韓國水資源學會論文集
第43卷 第1號 · 2010年 1月
pp. 81~94

DOI: 10.3741/JKWRA.2010.43.1.81

상호정보량 기법을 적용한 인공신경망 입력자료의 선정

Input Variables Selection of Artificial Neural Network Using Mutual Information

한 광 희* / 류 용 준** / 김 태 순*** / 허 준 행****

Han, Kwanghee / Ryu, Yongjun / Kim, Taesoon / Heo, Jun-Haeng

Abstract

Input variable selection is one of the various techniques for improving the performance of artificial neural network. In this study, mutual information is applied for input variable selection technique instead of correlation coefficient that is widely used. Among 152 variables of RDAPS (Regional Data Assimilation and Prediction System) output results, input variables for artificial neural network are chosen by computing mutual information between rainfall records and RDAPS' variables. At first the rainfall forecast variable of RDAPS result, namely APCP, is included as input variable and the other input variables are selected according to the rank of mutual information and correlation coefficient. The input variables using mutual information are usually those variables about wind velocity such as D300, U925, etc. Several statistical error estimates show that the result from mutual information is generally more accurate than those from the previous research and correlation coefficient. In addition, the artificial neural network using input variables computed by mutual information can effectively reduce the relative errors corresponding to the high rainfall events.

Keywords : mutual information, artificial neural network, input variable selection, RDAPS

요 지

본 연구는 인공신경망의 성능을 향상시키기 위한 여러 가지 방법들 중의 하나인 입력변수 선정기법에 관한 연구로서, 일반적으로 널리 사용되고 있는 상관계수를 이용한 입력변수 선정기법 외에 상호정보량을 활용한 방법을 적용하여 인공신경망의 성능을 향상시키고자 하였다. 대상자료는 기상청에서 제공하는 RDAPS자료의 152개 출력값으로 지상강우량의 예측값인 APCP를 포함하고 있으며, 강우관측값간의 상호정보량을 구해 가장 영향력이 큰 변수를 입력변수로 사용하였다. 기존연구결과, 그리고 상관계수만을 이용해서 입력변수를 선정한 결과와 비교해볼 때, 상호정보량을 적용한 경우 입력변수는 주로 바람과 관련된 변수들이 선정되었으며, 평균제곱근오차, 평균제곱근상대오차, 그룹별로 구분한 경우의 절대오차, 그리고 구간별로 구분한 경우의 상대오차를 비교한 경과 상호정보량을 이용한 입력변수 선정방법의 정확도가 전반적으로 높은 것으로 나타났으며, 특히 강우량이 상대적으로 큰 경우의 오차를 많이 감소시킬 수 있는 것으로 나타났다.

핵심용어 : 상호 정보량, 인공신경망, 입력변수선정기법, RDAPS

* 연세대학교 대학원 토목공학과 석사과정
Graduate student, School of Civil and Environmental Engineering, Yonsei Univ., Seoul 120-749, Korea
** 연세대학교 대학원 토목공학과 석사과정
Graduate student, School of Civil and Environmental Engineering, Yonsei Univ., Seoul 120-749, Korea
*** 연세대학교 대학원 토목공학과 사회환경시스템공학부 연구교수
Lecturer, School of Civil and Environmental Engineering, Yonsei Univ., Seoul 120-749, Korea
**** 교신저자 · 연세대학교 사회환경시스템공학부 토목환경공학과 교수
Professor, School of Civil and Environmental Engineering, Yonsei Univ., Seoul 120-749, Korea
(e-mail: jhheo@yonsei.ac.kr)

1. 서 론

인공신경망은 인간의 뇌에서 이루어지는 생물학적인 학습과 지식전달절차를 컴퓨터 공학적으로 응용한 분야로서, 인간의 뇌가 특정 작업을 하는 절차를 모의(model)하도록 설계된 신경망(neural network)을 의미한다(Haykin, 1999). 인공신경망은 수문학과 관련된 여러 분야에서 상당히 널리 사용되고 있으며 특히 비선형성을 갖는 유출량이나 수위를 예측하거나, 주어진 입력값에 대응하는 출력값을 얻는데 많이 사용되고 있다(ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000). 그러나 인공신경망을 적용하는데 있어서는, 인공신경망의 형태를 결정하는 것과 함께 신경망의 학습을 위해서 사용되는 자료가 얼마나 전체 자료의 특성을 잘 반영하고 있느냐가 중요한 요소가 되며, 특히 입력자료와 상이한 자료를 이미 학습이 완료된 신경망에 사용할 때 얼마나 적절한 출력값을 얻느냐에 관련된 다양한 연구결과가 발표되고 있다.

신경망의 학습에 사용할 적절한 입력변수 선정을 위한 다양한 시도 중에서 가장 널리 사용되는 것은 입력변수 선택기법(input variables selection technique)으로 알려진 wrapper 기법과 filter 기법이다. Wrapper 알고리즘에는 처음에 하나의 입력변수를 사용하여 가장 좋은 변수를 골라내고, 다시 나머지 변수들과의 쌍에서 가장 좋은 변수쌍을 골라내는 식으로 입력변수를 하나씩 늘려가는 forward selection과 반대로 처음부터 가능한 모든 입력변수를 이용하여 결과를 구하고 다시 하나씩 변수를 줄여가면서 최적의 입력변수 쌍을 골라내는 backward elimination이 있다.

Forward selection과 backward elimination은 모델이 최적화 될 때까지 계속되므로 만약 d개의 가능 입력변수가 존재한다면 적절한 입력변수 선택을 위해 총 2^d-1번의 수행이 이루어져야하고 모델의 구조에 따라 입력변수의 선택이 달라져야 하는 단점(Kwak and Choi, 2002)이 있는 반면, filter 알고리즘은 가능 입력변수와 목표값 사이의 통계적인 분석을 통하여 입력변수 선정이 이루어지므로 wrapper 알고리즘에 비해 구조가 간단하며 모델의 구성방법 변동에 대한 영향을 전혀 받지 않는다는 장점이 있다. Filter 알고리즘으로 주로 사용되는 통계적 분석방법은 상관성 분석(correlation coefficient)과 주성분 분석(principal component analysis) 등이 있으며 실제로 인공신경망 모델의 입력층을 구성하는데 자주 사용되고 있다(신주영 등, 2008).

그러나 앞의 두 방법은 데이터 변형에 따른 차이가 민감하고, 자료간의 선형관계를 고려하는데 주로 사용하기 때문에 신경망 모형과 같이 비선형 자료의 특성을 모의하는데 사용되기에는 무리가 있으므로, 이런 단점을 극복하기 위하여 최근에는 개개의 변수가 가지는 고유한 상호정보량(mutual information)을 바탕으로 한 인공신경망 입력변수 선택기법이 많이 사용되고 있다(May *et al.*, 2008). 상호 정보량 기법은 두 무작위 변수의 정보량을 계산하여 변수간의 관련성을 측정하는 방법이며(Cover and Tomas, 2006), 두 변수간의 독립성 구조에 관한 가정이 없고 데이터 변형이나 잡음(noise)에 대한 영향이 적어 다른 기법에 비해 신뢰도가 높다고 알려져 있다(Peng *et al.*, 2005).

