Very Fast Decision Tree 기반 Naive Bayesian 알고리즘의 Weight 부여 기법

김세준[O], 유승언[*], 이병준[*], 김경태[*], 윤희용[**]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {ksj105,seyoo90,byungjun}@skku.edu[O*], kyungtaekim76@gmail.com[*], youn7147@skku.edu[**]

An Attribute Weighting Approach for Naive Bayesian based on Very Fast Decision Tree

Se-Jun Kim[O], 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 ●

본 논문에서는 지도 기계 학습 알고리즘 중 하나인 Naive Bayesian (NB) 알고리즘의 데이터 분류 정확도를 향상시키기 위하여 데이터 속성에 Weight를 부여하는 새로운 기법을 제안하였다. 기존에 Decision Tree(DT) 알고리즘의 깊이를 이용하여 Weigth를 부여하는 방법이 제안되었으나, DT를 구축하는데 오버헤드가 크기 때문에 데이터의 실시간 분석이나 자원 제한적인 환경에서의 적용은 어렵다는 단점이 있다. 이를 해결하기 위하여 본 논문에서는 최소한의 데이터를 사용하여 신속하게 DT를 구축하는 Very Fast Decision Tree (VFDT) 알고리즘 기반의 Weight 부여 기법을 제안함으로써 적은 오버헤드로 NB의 정확도를 향상시킨다.

키워드: 기계학습(Machine learning), 데이터 분류(Data Classification)

I. Introduction

최근 컴퓨터 하드웨어의 발전으로 기존에 이론적으로만 제시되어왔던 기계 학습 알고리즘들이 실제 환경에서 구현 가능해지며 다양한 접근 방법이 제안되고 있다. NB는 기계학습 중에서도 가장 기초적인 알고리즘으로, 계산이 복잡하기 않으면서 적절한 데이터 분류 정확도를 보여주기 때문에 IoT 디바이스 등 자원제한적인 환경에서 다양하게 사용될 것으로 전망된다.

그러나 NB의 정확도는 타 기계학습 알고리즘보다 낮은 수준으로, 이를 해결하기 위한 방법으로 데이터의 속성마다 Weight를 부여하여 노이즈 데이터의 영향을 줄이는 많은 방법이 제시되었다. 이 중 한 방법으로 DT의 데이터 속성에 대한 트리 내부 깊이를 기반으로 NB의 Weight를 결정하는 기법이 제안되었다[1]. 그러나 DT를 구축하는 데는 큰 시공간적 오버헤드가 발생하기 때문에 자원제한적인 환경에서 적용되기에는 한계가 있다.

본 논문에서는 이를 해결하기 위하여 원하는 정보량을 가지는 데이터만으로 신속하게 DT를 구축하는 VFDT 알고리즘을 이용하여 NB의 데이터 속성에 대한 Weight를 부여하는 기법을 제안한다.

II. Preliminaries

1. Related works

1.1 Very Fast Decision Tree (VFDT)

VFDT는 Hoeffding Bound를 이용하여 수집된 데이터가 특정 정보량에 도달하면 DT를 구축하는 기법으로, 다중 속성을 가지는 데이터의 Hoeffding Bound ε는 아래의 식과 같이 나타난다.

$$\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}, where\, R = \log(n\, of\, classes)$$

VFDT는 데이터 셋에서 데이터를 읽어옴과 동시에 각 속성에 대하여 가장 높은 정보량을 가지는 속성 G_1과 G_2를 실시간으로 추출한다. G_1, G_2에 대하여 각 정보량의 차가 Hoeffding Bound를 초과하면 현재 읽어온 데이터만을 이용하여 DT를 구축한다. 이를 통하여 기존의 DT보다 신고하게 합리적인 정확도로 DT를 구축할 수 있다.

III. The Proposed Scheme

본 논문에서 제안하는 기법은 VFDT를 구축하는 것으로 시작한다. NB를 이용하여 데이터 분류를 하고자 할 때, 데이터 셋을 순서대로 읽어오게 된다. 이 때 VFDT 알고리즘을 이용하여 각 속성에 대한 정보량의 변화를 모니터링하고, 가장 높은 정보량을 가지는 속성 G_1과 G_2에 대하여 각 정보량의 차가 Hoeffding Bound를 초과하면 현재 읽어온 데이터만을 이용하여 DT를 구축한다. DT를 구축한 후 각 데이터 속성에 대하여 트리의 깊이를 추출한다. 트리의 깊이는 Root 노드로부터 분기점이 되는 속성까지의 거리를 나타낸 것으로, DT의 특성상 정보량이 적은, 즉 데이터를 분류하는데 크게 영향을 주지 않는 데이터일수록 깊이가 크게 나타난다. 제안하는 기법에서는 트리의 깊이에 따라 데이터 속성의 Weight가 결정되는데, 이는 아래의 식과 같다.

$$w_i = \frac{1}{\sqrt{d_i}}$$

위 과정을 통하여 Weight가 결정되고 난 후, 데이터 분류는 NB를 이용하여 실시하는데, 먼저 기본 NB의 분류를 위한 확률식은 데이터 값 $X = \{x_1,x_2,...,x_n\}$ 이고, 데이터 분류결과 라벨은 $C = \{C_1,C_2,...,C_n\}$ 일 때, 아래와 같다.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

본 논문에서 제안하는 기법은 NB를 통한 데이터 분류를 할 때, 분류에 정확도를 떨어트리는 노이즈 데이터의 영향을 줄이기 위하여 각 속성마다 Weight를 부여한다. 속성 $A=\{A_1,A_2,...,A_n\}$에 대하여 Weight를 부여한 확률식은 아래와 같다.

$$P(x_i|C_i) = \prod_{j=1}^{n} P(A_{ij}|C_i)^{w_i}$$

이를 통한 최종적인 데이터 분류는 각 결과 라벨에 대한 중 가장 높은 확률을 가지는 라벨로 이루어진다.

IV. Conclusions

본 논문에서는 NB 알고리즘의 정확도를 향상시키기 위한 방법으로 VFDT의 데이터 속성에 대한 깊이를 이용하여 Weight를 부여하는 기법을 제안하였다. 제안한 기법은 기존의 DT를 이용한 Weight 부여 기법보다 시공간적 오버헤드가 적고, 합리적인 데이터 분류 정확도를 나타낼 것으로 기대된다. 향후 연구로는 제안한 기법을 자원제한적인 환경에서 구현하여 성능을 평가한다.

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] D. M. Farid et al., “Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks”, Expert Systems with Applications 41, pp.1937-1946, 2014