ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 64, No. 1, pp. 99~106, 2015
http://dx.doi.org/10.5370/KIEE.2015.64.1.099

방대한 기상 레이더 데이터의 원할한 처리를 위한 순환 가중최소자승법 기반 RBF 뉴럴 네트워크 설계 및 응용

Design of RBF Neural Networks Based on Recursive Weighted Least Square Estimation for Processing Massive Meteorological Radar Data and Its Application

강 전 성* · 오 성 권†
(Jeon-Seong Kang · Sung-Kwun Oh)

Abstract - In this study, we propose Radial basis function Neural Network(RBFNN) using Recursive Weighted Least Square Estimation(RWLSE) to effectively deal with big data class meteorological radar data. In the condition part of the RBFNN, Fuzzy C-Means(FCM) clustering is used to obtain fitness values taking into account characteristics of input data, and connection weights are defined as linear polynomial function in the conclusion part. The coefficients of the polynomial function are estimated by using RWLSE in order to cope with big data. As recursive learning technique, RWLSE which is based on WLSE is carried out to efficiently process big data. This study is experimented with both widely used some Machine Learning (ML) dataset and big data obtained from meteorological radar to evaluate the performance of the proposed classifier. The meteorological radar data as big data consists of precipitation echo and non-precipitation echo, and the proposed classifier is used to efficiently classify these echoes.

Key Words : RBFNN, Pattern Classifier, Recursive Weighted Least Square Estimation(RWLSE), Fuzzy C-mean(FCM), Big Radar Data

1. 서 론

레이더를 이용한 강수 관측 시 레이더의 위치, 대기의 상태, 주위 지형 등에 따라 강수 이외의 에코가 관측되는데 이러한 에코를 비강수에코(Non-precipitation echos)라 한다. 비강수에코의 제거는 레이더를 이용한 강수에코만을 추정하기 위해 반드시 선행 되어야 하는 과정이다. 비강수 지형 에코란 레이더 빔의 전파 효과, 과대굴절, 대기의 이상에 의해 다양하게 또는 넓은 지역에 걸쳐 나타나게 된다. 비강수 Clear 에코는 강수는 발생하지 않지만 대기 중의 곤충이나 작은 부유입자, 난류 등에 의해 레이더 파가 반사되어 발생하는 에코이다. 이러한 비강수에코의 특성으로 레이더 데이터를 이용하여 기상 예측 및 관측이 어렵기 때문에 강수에코 및 비강수에코의 특성을 파악해야 하며, 레이더 자료에서 비강수에코를 제거해야 할 필요가 있다. 또한 기상 레이더 데이터의 용량은 빅데이터 수준으로 방대하기 때문에 학습 데이터로써 몇 년 동안 모아온 데이터를 사용하기 어렵다. 위와 같이 기상레이더에서 취득한 빅데이터를 처리하기 위해서 순환동정을 이용한 RBFNN을 제안한다. 본 논문에서는 기상 레이더 빅데이터인 UF(Universal Format)데이터를 사용하였으며, 다항식 방사형 기저 함수 신경회로망(Radial Basis Function Neural Networks)을 이용하여 패턴 분류기를 설계하였다. RBFNN의 조건부에서는 Fuzzy C-Means(FCM)클러스터링을 사용하여 입력 데이터의 특성을 고려한 적합도를 구하였으며, 결론부에서는 다항식의 형태를 결정하고 기상 레이더에서 취득한 빅데이터를 처리하기 위해 순환동정 알고리즘을 적용한 최소자승법을 사용하여 연결가중치를 동조 하였다. 이 알고리즘은 데이터가 업데이트됨에 따라 계수를 추정할 수 있으며, 낮은 용량의 메모리로도 효율적으로 데이터를 처리하는 것을 보인다. 추론부에서는 조건부에서 구한 적합도와 결론부에서 구한 다항식을 결합하여 최종출력 및 패턴분류율을 구하였다. 2장에서는 기상레이더 데이터의 특성, 3장에서는 패턴 분류를 위한 지능형 알고리즘인 다항식 기반 방사형 기저함수 신경회로망에 대한 내용[1], 4장에서는 대용량 빅데이터의 후반부 파라미터 동정을 위한 Recursive 가중 최소자승법에 대해 설명하며, 5장에서는 사용된 패턴판단모듈을 설명한다. 6장에서는 Machine Learning Data를 사용하여 RBFNN모델에 사용되는 WLSE와 LSE의 패턴분류율을 비교한 후, 각 규칙에 대해 비교한 패턴분류율을 그래프를 보여준다. 또한, RWLSE 알고리즘을 접목시킨 패턴분류기로 기상레이더 데이터를 분류한 패턴분류율을 나타낸 후[2], 이미지를 통해 보여준다. 마지막으로 7장에서 실험에 대한 결론에 대하여 설명한다.

† Corresponding Author : Dept. of Electrical Engineering, The University of Suwon, Korea
E-mail : ohsk@suwon.ac.kr
* Dept. of Electrical Engineering, The University of Suwon, Korea
Received : December 05, 2014; Accepted : December 21, 2014

2. 기상레이더 데이터의 특성

기상레이더 자료인 UF데이터를 분석하여 강수에코와 비강수에코의 특성을 파악할 수 있었다. 강수 학습데이터로는 각 강수사례들(대류형, 층운형, 기타 등등)의 하나씩을 샘플링 하여 학습데이터를 구성하였으며, 비강수 학습데이터로는 맑은 날 중에서 지형에코, Clear 에코를 샘플링 하여 구성하였다. 샘플링 기준은 각 사례들 중 중복되지 않는 데이터로 구성하였다. 학습데이터의 수는 Null값이 포함되지 않은 강수에코 858,092개, 비강수에코 929,729개로 총 1,787,821개 이며, 이를 입력 데이터로 사용하였다. 또한 테스트데이터로는 학습데이터에 포함되지 않은 다른 날짜, 다른 시간대의 강수에코 사례, 지형에코 사례 및 Clear 에코 사례를 테스트 해보았으며 테스트데이터의 수는 344,520개 이다. Radar는 전자기파를 이용하여 어떤 물체를 감지하고 그 물체가 관측자로부터 어떤 상대적인 위치에 있는가를 분석해내는 일종의 원격탐사 장비이다. 이러한 기상레이더에 의한 자료는 다른 어떤 관측수단으로 제공하기 힘든 실제적인 자료를 제공하는 원천이 되고 있다. 본 연구에서 사용된 UF(Universal Format) 기상레이더 빅데이터는 Radar site에서 관측되는 자료들을 바이너리(binary)형태 자료로 저장된 데이터를 말한다. 저장되는 값들은 필터링 후 반사도(CZ), 필터링 후 시선속도(VR), 필터링 후 스펙트럼 폭(SW), 필터링 전 반사도(DZ)가 저장된다.