따라서 본 논문에서는 그동안 널리 사용되어왔던 입력자료 간의 상관계수(correlation coefficient)를 이용한 입력변수 선택기법 대신에 입력변수간의 상호정보량(mutual information)을 사용한 입력변수 선택법을 적용하여 신경망을 구성한 후 기존방법으로부터 도출된 결과와 비교하였다. 연구에 사용된 입력자료는 기상청에서 사용하고 있는 단기 수치예보모델인 RDAPS (regional data assimilation and prediction system)의 출력자료이며 각 출력자료들 간의 상호정보량을 이용하여 입력변수를 선택하여 인공신경망을 구성하였다.

2. 기본이론

2.1 입력변수 선정기법(Input Variable Selection Technique)

2.1.1상호 정보량(mutual information)

Shannon의 정보이론(information theory)은 정보량이라는 개념을 정량적으로 수식화 하는데 사용되는 것으로(Shannon, 1949), 각 변수의 정보량을 엔트로피(entropy)라는 측정기준을 이용하여 정량화 한 것이다. 만약 엔트로피가 크다면 이는 해당 변수에 그만큼 정보량이 많다는 것을 의미하고, 변수의 정보량이 많을수록 불확실성은 줄어들게 된다.

변수 X의 확률밀도함수(probability density function)를 $p(x)=\Pr\{X=x\}$로 정의한다면 변수 X의 엔트로피는 Eq. (1)로 정의된다. 그리고 변수 X, Y 중 한 변수의 정보량을 알고 있을 때 나머지 변수의 정보량은 조건 엔트로피(conditional entropy)인 Eq. (2)를 이용하여 산정할 수 있다.

$$H(X)=-\sum_{x\in X}p(x)\log p(x) \tag{1}$$

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \quad (2)$$

일반적으로 조건 엔트로피는 고유의 엔트로피보다 작거나 같다. 이 두 값이 같을 경우 두 변수는 독립적이라고 말할 수 있다. X, Y 두 변수간의 상호 정보량은 다음 Eq. (3a)와 Eq. (3b)로 정의하고 Fig. 1의 형태로 관계를 나타낼 수 있다(May *et al.*, 2008).

$$I(X;Y) = H(Y) - H(Y|X) \quad (3a)$$

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (3b)$$

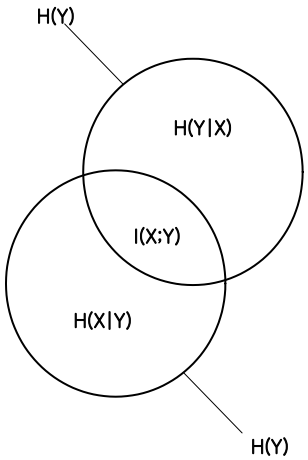

Fig. 1. Relationship Between Mutual Information and Entropy

상호 정보량은 '0'부터 '1'사이의 값으로 도출되는데 '0'에 가까울수록 두 변수는 서로 독립성을, '1'의 값에 가까울수록 종속성을 가지는 것으로 판단된다. 만약 변수가 불연속적(discrete)이라면 Eq. (3a) 그리고 Eq. (3b)로 산출이 가능하지만 연속(continuous)변수일 경우, 엔트로피와 상호 정보량은 Eq. (4a)와 Eq. (4b)를 이용하여 구해야 한다.

$$H(X) = -\int p(x)\log p(x)dx \quad (4a)$$

$$I(X;Y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (4b)$$

Eq. (4a)와 Eq. (4b)에서 확률밀도함수인 $p(x)$, $p(y)$, $p(x,y)$를 찾고 적분을 구하는 것이 어렵기 때문에 이에 대한 해결책으로 확률밀도추정법(density estimation method, Kwak and Choi, 2002)를 이용해 변수가 가지는 연속적인 정보를 몇 개의 부분으로 이산화시켜 엔트로피와 상호 정보량을 구하였으며, 이때 발생하는 오차를 최소화하기 위해 Parzen window density를 이용하여 확률밀도함수 $p(x)$를 근사화하였다.

Parzen window에서 N개의 자료를 가진 변수 x가 주어졌을 때, 근사화 된 확률분포함수 $\hat{p}(x)$는 다음 Eq. (5)의 형태를 따른다.

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x^{(i)}, h) \quad (5)$$

여기서 $\delta(.)$는 Parzen window 함수이고, $x^{(i)}$는 i번째 자료값, h는 window 의 폭(width)을 말한다. 적당한 $\delta(.)$, h의 값이 선택되어지고 자료값(N)이 충분히 클 때, $\hat{p}(x)$는 실제 확률분포함수인 $p(x)$와 가깝게 근사화 되며, 일반적으로 $\delta(.)$값은 Eq. (6)의 Gaussian window를 이용하여 구하게 된다.

$$\delta(z,h) = \exp\left(-\frac{z^T\Sigma^{-1}z}{2h^2}\right) \Big/ \left[(2\pi)^{d/2}h^d|\Sigma|^{1/2}\right] \quad (6)$$

여기서 $z = x - x^{(i)}$이고, d는 변수 x의 차원(dimension)을 나타내므로 이변수의 확률밀도함수를 구하기 위해서는 $d=2$가 되어야 한다. 또한, Σ는 z의 공분산(covariance)이며 대각 행렬(diagonal matrix)로 이루어져 있다.

2.2 인공신경망(Artificial neural network)

2.2.1 인공신경망의 구조

인공신경망 구조는 입력층과 출력층, 그리고 이 사이에 하나 또는 그 이상의 은닉층으로 구성되고 모든 층이 서로 연결이 되어 있어 각각의 연결마다 연결강도가 결정된다. 이 연결강도는 가중치(weight)가 되며, feedback으로 보정이 가능하다(Bishop, 1995). 인공신경망의 기본적인 구조는 Fig. 2와 같다.

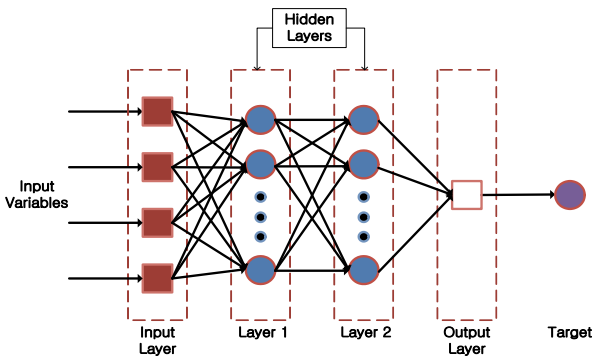

Fig. 2. A Structure of Artificial Neural Network

2.2.2 인공신경망의 학습

학습(learning)은 입력층과 출력층을 거쳐 나온 학습 자료에 대응하여 일정한 학습규칙을 통해 연결 가중치가 보정되는 과정이다. 또한 반복적인 학습과정을 통해 네트워크 안의 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 되며 학습시간 t에 대한 오차함수는 다음 Eq. (7)과 같다.

$$E(t) = \frac{1}{2}\sum(y_i(t) - d_i(t))^2 \tag{7}$$

학습의 최종적인 목적은 오차함수의 값을 최소로 하는 연결 가중치를 구하는 것이다. 최초에 가중치의 값은 임의로 결정되며 학습과정을 통해 일정한 학습규칙을 가지고 가중치가 변화하게 된다. Eq. (8)에 가중치 변환 함수를 나타내었다.

