빅데이터 환경에서 학습 정확도 향상을 위한 의미 계층 기반 속성 집단화 기법

이건선[1], 이건수[2], 강병권[3*]
[1]순천향대학교 ICT융합 재활 공학과
[2](주)솔루게이트 기업부설연구소
[3*]순천향대학교 정보통신공학과
e-mail: kslee005@naver.com[1], keonsoo@soluagete.com[2], bgkang@sch.ac.kr[3*]

A Method of Grouping Features from Big Data based on Semantic Hierarchy for Accuracy Enhancement

Keonsun Lee[1], Keonsoo Lee[2], Byeong-G Kang[3*]
[1]Dept of ICT Convergence Rehabilitation Engineering, Soonchunhyang University
[2]Research and Development Laboratory at Solugate Ltd.
[3*]Dept. of Information and Communication Engineering, Soonchunhyang University

요 약

빅데이터 기반의 기계학습은 대규모 데이터를 이용하여, 숨겨진 패턴을 찾아내는 학습과정과, 그렇게 찾아낸 패턴을 이용하여 새로운 데이터를 해석하는 추론과정으로 이루어진다. 이 과정을 통해 학습된 패턴은 데이터를 구성하는 속성들과 긴밀한 연관성을 갖고 있다. 학습에 사용된 데이터의 원 데이터를 구성하는 각각의 속성과 추론 결과가 동일한 계층 관계를 갖고 있다면, 모든 속성을 동일하게 처리할 수 있지만, 그렇지 않은 경우, 속성들 사이의 계층 정보를 고려하는 것이, 추론 결과의 정확도를 높일 수 있다. 이에 본 연구에서는 속성들 사이의 계층 관계를 고려한 추론 기법을 제안하고, 사례연구를 통해 제안 방법을 실제 상황에 적용하는 방법을 제시한다.

1. 서론

4차 산업혁명을 이끄는 대표적인 핵심 기술로는 빅 데이터와 인공지능을 꼽을 수 있다 [1]. 빅 데이터와 인공지능이 서로 밀접하게 연결되어 있는 까닭은 인공지능의 목표는 시스템이 지능적으로 행동하기 위한 지식베이스를 만드는 것이고, 이 지식베이스를 만드는 방법 중 하나인 기계학습을 이용하기 위해서는 많은 학습용 데이터가 필요하다. 빅 데이터는 대량의 학습용 데이터를 제공해줌으로, 기계학습을 통한 지식 베이스 구축을 가능하게 해 준다 [2]. 즉, 빅 데이터로부터 얻은 대규모 데이터를 기반으로 데이터에 숨어있는 암묵지(Implicit Knowledge)를 자율적으로 찾아내서, 지식베이스를 만들고, 그 지식베이스를 활용해 지능적인 행동을 수행하는 것이 빅 데이터와 인공지능의 시너지 효과의 근간이라고 할 수 있다.

더 많은 데이터는 보다 좋은 지식을 만들기 위한 기반이 되지만, 동시에 더 많은 데이터가 항상 좋은 지식을 보장하는 것은 아니다. 소위 '차원의 저주(Curse of Dimensionality)'라고 불리는 제약이 존재하며, 동시에 'GIGO(Garbage-In Garbage-Out)' 같은 격언 또한 존재한다 [3, 4]. 차원의 저주는 정보의 양이 많아질수록, 그것을 처리하기 위한 계산 복잡도가 높아지며, 불확실성 또한 증가한다는 것이고, GIGO는 입력 데이터가 좋지 못하면, 아무리 좋은 알고리즘을 적용한다고 해도, 좋은 결과가 나올 수 없다는 것을 의미한다. 본 연구에서는, 이 두 가지 문제를 해결하기 위해, 빅데이터의 효율적인 활용을 위해, 데이터의 속성을 관리하는 기법을 제안한다. 제안 방법은 데이터의 차원수를 줄이기 위해서, 데이터의 속성들 사이의 계층 관계를 찾아내어, 하위 속성 그룹을 대표하는 상위 속성을 제시한다. 이 제안 방법은 각 속성의 영향이 미치는 범위를 그 속성의 의미 그룹에 한정시켜 무분별한 속성 관계를 방지하고, 속성 값을 동적으로 검증하여, 학습 및 추론 결과의 정확성을 높일 수 있다.