저장 되는 반사도는 dBZ(decibel Z)라는 단위를 사용하는데, 이는 반사도 인자의 대수(logarithm)로써 다음 식과 같다.

$$dBZ = 10 \cdot \log_{10}\left(\frac{Z}{1mm^6/m^3}\right) \tag{1}$$

즉, $1mm^6/m^3$에 대한 비 값을 대수로 나타낸 것을 의미한다. 예를 들면, 단위부피 $1m^3$안에 직경 $1mm$인 물방울이 한 개 있으면 0[dBZ], 10개 있으면 10[dBZ], 100개 있으면 20[dBZ], 1000개 있으면 30[dBZ]가 되는 것이다. 또한 Z는 레이더 방정식을 이용하여 다음과 같이 식 (2)로 계산될 수 있다.

$$Z = \frac{2^{10}(\ln 2)}{\pi^3 c}\left[\frac{\lambda^2}{P_t \gamma G^2 \theta_{3dB}^2}\right]\left[\frac{\gamma^2 \overline{P_r}}{|K|^2}\right] \tag{2}$$

여기서, $\overline{P_r}$는 평균 반사전력 (Watt), P_t는 최대 송신 출력 (Watt), G는 안테나 이득 (무 차원), λ는 레이더 파장 (m), θ_{3dB}는 안테나 빔 출력 반치 폭 (radian), γ는 펄스 지속시간 (펄스 폭)(sec), c는 전자기파의 전파속도, 상수(빛의 속도) = $3\times10^8 msec^{-1}$, γ는 레이더와 목표물간의 거리 (m), K는 복소 굴절율, 일반적으로 $|K|^2$값은 물일 경우 0.93, 얼음일 경우 0.2로 취해진다. Z는 레이더 반사도 인자를 의미한다.

기상청에서 보유하고 있는 기상레이더의 산출물은 CZ, VR, SW, DZ가 있지만 비강수에코와 강수 에코의 정보가 남아있는 CZ와 DZ만을 사용하였으며 CZ는 dBZ값이 필터링을 거친 후의 자료이기 때문에 값이 작아지는 현상을 보여 참조용으로 사용하였다. 본 연구과제에서는 반사도의 표준편차(SDZ), 반사도의 연직기울기(VGZ), 변곡점 개수의 백분율(SPN), 반사도의 빈도수(FR)를 입력변수로 사용하였다.

입력데이터 구성 과정은 다음과 같다.

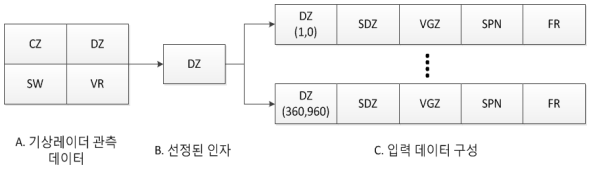


그림 1 입력데이터 구성

Fig. 1 Construction of input data

레이더 데이터 DZ, CZ, VR, SW 중 필터링 전 반사도 값으로 비강수에코에 대한 정보가 많은 DZ를 선택한 후, DZ의 각 소속변수로는 표준편차(SDZ), 연직기울기(VGZ), 변곡점 개수의 백분율(SPN), 빈도수(FR)을 사용한다.

이러한 기상레이더 데이터의 특성은 빅데이터의 3가지 특성인 규모(Volume), 다양성(Variety), 속도(Velocity)와 부합된다.

3. 패턴 분류기의 구조 및 학습방법

3.1 다항식 방사형 기저함수 신경회로망(RBFNN)

본론에서는 패턴 분류를 위한 제안된 다항식 기반 RBFNN 패턴 분류기[3]에 대하여 설명한다. RBFNN 구조는 FCM 클러스터링에 기반 한 분할함수를 활성함수로 사용하며, 다항식 함수로 구성된 연결가중치를 사용함으로 서 식 (3)의 퍼지 규칙 표현과 같이 언어적인 관점에서 해석 될 수 있다.

$$If\ x\ is\ A_i\ Then\ f_{ji}(x) \tag{3}$$

x는 입력 벡터, A_i는 FCM 클러스터링에 의한 $i(i=1,\cdots,c)$번째 그룹의 소속 함수, $f_{ji}(x)$는 $j(j=1,\cdots,s)$번째 출력에 대한 i

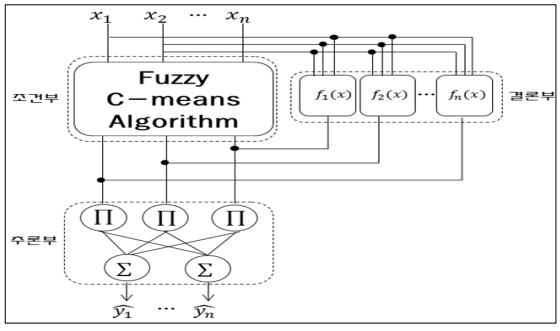


그림 2 조건부, 결론부, 추론부의 세 가지 모듈로 표현된 RBFNN의 구조

Fig. 2 Structure of RBFNN expressed as conditional phase, conclusion phase and Inference phase

번째 퍼지 규칙의 다항식이다. FCM 클러스터링을 이용함으로써 네트워크 측면에서는 활성함수를 언어적 측면에서는 소속 함수의 기능을 수행한다. "*Then*"이후 결론부의 다항식은 네트워크 연결 가중치로, 퍼지 규칙의 로컬 모델로 동작된다. 추론부에서 네트워크의 최종출력이 퍼지 규칙의 추론 결과로서 구해진다. 이와 같이 제안된 RBFNN 구조는 퍼지 규칙에 기반 한 네트워크 구조를 가지며, 조건부, 결론부, 추론부와 같이 세 가지 기능적 모듈로 분리되어 동작한다. 그림 4은 기능적 모듈로서의 RBFNN 구조[4]를 보여준다.

