인공 신경망에서 은닉 유닛 명확화를 이용한 효율적인 규칙추출 방법

이헌주† · 김현철††

요 약

인공 신경망은 최근 다양한 분야에서 뛰어난 성능을 보여주고 있다. 하지만 인공 신경망이 학습한 지식이 정확히 어떤 내용인지를 사람이 파악하기 어렵다는 문제점이 존재하는데, 이를 해결하기 위한 방법 중 하나로 학습된 인공 신경망에서 규칙을 추출하는 방법들이 연구되고 있다. 본 연구에서는 학습된 인공 신경망으로부터 규칙을 추출하는 방법 중 하나인 ordered-attribute search(OAS) 알고리즘을 사용하여 인공 신경망으로부터 규칙을 추출해보고, 추출된 규칙을 개선하기 위해 규칙들을 분석하였다. 그 결과로 은닉 층의 출력값 분포가 OAS 알고리즘을 이용해 추출된 규칙의 정확도에 영향을 주는 것을 파악하였고, 은닉 유닛 명확화 기법을 통해 은닉 층 출력값을 이진화하여 효율적인 규칙을 추출할 수 있음을 제시하였다.

주제어 : 인공 신경망, 규칙추출, 은닉 유닛 명확화

A Efficient Rule Extraction Method Using Hidden Unit Clarification in Trained Neural Network

Hurn-joo Lee† · Hyeoncheol Kim††

ABSTRACT

Recently artificial neural networks have shown excellent performance in various fields. However, there is a problem that it is difficult for a person to understand what is the knowledge that artificial neural network trained. One of the methods to solve these problems is an algorithm for extracting rules from trained neural network. In this paper, we extracted rules from artificial neural networks using ordered-attribute search(OAS) algorithm, which is one of the methods of extracting rules, and analyzed result to improve extracted rules. As a result, we have found that the distribution of output values of the hidden layer unit affects the accuracy of rules extracted by using OAS algorithm, and it is suggested that efficient rules can be extracted by binarizing hidden layer output values using hidden unit clarification.

Keywords : Artificial Neural Network, Rule Extraction, Hidden Unit Clarification

† 정 회 원: 고려대학교 정보대학 컴퓨터학과
†† 종신회원: 고려대학교 정보대학 컴퓨터학과 교수(교신저자)
논문접수: 2018년 1월 11일, 심사완료: 2018년 1월 23일, 게재확정: 2018년 1월 29일
* 이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4003558)

1. 서론

인공 신경망(artificial neural network)은 최근 계층을 깊게 쌓는 딥 러닝(deep learning)으로 진화하면서 양상 및 음성인식이나 번역 등의 분야에서 뛰어난 성능을 보여주고 있다.

하지만 이러한 뛰어난 성능에도 불구하고 인공 신경망이 학습한 지식이 어떠한 내용인지를 사람이 이해하기 어려워, 의사결정 오류가 치명적인 결과를 가져올 수 있는 높은 신뢰도 검증을 요구하는 분야에서 사용하기에는 아직 위험성이 있다는 문제점이 있다[1].

이와 같은 인공 신경망의 문제점을 해결하기 위한 방법 중 하나로 인공 신경망으로부터 인간이 이해할 수 있는 형태의 규칙을 추출하는 방법들이 고안되었다. 규칙추출(rule extraction)은 decompositional/pedagogical/eclectic 총 3가지 접근법을 통해 주로 이루어져 왔다[1][2]. 이러한 규칙추출 접근법 중 decompositional approach는 인공 신경망을 화이트 박스로 보고 규칙을 추출하는 접근 방법이다.