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t+1) + \mu \Delta w_{ij}(t) \tag{8}$$

여기서, μ는 모멘텀이고 일반적으로 경사하강법(gradient descent method)을 통해 가중치 증가량 Δw_{ij}를 구하게 된다.

3. 분석방법

본 연구에서는 기상청에서 생산되는 수치예보자료중 하나인 30 km 격자기반의 RDAPS (Regional Data Assimilation and Prediction System) 격자 중앙지점과 가장 근접한 기상청 유인관측지점 네 곳(장수, 이천, 천안, 전주)의 강우 관측치를 인공신경망 목표값으로 두고 입력변수 선택기법을 달리하여 다음 세 가지의 인공신경망 모델을 구성하였다.

1) 신주영 등(2008)의 연구에서 사용되었던 입력변수로 구성된 인공신경망
2) 상관성 분석을 통해 입력변수가 선택 된 인공신경망
3) 상호정보량을 통해 입력변수가 선택 된 인공신경망

위에서 1)은 입력변수로 모두 4가지를 선택하였는데, 지상강우량 예측값인 ASFC는 지상강우량과의 직관적인 상관성을 고려하기 위해서 선정되었고 850 mb의 혼합비인 M850은 상관계수가 가장 높게 나타난 변수이며 700 mb의 동서 그리고 남북방향 풍향인 U700과 V700은 김광섭(2006)의 연구결과를 인용하여 선정한 것이다. 이에 반하여 본 연구에서 구한 2)와 3)은 각각 상관계수와 상호정보량만을 이용하여 입력변수를 선택한 것이다.

Fig. 3에는 RDAPS 지점과 기상청 지점을 원 표식, 세모 표식으로 각각 나타내었으며 Table 1에는 실험에 적용된 지점 네 곳의 위·경도와 3시간 누적 강우자료의 평균, 그리고 최대값을 나타내었다. 마지막으로 인공신경망 학습 결과인 출력값의 예측 정확도를 출력하기 위해 평균제곱근오차(Root Mean Square Error, RMSE)와 평균제곱근상대오차(Relative RMSE, RRMSE), 그리고 결정계수를 사용하였다.

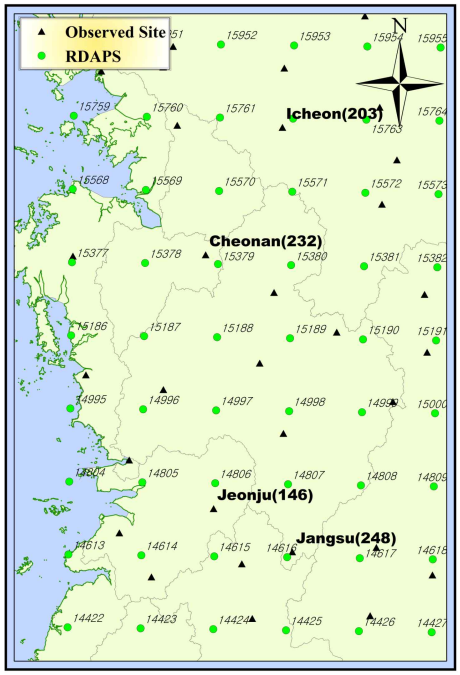

Fig. 3. RDAPS Grid Number and Rainfall Recording Site of Korea Meteorological Administration

Table 1. Information of Rain Gauge Stations

Station code	Station name	Basin	Location		Beginning of observation	Average rainfall (mm/3hr)	Maximum rainfall (mm/3hr)
			Longitude	Latitude			
30011248	Jangsu	Geum gang	127-31	35-39	1988-01-01	5.4	66.5
33011146	Jeonju	Mankyung gang	127-09	35-49	1918-06-15	5.9	103.0
31011232	Cheonan	Sapgyo cheon	127-07	36-46	1973-01-01	5.5	105.5
10071203	Icheon	Han gang	127-29	37-16	1973-01-01	5.6	54.0

Table 2. Description of RDAPS Data

Variable name	Description	Dimension	X	Y	Z	Unit
APCPsfc	3 hours accumulated rainfall	2	191	171	1	mm/3hr
HGTsfc	Geopotential height on surface	2	191	171	1	gpm
HGTprs	Geopotential height	3	191	171	24	gpm
MIXRsfc	Mixing ratio on surface	2	191	171	1	kg/kg
MIXRprs	Mixing ratio	3	191	171	24	kg/kg
PRMSmsl	Sea-level pressure	2	191	171	1	Pa
TMPsfc	Temperature on surface	2	191	171	1	K
TMPprs	Temperature	3	191	171	24	K
UGRDsfc	Wind velocity of East-West direction on surface	2	191	171	1	m/s
UGRDprs	Wind velocity of East-West direction	3	191	171	24	m/s
VGRDsfc	Wind velocity of North-South direction on surface	2	191	171	1	m/s
VGRDprs	Wind velocity of North-South direction	3	191	171	24	m/s
DZDTsfc	Wind velocity of vertical direction on surface	2	191	171	1	m/s
DZDTprs	Wind velocity of vertical direction	3	191	171	24	m/s

3.1 RDAPS 자료

본 연구에서 사용한 RDAPS는 총 152개의 자료를 가지고 있고, 총 66시간 예보자료가 3시간 간격으로 하루 2번(00UTC, 12UTC)에 걸쳐 생성된다. RDAPS 자료의 종류와 간단한 설명은 Table 2에 나타냈다.

3.1.1 인공신경망 입력자료 선정

RDAPS에 포함되어 있는 152개의 모든 자료가 인공신경망 가능 입력 변수가 되지만 152개의 자료를 모두 인공신경망의 입력층에 넣어 수행을 할 경우 모형이 복잡해지고 많은 메모리를 소요하게 된다. 따라서 신주영 등(2008)은 상관성 분석을 통해 가장 상관계수가 높은 M850 (850 mb에서의 혼합비), 보정 될 자료인 지표면 강우량(APCP), 그리고 김광섭(2006)의 연구를 통해 강우이동을 가장 잘 나타내리라 판단되었던 700 mb의 풍향자료 U700(동서방향의 풍속)과 V700(남북방향의 풍속)을 이용하여 인공신경망의 입력 변수를 구성하였다.

본 연구에서 사용한 RDAPS 자료는 2005년도 자료로 상관성 분석과 상호 정보량을 통해 두 종류의 입력변수를 각각 채택하였다. 먼저 기상청 지점의 관측강우를 RDAPS 자료의 생성값과 맞추기 위해 3시간씩 누적시켰다. 또한 RDAPS 자료 중 2005년 7월 5일의 00UTC와 12UTC 자료의 결측에 맞추어 관측강우의 강우자료를 삭제하였고, 비교적 우기인 4월~10월의 자료만 남겨 각각의 변수마다 총 1706개의 자료를 획득하였다. 이중에서 각 지점별로 무강우일을 제외한 자료만을 대상으로 인공신경망을 적용하였으며 4개 지점 모두 210여개의 자료가 사용되었다. 최종적으로 기상청 지점의 강우 관측값과 RDAPS의 출력값인 152개변수간의 상호 정보량과 상관계수 중 상위 15번째까지의 값을 Table 3에 나타내었다.