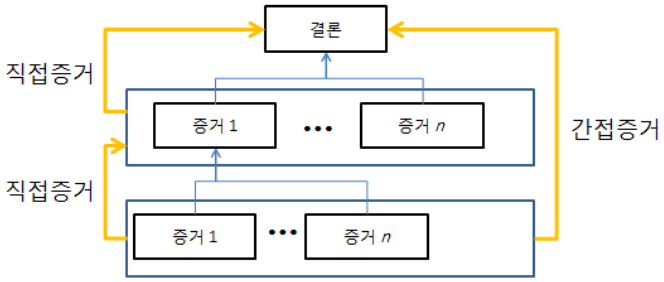
2. 연구 배경

지식을 생성하는 방법은 크게 연역추론(Deduction), 귀납추론(Induction), 그리고 가추론(Abduction)으로 분류할 수 있다 [5]. 연역추론의 경우, 그 전제가 참이라면, 추론 결과로 획득한 지식 역시 논리적으로 참의 진리 값을 확신할 수 있지만, 추론 결과는 전제의 하위 지식이라는 점에서 그 한계를 갖는다. 귀납추론의 경우, 새로운 지식을 획득할 수 있지만, 추론 결과로 획득한 지식의 진리 값이 거짓일 수 있다는 한계를 갖는다. 반면 가추론의 경우, 특정 조건하에서, 새롭게 얻은 지식의 진리 값을 논리적으로 확신할 수 있다는 점에서 연역과 귀납의 장점을 모두 갖고 있다고 할 수 있다. 가추론은, 귀납을 통해 가설을 생성하고, 그 가설을 전제로 삼아, 연역을 수행한 뒤, 연역의

결과의 진리 값을 검증하여, 참이라면, 가설을 참이라고 증명하는 방식을 따르고 있다. 이 과정에서, 현상에 대한 설명력을 획득할 수 있고, 그 설명력을 통해 가설의 진리치를 확인할 수 있게 된다. 그렇기 때문에, 현상에 대한 유일한 가설이라면, 그 조건하에서 그 가설은 논리적으로 참의 진리값을 얻을 수 있다. 본 연구에서는 가추론을 이용해, 속성 간 계층 관계를 유도하는 방법을 제안한다.

3. 제안 방법

추론을 통해 어떤 결과를 도출할 때는 그 결과를 지지하는 증거 혹은 근거가 존재해야 한다. 또한 그 각각의 증거 역시 그것을 지지하기 위한 또 다른 증거가 존재한다. 증거와 결론이 직접적으로 연결되어 있다면, 그것을 직접 증거라고 한다. 증거와 결론 사이에 또 다른 과정이 포함되면, 그 증거는 결론을 위한 간접증거(정황증거)가 된다. 그림 1은 이러한 증거와 결론 사이의 관계를 보여준다.



(그림 1) 직접증거와 간접증거의 개념

최종 결론을 내리기 위한 직접 증거들을 갖고 있다면, 결론의 정확성은 보장할 수 있지만, 현실에서, 결론을 위한 직접 증거를 바로 찾기 보다는 결론을 내리기 위한 정황 증거들을 수집되는 경우가 즐비하며, 또한 이러한 정황 증거들에 대한 결측 역시 빈번하게 발생된다. 이런 경우, 정황증거를 통해 무리하게 결론을 내리려고 하면, 결론에 대한 정확도가 하락하게 되는 결과를 빚게 된다. 이에 본 연구에서는 추론은 직접 증거에 의해서만 이뤄지도록 하기 위해, 속성들의 계층을 생성하고, 이 계층 관계를 통해, 직접 증거 관계를 제한하는 방법을 제안한다.