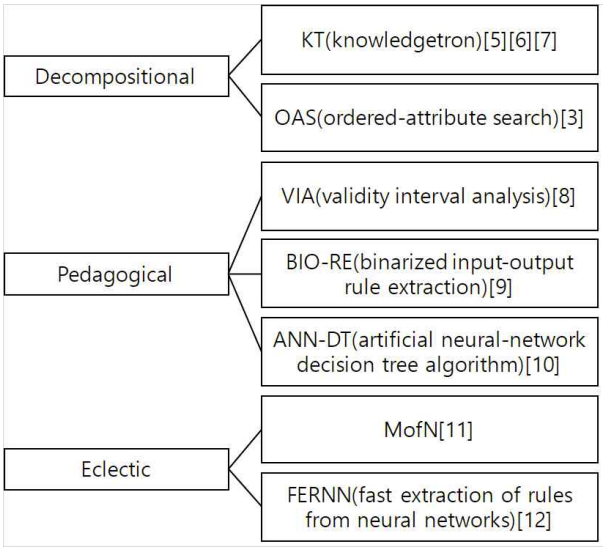
본 연구에서는 decompositional approach 중 하나인 ordered-attribute search(OAS) 알고리즘[3]을 사용하였을 때 어떠한 규칙이 나오는지 살펴보고 그 결과를 향상시킬 방안에 대해 조사하였다. OAS 알고리즘과 같은 decompositional approach의 단점 중 하나가 은닉 유닛의 활성화 함수인 sigmoid를 통해 나온 출력값을 이진화 하는 과정에서 정확도가 떨어질 수 있다는 것인데, 본 연구에서는 은닉 유닛의 출력값을 분석하여 이에 대해 명확하게 파악을 해보고, 해결 방안으로 구조적 학습법 중 은닉 유닛의 명확화 기법[4]을 적용하여 OAS 알고리즘의 결과에 어느 정도의 향상이 발생하는지를 실험을 통해 확인하였다.

본 논문의 구성은 다음과 같다. 2장 이론적 배경에서는 각 규칙추출 알고리즘과 은닉 유닛 명확화 알고리즘에 대해 소개를 하였다. 3장 연구방법에서는 연구에 사용된 데이터와 데이터 전처리에 대한 설명과, OAS 알고리즘과 은닉 유닛 명확화 알고리즘의 적용 절차에 대해 설명하였다. 4장에서는 실험 결과 분석 및 은닉 유닛 명확화 적용으로 결과가 어떻게 바뀌었는지를 설명하였고, 5장에서는 연구 결과에 대한 요약 및 앞으로의 연구방향에 대해 서술하였다.

2. 이론적 배경

2.1 규칙추출 알고리즘

학습된 인공 신경망으로부터 규칙을 추출하는 연구는 [그림 1]과 같이 decompositional approach와 pedagogical approach, 그리고 eclectic approach 총 3가지 접근법이 있다.



[그림 1] Rule Extracion 알고리즘

2.1.1 Decompositional Approach

decompositional approach는 인공 신경망을 화이트박스로 보고 규칙을 추출하는 접근 방법이다. decompositional approach는 계산 비용이 많이 들고 검색 공간을 많이 사용하는 반면 각 층의 유닛을 하나하나 들여다보고 규칙을 추출하기 때문에 투명성 측면에서 다른 접근법 보다 뛰어나다[1]. decompositional approach는 다음과 같은 단계를 포함하게 된다[3].

1. 인공 신경망의 개별 유닛의 수준에서 중간규칙을 추출한다.
2. 각 유닛으로부터 추출된 중간규칙은 재 작성(rewrite)단계를 거쳐 심볼을 제거해 준다. 재 작

성 과정 중에는 다른 규칙과 중복되거나, 다른 규칙에 포함되거나 서로 불일치되는 규칙들이 제거된다.

최초로 제안된 decompositional approach들 중 하나는 Fu(1993)가 제안한 KT(knowledgetron) 알고리즘이다[5][6][7]. KT 알고리즘은 인공 신경망의 구조를 가장 직관적인 방법으로 사용하여 규칙을 추출하는 방법이라고 할 수 있다. 모든 뉴런을 If-Then 규칙으로 나타내는 것으로, 뉴런의 임계 값을 초과하게 만드는 입력값의 조합을 찾아낸다. KT 알고리즘이 만들어 내는 규칙은 confirm rule/disconfirm rule 두 가지 종류가 존재하는데 confirm rule의 경우는 출력을 활성화되는 규칙이고 disconfirm rule은 그 반대에 해당한다.

본 연구에서 사용한 OAS 알고리즘 역시 decompositional approach를 기반으로 만들어진 알고리즘이다. 기존의 다른 decompositional approach가 갖는 문제점 중 하나로 계산하는데 비용이 많이 들고 규칙 검색 공간이 입력 속성의 개수에 따라 지수적으로 증가하는 문제점이 있었는데, OAS 알고리즘에서는 규칙 검색 공간을 줄이고 계산적으로 효율적인 방법을 제안하였다.

