions

향후 연구에서는 본 논문에서 제안된 최대 엔트로피 이론의 효율성을 입증하기 위해 기존의 Decision trees, Naive Bayes 등과 같은 다른 분류 알고리즘들과 비교하여 평가를 진행할 것이다. 또한 분류에 대해 더 높은 성능을 얻기 위해 변수를 자동으로 조정하는 방법에 초점을 두고 정확도와 정량적 결과 값을 도출하기 위해 문제점을 보완해 줄 알고리즘을 도입할 것이다. 추가로 비교하게 될 알고리즘들이 어떻게 높은 정확도를 달성할 수 있었는지 명확히 조사하기 위해 이론적인 측면을 더 연구할 것이다.

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] K. Wei, R. Iyer, S. Wang, W. Bai, and J. Bilmes, “Mixed robust/average submodular partitioning: Fast algorithms, guarantees, and applications“, NIPS 2015 Extended Supplementary.

[2] Gideon Mann, Ryan McDonald, Mehryar Mohri, Nathan Silberman, Daniel D. Walker, “Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models”, Advances in Neural Information Processing Systems 22 (NIPS), 2009

[3] Yi Yin, Dan Feng, Yue Lid, Shuifang Yine and Zhan Shi, “Text prediction method based on multi-label attributes and improved maximum entropy model”, Journal of Intelligent & Fuzzy Systems, pp. 1097-1109, 2018

[4] Hieu X. Phan, Minh L. Nguyen, S. Horiguchi, Bao T. Ho, Y. Inoguchi, “Classification with Maximum Entropy Modeling of Predictive Association Rules”, European Conference on Macine Learning : ECML 2005, LNAI 3720, pp. 682-689, 2005