제안 방법은 다음의 다섯 단계로 이루어진다. 첫 번째 단계에서는 최종 추론을 수행하기 위해 수집 가능한 모든 속성들을 나열한다. 두 번째 단계에서는 도메인 지식을 이용해, 속성들 사이의 상/하위 개념을 묶는다. 이 과정은 다시 유사 속성의 집합을 만드는 과정과, 그렇게 생성된 그룹 속성을 대표하는 신규 속성을 할당하는 과정으로 세분화된다. 세 번째 단계에서는 집단화된 속성이 하위 속성들을 올바르게 포함하고 있는지를 검증한다. 검증 과정은 실제 데이터를 기준으로, 하위 속성의 데이터를 상위 속성이 모두 포함할 수 있는지에 대한 판별을 통해 진행된다. 이 검증과정에서 실패한 그룹 속성은 폐기되고, 두 번째 단계로 되돌아가서, 그룹 속성을 새롭게 생성한다. 네 번째 단계에서는 그룹 속성의 수준을 맞추는 작업을 수행한다. 동일 수준의 속성들의 합집합은 상위 속성과 일치하고, 어느 하위 속성에도 포함되지 않는 데이터는 존재할 수 없으며, 동일 수준의 속성들 사이의 교집합은 없어야 하고, 만약, 존재하는 경우, 그 교집합 부분은 새로운 속성이 되며, 하위 수준의 속성으로 정의된다. 마지막 단계에서는 계층구조에 맞춰 결론을 내기 위한 증거를 제공해주는 과정으로, 각 그룹 속성은 하위 속성의 결과를 취합하여, 각 그룹 속성의 의미에 부합하는 결과를 정리해 상위 속성으로 전달해주는 작업을 수행한다. 제안 방법의 각 단계별 흐름은 표 1과 같다.

<표 1> 속성 계층 생성을 위한 단계

단계	내용
1	데이터의 속성 목록 생성
2	도메인 지식에 기반한 속성 집합 생성
3	생성된 그룹 속성의 유효성 검증
4	그룹 속성 간 구조 생성
5	속성 계층을 이용한 추론 수행

4. 사례 연구

본 연구에서 제안하는 방법을 적용하기 위해, 보건복지부와 사회정보보장원에서 운영하는 복지사각지대 발굴 관리 시스템을 적용 대상으로 선정하였다 [6, 7, 8]. 복지사각지대 발굴은 “수급자가 되어야 함에도 불구하고 다양한 이유로 수급자로 선정되지 못한 경우”를 사전에 탐지하여, 복지 서비스를 제공하는 것으로 현재, 29개의 외부 연계 변수를 이용하고 있으며, 각각의 변수를 구성하는 하위 속성들은 모두 200여개가 존재한다.

실제 발굴 데이터에 대한 접근이 제한된 관계로 본 사례 연구 단계에서는 다섯 단계의 제안 방법 중, 세 번째와 다섯 번째의 단계를 제외된 상태로 적용하였다. 그림 2는 복지사각지대 발굴 모델에 제안 방법을 적용한 결과로, 외부 연계 정보에 대한 계층 정보가 보여진다.

5. 결론

기계학습을 적용함에 있어서, 어떤 속성을 사용할 것인지를 정하는 것은 학습 과정에서의 계산 복잡도에 영향을 미칠 뿐만 아니라, 학습 결과의 정확도에도 영향을 미친다. 보다 정확한 학습과 추론을 수행하기 위해서는 올바른 데이터의 속성의 정의와 선택이 필요하다. 본 연구에서는 속성 선택을 위해, 속성들의 의미 기반 계층 구조를 생성하여, 학습 목적에 맞는 속성 선택을 지원하기 위한 기법을 제안하였다. 사례 연구에서는 제안된 방법을 복지사각지대 발굴 모델에 적용해, 기존에 사용된 200여개의 발굴 속성을 계층화 시켰다. 본 방법은 추론 결과에 대한 설명력을 증대시키고, 추론 과정에서의 계산 복잡도를 낮출 수 있다는 장점을 갖고 있지만, 이를 위해 속성 간 계층 관계를 도출하기 위한 도메인 전문가의 지식을 필요로 한

다. 이 부분을 보완하기 위해, 속성들 사이의 의미 관계를 자동으로 추출하기 위한 방법에 추후 연구로 진행할 예정이다.

