RDAPS 입력자료 선정을 위한 Mutual Information기법 적용

Mutual Information Technique for Selecting Input Variables of RDAPS

한광희*, 류용준, 김태순***, 허준행****

Kwanghee Han, Yongjun Ryu, Tae-Soon Kim, Jun-Haeng Heo

요 지

인공신경망(artificial neural network) 기법은 인간의 두뇌 신경세포의 활동을 모형화한 것으로 오랜 시간 동안 발전해 왔으며 여러 분야에서 활용되고 있고 수문분야에서도 인공신경망을 이용한 연구가 활발히 진행되어 왔다. RDAPS와 같은 단기수치예보 자료는 강우의 유무 판단과 같은 정성적인 분석에서 비교적 정확도가 높지만 정확한 강우량의 추정과 같은 정량적인 부분에서는 정확도가 매우 낮으므로 인공신경망 기법과 같은 후처리 기법을 통해서 정확도를 높이게 된다. 인공신경망 기법을 수행할 때, 가장 중요한 것은 입력변수선택(input variable selection)으로 입력 변수의 적절한 선택이 결과값에 큰 영향을 주게 된다. 본 연구에서는 mutual information을 입력 변수 선택 기법으로 채택하여, 인공신경망의 입력변수 선정의 정확도를 알아보고자 한다. Mutual information은 주어진 자료의 엔트로피값을 이용하여 변수들 간의 독립과 종속의 관계를 나타내는 기법으로서, MI값은 '0'에서 '1'의 값을 가지며 '0'에 가까울수록 변수들 간의 관계가 독립적이고 '1'에 가까울수록 종속적인 관계를 나타낸다. 인공신경망의 입력변수선정에 대한 mutual information의 정확도를 알아보기 위해, 기존 입력변수선택 기법과 mutual information을 이용했을 경우의 인공신경망의 처리능력, 정확도를 비교 검토하였다.

핵심용어: Mutual Information, 인공신경망 기법, RDAPS

1. 서론

근래 세계 곳곳에서는 이상 기후변화로 인한 천재지변이 빈번하게 일어나고 있으며 이로 인한 여러 가지 환경문제가 대두되고 있다. 우리나라의 경우만 해도 매년 수해와 가뭄으로 인한 피해의 복구와 예방에 많은 예산을 들이고 있다. 점점 심각해지는 이상 기후변화와 강우 예보의 부정확성은 사회간접자본과 관련된 사업 전반에 걸친 위기가 아닐 수 없다. 따라서 정확한 강우 예측 시스템을 개발하는 것은 매우 중요하다.

우리나라 기상청에서는 장·단기 수치예보를 위해 GDAPS(global data assimilation and prediction system)와 RDAPS(regional data assimilation and prediction system) 모델을 각각 이용하여 자료를 생산하고 있다. 하지만 RDAPS 모델의 자료는 강우예측 부분에서 비교적 큰 오차를 보이고 있어 정확도 향상을 위한 후처리 과정이 필요하다. 이러한 후처리 기법으로 신주영 등(2008)은 인공신경망(artificial neural network) 기법을 이용하여 RDAPS 모델 예측 자료의 정확도를 높이는 연구를 하였다.

인경신경망 기법을 수행할 때, 가장 중요한 것은 입력 변수 선택(input variable selection)에 있다. 적절한 입력변수의 선택이 인공신경망의 결과값에 큰 영향을 주게 된다. Mutual information은 최근에 연구 된 입력변수 선택 기법(input variable selection technique) 중 하나로 인공신경망 수행 시, 입력 변수간 독립성의 척도를 나타내어 준다. 이 기법은 두 변수간의 독립성 구조에 관한 가정이 없고 데이터 변환이나 noise에 대한

* 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 석사과정, E-mail : darksea@yonsei.ac.kr

** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 석사과정, E-mail : ryj@yonsei.ac.kr

*** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 BK연구교수, E-mail : chaucer@yonsei.ac.kr

**** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 교수, E-mail : jhheo@yonsei.ac.kr

영향이 적어 다른 기법에 비해 신뢰도가 높다(Hanchuan Peng 등 2005). 또한, mutual information의 결과값은 '0'부터 '1'사이의 값으로 도출되는데 0에 가까울수록 각각의 변수는 서로 독립성을, 1의 값에 가까울수록 종속성을 띈다.

2. 인공신경망(Artificial neural network)

2.1 인공신경망의 구조

인공신경망 구조는 층(layer)으로 뼈대를 이룬다. 입력층과 출력층, 그리고 이 사이에 하나 또는 그 이상의 은닉층으로 구성된다. 모든 층은 서로 연결이 되어 있어 각각의 연결마다 연결강도 즉, 가중치(weight)가 결정되며 연결 가중치는 feedback으로 보정이 가능하다. 인공신경망의 기본적인 구조는 그림 1과 같다.