OAS 알고리즘을 통해 규칙을 추출할 때 각 노드의 출력값이 sigmoid를 통하게 되는데, 이때 출력이 명확하게 활성화 또는 비활성화 되는 것이 아닌 그 사이의 값을 가지게 되고, 이로 인해 규칙의 정확도가 떨어지는 결과를 갖게 될 수 있다.

2.1.2 Pedagogical Approach

pedagogical approach는 앞에서 설명한 decompositional approach와 달리 인공 신경망의 내부 구조를 고려하지 않고 인공 신경망을 블랙박스로 보고 입력과 출력만의 관계를 통해 규칙을 얻어내는 접근법이다. pedagogical approach는 모든 층의 유닛에 대해 개별로 알고리즘을 적용해 재작성 과정을 거치는 decompositional approach에 비해 검색 공간과 계산 시간 측면에서 효율적이지만 입력과 출력만의 관계를 살펴보는 블랙박스 구조이다 보니 투명성 측면에서 비교적 떨어진다고 볼 수 있다[1].

pedagogical approach는 VIA(validity interval analysis)[8], BIO-RE(binarized input-output rule extraction)[9], ANN-DT[10] 와 같은 방법들이 존재한다.

2.1.3 Eclectic Approach

eclectic approach은 앞에서 설명한 decompositional approach와 pedagogical approach를 둘 다 혼합해서 사용하는 접근법이다. 대표적인 알고리즘으로는 MofN[11], FERNN(fast extraction of rules from neural networks)[12]와 같은 방법들이 존재한다.

2.2 은닉 유닛 명확화 알고리즘

앞에서 언급한 바와 같이 학습된 인공 신경망의 각 노드의 출력값을 명확하게 이진화 시킬 수 없는 문제가 존재하는데, 본 연구에서는 이를 해결하기 위해 은닉 유닛 명확화 알고리즘[그림 2]을 적용하여 이 문제의 개선을 시도하였다. 은닉 유닛 명확화 알고리즘은 구조적 학습(structural learning with forgetting)[4]에서 언급된 3가지 알고리즘 중 하나로 은닉 유닛의 출력값이 최대한 활성화되거나 최대한 비활성화되는 방향으로 인공 신경망을 학습해주는 알고리즘이다. 은닉 유닛 명확화를 이용한 학습은 다음과 같이 나타낼 수 있다.

$$J_h = J + c\sum_i \min\{1-h_i,\ h_i\}$$

[그림 2] 은닉 유닛의 명확화

위 식에서 J는 기존의 비용함수이고, hi는 0~1 사이의 값을 갖는 i번째 은닉 유닛의 출력값이다. J에 더해지는 값은 페널티 항목이 되는데 c는 이 페널티 항목의 가중치가 된다. 페널티 항목은 모든 hi의 값이 0이나 1에 가까워질 때 작은 값을 갖게 된다. 이 페널티 항목이 더해진 새로운 비용함수 Jh로 인해 인공 신경망의 각 은닉 유닛의 출력값이 완전 활성화 또는 완전 비활성화에 가

까워지도록 학습되어진다.

3. 연구방법

3.1 연구 자료

본 연구에서는 실험을 위해 비교적 간단한 공개 데이터인 IRIS 도메인을 적용하였다. 간단한 데이터 셋을 사용한 이유는 규칙추출 알고리즘인 OAS의 계산 복잡도가 높아 복잡한 데이터 셋을 사용했을 때 결과가 나오기까지 시간이 오래 걸리는 문제점이 있고, 인공 신경망에 은닉 유닛 명확화 기법을 적용했을 때 OAS 알고리즘을 통한 규칙추출에 어떠한 영향이 발생하는지를 확인 하는 것이 이 연구의 주 목적이기 때문에 간단한 데이터를 사용해 실험을 진행하였다.

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

3.2 연구 절차

본 연구에서는 인공 신경망에 OAS 알고리즘을 적용해 만들어진 규칙을 분석해보고, 그 결과가 은닉 유닛 명확화 알고리즘을 통해 향상되는지를 확인하기 위해 다음과 같은 절차를 통해 연구를 진행하였다.