Acknowledgement

본 연구는 문화체육관과부 및 한국콘텐츠진흥원의 2019년도 문화기술연구개발 지원사업으로 수행되었음

참고문헌

[1] Chun, K. W., Kim, H., and Lee, K. "A Study on Research Trends of Technologies for Industry 4.0; 3D Printing, Artificial Intelligence, Big Data, Cloud Computing, and Internet of Things". In Advanced Multimedia and Ubiquitous Engineering (pp. 397-403). Springer, Singapore. 2018

[2] Russell, S. J., and Norvig, P. "Artificial intelligence: a modern approach". Malaysia; Pearson Education Limited,.2016

[3] Bishop, C. M. "Pattern recognition and machine learning". springer. 2006

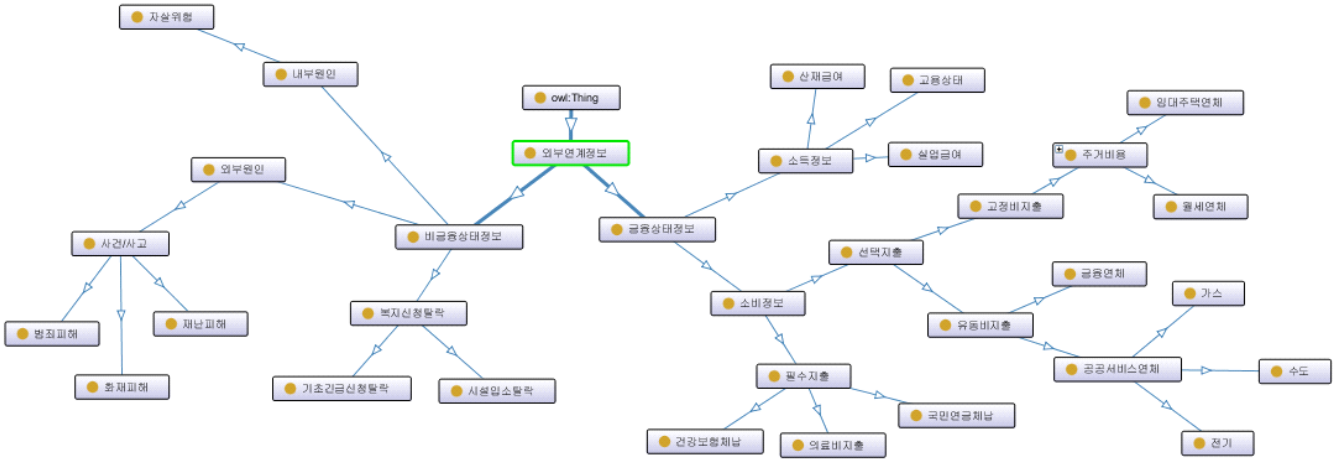
[4] Grimes, D. A. "Epidemiologic research using administrative databases: garbage in, garbage out". Obstetrics & Gynecology, 116(5), 1018-1019. 2010

[5] Reichertz, J. "Abduction: The logic of discovery of grounded theory (pp. 214-228)". London: Sage. 2007

[6] 이인수 "취약계층 발굴지원II: 복지 사각지대 발굴 관리 시스템 운영 현황" 사회보장정보 브리프 8호, 2018

[7] 김은하·추병주·김경준·최현수·오미애·김용대·김경범·이창수·신동민·주명원·위세아, "사회보장정보시스템을 활용한 복지 사각지대 발굴방안 연구" 사회보장정보원 정책보고서, 2015

[8] 이우식, 이인수 "인공지능 Deep Learning을 기술을 활용한 복지사각지대 발굴 개념 연구" 사회보장정보원 정책연구보고서 18-07, 2018



(그림 2) 외부 연계 변수에 대한 계층 관계도