본 연구에서는 인공신경망을 구성하기 위한 자료의 수가 충분하지 않아서 학습시킨 결과만을 이용하여 모형의 적용성을 평가하였다. 인공신경망의 경우 전체 자료를 모두 3개의 구간으로 나누어서 적용하는 절차가 필요한데, 가중치를 구하기 위한 learning set과 이때 구한 가중치를 검증하기 위한 validation set, 그리고 최종적으로 학습시에 적용하지 않은 unseen data를 이용한 testing set이 모두 충분한 개수만큼 확보되어야만 최적의 인공신경망 모형을 구성할 수 있다. 또한, 전체 자료를 학습자료와 검증자료로 구분할 경우에도 단순히 특정 시간을 중심으로 한 구분방법을 통해서는 전체 자료군을 대표하는 learning / testing set을 구성하는데는 한계가 있다고 할 수 있다. 여기서는, 이런 오류들을 줄이기 위해서 전체 210여개 자료 전부를 이용해서 인공신경망을 학습시켰으며 학습된 인공신경망으로부터 얻어낸 결과를 이용하였다.

전체적으로 상호정보량을 이용했을 경우에는 RDAPS 자료 중, 층은 다르지만 바람과 관련이 있는 변수

(DZDT, UGRD, VGRD)가 강우에 많은 영향을 미치고 있는 것으로 나타났다. 반면 상관성 분석의 경우, 강우와 관련이 높다고 측정 된 변수가 지점마다 매우 상이하게 나타났다.

3.2 인공신경망 구성

인공신경망 구조는 층(layer)으로 이루어져 있다. 본 연구에서의 인공신경망은 전방향 다층퍼셉트론구조로 하였고, 역전파(backpropagation) 알고리즘의 학습방법을 채택하였다. 입력층에는 총 4개의 입력변수가 들어가도록 동일한 수의 노드(node)를 주었고, 은닉층은 3개의 층으로 구성되며 각각 10개의 노드를 주었다. 역전파 알고리즘의 인공신경망은 각각의 은닉층과 출력층의 노드에서 전이함수(transfer function)가 사용된다. 일반적으로 값의 범위가 $0 \le a \le 1$로 출력되는 로그 시그모이드(log sigmoid) 전이함수(Eq. (9a))가 사용되나, 본 연구에서는 출력층의 노드에서만 로그 시그모이드 전이함수를 이용하였고 나머지 노드에서는 $-1 \le a \le 1$의 범위를 가지는 쌍곡선 탄젠트 시그모이드(hyperbolic tangent sigmoid) 함수(Eq. (9b))를 이용하였다.(Fig. 4) 쌍곡선 탄젠트 시그모이드 전이함수를 사용하는 이유는 입력층에서 은닉층으로, 은닉층에서 또 다른 은닉층으로 전이가 될 때 유연성(flexibility)을 높여주기 위함이며, 출력층에서 로그 시그모이드 전이함수를 사용하는 이유는 강우값을 출력값(output)으로 나타낼 때 음(-)의 값을 가지지 않게 하기 위함이다.

$$a = \log sig(n) = 1/(1+e^{-n}) \tag{9a}$$

$$b = \tan sig(n) = 2/(1+e^{-2n}) - 1 \tag{9b}$$

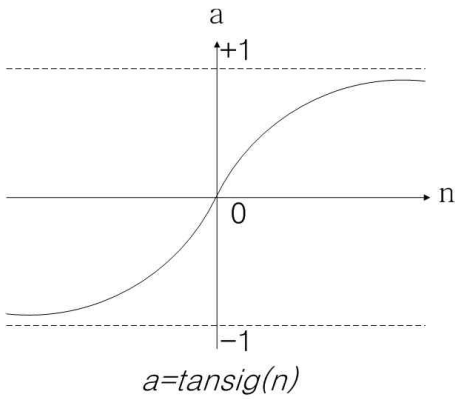

Fig. 4. Hyperbolic Tangent Sigmoid Transfer Function

Table 3. Mutual Information and Correlation Coefficient Between the Observed Rainfall and the Variables of RDAPS Output

Site / Rank	Jangsu				Jeonju			
	MI		Correlation Coefficient		MI		Correlation Coefficient	
	Variance	Value	Variance	Value	Variance	Value	Variance	Value
1	D300	0.6569	D100	0.3096	U925	0.9894	M450	0.2598
2	D100	0.6246	D150	0.1890	D925	0.9615	M300	0.2561
3	D050	0.6111	D250	0.1809	D300	0.9292	V450	0.2442
4	U000	0.6094	D200	0.1769	VGRD	0.9289	MIXR	0.2384
5	V550	0.5952	D050	0.1727	V600	0.9285	V700	0.2308
6	V350	0.5875	U150	0.1643	U650	0.9200	M150	0.2296
7	V400	0.5844	M050	0.1417	U800	0.9166	T450	0.2286
8	U650	0.5731	D350	0.1389	U700	0.9115	T925	0.2285
9	U400	0.5672	D500	0.1375	U450	0.9065	T300	0.2247
10	V450	0.5630	M100	0.1325	U350	0.8995	V925	0.2223
11	D150	0.5625	D300	0.1295	U950	0.8968	V600	0.2199
12	U950	0.5589	U925	0.1263	V850	0.8967	V850	0.2172
13	U450	0.5569	APCP	0.1252	U900	0.8956	T700	0.2063
14	V500	0.5547	M150	0.1249	D070	0.8948	T850	0.2053
15	V900	0.5544	V925	0.1238	D600	0.8945	D925	0.2042

Table 3. Mutual Information and Correlation Coefficient Between the Observed Rainfall and the Variables of RDAPS Output ( )

Site / Rank	Cheonan				Icheon			
	MI		Correlation Coefficient		MI		Correlation Coefficient	
	Variance	Value	Variance	Value	Variance	Value	Variance	Value
1	D150	0.9055	M750	0.1767	V050	0.7002	D300	0.3555
2	V300	0.8790	M100	0.1730	D050	0.6669	M250	0.2681
3	U150	0.8568	M250	0.1626	U300	0.6622	V050	0.2137
4	UGRD	0.8499	VGRD	0.1601	D150	0.6619	M100	0.2127
5	V450	0.8406	D100	0.1488	V200	0.6610	D450	0.2097
6	U050	0.8352	M300	0.1327	D250	0.6596	TMP	0.2042
7	U350	0.8336	T750	0.1324	U150	0.6313	D250	0.1929
8	V875	0.8309	M150	0.1253	U600	0.6266	M300	0.1870
9	D875	0.8298	D150	0.1228	D300	0.6227	M050	0.1811
10	D250	0.8277	V150	0.1208	V875	0.6020	M400	0.1789
11	V975	0.8274	H500	0.1084	V600	0.5929	MIXR	0.1780
12	V800	0.8253	V975	0.1079	U450	0.5879	M450	0.1754
13	DZDT	0.8238	H950	0.1069	U070	0.5875	M150	0.1734
14	U800	0.8159	H650	0.1069	V350	0.584	D050	0.1699
15	V070	0.8143	H350	0.1063	U925	0.5729	T400	0.1681