3.2 RBFNN의 조건부

RBFNN의 조건부는 FCM 클러스터링 알고리즘을 사용한다. 이는 학습 데이터의 특성 반영을 위해 입력 공간을 c개의 클러스터 수(퍼지 규칙 수)만큼의 로컬 영역으로 분리하고 각 로컬 영역의 소속정도를 퍼지 집합으로서 출력한다. FCM(Fuzzy C-Means)클러스터링 알고리즘은 비슷한 패턴, 속성, 형태 등의 기준을 통해 데이터를 분류하는 알고리즘으로, 데이터와 각 클러스터와의 거리를 기준으로 소속정도를 측정하여 데이터를 분류한다. 이를 이용하여 다항식 기반 RBFNN의 조건부 활성함수 형태를 표현하였으며, 아래 단계를 통해 수행된다.

Step 1) 클러스터의 개수, 퍼지화 계수를 선택하고 소속함수($U^{(0)}$)를 초기화 한다.

$$U^{(0)} = \left\{ u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1 \forall k, 0 < \sum_{k=1}^{n} u_{ik} < n \forall i \right\} \tag{4}$$

Step 2) 각 클러스터에 대한 중심 벡터를 구한다.

$$u_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m X_{kj}}{\sum_{k=1}^{n} (u_{ik})^m} \tag{5}$$

Step 3) 중심과 데이터와의 거리를 계산하며, 이를 통해 새로운 소속함수(U(1))를 계산한다.

$$d_{ik} = d(x_k - v_i^{(r)}) = \left[\sum_{j=1}^{l} \left(x_{kj} - v_{ij}^{(r)} \right)^2 \right]^{1/2} \tag{6}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)}} \tag{7}$$

Step 4) 오차가 허용범위 안에 도달하면 알고리즘을 종료하고, 그렇지 않으면 Step 2로 돌아간다.

$$\| U^{(r+1)} - U^{(r)} \| \leq \epsilon \tag{8}$$

3.3 RBFNN의 결론부

RBFNN 구조의 결론부는 조건부에서 분리한 각 로컬 영역을 다항식 함수의 로컬 회귀모델로서 표현하여 식 (3)의 "*Then*" 이후의 규칙을 형성한다. $f_i(x)$는 상수항, 선형식, 2차식, 변형된 2차식의 네가지 타입의 함수가 있으며, 그중 하나인 상수항은 다음과 같은 형태를 갖는다[8].

[Type 1] 상수항(Constant)

$$f_j(x) = a_{i0} \tag{9}$$

여기서 $x = [x_1, x_2, ..., x_k]$, $f_j = (x_1, ..., x_k)$는 j번째 규칙에 대한 후반부로써 j번째 퍼지 규칙에 대한 로컬 모델이다.

제안된 RBFNN의 구조는 위에서 언급한 바와 같이 조건부를 FCM 클러스터링을 통한 퍼지 공간 분할, 결론부를 다항식으로 로컬 영역을 표현하는 로컬 회귀모델로 이해할 수 있다.

3.4 RBFNN의 추론부

추론부에서는 "*If−then*" 퍼지 규칙 기반의 퍼지 추론에 의해 네트워크의 최종출력을 구하게 된다. 그림 2을 보면 각 퍼지 규칙의 소속 함수와 다항식 로컬모델이 곱하여진 후 그 합이 출력층 뉴런의 최종 출력으로 되는 것을 볼 수 있다. 이와 같은 일련의 과정은 퍼지 추론 과정과 같다. 결론적으로 RBFNN에서 $j(j=1,\cdots,s)$번째 출력의 최종 출력은 퍼지 추론에 의한 식 (10)와 같이 표현된다.

$$\hat{y} = \sum_{j=1}^{c} u_j f_j(x_1, ..., x_k) \tag{10}$$

이와 같이 다항식 형태의 연결가중치를 사용함으로서 (3)식의 퍼지 규칙 표현과 같은 언어적 관점에서의 해석이 가능해졌다.

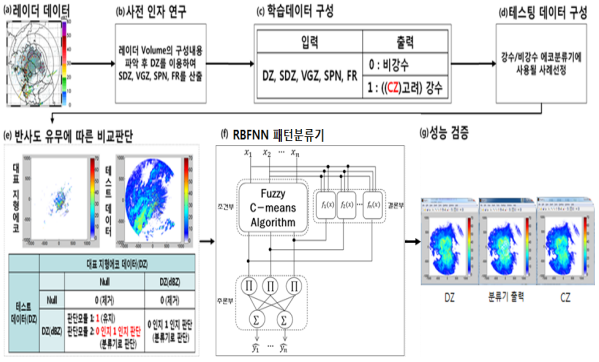


그림 3 RBFNN 기반 패턴 분류 시스템 설계 과정

Fig. 3 Design procedure of RBFNN based pattern-classification system

4. Fuzzy Recursive Weighted Least Square Estimation

결론부의 연결가중치는 다항식으로 구성되며 파라미터 계수는 순환적 가중최소자승법(RWLSE)을 통해 동조된다[5]. 기상 레이더 자료의 경우 데이터의 수가 수백만 개에 달하기 때문에 기존의 LSE를 사용할 경우 슈퍼컴퓨터가 아닌 이상, 메모리상에서 행렬의 저장 공간이 부족하여 연산을 수행하지 못하는 문제가 발생한다. 순환적 가중최소자승법의 목적은 빅데이터를 처리함에 있어서 연산 한계의 부족함을 해결하며 연산상의 메모리 부족문제를 해결함으로써 적은 메모리로도 빅데이터를 처리하여 효율성을 증대시키는 것이다.