첫 번째로, IRIS 데이터에 OAS 알고리즘을 적용하기 위해 <표 1>처럼 연속된 값을 갖는 입력 속성을 3개의 간격을 갖는 속성으로 이산화 시켜, 총 12개의 속성을 갖는 데이터로 변환하였다[3].

두 번째로, IRIS 데이터를 학습시키기 위해 입력 층, 한 개의 은닉 층, 출력 층으로 구성된 인공 신경망을 구성하였다. 인공 신경망은 <표 2>에서처럼 입력 층은 이산화된 12개의 속성을 입력값으로 갖기 위해 12개의 노드로 구성이 되고, 은닉 층은 4개의 노드, 출력 층은 3개의 노드로 구성이 되었다. 은닉 층과 출력 층의 활성함수는 sigmoid를 사용하였고, 비용함수로는 크로스 엔트로피를 사용하였다.

세 번째로, OAS 알고리즘을 적용하여 학습된 인공 신경망으로부터 규칙을 추출해보고, 은닉 층 출력값 분석을 통해 그 규칙의 정확도를 떨어뜨리는 원인을 확인하였다.

<표 1> 이산화된 IRIS 도메인 데이터

원본 데이터 속성	이산화된 속성
sepal-length	sepal-length <= 5.4
	5.4 < sepal-length <= 6.3
	6.3 < sepal-legnth
sepal-width	sepal-width <= 2.8
	2.8 < sepal-width <= 3.1
	3.1 < sepal-width
petal-length	petal-length <= 2.7
	2.7 < petal-length <= 5
	5 < petal-length
petal-width	petal-width <= 0.7
	0.7 < petal-width <= 1.6
	1.6 < petal-length

네 번째로, 은닉 유닛 명확화 기법을 적용하여 은닉 층 출력값과 규칙의 결과가 어떻게 향상되는지 확인하였다.

<표 2> 인공 신경망 구조

	입력 층	은닉 층	출력 층
노드 개수	12	4	3

4. 연구결과

4.1 학습된 인공 신경망의 분석

IRIS 데이터를 갖고 인공 신경망을 2000 epoch를 학습시켰을 때 인공 신경망의 은닉 층의 출력값 분포는 [그림 3]과 같다. 활성화 함수가 sigmoid이기 때문에 은닉 층의 출력값은 0~1사이의 값을 갖게 된다. 대부분 0~0.15사이와 0.7~1사이의 값을 갖는 것을 확인할 수 있지만 그 사이의 값도 일부 존재하는 것을 확인할 수 있다.

