CDRPs 를 이용한 인공 신경망에서 추출된 규칙 개선방법

이헌주*, 김현철**
*고려대학교 컴퓨터학과
**고려대학교 컴퓨터학과
e-mail : boxerlee@korea.ac.kr

Improved rule extraction from artificial neural network using CDRPs

Hurn-Joo Lee*, Hyeoncheol Kim**
*Dept. of Computer Science, Korea University
** Dept. of Computer Science, Korea University

요 약

최근 인공 신경망은 다양한 분야에서 뛰어난 성능을 보여주고 있지만 인공 신경망이 학습한 지식이 어떠한 내용인지를 사람이 이해하기 어렵다는 문제점이 있다. 이와 같은 문제점을 해결하기 위한 방법 중 하나로 인공 신경망으로부터 인간이 이해할 수 있는 형태의 규칙을 추출하는 방법들이 고안이 되었다. 본 연구에서는 규칙추출 알고리즘 중 하나인 OAS 알고리즘을 이용해 규칙을 추출해보고 CDRPs(Critical Data Routing Paths)를 활용하여 추출한 규칙의 품질을 개선하는 방법을 제시하였다.

1. 서론

최근 인공 신경망이 딥러닝으로 진화하면서 다양한 분야에서 활용되며 뛰어난 성능을 보여주고 있다. 하지만 인공 신경망의 이러한 뛰어난 성능에도 불구하고 인공 신경망이 학습한 지식이 어떠한 내용인지를 사람이 이해하기 어려워 의사 결정 오류가 치명적인 결과를 가져올 수 있는 높은 신뢰도 검증을 요구하는 분야에서 사용하기에는 아직 위험성이 있다는 문제점이 있다[1].

이와 같은 문제점을 해결하기 위한 방법 중 하나로 인공 신경망으로부터 인간이 이해할 수 있는 형태의 규칙을 추출하는 방법들이 고안이 되었다.

본 연구에서는 decompositional 접근법을 사용하는 규칙추출 알고리즘 중 하나인 OAS(ordered-attribute search)알고리즘을 사용하여 학습된 인공 신경망으로부터 규칙을 추출해보고, 학습된 인공 신경망으로부터 distillation guided routing method 를 이용해 CDRPs(Critical Data Routing Paths)를 추출하여 각 입력 샘플별로 인공 신경망의 중요한 노드와 중요하지 않은 노드를 구별하여 이를 추출된 규칙의 품질 개선에 적용시켜 보았다.

그 결과 IRIS 도메인 데이터에서 추출한 규칙의 정확도와 커버리지를 많이 감소시키지 않으면서 규칙의 개수를 평균 99 개에서 22 개까지 줄여 규칙의 품질을 향상시킬 수 있었다.

2. 관련 연구

2.1. 규칙추출 알고리즘

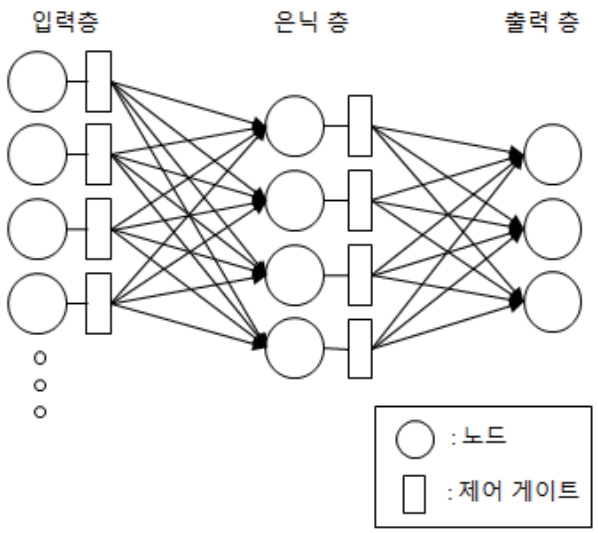
학습된 인공 신경망으로부터 규칙을 추출하는 연구는 decompositional 접근법과 pedagogical 접근법, 그리고 eclectic 접근법이 있다[1][2]. 그 중 decompositional 접근법은 인공 신경망을 화이트박스로 보고 규칙을 추출하는 접근 방법이다. decompositional 접근법은 계산 비용이 많이 들고 검색 공간을 많이 사용하지만 투명성 측면에서 다른 접근법 보다 뛰어나다고 할 수 있다.

본 연구에서 사용한 OAS 알고리즘 역시 decompositional 접근법을 기반으로 만들어진 알고리즘으로 기존의 decompositional 접근법 기반의 알고리즘이 갖는 문제점인 계산비용과 검색공간이 많이 필요한 부분을 개선한 알고리즘이다[3].