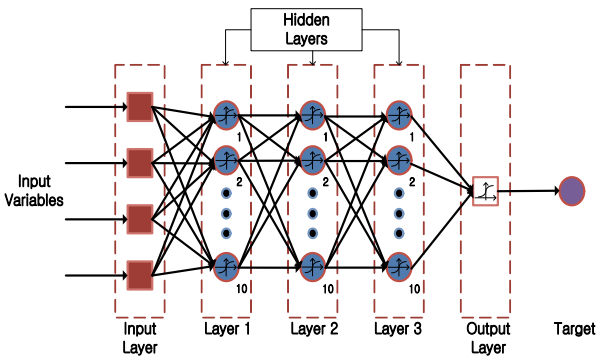

Fig. 5. Modeling the Artificial Neural Network

최종적으로 모델링 된 인공신경망의 모습은 Fig. 5와 같다.

4. 적용 및 결과

본 연구에서는 신주영 등(2008)이 인공신경망 입력자료로 사용한 변수를 이용해 ANN_1의 인공신경망을 모델링 하였다. 더불어 기상청 지점 네 곳(장수, 이천, 천안, 전주)에 대한 상호 정보량, 상관성 계수에 따라 목표값인 관측강우와 관련성이 높다고 판단 된 변수를 각각 채택하여 ANN_2와 ANN_3을 각각 4개씩, 모두 9개의 인공신경망을 모델링 하였다(Table 4).

그리고 인공신경망 수행결과의 정확도를 나타내기 위해 다음의 Eqs. (10)과 (11)로 정의된 평균제곱근오차(RMSE, root mean square error)와 평균제곱근상대오차(RRMSE, relative RMSE)를 적용하였다.

$$\text{평균제곱근오차}(RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - D_i)^2} \quad (10)$$

$$\text{평균제곱근상대오차}(RRMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - D_i}{D_i}\right)^2} \quad (11)$$

여기서, n은 전체 자료수, Y_i는 i번째 출력값, D_i는 i번째 목표값을 나타낸다. ANN_1~3의 수행결과에 따른 정확도 비교를 Table 5에 나타내었다.

Table 4. Types of Artificial Neural Network

Method	Shin et al. (2008)	Correlation coefficient				Mutual information			
Site	All	Jangsu	Jeonju	Cheonan	Icheon	Jangsu	Jeonju	Cheonan	Icheon
Name	ANN_1	ANN_2-1	ANN_2-2	ANN_2-3	ANN_2-4	ANN_3-1	ANN_3-2	ANN_3-3	ANN_3-4
Selected input variables	APCP	APCP	APCP	APCP	APCP	APCP	APCP	APCP	APCP
	M850	D100	M450	M750	D300	D300	U925	D150	V050
	U700	D150	M300	M100	M250	D100	D925	V300	D050
	V700	D250	V450	M250	U350	D050	D300	U150	U300

Table 5. RMSE and RRMSE Results

Site	Type	RMSE (mm/3hr)	RRMSE
Jangsu	ANN_1	8.8306	3.2150
	ANN_2-1	4.6854	3.1428
	ANN_3-1	3.4827	3.3212
Jeonju	ANN_1	6.3948	9.5635
	ANN_2-2	7.1798	10.1424
	ANN_3-2	4.5806	6.1710
Cheonan	ANN_1	8.3718	14.0271
	ANN_2-3	6.7866	11.9825
	ANN_3-3	3.8955	6.1983
Icheon	ANN_1	3.3790	2.4732
	ANN_2-4	4.4630	1.7265
	ANN_3-4	2.4053	1.9313

Table 5에서 보는 바와 같이 상호 정보량을 이용하여 입력변수가 선택된 ANN_3 모형의 평균제곱근오차(RMSE)결과는 장수(3.4827), 전주(4.5806), 천안(3.8955), 이천(2.4053)으로 모든 지점에서 다른 인공신경망에 비해 비교적 오차가 작은 것으로 나타났다. 평균제곱근 상대오차(RRMSE)의 경우, 장수와 이천 지점에서 ANN_2(장수 3.1428, 이천 1.7265)의 오차가 ANN_3(장수 3.3212, 이천 1.9313) 보다 작게 나타났다.

또한, 전주와 천안의 경우 상관계수를 이용해서 입력변수를 선정했을 때 대부분 혼합비와 관련된 변수들이 선정되었으며 이 경우 상호정보량을 이용해서 입력변수를 선정한 결과의 오차를 비교해보면 상대적으로 오차가 상당히 크게 나타났으며, 장수와 이천 지점의 경우 상관계수나 상호정보량을 이용한 경우 모두 바람과 관련된 변수들이 선정되었으며 이에 따른 결과값의 오차의 차이가 상대적으로 전주나 천안보다 적은 것으로 나타났다.

다음의 Fig. 6에 평균제곱근 오차가 제일 큰 천안지점의 관측강우량과 보정된 예측강우량을 각각 X축과 Y축에 도시한 후 기울기를 구하여 나타낸 것이다.

여기서 마름모꼴 모양의 점은 관측강우량에 대응하는 보정된 예측강우량의 값을 그린 것이며 파란색 점선은 이를 직선에 회귀시킨 것이다. 그리고 빨간 실선은 예측강우량의 값이 관측값과 일치하는 경우를 나타내는