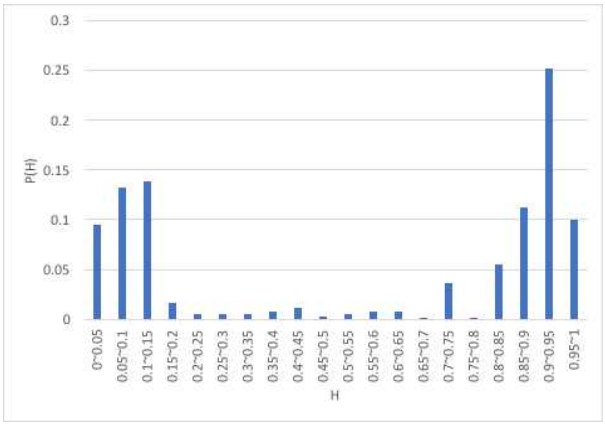
4.2 OAS 알고리즘 결과 분석 결과

학습한 인공 신경망에 OAS 알고리즘을 적용하여 규칙을 추출하는 실험을 진행하였다. 총 100회

<표 3> IRIS도메인으로부터 추출한 규칙중 일부

추출한 규칙
If (petal-length <= 2.7) and (petal-width <= 0.7) then setosa 1.000
If (sepal-length <= 5.4) and not (2.7 < petal-length <= 5) and (petal-width <= 0.7) then setosa 1.000
If not (6.3 < sepal-length) and not (2.7 < petal-length <= 5) and (petal-width <= 0.7) then setosa 1.000
If not (2.7 < petal-length <= 5) and not (5 < petal-length) and (petal-width <= 0.7) then setosa 1.000
If (2.7 < petal-length <= 5) and (0.7 < petal-width <= 1.6) then versicolor 0.979
If (2.7 < petal-length <= 5) and not (petal-width <= 0.7) and not (1.6 < petal-length) then versicolor 0.979
If not (sepal-length <= 5.4) and (2.7 < petal-length <= 5) and not (petal-width <= 0.7) then versicolor 0.843
If not (petal-length <= 2.7) and not (5 < petal-length) and (0.7 < petal-width <= 1.6) then versicolor 0.979
If (5 < petal-length) and (1.6 < petal-length) then virginica 1.000
If (sepal-width <= 2.8) and (1.6 < petal-length) then virginica 1.000
If not (petal-length <= 2.7) and not (2.7 < petal-length <= 5) and (1.6 < petal-length) then virginica 1.000
If not (3.1 < sepal-width) and not (2.7 < petal-length <= 5) and (1.6 < petal-length) then virginica 1.000

의 실험을 반복하였고, 반복 실험한 결과의 평균을 구하였다. 그 결과로 평균 82.91개의 규칙이 추출되었으며, 추출된 규칙은 0.967의 정확도를 나타내었다.



[그림 3] 은닉 층 출력값 분포

<표 3>에 표기된 규칙은 OAS 알고리즘을 IRIS 도메인에 적용하여 If-Then 규칙으로 추출한 결과 중 일부이다. 위의 결과를 좀 더 개선할 방안을 찾아보기 위해 추출된 규칙들을 정확도가 높은 것과 떨어지는 것으로 분류해 분석을 시도하였다. 먼저 정확도가 떨어지는 규칙부터 분석을 진행하였는데, 아래에 표시된 규칙은 정확도가 20%에 불과한 규칙이다.

* If (2.7 < petal-length <= 5) and (1.6 < petal-length) then versicolor 0.2

위 규칙에 해당하는 데이터들이 인공 신경망을 통과할 때 은닉 층의 출력값을 뽑아본 결과 <표 4>와 같은 결과를 얻을 수 있었다.

<표 4> 정확도가 20%인 규칙에 해당하는 데이터의 은닉 층 출력값

Node 1	Node 2	Node 3	Node 4
0.103	**0.423**	**0.539**	**0.485**
0.059	0.295	**0.692**	**0.538**
0.128	**0.340**	**0.674**	**0.317**
0.05	0.229	0.785	**0.421**
0.05	0.229	0.785	**0.421**
0.05	0.229	0.785	**0.421**
0.05	0.229	0.785	**0.421**
0.069	**0.310**	**0.691**	**0.458**
0.069	**0.310**	**0.691**	**0.458**
0.05	0.229	0.785	**0.421**

<표 4>를 분석했을 때 0.3~0.7에 해당하는 값이 많이 존재하는 것을 알 수 있었다.

이번에는 위와 반대되는 경우를 살펴보기 위해 정확도가 100%인 규칙도 살펴보았다.

* If (6.3 < sepal-length) and (2.7 < petal-length <= 5) and not (1.6 < petal-length) then versicolor 1.00

<표 5>는 위 규칙에 해당하는 데이터들이 인공 신경망을 통과할 때 은닉 층의 출력값인데, 출력값을 살펴보면 앞에 <표 4>와 다르게 0.3~0.7 사이에 해당하는 값이 전혀 존재하지 않는 것을

확인할 수 있었다.

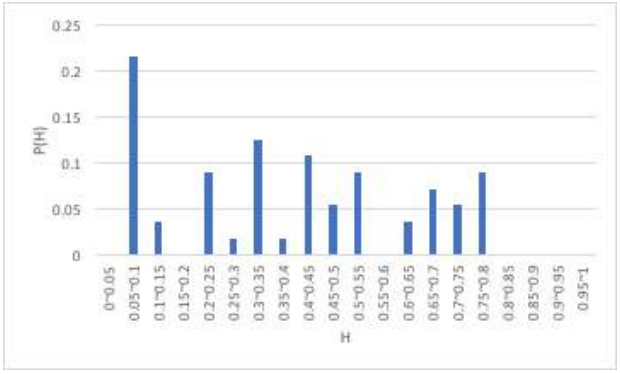
<표 5> 정확도가 100%인 규칙에 해당하는 데이터의 은닉 층 출력값

Node 1	Node 2	Node 3	Node 4
0.067	0.859	0.087	0.950
0.067	0.859	0.087	0.950
0.032	0.788	0.154	0.944
0.044	0.711	0.229	0.936
0.032	0.788	0.154	0.944
0.044	0.788	0.154	0.944
0.044	0.788	0.154	0.944
0.044	0.788	0.154	0.944
0.032	0.711	0.229	0.936
0.044	0.788	0.154	0.944