4.1 퍼지 순환적 가중 최소자승법(Fuzzy RWLSE)의 과정

선형 모델인 퍼지 시스템의 규칙 베이스[6]는 다음과 같다.

$$R^i: If\ x_1 is\ A_{i1},\cdots,\ \text{and} x_k is\ A_{ik},\\ then\ y_i = a_{i0} + a_{i1}x_1 + \cdots + a_{ik}x_k \quad (11)$$

초기 몇 개의 샘플링데이터(N개)를 가지고 후반부 파라미터를 결정한다.

선형 모델인 퍼지 시스템의 규칙 베이스를 행렬로 표현하면 다음과 같다. y는 출력데이터, x는 입력데이터, w는 뉴로-퍼지 기반 알고리즘의 조건부에서 FCM클러스터링으로 구한 적합도이다.

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix} \begin{bmatrix} a_{10} \\ a_{11} \\ \vdots \\ a_{1k} \end{bmatrix} \quad (12) \qquad \mathbb{W}(N) = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_N \end{bmatrix} \quad (13)$$

$$w_{ji} = \frac{A_{j1}(x_{1i}) \wedge \cdots \wedge A_{jk}(x_{ki})}{\sum_{j=1}^{l} A_{j1}(x_{1i}) \wedge \cdots \wedge A_{jk}(x_{ki})} \quad (14)$$

Step 1) 첫 번째 과정에서 구한 행렬을 식으로 표현하면 다음과 같다.

$$\mathbb{Y}(N) = \mathbb{C}(N)\mathbb{X}(N) \quad (15)$$

여기서, N=데이터 번호, k=입력변수의 수

$$\mathbb{Y}(N) = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} \quad \mathbb{C}(N) = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix} \quad \mathbb{X}(N) = \begin{bmatrix} a_{10} \\ a_{11} \\ \vdots \\ a_{1k} \end{bmatrix} \quad (16)$$

Step 2) N번째 데이터까지는 가중 최소 자승법으로 다음 식에 의해 구해진다.

$$\mathbb{X}(N) = (\mathbb{C}(N)^T\mathbb{W}(N)\mathbb{C}(N))^{-1}\mathbb{C}(N)^T\mathbb{W}(N)\mathbb{Y}(N)) \quad (17)$$

$$\mathbb{W}(N) = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_N \end{bmatrix} \quad (18)$$

Step 3) N+1번째부터 입력 데이터에 대해서 순환동정 알고리즘 공식을 이용하여 계수를 추정한다.

$$\mathbb{P}(N) = (\mathbb{C}^T(N)\mathbb{W}(N)\mathbb{C}(N))^{-1} \quad (19)$$

$$\begin{cases} \mathbb{P}(N+1) = \mathbb{P}(N) - \dfrac{\mathbb{P}(N)c(N+1)c(N+1)^T\mathbb{P}(N)}{1+c(N+1)^T\mathbb{P}(N)c(N+1)}, \\ a(N+1) = a(N) + \mathbb{P}(N+1)c(N+1)(y(N+1) - c(N+1)^T a(N)) \end{cases} \quad (20)$$

$$c(N+1) = \begin{pmatrix} 1 & x_{(N+1)1} & x_{(N+1)2} & \cdots & x_{(N+1)k} \end{pmatrix} w_{N+1} \quad (21)$$

$a(N+1)$은 업데이트 되는 N+1번째의 계수이며, $c(N+1)$은 업데이트된 데이터로 각 규칙에 대한 적합도를 곱한 값이다.

파라미터를 동시에 구하는 전역학습 방법인 RLSE와 달리 위의 식처럼 RWLSE는 각 규칙에 대한 파라미터를 독립적으로 구하는 지역학습방법으로 각 규칙에 대해 해석 할 수 있으며, 행렬의 크기가 줄어들어 계산 부하를 줄이며 연산시간 또한 단축시킬 수 있다[7]. 하지만, 규칙 수가 적을 때는 전역학습방법인 RLSE에 비해 오히려 크기가 작은 행렬을 여러 번 연산해야 하는 RWLSE의 연산시간이 오래 걸릴 경우도 있다.

이러한 방법은 사용할 수 있는 메모리 용량까지는 기존의 WLSE로 계수를 추정한 뒤, 다음에 추정될 계수는 이미 얻어진 데이터들로부터 하나씩 업데이트하며 다항식의 계수를 얻는 프로세스이다. 이 알고리즘의 특징은 빅데이터를 처리함에 있어서 연산 한계의 부족함 해결과 실시간으로 데이터를 처리하여 학습에 용이하며 연산상의 메모리 부족을 해결함으로써 적은 메모리로도 빅데이터를 처리 할 수 있다는 것이다[8].

그림 4은 위의 과정을 도식화 한 것으로 선택한 데이터(n개)까지를 LSE 또는 WLSE를 사용하여 구해진 n번째의 파라미터를 Recursive 수식에 대입한 후, n+1번째 데이터의 파라미터를 구하는 것이다.

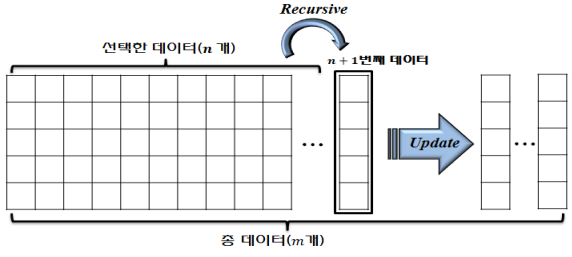


그림 4 순환동정 알고리즘의 데이터 갱신과정

Fig. 4 Update process of recursive identification algorithm

5. 패턴 판단모듈

본 연구에서는 Fuzzy-recursive 동정을 도입한 RBFNN 모델

을 사용하였으며, 기상청 레이더 사이트에서 데이터를 획득한 뒤에 강수에코와 비강수에코의 좀 더 확실한 분류를 위하여 패턴 판단모듈을 개발하게 되었다. 패턴 판단모듈은 대표 지형에코와 테스트 데이터의 패턴을 비교하여 Null값을 제거하는 모듈이다[9]. 판단모듈은 강수 사례시 현재 기상예보에 사용하는 QC(Quality Control)데이터로서, 성능 검증을 위해 비교대상으로만 사용되었다. QC에서 제거되는 반사도 값이 0~10dBZ인 약한 강수에코를 보존하며, 비강수 사례 테스트 시 반사도 정보가 없는 Null값, 확실한 비강수에코들을 뉴로-퍼지 알고리즘에 상관없이 제거한다.