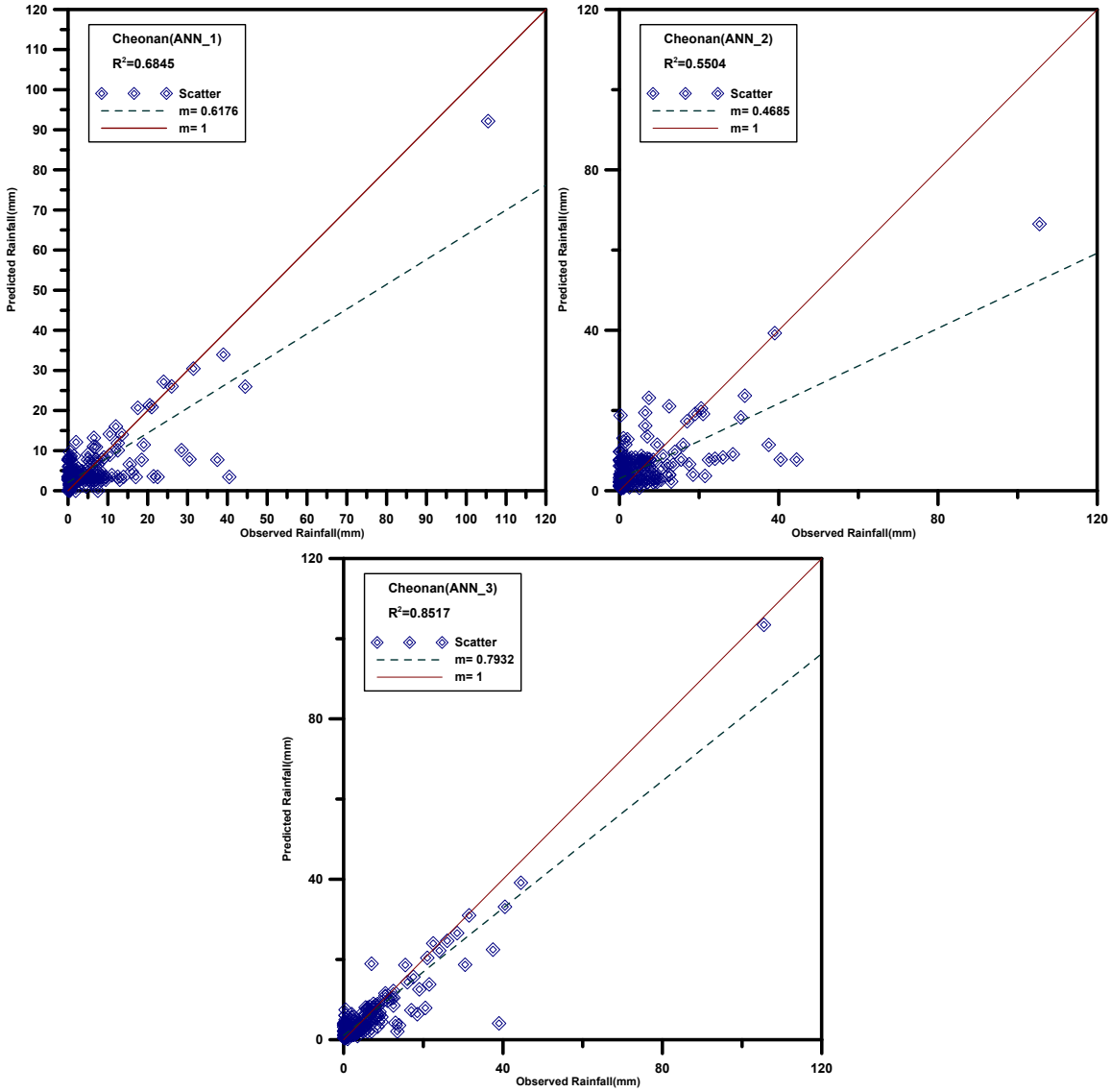

Fig. 6. Scatter plot of observed and predicted rainfall

것으로 파란색 점선이 빨간 실선에 가까울수록 정확도가 높은 것을 나타낸다. 그림에서 볼 수 있는 것과 같이, ANN_3로 예측된 강우량의 정확도가 가장 높은 것으로 나타났으며, 세 가지 경우 모두 보정된 예측강우의 대부분이 하향 추정된 것으로 나타났다.

Fig. 7은 전주, 이천, 천안, 장수 4개 지점별로 관측된 강우량과 인공신경망을 이용해서 구한 강우량을 도시한 것으로, 세로축은 3시간 누적강우량을 나타내고 가로축은 연속된 시간이 아닌 본 연구에서 사용한 무강우를 제외한 강우사상 발생시각을 나타낸다. 연속된 시간이 아니므로 각각을 점으로 도시해야하나 편의상 선으로 연결해서 도시하였다.

각각의 그림은 위에서부터 1) 관측강우량, 2) 상호정보량을 이용해서 입력변수를 선정한 인공신경망(ANN_3으로 시작)으로부터 구한 강우량, 3) RDAPS에서 예측한 지상강우량인 APCP값, 4) 신주영 등(2008)에서 사용한 입력변수를 사용한 인공신경망(ANN_1)으로부터 구한 강우량, 그리고 5) 상관계수를 이용해서 입력변수를 선정한 인공신경망(ANN_2로 시작)으로 구한 강우량을 나타낸다. 4개의 지점에 대해서 도시한 결과를 살펴보면 모두 상호정보량을 이용해서 입력변수를 구성한 경우의 강우량이 실제 관측된 강우량과 가장 근접하게 산정된 것을 볼 수 있으며, 특히 각 지점의 첨두강우량을 살펴보면 다른 인공신경망으로 구한 결과나 RDAPS의 APCP보다 상호정보량을 이용한 경우의 결과값이, 특히 비교적 큰 강우량에 해당하는 값에 대해서 더 정확한 결과를 나타냈다. 따라서, 상호정보량을 이용한 경우 상대적으로 큰 강우량에 대한 정확도를 높일 수 있을 것으로 기대된다.