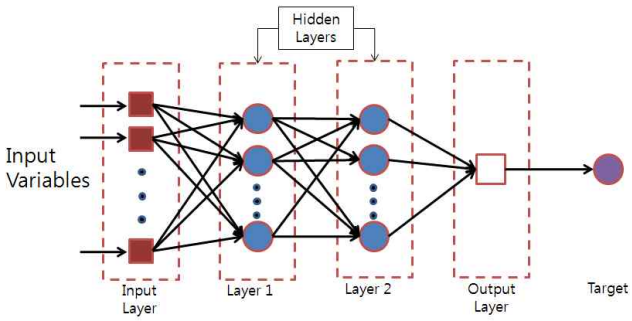

그림 1. 인공신경망 구조 (One input layer and output layer, two hidden layers)

2.2 인공신경망의 학습

학습(learning)은 입력층과 출력층을 거쳐 나온 학습자료에 대응하여 일정한 학습규칙을 통해 연결 가중치가 보정되는 과정이다. 또한 반복적인 학습과정을 통해 네트워크안의 입력자료와 출력자료의 진행을 최적화하도록 가중치를 결정해 나간다. 이 때 학습의 결과로 나온 출력값($y_i(t)$)과 예상결과인 목표값($d_i(t)$)의 차이를 비교하여 평균제곱오차(mean square error : MSE)를 구하게 된다. 학습시간 t에 대한 오차함수는 다음식 (1)과 같다.

$$E(t) = \frac{1}{2}\sum (y_i(t) - d_i(t))^2 \quad (1)$$

학습의 최종적인 목적은 오차함수의 값을 최소로 하는 연결 가중치를 구하는 것이다. 최초에 가중치의 값은 임의로 결정되며 학습과정을 통해 일정한 학습규칙을 가지고 가중치가 변화하게 된다. 가중치 변환 함수는 다음과 같다.

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t+1) + \mu \Delta w_{ij}(t) \quad (2)$$

여기서, μ는 모멘텀이고 일반적으로 gradient descent method를 통해 가중치 증가량 Δw_{ij}를 구하게 된다. 인공신경망은 정확하게는 아니지만 학습을 통해 입력자료와 출력결과의 관계를 대략적으로 결정할 수 있다.

3. 입력 변수 선택 기법

입력변수선택기법(input variable selection techniques)은 인공신경망 등과 같은 모델에 입력 가능한 모든 변수들 중에 가장 관련성이 높은 변수를 찾아내는 방법이다. 적절한 입력 변수를 찾는 것은 모델의 결과에 상당한 영향을 주게 된다. 만약 부적절한 변수를 입력한다면 모델의 결과는 과대 또는 과소추정 될 수 있다.

3.1 Mutual information

Mutual information은 변수간의 관련성을 결정해주는 좋은 척도가 된다. Mutual information을 계산하기 위해서는 각 변수의 엔트로피값이 결정되어야 하는데 엔트로피는 변수의 불확실성을 말한다. 변수는 X, 확률밀도함수(probability density function)를 $p(x)=\Pr\{X=x\}$로 정의한다면 X의 엔트로피는 다음 식 (3)으로 정의된다.

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

그리고 X, Y 두 변수간의 mutual information은 다음 식 (4)과 같다.

$$H(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \tag{4}$$

Mutual information의 값이 크다면 두 변수간은 밀접한 관련을 가지고 있는 것이며, 반대로 값이 작다면 두 변수의 관련성은 적다는 것을 나타낸다. 만약 데이터가 연속성을 가지고 있다면 mutual information은 식 (5)로 구한다.

$$H(X;Y) = \int \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5}$$

확률밀도함수인 $p(x)$, $p(y)$, $p(x,y)$을 찾고 이중적분을 하는 것은 매우 어렵다. 그래서 일반적으로 데이터를 몇 개의 부분으로 나눠 이산화 시켜 엔트로피와 mutual information을 구하게 된다. 하지만 이때 오차가 발생할 수 있다. 따라서 이번 연구에서는 parzen window density를 이용하여 입력 분포를 구하였다. Parzen window density를 이용하면 다른 방법에 비해 오차가 적고 더 나은 결과값을 구할 수 있다(Hanchuan Peng 등 2005).