판단모듈을 개발하기 위하여 먼저 대표 지형에코를 제작하였다. 대표 지형에코는 총 10시간 동안의 지형에코만을 나타내는 61개 UF자료 중 Sweep 0.5°일 때 중복되는 (Ray, Bin) 좌표에서의 DZ(필터링 되기 전 반사도 값)가 40개 이상(61개 자료 중 40개 이상) 존재하는 DZ 값들을 평균으로 취합하여 제작하였다.

패턴 판단모듈은 대표 지형에코를 이용하여 강수 사례 시 강수에코와 지형에코의 중복영역을 뉴로-퍼지 알고리즘으로 강수에코인지 지형에코인지 분류하며, 강수에코들 중 지형에코의 특성을 갖는 약한 강수들을 살리는데 목적이 있다. 다음 표와 같이 테스트 데이터와 대표 지형에코의 (Ray, Bin)좌표 정보를 비교하여 에코 판단모듈 경우의 수로 강수에코인지 Null값인지 판단한다.

표 1 패턴판단 모듈
Table 1 Pattern-judgement modules

		대표 지형에코 데이터(좌표)	
		Null	DZ(dBZ)
테스트 데이터	Null	① 0(제거)	② 0(제거)
	DZ(dBZ)	③ 1(존재)	④ 0 or 1 (분류기로판단)

패턴 판단모듈의 조건 ①, ②는 테스트 데이터가 Null값이라면 뉴로-퍼지 알고리즘에 상관없이 제거를 한다. 이는 Null값은 쓰레기 값으로 에코가 없다는 의미이기 때문에 굳이 알고리즘으로 분류할 필요가 없다 판단하고, 출력한다[10]. 패턴 판단모듈의 조건 ③은 테스트 데이터의 좌표(Ray, Bin)의 반사도 정보가 있지만 대표 지형에코에서는 Null값이 있는 경우로, 지형에코는 비강수 시 변화가 없다는 특성을 이용하여 대표 지형에코 이외에 생기는 반사도 값을 강수에코로 판단하는 것이다. 패턴 판단모듈의 조건 ④는 대표 지형에코의 좌표정보와 테스트 데이터의 좌표정보가 겹쳐지는 곳에서 둘 다 DZ값이 존재 할 경우, 뉴로-퍼지 알고리즘을 사용하여 테스트 데이터가 강수에코인지 지형에코인지를 판별하게 된다. 패턴 판단모듈은 강수 사례 시 QC의 단점인 0~10dBZ를 제거하며 전체적으로 DZ값을 낮추는 것을 보완해주는 역할을 한다. 또한 비강수 사례에서는 비강수에코가 확실한 부분은 DZ의 반사도 값이 음수이거나, Null 값이 있으며 레이더 지도에 나타나지 않는다. DZ의 dBZ 값이 60 이상인 부분은 비강수에코로 판단하여 제거한다. 현재 한반도에서 발생한 시간당 최고 강수량은 서울지역의 1937년 7월 30일 146.9mm이다. 60 dBZ 이상인 값들을 지형에코로 지정한 이유는 60dBZ를 시간당 강수량으로 계산을 하면 205.04mm 가 된다. 이는 한반도 시간당 최대강수량을 기준으로 어느 정도 오차를 두어서 한계 값을 정해주는 것이다. 따라서 위와 같은 값을 갖는 것들을 지형에코로 바로 판단 할 수 있게 모듈을 지정하였다. 식 (22)은 dBZ 값을 시간당 강수량으로 계산하는 식이며, 식 (23)은 판단 모듈을 사용할 시 RBFNN을 사용할 것인지를 판단하는 식이다.

$$dBZ = 10\log_{10} Z,\ \ R = \left(\frac{Z}{200}\right)^{\frac{1}{1.6}} \tag{22}$$

$$\begin{array}{ll} If & dBZ > 60 \text{ and } dBZ < 0,\ then\ Y = 0(\text{비강수}) \\ else & RBFNN(\text{강수인지 비강수인지 판단}) \end{array} \tag{23}$$

여기서 Z는 dBZ를 로그스케일로 바꾸기 전 값이며, R은 시간당 강수량이다. 위의 식은 패턴 판단모듈 조건 ③, ④에 적용된다.

6. 실험 결과 및 고찰

6.1 실험의 전체 개요

본 연구에서는 먼저, WLSE의 특성을 살려 Recursive에 접목시킨 이유를 설명하기 위해 RBFNN기반 LSE와 WLSE의 패턴분류율을 비교한다. Machine Learning Data인 WDBC, PIMA, Heart를 사용하였으며, 퍼지화 계수는 2, 다항식은 선형식(Linear)을 사용하였다. 또한, 다항식 기반 기저함수 신경회로망(Radial Basis Function Neural Networks : RBFNN) 알고리즘을 기반으로 기상 레이더 데이터를 이용하여 강수에코와 비강수에코(지형에코, 청천 에코)인 대용량 데이터를 학습시킨 후에, 비강수에코는 제거하며 강수에코는 분류하여 에코 맵으로 나타내는 방법을 수행한다.