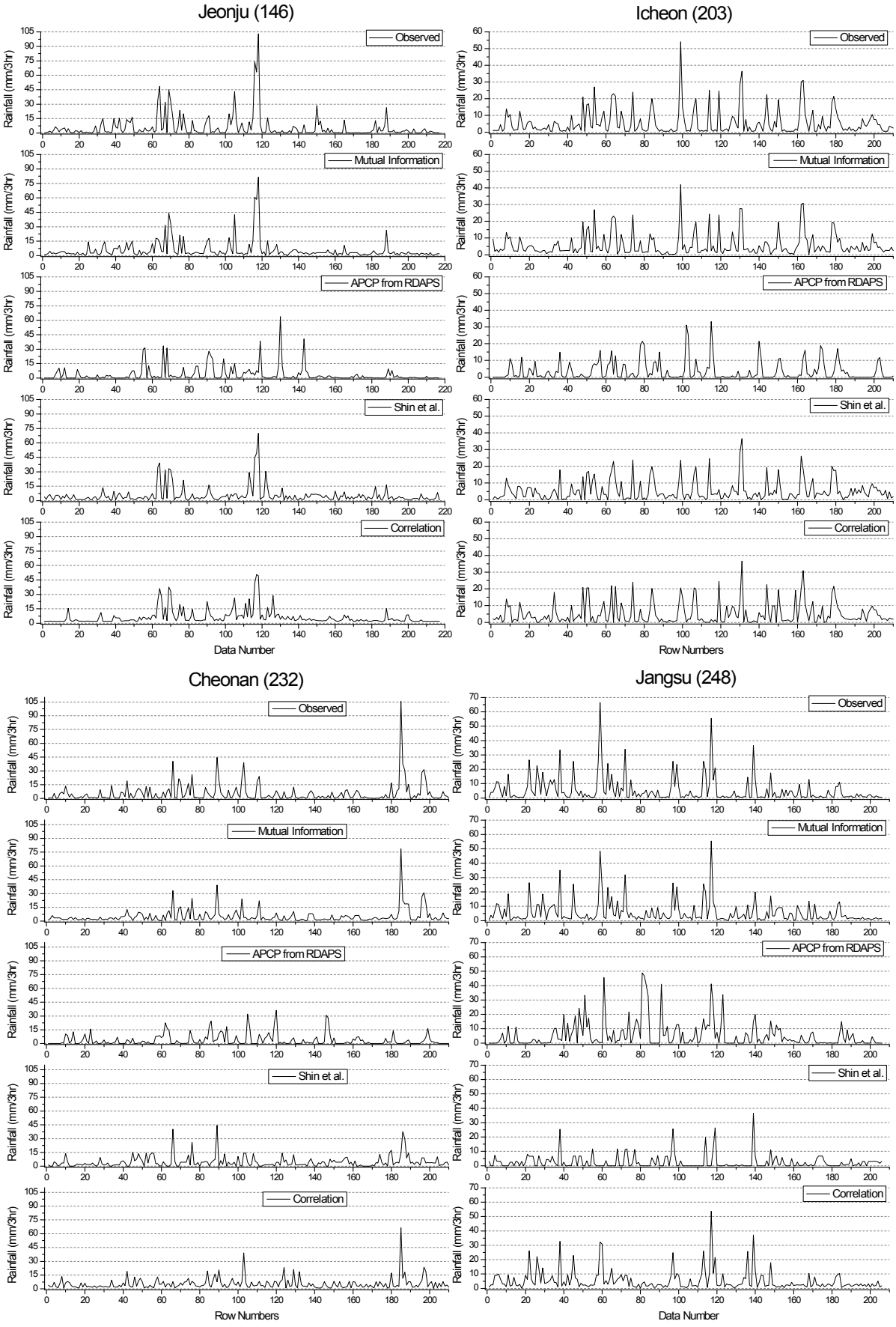

Fig. 7. Comparisons Between Observed and Estimated Rainfall

Table 6. Absolute Errors According to Group and Type (mm/3h)

Site	Group / Type	1	2	3	4
Jangsu	ANN_1	**70.49**	74.46	128.29	582.88
	ANN_2-1	94.62	86.82	90.86	241.24
	ANN_3-1	100.01	**61.99**	**64.22**	**143.66**
Jeonju	ANN_1	138.49	166.05	100.67	396.94
	ANN_2-2	180.70	204.34	126.58	425.43
	ANN_3-2	**107.07**	**95.57**	**94.56**	**244.87**
Cheonan	ANN_1	149.04	139.49	97.92	454.91
	ANN_2-3	186.70	147.29	129.43	471.83
	ANN_3-3	**118.15**	**75.09**	**74.93**	**240.52**
Icheon	ANN_1	77.09	96.46	97.91	176.46
	ANN_2-4	**44.46**	**56.96**	**65.07**	172.36
	ANN_3-4	82.02	61.25	72.76	**105.70**

Table 7. Relative Errors According to Class

Standard		1mm/3hr		3mm/3hr		5mm/3hr		10mm/3hr		20mm/3hr	
Class		−	+	−	+	−	+	−	+	−	+
Jangsu	ANN_1	**188.62**	124.75	**255.24**	58.13	**270.30**	43.06	**291.24**	22.13	**302.70**	10.67
	ANN_2-1	224.56	119.47	314.48	29.55	322.98	21.04	335.12	8.91	340.05	3.98
	ANN_3-1	223.64	**81.90**	284.00	**21.54**	290.86	**14.68**	300.15	**5.39**	302.82	**2.72**
Jeonju	ANN_1	821.57	126.05	908.89	38.73	922.71	24.91	932.65	14.97	942.50	5.13
	ANN_2-2	991.10	145.91	1090.32	46.69	1112.23	24.79	1121.20	15.81	1132.01	5.01
	ANN_3-2	**573.43**	**83.86**	**629.76**	**27.54**	**638.69**	**18.60**	**648.03**	**9.27**	**654.85**	**2.44**
Cheonan	ANN_1	695.38	151.32	805.95	40.75	814.30	32.40	831.49	15.20	838.97	7.73
	ANN_2-3	855.54	175.43	984.32	46.65	994.00	36.97	1016.20	14.78	1024.01	6.96
	ANN_3-3	**525.07**	**77.45**	**575.62**	**26.89**	**581.79**	**20.73**	**594.09**	**8.43**	**599.24**	**3.28**
Icheon	ANN_1	150.52	111.35	226.43	35.44	241.73	20.14	255.10	6.77	258.73	3.14
	ANN_2-4	**80.77**	**63.68**	**121.47**	**22.98**	**131.52**	**12.93**	**138.33**	6.12	**139.89**	4.56
	ANN_3-4	140.21	86.57	201.68	25.10	213.71	13.06	223.04	**3.74**	224.97	**1.81**

일반적으로 3시간 누적강우가 10 mm~30 mm일 경우에 인공신경망으로 보정 된 예측 강우가 목표값과 잘 들어맞는 것을 확인할 수 있지만 0mm에 가까운 강우나, 50 mm~100 mm에 이르는 극치값은 정확도가 상대적으로 떨어지는 것으로 나타났다. 이를 수치적으로 알아보기 위해 각 지점의 관측 강우자료를 오름차순으로 정렬한 후 25 % 구간별 구분하여 4개의 구간으로 나눈 값에 대해서 Eq. (12)를 통해 각 그룹의 절대오차를 구하였다(Table 7).

$$\text{절대오차}(Absolute\ Error) = \sum_{i=1}^{n} |D_i - Y_i| \qquad (12)$$

다음의 Table 6은 각 그룹의 절대오차를 기법별로

나타낸 표이며 여기서 진하게 밑줄 친 값들이 각 지점과 그룹별로 가장 적은 절대오차를 나타낸다. 표를 살펴보면 장수지점의 최하위 25 % 그룹과 이천지점의 1~3 그룹의 경우를 제외하고는 모든 경우에 대해서 상호정보량을 이용한 인공신경망(ANN_3)의 절대오차가 가장 적은 것으로 나타났다. 특히 최상위 25 %에 해당하는 그룹 4의 경우 다른 경우보다 절대오차. 갭소가 두드러지는 것으로 나타나서 상호정보량을 이용한 인공신경망을 이용하면 강우량이 큰 경우의 오차를 상당부분 감소시킬 수 있는 것으로 보인다.

강우량의 크기에 따른 오차를 좀 더 자세히 살펴보기 위해서 Eq. (13)로 정의되는 상대오차를 이용하여 각각 1, 3, 5, 10, 20 mm/3hr에 해당하는 강우량을 기준으로, 기준값보다 적은 강우량(표에서 -로 표시된 행)과 큰 강우량(표에서 +로 표시된 행)만을 대상으로 상대오차를 구한 값이 Table 7이다.

$$\text{상대오차}(\text{a}^{tive}\ error) = \sum_{i=1}^{n}\left|\frac{D_i - Y_i}{D_i}\right| \tag{13}$$

각 지점별로 기준강우량보다 큰 강우량의 경우(+), 이천지점을 제외하고는 모든 지점과 모든 기준강우량에 대해서 상호정보량을 이용한 인공신경망(ANN_3)의 결과가 가장 적은 상대오차를 보여주었지만, 기준강우량보다 적은 경우(-)를 살펴보면 전주와 천안 지점에 대해서는 상호정보량을 이용한 결과가 가장 적은 오차를 보여준 반면, 장수와 이천 지점에서는 ANN_1이나 ANN_2에 의한 결과의 오차가 더 적은 것으로 나타났다.

결과적으로 상호정보량을 이용한 인공신경망을 이용하면, 강우량이 상대적으로 큰 경우의 오차를 많이 감소시킬 수 있다고 말할 수 있다.