3.2 입력 자료(RDAPS)

RDAPS는 총 152개의 자료를 가지고 있고 자료의 생산은 하루 2번(00UTC 12UTC)에 걸쳐 이루어지며, 3시간 간격으로 총 48시간 예보자료를 포함하고 있다(한국수자원공사, 2005). 152개의 모든 자료가 인공신경망 가능 입력 변수가 되지만 152개의 자료를 모두 인공신경망의 입력값에 넣어 수행을 할 경우 모델이 복잡해지고 많은 메모리를 소요하게 된다. 따라서 신주영 등(2008)은 상관성 분석을 통해 가장 상관성 계수가 높은 M850(850mb에서의 혼합비), 우리가 보정하고자 하는 지표면 강우량값(ASFC), 그리고 강우이동을 가장 잘 나타내리라 판단되는 700mb의 풍향자료 U700(동서방향의 풍향)과 V700(남북방향의 풍향)을 이용하여 인공신경망의 입력 변수를 구성하였다.

4. 연구 결과 및 분석

4.1 Mutual information 결과 및 분석

본 연구에서는 금강유역 두 지점(전주, 천안)의 지점 관측 강우량(기상청 자료)을 목표자료값으로 하여 실험을 진행하였다. 지점별 mutual information 결과로 나온 입력 변수는 다음의 표 1과 같다.

표 1. 지점 별 입력변수

번 호	관측소명	순 위	입력변수	내 용
1	전주	1	T350	350mb 기압대의 기온
		2	T400	400mb 기압대의 기온
		3	H050	50mb 기압대의 지오포텐셜 고도
		고정	ASFC	지표면 강우량값
2	천안	1	M800	800mb 기압대의 혼합비
		2	M750	750mb 기압대의 혼합비
		3	U800	800mb 기압대의 동서방향 풍속
		고정	ASFC	지표면 강우량값

Mutual information을 계산 한 결과 각각의 지점별로 관련이 높은 입력변수가 모두 다르게 나타난다. 기온, 지오포텐셜 고도, 동서방향 풍속, 혼합비 등이 지점 강우량값과 밀접한 관련이 있는 것을 알 수 있다.

4.2 인공신경망 학습 결과 및 분석

본 연구에서는 상관성 계수와 mutual information의 결과로 나온 각각의 입력변수를 인공신경망에 입력값에 넣고 학습을 시켰다. 학습 결과는 표 2에 직접 비교하였다.

표 2. 상관성 분석과 mutual information 입력변수의 인공신경망 수행 결과 비교

번 호	관측소명	상관성 분석		Mutual information	
		MSE	상관계수(R)	MSE	상관계수(R)
1	전주	0.0060	0.5464	0.0049	0.6765
2	천안	0.0108	0.3019	0.0007	0.6960

여기서 MSE는 최종적인 오차의 크기와 같고 상관계수(R)는 결과값과 목표값의 양을 선형 회기분석 했을 때의 기울기 값이다. 결과적으로 mutual information으로 구한 입력변수를 활용했을 때 상관성분석 보다 모든 지점에서 오차가 적었고 상관계수는 높은 것으로 나타났다.

5. 결 론

본 연구에서는 RDAPS 기상수치모델 자료를 이용해 인공신경망을 수행하였고, 152개의 RDAPS 자료 중 인공신경망 입력값에 적용할 적절한 입력변수 선택을 위해 mutual information기법을 채택하였다. 변수간의 관련성을 결정하는데 있어서 좋은 척도가 되는 mutual information은 '0'부터 '1'사이의 값으로 나타나며 '1'에 가까울수록 변수간의 관련성이 높고 '0'에 가까울수록 관련성이 낮다. 금강유역의 두 지점에서 Mutual information을 구한 결과, 전주 지점에서 관측된 강우량은 RDAPS 자료 중 350mb 기압대의 기온, 400mb 기압대의 기온, 50mb 기압대의 지오포텐셜 고도가 가장 밀접한 관련이 있는 것으로 나타났으며, 천안 지점의 관측 강우량은 800mb 기압대의 혼합비, 750mb 기압대의 혼합비, 800mb 기압대의 동서방향 풍속과 관련성이 높아 인공신경망의 입력변수로 채택되었다. 두 지점에서 상이한 입력변수가 채택된 것은 지점마다 또 다른 지역적인 인자가 존재하기 때문이라고 생각된다. 또한 기존의 방법보다 mutual information으로 구한 입력변수를 활용했을 때, 모든 지점에서 MSE의 값은 작았으며 상관계수는 높은 것으로 나타났다. 결과적으로 mutual information이 상관성 분석보다 변수간의 관련성을 나타내는 지표로서 더욱 뛰어난 것으로 분석되었다.

참고문헌

1. 신주영 (2008). "인공신경망을 이용한 RDAPS 강수량 예측 정확도 향상." 2008년 한국수자원학회 학술발표회 논문집, pp.1013-1017.

2. 한국수자원공사 (2005). 유역 물 관리 운영 기술 개발.

3. Hanchuan Peng, Fuhui Long, and Chris Ding (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238.

4. N. Kwak and C.H. Choi (2002). "Input Feature Selection by Mutual Information Based on Parzen Window." IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1667-1671.