표 2 사용된 기계 학습 데이터
Table 2 Used machine learning data

	입력변수 개수	전체 데이터 수	학습 데이터 수	테스트 데이터 수
WDBC	30	569	456	113
PIMA	8	768	614	154
Heart	13	270	216	54

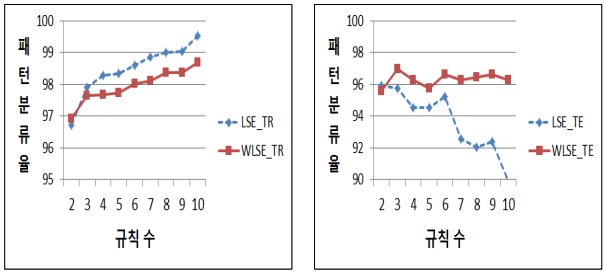
표 3 RBFNN의 파라미터 설정 값
Table 3 RBFNN's parameter setting values

Parameter	Value
Fuzzification coefficient	2
Polynomial type	Linear

6.2 LSE기반 RBFNN과 WLSE기반 RBFNN의 패턴분류율 비교

아래의 표는 LSE와 WLSE기반 RBFNN의 패턴분류율을 비교한

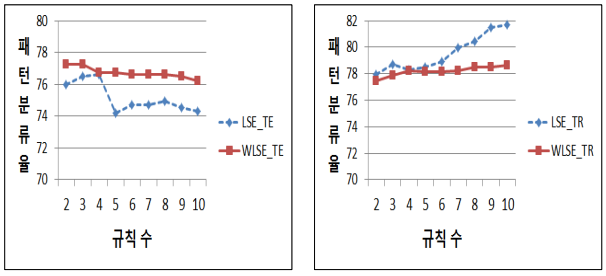
결과이며 데이터의 학습과 테스트는 4:1의 비율로 5-fold cross validation을 사용하여 데이터로 구성하였으며, 각 데이터의 패턴분류율을 평균으로 취하여 최종 패턴분류율로 표시하였다. 사용한 데이터는 Machine Learning Data인 WDBC, PIMA, Heart를 사용하였으며 총 WDBC의 학습데이터의 수는 456개, 테스트데이터의 수는 114개이다. PIMA데이터 학습데이터의 수는 614개, 테스트데이터의 수는 154개이다. Heart데이터 학습데이터의 수는 216개, 테스트데이터의수는 54개이다. 각각의 데이터에 대해 규칙수를 2~10으로 늘려가며 실험하여 그래프로 나타낸다. 결과적으로, LSE와



(a) In case of Training (b) In case of Testing

그림 5 WDBC데이터의 패턴분류율 비교 그래프

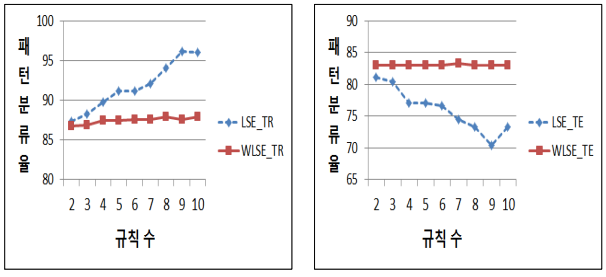
Fig. 5 Comparison graph of pattern classification rate of WDBC data



(a) In case of Training (b) In case of Testing

그림 6 PIMA데이터의 패턴분류율 비교 그래프

Fig. 6 Comparison graph of pattern classification rate of PIMA data



(a) In case of Training (b) In case of Testing

그림 7 Heart데이터의 패턴분류율 비교 그래프

Fig. 7 Comparison graph of pattern classification rate of Heart data

WLSE의 패턴분류율은 큰 차이가 없지만 규칙 수가 늘어남에 따라 LSE기반 RBFNN의 Testing 데이터의 패턴분류율은 Over-fitting이 일어날 가능성이 있는 데에 반해, WLSE기반 RBFNN의 Testing 데이터의 패턴분류율은 안정적으로 유지된다.

아래의 그림은 Machine Learning(ML) 데이터의 각 패턴분류율을 x축에는 규칙 수, y축에는 패턴분류율로 표기하여 그래프로 나타낸 것이다. 점선은 LSE의 패턴분류율이며, 실선은 WLSE의 패턴분류율이다.

6.3 Recursive 동정 알고리즘을 사용한 계수 추정

다음 그래프는 WDBC데이터를 사용하여 파라미터를 추정한 그래프의 이미지이다. 453개의 학습데이터 중에서 150개까지의 데이터를 기존의 LSE(Least Square Estimation)으로 추정을 한 뒤에, 151번째 데이터부터 RLSE(Recursive Least Square Estimation)으로 453번째의 데이터까지 추정한다. 데이터 한 개당 248개의 계수를 이전 LSE로 나온 데이터로부터 업데이트 하며 추정한다. 그래프의 x축은 전체 계수를 보여주며 y축은 추정된 계수 값을 보여준다. RLSE의 패턴분류율은 학습은 98.68, 테스트는 92.11으로 나오는 것을 확인 할 수 있다.

그래프의 실선은 LSE로 전체 453개의 데이터를 추정한 계수이며, 점선은 RLSE로 추정한 계수이다. RLSE로 전체 데이터를 추정했을 때와 LSE로 전체 데이터를 추정했을 때의 계수가 비슷해지는 것을 그래프로부터 확인 할 수 있다.

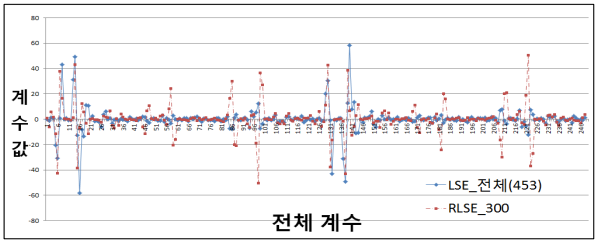


그림 8 Recursive 동정알고리즘을 사용한 WDBC 데이터의 계수추정

Fig. 8 Estimation of WDBC data using the recursive identification algorithm

6.4 학습데이터 구성 및 초기 파라미터 설정

UF데이터를 분석하여 강수에코와 비강수에코의 특성을 파악할 수 있었다. 각 사례 당 344,520개의 데이터이며, 강수 학습 데이터로는 각 5가지의 강수사례들(강설, 강우 밴드, 대류셀, 발달하는 대류셀)을 하나씩 샘플링 하여 중복되지 않는 데이터로 강수 학습데이터를 구성하였다. 총 강수 학습 데이터 수는 각 데이터의 Null값이 포함되지 않은 강수에코로 816,351개이다. 비강수 학습데이터로는 6가지 비강수 사례들(약한 청천에코, 파랑에코, 약한 파랑에코, 이상전파+파랑, 청천+파랑, 강한 청천에코)을 샘플링 하여 구성하였다. 각 사례 당 344,520개로 Null값을 제거한 총 비강수 학습데이터의 수는 848,130개이다. 강수와 비강수를 합친 총 학습데이터의 수는 1,664,481개이고, 이를 입력 데이터로 사용