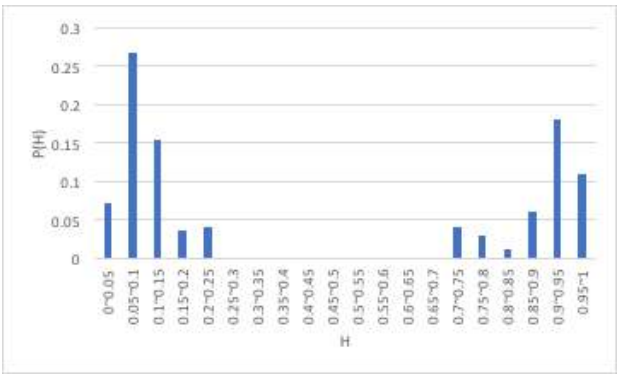
위 결과들을 좀 더 정확하게 확인해보기 위해 정확도가 75% 미만인 모든 규칙들에 해당하는 데이터가 인공 신경망을 통과할 때의 은닉 층 출력값과 그와 반대 경우의 출력값을 각각 뽑아 분포로 나타내어본 결과 [그림 4][그림 5] 과 같은 결과를 얻을 수 있었다.

[그림 4]는 주로 0.3~0.7 사이의 값에 출력값이 몰려 있는 반면 [그림 5]는 0.3~0.7 사이의 값이 완전히 제외된 것을 확인할 수 있다.

이 실험 결과로 은닉 층의 출력값이 명확하게 이진화 되지 않는 경우 잘못된 규칙을 뽑아낼 확률이 높다는 것을 확인할 수 있었고, 반대의 경우에는 정확도가 높은 규칙을 뽑아내는 것을 확인할 수 있었다.