다음의 Table 8은 ANN_1, ANN_2, 그리고 ANN_3으로 구한 강우량과 실제 관측된 값과의 정확도 및 오차를 나타낸 것으로 정확도(accuracy)는 전체 자료를 대상으로 하였으며, 오차(error rate)는 비교적 큰 강우에 대한 오차를 살펴보기 위해서 5mm/3hr 보다 큰 강우에 대해서 계산한 값이다. 여기서 정확도와 오차는 다음의 Eqs. (14), (15)를 이용한 것이고, 여기서 R은 실제로 관측된 강우량이고, A는 각각의 인공신경망을 이용해서 추정한 강우량이다.

$$Accuracy = \left(1 - \sum\frac{R-A}{R}\right)\times 100 \tag{14}$$

$$Error\ rate = \left(\sum\frac{R-A}{R}\right)\times 100 \tag{15}$$

상호정보량을 이용해서 입력자료를 구성한 ANN_3 모형은, 정확도의 경우 ANN_1과 비교해서 10.48 %~38.55 %의 정확도 향상을 보였으며, ANN_2와 비교해서는 1.42 %~35.40 %의 정확도 향상결과를 보였다. 또한 오차의 경우 ANN_1과 비교해서는 8.62 %~50.57 %의 오차를 감소시킬 수 있었으며, ANN_2와 비교해서는 7.06 %~23.61 %까지 감소시킬 수 있었다.

Table 8. Accuracy and Error Rate of Each ANN Model

Site	Group Type	Accuracy(%)	Error rate(%)
Jangsu	ANN_1	26.89	69.11
	ANN_2-1	55.58	29.60
	ANN_3-1	65.44	18.54
Jeonju	ANN_1	35.38	38.78
	ANN_2-2	24.51	41.46
	ANN_3-2	56.33	23.95
Cheonan	ANN_1	30.16	48.05
	ANN_2-3	22.37	49.73
	ANN_3-3	57.77	26.12
Icheon	ANN_1	62.80	22.48
	ANN_2-4	71.86	20.92
	ANN_3-4	73.28	13.86

5. 결 론

본 연구에서는 인공신경망을 구성하기 위한 입력변수 선택기법 중 하나인 상호정보량 기법을 적용한 인공신경망을 이용하여 RDAPS 예측강우를 보정하였다. 이산화 되지 않은 연속형인 자료의 경우 상호 정보량의 값을 도출하기가 매우 어렵기 때문에 Parzen window method를 사용하여 상호정보량 값을 근사화시켜 계산하였으며, 다음의 결론을 얻을 수 있었다.

1) RDAPS 예측지점과 가장 근접한 기상청 강우 관측지점 네 곳(장수, 전주, 천안, 이천)을 선택하여 상호정보량과 상관성분석을 실시한 결과, 전자는 바람과 관련이 있는 변수(DZDT, UGRD, VGRD)가 강우에 많은 영향을 미치고 있는 것으로 일관성 있게 나타났지만, 후자는 강우와 관련이 높다고 측정된 변수가 지점마다 매우 상이하게 나타났다.

2) 지점별, 기법별로 총 아홉 개의 인공신경망을 모델링하여 실험을 수행한 결과, 상호 정보량을 이용하여 입력변수를 구성한 인공신경망의 평균제곱근오차가 4개 지점 모두에서 제일 적게 나타났고, 평균제곱근 상대오차의 경우 전주와 천안에서는 상호정보량을 이용한 인공신경망의 오차가 제일 적게 나타났으나 장수와 이천지점에서는 상관계수를 이용한 경우가 더 적게 나타났다.

3) 전주, 천안지점과 같이 상관계수를 이용한 입력변수 선택 시 혼합비와 관련된 변수들이 선택된 경우 상호정보량을 이용해서 오차를 크게 줄일 수 있는 것으로 나타났으며, 장수, 이천지점과 같이 상관계수나 상호정보량 모두 바람과 관련된 변수를 선택한 경우 결과값의 오차에 큰 차이가 없는 것으로 나타났다.

4) 관측강우의 크기에 따라서 4개의 구간으로 나누어 오차를 검토한 결과 최상위 25%에 해당하는 그룹의 경우 상호정보량을 이용한 결과의 절대오차가 크게 감소하는 것으로 나타났으며, 1, 3, 5, 10, 20mm/3hr를 기준으로 자료를 상, 하위 값으로 나누었을 경우, 모든 상위값에 대해서 상호정보량을 이용한 경우의 상대오차가 가장 적은 것으로 나타났다.

결론적으로, 상호정보량을 이용하여 입력변수를 선택하면 일반적으로 바람과 관련된 입력변수가 주로 선택되며, 또한 비교적 큰 강우량의 추정에서 발생하는 절대오차 및 상대오차를 크게 줄일 수 있는 것으로 나타났다.

감사의 글

본 연구는 국토해양부가 출연하고 한국건설교통기술평가원에서 위탁 시행한 건설기술혁신사업(08기술혁신F01)에 의한 차세대홍수방어기술개발연구단의 연구비 지원에 의해 수행되었습니다.

참 고 문 헌

강부식, 이봉기 (2008). “가강수량과 인공신경망을 이용한 중규모수치예보의 강수확률예측 개선기법.” **한국수자원학회학술발표회논문집**, 한국수자원학회, pp. 1027-1031.

김광섭 (2006). “상층기상자료와 신경망기법을 이용한 면적강우 예측.” **한국수자원학회논문집**, 한국수자원학회, Vol. 39, No. 8, pp. 717-726.

신주영, 최지안, 정창삼, 허준행 (2008). “인공신경망을 이용한 RDAPS 강수량 예측 정확도 향상.” **한국수자원학회학술발표회논문집**, 한국수자원학회, pp. 1013-1017.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000). “Artificial neural networks in hydrology. II: hydrologic applications.” *Journal of Hydrologic Engineering*, Vol. 5, No. 2, pp. 124-137.

Bishop, C. M. (1995), *Neural networks for pattern recognition*, Oxford. pp. 140-148.

Cover, T. M. and Thomas, J. A. (2006), *Elements of information theory*, Wiley-Interscience.

Hall, T., Brooks, H. E. and Doswell, C. A. III (1999), “Precipitation forecasting using a neural network.” *Weather and forecasting*, Vol. 14, No. 12, pp. 338-345.

Haykin, S. (1999), *Neural Networks a comprehensive foundation*, Prentice Hall, Inc.

Kwak, N. J. and Choi, C. H. (2002), “Input Feature Selection by Mutual Information Based on Parzen Window.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 12, pp. 1667-1671.

Kwak, N. J. and Choi, C. H. (2002), “Input feature selection for classification problems.” *IEEE Transactions on Neural Networks*, Vol. 13, No. 1, pp. 143-159.

May, R. J., Maier, H. R., Dandy, G. C. and Fernando,

T. M. K. G. (2008), “Non-linear variable selection for artificial neural networks using partial mutual information.” *Environmental Modelling & Software*, Vol. 23, No. 10-11, pp. 1312-1326.

Peng, H., Long, F. and Ding, C. (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.

Shannon, C. E. (1949), *A Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

논문번호: 09-099	접수: 2009.08.31
수정일자: 2009.11.04/11.27	심사완료: 2009.11.27