하였다. 기존 WLSE에 순환 학습 방법을 추가한 RWLSE를 기반으로 빅데이터 처리를 위한 학습을 수행한다. 또한 테스트데이터로는 학습데이터에 포함되지 않은 다른 날짜, 다른 시간대의 강수에코 사례, 지형에코 사례 및 청천 에코 사례를 테스트 해보았으며 테스트데이터의 수는 344,520개이다. 테스트를 위한 초기 파라미터 설정 값은 표4와 같으며, 이는 각 파라미터들 중 제일 좋은 성능의 값을 나타내는 파라미터들로 실험을 하였다.

표 4 RWLSE-based RBFNN의 파라미터 설정 값
Table 4 Parameter setting values of RWLSE-based RBFNN

Parameter	Value
Fuzzification coefficient	2.5
Fuzzy rules(Number of cluster)	4
Polynomial type	Linear

6.5 강수에코 사례의 학습/테스트

강수 사례의 테스트로는 표6에서 볼 수 있으며, 강수 사례 중 2012년 04월 03일 08시 ~ 2012년 12월 14일 13시 총 8개의 레이더 자료를 테스트하였으며 분류 전, 분류 후, CZ 데이터를 비교 분석하여 그림 9에 나타내었다. 필터링 후의 반사도 값을 저장하는 CZ는 비강수에코 종류 중 하나인 지형에코를 어느 정도 제거해 주며, 일부 레이더로부터 찍힌 의미 없는 비강수에코를 제거하지만 완벽히 제거하지는 못한다. 이러한 CZ는 DZ와 성격이 유사하여 최종적으로 결과를 비교하는데 사용하였다. 분류 전의 지형에코들이 분류 후에는 제거되는 것을 볼 수 있으며, 제안된 패턴분류기의 분류 후 그림과 CZ를 비교해 보았을 때, 지형에코들이 조금 더 제거된 것을 볼 수 있다.

표 5 테스트한 강수에코 패턴분류율
Table 5 Pattern classification rate of precipitation echo for testing

에코 유형	날짜	패턴분류율	
		TR	TE
강수	2012년 04월 03일 08시 00분	88.91	90.79
강수	2012년 04월 21일 05시 00분		90.44
강수	2012년 10월 22일 09시 00분		94.38
강수	2012년 10월 27일 04시 00분		90.54
강수	2012년 11월 25일 20시 00분		92.38
강수	2012년 12월 05일 13시 00분		94.82
강수	2012년 12월 14일 06시 00분		90.90
강수	2012년 12월 14일 13시 00분		91.02

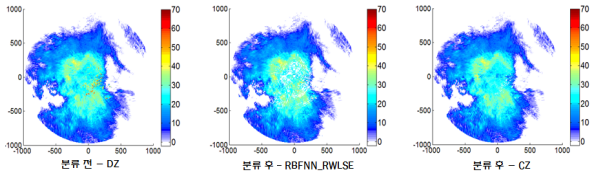


그림 9 강수 에코 지도(2012년 12월 14일 06시 00분)
Fig. 9 Precipitation echo map(December 14th, 2012, at 06:00 am)

6.6 비강수(지형, 청천)에코 사례의 학습/테스트

비강수 사례의 테스트로는 표6에서 볼 수 있으며, 비강수 사례 중 지형 에코로 나타나는 2012년 05월 15일 08시 ~ 2012년 05월 20일 19시 총 4개의 레이더 자료를 테스트 하였으며 분류 전, 분류 후, CZ 데이터를 비교 분석하여 그림 10에 나타내었다. 분류 전의 지형에코들이 분류 후에는 제거되는 것을 볼 수 있으며, 레이더에서 필터링을 거쳐 나온 반사도 값인 CZ와 비교해 볼 때, CZ보다 더 많은 지형에코가 제거됨을 보인다.

표 6 테스트한 비강수에코 패턴분류율
Table 6 Pattern classification rate of non-precipitation echo for testing

에코 유형	날짜	패턴분류율	
		TR	TE
지형(비강수)	2012년 05월 15일 08시 00분	88.91	99.96
지형(비강수)	2012년 05월 15일 11시 00분		100.00
지형(비강수)	2012년 05월 20일 08시 30분		100.00
지형(비강수)	2012년 05월 20일 19시 00분		100.00
청천(비강수)	2012년 05월 06일 20시 00분		99.88
청천(비강수)	2012년 05월 07일 21시 00분		94.92
청천(비강수)	2012년 10월 06일 19시 00분		99.19
청천(비강수)	2012년 10월 07일 23시 50분		94.12

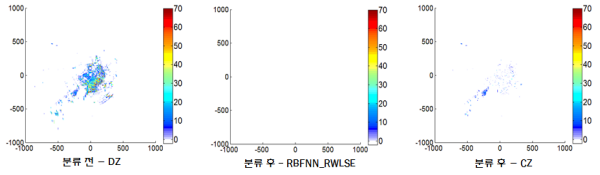


그림 10 비강수(지형) 에코 지도(2012년 05월 20일 19시 00분)
Fig. 10 Non-precipitation(ground) echo map(May 20th, 2012, at 07:00 pm)

비강수 사례중 청천 에코로 나타나는 2012년 05월 06일 20시 ~ 2012년 10월 07일 23시 50분의 레이더 자료를 분류 전, 분류 후, CZ데이터의 그림 11로 표현하였다. 그림을 보면 분류 전의 반사도 값들이 분류 후에는 제거 되는 것을 볼 수 있으며, CZ와 비교해 보아도 확실히 제거됨을 볼 수 있다. 이는 레이더 상에서 필터링을 거친 CZ보다 성능이 나은 것으로 판단된다.