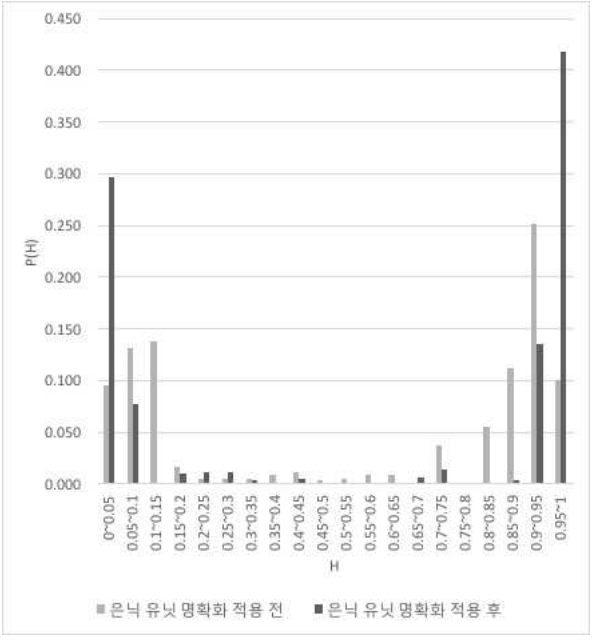
[그림 4] 정확도 75%미만 규칙의 은닉 유닛 출력값 분포



[그림 5] 정확도 75%이상 규칙의 은닉 유닛 출력값 분포

4.3 은닉함수 명확화 적용 결과 분석

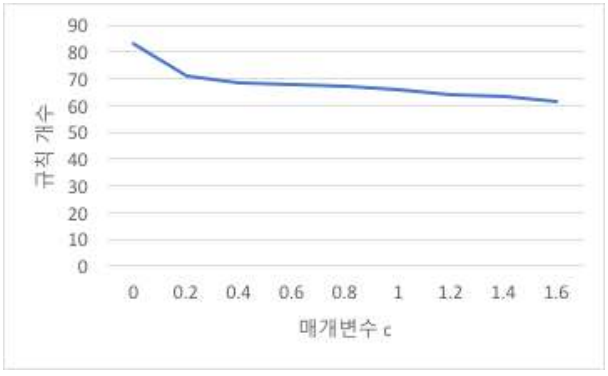
위 실험 결과로 은닉 층 출력값이 규칙의 정확도에 영향을 주는 것을 확인하였고, 이 부분을 개선하는 것이 규칙추출에 있어 중요한 부분임을 알 수 있었다. 이제 이 문제점을 개선해 보기 위해 은닉 함수 명확화를 적용하는 실험을 진행하였다. 앞에서 학습한 인공 신경망에 추가로 2000 epoch를 [그림 2]에 있는 식을 적용하여 학습 시켜 은닉 유닛 명확화를 진행하였다. 그 결과 [그림 6]과 같이 은닉 층 출력값의 분포가 변화하였다.



[그림 6] 은닉 유닛 명확화 적용 전후 비교

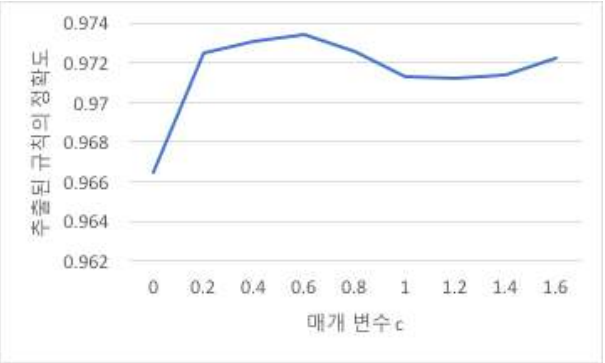
[그림 6]의 분포를 살펴보면 이전 결과와 비교했을 때 중간 값에 해당하는 0.3~0.7값이 줄어들고 완전한 활성화 혹은 완전한 비활성화 쪽으로 값이 몰려, 이전보다 더 이진화를 잘 시킨 것을 확인할 수 있다.

그다음으로는 비용함수의 페널티 항목 가중치인 매개변수 c값을 변화시켜 가면서 실험을 하여 추출된 규칙 개수와 정확도가 c값에 따라 어떻게 변화하는지를 살펴보았다. [그림 7]의 실험결과를 살펴보면 매개변수 c가 0인 경우는 은닉 유닛 명확화를 적용하지 않았을 때의 값인데 c가 조금씩 증가할수록 규칙의 개수가 점차 줄어드는 것을 그래프를 통해 확인할 수 있었다. c가 1.6이었을 때는 은닉 유닛 명확화를 적용하지 않았을 때 보다 규칙의 개수가 88.12개에서 61.62개로 30%가량 줄어든 것을 확인할 수 있었다. 만약 이렇게 규칙이 줄어들면서 정확도와 범위(coverage)가 같이 감소한다면 이 실험 결과가 의미가 없을 수도 있지만 범위는 100%에서 변화가 없었고, 정확도의 경우는 아래와 같이 c가 증가함에 따라 오히려 개선되는 것을 볼 수 있다.



[그림 7] c의 변화에 따른 규칙 개수 변화

[그림 8]을 살펴보면 은닉 유닛 명확화를 적용한 경우와 그렇지 않은 경우를 비교했을 때 적용한 경우가 오히려 정확도가 증가하는 것을 확인할 수 있었다. 이 실험 결과를 바탕으로 OAS 알고리즘을 사용할 때 은닉 유닛 명확화를 통해 추가 학습을 시키는 경우 더 정확하고 좋은 규칙을 만들 수 있다는 것을 실험 결과를 통해 확인할 수 있었다.



[그림 8] c의 변화에 따른 규칙 정확도 변화

5. 결론

본 연구에서는 인공 신경망을 사람이 이해할 수 있는 형태의 규칙으로 만들어 주는 규칙추출 알고리즘 중 하나인 OAS 알고리즘을 사용하였을 때의 결과에 대해 살펴보았다. OAS 알고리즘을 통해 학습된 인공 신경망으로부터 규칙을 추출해보고, 그 결과 중 정확도가 떨어지는 규칙의 특징을 분석하였다. 그 결과 정확도가 떨어지는 규칙에 해당하는 데이터가 학습된 인공 신경망을 통과할 때 각 은닉 층 출력값의 분포가 0.3~0.7사이에 많이 존재한다는 결과를 얻을 수 있었다. 이러한 중간 값이 많이 존재한다는 것은 출력값의 이진화가 잘 되지 않아 OAS 알고리즘을 사용해 규칙을 생성할 때 문제를 일으킬 가능성이 있다고 볼 수 있다.