2.1. CDRPs(Critical Data Routing Paths)

해석 가능한 인공 신경망(Interpretable Artificial Neural Network)에 대한 방법 중 하나로 각각의 입력 샘플에 대한 데이터 라우팅 경로상의 중요한 노드를 찾는 방법이 있다[4]. 이 방법은 (그림 1)과 같이 학습된 인공 신경망 모델에 distillation guided routing method 를 이용해 각 층의 출력 채널에 제어 게이트를 두고 이 모델이 기존의 모델과 출력 값이 비슷해지도록 제어 게이

트를 학습시켜 CDRPs 를 추출하게 된다. 이렇게 추출되어진 CDRPs 는 각각의 입력 데이터에 대해 인공신경망에서 중요한 노드가 어느 것인지를 알 수 있게 해준다.



(그림 1) 제어 게이트가 결합된 인공 신경망

3. 연구 방법

3.1. 연구 자료

본 연구에서는 실험을 위해 비교적 간단한 공개 데이터인 IRIS 도메인을 적용하였다. 그 이유는 데이터가 복잡한 경우 규칙을 추출하는데 시간이 오래 걸리는 문제점이 있기 때문이다. 그리고 본 연구의 목적이 추출한 규칙에 CDRPs 를 활용해 규칙을 개선했을 때 어떠한 영향이 발생하는지를 확인하는 것이기 때문에 간단한 데이터로도 실험이 가능하였다.

IRIS(붓꽃) 도메인은 통계학자인 피셔가 소개한 데이터로 붓꽃의 3 가지 종에 대해 꽃받침과 꽃잎의 너비와 길이를 정리한 데이터다.

3.2. 연구 절차

본 연구에서는 OAS 알고리즘을 적용해 규칙을 추출해보고 CDRPs 를 활용했을 때의 결과를 확인해보기 위해 다음과 같은 절차를 통해 연구를 진행하였다.

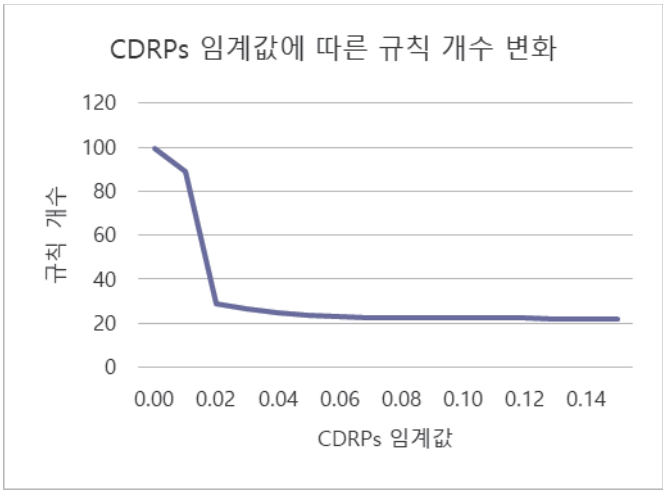
IRIS 데이터를 인공 신경망에 학습시킨 다음 OAS 알고리즘을 통해 규칙을 추출하는 방법은 이전에 OAS 알고리즘 연구에서 사용했던 모델과 동일한 방법을 이용하였다[3]. IRIS 데이터에 OAS 알고리즘을 적용하기 위해 입력 속성값을 3 개의 간격을 갖는 속성으로 이산화 시켜, 총 12 개의 속성을 갖도록 변환하였고, 12 개의 입력 값으로 구성된 입력 층과 4 개의 노드로 구성된 은닉 층, 그리고 3 개의 노드로 구성된 출력 층으로 이루어진 인공 신경망을 구성하였다.

이렇게 구성된 인공 신경망에서 OAS 알고리즘을 이용해 규칙을 추출하고, distillation guided routing method 를 이용해 CDRPs 를 추출하였다. 원래 CDRPs 를 추출하는 방법은 각 입력 샘플에 대해 CDRPs 를 추출하지만 본 연구에서는 각 입력 샘플 별 CDRPs 를 구하지 않고 데이터를 카테고리별로 나누어 각 카테고리별 CDRPs 를 추출하였다. 그리고 CDRPs 를 추출할 때 중요한 노드인지 중요하지 않은 노드인지의 기준이 되는 CDRPs 임계값을 0 에서 0.15 까지 변경을 하면서 규칙의 결과에 어떻게 영향을 주는지 실험을 통해 관찰하였다.

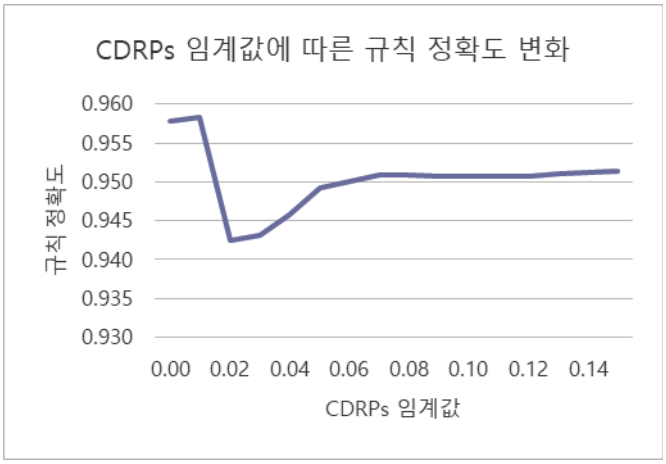
마지막으로 CDRPs 임계값을 기준으로 이진화된 CDRPs 정보를 이용해 중요하지 않은 노드(negligible node)와 관련이 있는 규칙은 없어도 되는 규칙으로 간주하고 규칙 목록에서 제거하는 과정을 진행하였다.