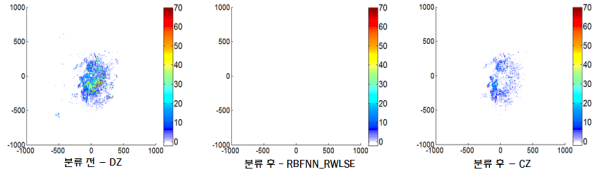


그림 11 비강수(청천) 에코 지도(2012년 05월 06일 20시 00분)
Fig. 11 Non-precipitation(clear) echo map(May 6th, 2012, at 08:00 pm)

7. 결 론

본 논문에서는 빅데이터를 처리하기 위한 Fuzzy RWLSE 알고리즘을 제안하였으며, 이 알고리즘의 다양한 분야에서의 필요성을 보였다. 다양하고 방대한 기상 레이더 빅데이터의 구성 분석 및 데이터 전처리 과정을 구축 하였으며, DZ를 이용하여 SDZ, VGZ, SPN을 산출 및 각 소속변수의 특성을 분석하였다. 다항식 기반 방사형 기저함수 신경회로망(RBFNN) 구조를 사용 하였고, 강수에코와 비강수에코인 대용량 데이터를 학습시킨 후, 비강수와 강수를 분류하여 비강수에코를 제거한 뒤 에코 맵으로 나타내는 방법을 제시하였다. 빅데이터를 처리하기 위한 퍼지 순환가중동정 알고리즘을 제안하였고 RBFNN에 순환적 가중최소자승법(RWLSE)을 접목시켰다. 기상 레이더 빅데이터인 UF데이터는 매우 방대하기 때문에 학습데이터를 만들 시에 메모리 용량의 부족으로 학습시키지 못하는 문제가 발생한다. 하지만 RWLSE를 사용함으로써 LSE를 사용함보다 더 나은 성능을 나타내며, 데이터가 추가됨에 따라 후반부 동정의 다항식의 계수를 업데이트 하는 성향을 보여 앞으로 사용될 빅데이터의 연산을 가능하게 하는 연산으로 볼 수 있다. 마지막으로 RWLSE를 사용한 후의 자료와 분류되기 전 DZ, 분류 후 CZ자료를 에코지도로 비교하여 RWLSE를 사용한 분류기가 더 확실히 분류되는 것을 그림으로 확인할 수 있었다.

감사의 글

This work was supported by GRRC program of Gyeonggi province [GRRC Suwon 2014-B2, Center for U-city Security & Surveillance Technology] and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1B3003568)

References

[1] S. K. Oh, W. D. Kim, and W. Pedrycz, "Fuzzy Radial Basis Function Neural Networks with information granulation and its parallel genetic optimization." Fuzzy Sets and Systems, Vol 237, pp 96-117, February 2014

[2] Kuo, Yi Ming, Hone Jay Chu, and Tsung Yi Pan. "Temporal precipitation estimation from nearby radar reflectivity using dynamic factor analysis in the mountainous watershed‐a case during Typhoon Morakot." Hydrological Processes 28.3 2014

[3] S-K. Oh, W-D. Kim, and W. Pedrycz, "Polynomial based radial basis function neural networks (RBFNN) realized with the aid of particle swarm optimization," Fuzzy Sets and Systems, Vol. 163, No. 1, pp. 54-77, 2011

[4] S. B. Roh, S. K. Oh, and W. Pedrycz. "Design of fuzzy radial basis function-based polynomial neural networks." Fuzzy sets and systems Vol 185, pp 15-37. December 2011

[5] Wang, Cheng, and Tao Tang. "Recursive least squares estimation algorithm applied to a class of linear-in-parameters output error moving average systems." Applied Mathematics Letters 29 (2014): 36-41.

[6] W.-D. Kim, S.-K. Oh, K.-S. Seo, and W. Pedrycz, "Growing Rule-based Fuzzy Model Developed with the Aid of Fuzzy Clustering ", IFSA World Congress & NAFIPS Annual Meeting, pp. 573-578, June 24-28, 2013.

[7] Ding, Shifei, and Xiaopeng Hua. "Recursive least squares projection twin support vector machines for nonlinear classification." Neurocomputing 130 (2014): 3-9.

[8] Rinnan, Åsmund, et al. "Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS." Journal of Chemometrics 28.5 (2014): 439-447.

[9] Walther, A., Schröder, M., Fischer, J., & Bennartz, R. (2009). Comparison of precipitation in the regional climate model BALTIMOS to radar observations. Theoretical and Applied Climatology, 1-14.

[10] Berenguer, M., Sempere-Torres, D., Corral, C., & Sánchez-Diezma, R. (2006). A fuzzy logic technique for identifying nonprecipitating echoes in radar scans. Journal of Atmospheric and Oceanic Technology, 23(9), 1157-1180.

저 자 소 개

강 전 성(Jeon-Seong Kang)

2012～현재 수원대학교 전기공학과 학부과정, 관심분야는 퍼지추론 시스템, 퍼지 순환 동정 알고리즘, 자동화 시스템

오 성 권(Sung-Kwun Oh)

1981년 연세대학교 전기공학과 졸업, 동 대학원 석사(1983), 박사(1993). 1983-1989년 금성산전연구소(선임연구원). 1996-1997년 캐나다 Manitoba 대학 전기 및 컴퓨터공학과 Post-Doc. 1993-2004년 원광대학교 전기전자 및 정보공학부 교수. 2005～현재 수원대학교 전기공학과 교수, 2002～현재 대한전기학회, 제어로봇시스템학회, 퍼지및지능시스템학회 편집위원. 2012～현재 Information Sciences 편집위원. 관심분야는 퍼지 시스템, 퍼지-뉴럴 네트워크, 자동화 시스템, 고급 computational intelligence, 지능 제어 등.