그리고 이 문제를 해결하기 위한 방법을 찾기 위해 은닉 유닛 명확화 기법을 사용해 인공 신경망을 추가로 학습하였을 때, 은닉 층 출력값의 결과가 어떻게 변화하는지 와 규칙의 품질이 개선이 되는지에 대한 실험을 진행하였다. 그 결과로 은닉 유닛 명확화 기법이 은닉 유닛으로부터 나오는 출력값을 완전 활성화 혹은 완전 비활성화에 가까운 값으로 만들어, 더욱 명확하게 이진화시킬 수 있었고, 그로 인해 불필요한 규칙의 개수는 줄이면서 더욱 정확한 규칙을 만들어 낼 수 있었다. 이러한 결과들을 바탕으로 은닉 유닛 명확화를 사용해 인공 신경망을 추가로 학습하는 경우 OAS 알고리즘으로부터 규칙을 효율적으로 추출할 수 있음을 확인할 수 있었고, 인공 신경망이 좀 더 해석하기 쉬운 형태로 학습된다는 것을

확인할 수 있었다.

본 연구의 향후 계획은 크게 2가지이다. 첫 번째로는 본 연구에서는 하나의 은닉 층을 갖고 실험을 하였는데, 향후에는 더욱 깊은 은닉 층을 갖고 있는 인공 신경망의 규칙을 추출하는 데에도 실험을 해보고 이에 대한 개선 방향을 찾아보고자 한다. 두 번째로는 본 연구에서는 OAS 알고리즘과 은닉 유닛 명확화 알고리즘을 갖고 실험을 진행하였었는데, 더 다양한 알고리즘 조합에 대한 실험도 진행하여 개선 가능성을 연구할 계획이다.

참 고 문 헌

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

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

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

[4] Ishikawa, M. (1996). Structural Learning with Forgetting. *Neural Networks, vol.9, no.3,* 509-521.

[5] L. M. Fu. (1994). Rule Generation from Neural Networks. *Systems, Man and Cybernetics, 24*(8), 1114 - 1124.

[6] L. M. Fu. (1991). Rule Learning by Searching on Adapted Nets. *AAAI,* 590-595.

[7] L. M. Fu. (1993). Knowledge-based connectionism for revising domain theories. *IEEE Transactions on Systems, Man, and Cybernetics 23*(1), 173-182.

[8] S. Thrun. (1995). Extracting rules from artificial neural networks with distributed representations. *Advances in neural information processing systems,* 505-512.

[9] I. A. a. J. G. Taha. (1999). Symbolic interpretation of artificial neural network. *IEEE Transactions on knowledge and data engineering, 11*(3), 448-463.

[10] G. P. C. A. a. F. S. G. Schmitz. (1999). ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks, 10*(6), 1392-1401.

[11] G. G. a. j. W. S. Towell. (1993) Extracting refined rules from knowledge-based neural networks. *Machine learning, 13*(1), 71-101.

[12] R. Setiono, W. K. Leow (2000). FERNN: An algorithm for fast extraction. *Applied Intelligence, 12*(1-2), 15-25.

이 헌 주

2009 경희대학교 컴퓨터공학과(학사)
2009 ~ 현재 삼성전자 SR(Samsung Research) Senior engineer
2017~현재 고려대학교 컴퓨터학과 석사과정
관심분야: 딥러닝
E-Mail: boxerlee@korea.ac.kr

김 현 철

1988 고려대학교 전산학과(이학학사)
1990 Univ. of Missouri - Rolla 졸업(석사)
1998 Univ. of Florida 졸업(박사)
1999~현재 고려대학교 정보대학 컴퓨터학과 및 사범대학 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 기계학습, 규칙추출
E-Mail: harrykim@korea.ac.kr