4. 연구 결과

실험은 총 100 번을 반복해서 진행한 다음 평균값을 계산하였고, CDRPs 임계값이 0 인 경우는 CDRPs 를 사용하지 않은 경우이다.



(그림 2) CDRPs 임계값에 따른 규칙 개수 변화

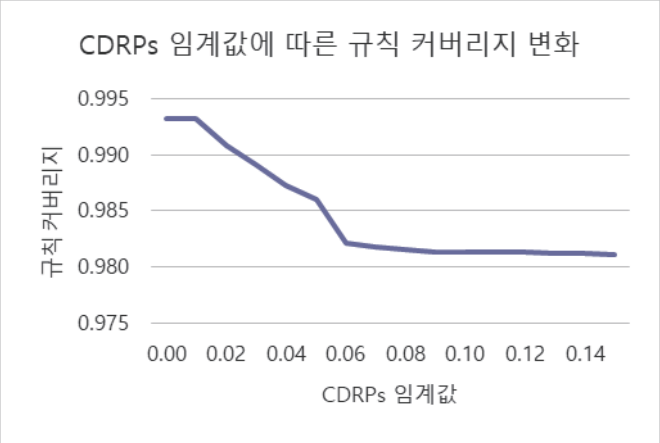


(그림 3) CDRPs 임계값에 따른 규칙 정확도 변화

(그림 2)에서는 CDRPs 임계값이 증가함에 따라 규칙의 수가 99 개에서 22 개로 규칙의 개수가 기존대비

22% 수준으로 줄어드는 것을 실험결과에서 볼 수 있다. CDRPs 임계값이 0.01 에서 0.02 로 갈 때 규칙 개수의 감소폭이 가장 컸었다.

(그림 3)의 결과에서는 CDRPs 가 증가함에 따라 처음에는 규칙 정확도가 평균 0.958 에서 0.942 까지 떨어졌으나 이후에 다시 0.951 까지 상승하였다. 0.007 정도가 감소하여 크게 정확도가 떨어지지 않는 것을 관찰할 수 있었다.



(그림 4) CDRPs 임계값에 따른 규칙 커버리지 변화

(그림 4)의 결과에서는 CDRPs 임계값에 따라 규칙의 커버리지가 0.993 에서 0.981 까지 약간 떨어졌는데 정확도의 감소폭이 0.012 로 크지 않은 것을 관찰할 수 있었다.

종합적으로 결과를 살펴보면 CDRPs 임계값에 따라 규칙의 정확도와 커버리지가 약간씩 줄어드는 현상이 보이긴 했지만 정확도의 감소폭은 0.007 이고 커버리지의 감소폭은 0.012 로 감소폭이 크지 않았고, 규칙의 평균 개수가 99 개에서 22 개까지 줄어들어 적은 규칙만으로도 인공 신경망을 나타낼 수 있어 어느정도 유의미한 결과를 만들어 냈다고 할 수 있다. IRIS 도메인의 경우 CDRPs 임계값이 0.07 일 때 전체적으로 가장 좋은 결과가 도출되었다.

5. 결론

본 연구에서는 학습된 인공 신경망으로부터 사람이 이해할 수 있는 형태의 규칙을 만들어주는 OAS 알고리즘을 적용해 규칙을 추출해 보았다. 그리고 학습된 인공 신경망으로부터 CDRPs 를 추출해 중요하지 않은 노드를 찾아내어 그에 해당하는 규칙을 제거하는 방법을 시도 하였다. 그 결과 규칙의 정확도와 커버리지를 많이 떨어뜨리지 않으면서도 규칙의 개수를 평균 99 개에서 22 개로 줄이는 결과를 얻을 수 있었다.

앞으로의 계획은 본 연구에서는 하나의 은닉 층을 갖고 단순한 데이터를 기반으로 실험을 진행하였는데, 향후에는 2 개 이상의 은닉 층을 갖는 깊은 인공 신경망의 규칙을 추출하는 시도를 해보고 개선 방향을 찾아볼 계획이다.

참고문헌

[1] T. Hailesilassie. “Rule extraction algorithm for deep neural networks: A review.” International Journal of Computer Science and Information Security, 14, 7, 2016, p. 376

[2] Andrews, R. Diederich, J. & Tickle, A. B. “Survey and critique of techniques for extracting rules from trained artificial neural networks.” Knowledge-Based System, 8(6), 1995, 373-389,

[3] Kim H. “Computationally Efficient Heuristics for If-Then Rule Extraction from Feed-Forward Neural Networks.” Lecture Notes in Computer Science, vol 1967. Springer, Berlin, Heidelberg, 2000

[4] Yulong Wang, Hang Su, Bo Zhang, Xiaolin Hu. “Interpret Neural Networks by Identifying Critical Data Routing Paths” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8906-8914