

퍼지 kNN과 conditional FCM을 이용한 퍼지 RBFNN의 설계

노 석범*, 오 성권**

* 원광대학교 전기 전자 정보 공학부

** 수원대학교 전기공학과

Design of Fuzzy RBFNN
Realized by Fuzzy kNN and Conditional FCM

Seok-Beom Roh and Sung-Kwun Oh

* Dept. of Electrical Electronic and Information Engineering, Wonkwang University

** Dept. of Electrical Engineering, The University of Suwon

Abstract - 퍼지 RBFNN의 설계에 있어 가장 중요한 과정인 Radial Basis Function의 결정은 퍼지 RBFNN의 모델링 성능을 좌우한다. 기존에는 FCM을 이용하여 Radial Basis Function의 초기 위치를 결정하고 오류 역전파 알고리즘과 같은 최적화 알고리즘을 이용하여 최적의 Radial Basis Function을 결정하였다. 근래에는 Conditional FCM을 이용하여 출력공간에 정의된 정보 입자의 정보를 이용하여 입력공간상에서 Radial Basis Function의 위치를 결정하여 퍼지 RBFNN의 성능을 개선시키고자 하는 연구 수행되어졌다. 그러나 출력공간상에서 얻은 정보입자를 입력공간상으로 정보 손실없이 전달할 수 없어서 기대한 만큼의 성능 개선을 이룰 수 없었다. 이를 개선하기 위해 출력 공간에서 정의된 정보 입자를 정보 손실없이 입력 공간에 투영하기 위하여 퍼지 kNN기법을 도입하여 새로운 퍼지 RBFNN 설계 방법을 제안한다.

1. 서 론

비선형성이 강하고, 복잡하며, 불명확한 시스템을 모델링하기 위하여 인간의 언어와 경험적 지식을 이용하여 쉽게 시스템을 묘사 할 수 있다. 위에 설명한 바와 같이 인간의 언어와 경험적 지식으로 묘사되는 시스템을 수학적으로 표현하기 위하여 퍼지 논리(Fuzzy logic)가 이용된다[1]. 퍼지 모델 분야의 대부분의 연구자들은 입력변수와 출력변수의 관계를 퍼지집합을 이용하여 표현하거나 근사화는 방법에 관심을 가지고 있다. 교사 신호가 없는 unsupervised learning에 기반을 둔 다양한 클러스터링 기법들을 이용하여 퍼지 규칙을 추출하는 모델들이 여러 연구자들에 의해 제시되고 있다. 다양한 클러스터링 기법들 중에서 Bezdeck이 제안한 fuzzy c-means (FCM) clustering [2] 기법이 가장 잘 알려진 퍼지 집합에 기반을 둔 대표적인 클러스터링 알고리즘이다. 기존의 FCM과 달리 Pedrycz에 의해 제안된 conditional fuzzy clustering은 출력공간에서 추출된 정보를 교사 신호로 이용하여 입력공간의 데이터들의 분포를 분석한다. Conditional fuzzy clustering 기법의 학습이 모두 끝난 후에는 입력공간상에서 각 cluster의 prototype (중심 값)을 얻을 수 있다. Conditional fuzzy clustering 기법에 의해 얻어진 prototype들이 오로지 입력공간상에서 정의되어지고 각각의 cluster에 속하는 소속정도를 구하기 위해서는 출력공간상에서 데이터 분포를 분석한 부가적인 정보가 필요하게 된다. 그러나 학습 데이터가 아닌 테스트 데이터의 경우에는 입력변수들의 값만이 알려져 있고 출력변수들의 값은 알려져 있지 않다고 가정하기 때문에 출력공간상에서 얻어야 할 부가적인 정보를 얻을 수 없게 된다. 출력공간상에서 얻어야 할 부가적인 정보의 부재는 conditional fuzzy clustering[3] 기법의 장점을 완벽하게 이용할 수 없게 만든다. 본 논문에서는 앞서 언급한 conditional fuzzy clustering의 단점을 보완하기 위하여 fuzzy K-nearest neighbors 기법을 사용하여 출력공간상의 부가적인 정보를 추정하고, 추정된 정보를 이용하여 각 cluster에 새로운 데이터가 속하는 소속정도를 구하게 된다.

2. 본 론

2.1 Conditional Fuzzy Clustering 기법

Radial Basis Function (RBF) 신경회로망을 설계하기 위하여 Radial Basis Function들의 위치와 모양을 정의 하는 것이 RBF 신경회로망의 모델링 성능을 결정하는 중요한 요소가 된다. Radial Basis Function의 위치와 모양을 정의하기 위해 clustering 기법을 사용하는 것은 잘 알려진 사실이다. RBF 신경회로망을 설계한다는 측면에서는 출력공간상에서 얻을 수 있는 부가적인 정보를 사용하는 것이 RBF 신경회로망의 성능 개선에 유리하다고 할 수 있다. 그러나 기존의 fuzzy c-means (FCM) clustering 기법은 전적으로 unsupervised learning에 기반을 둔 방법으로서 앞서 언급한 출력공간상의 부가적인 정보를 전혀 사용할 수 없다는 단점이 있다. 이러한 단점을 극복하기 위한 방법으로 conditional fuzzy c-means (CFCM) clustering 기법이 소개되었다. Conditional clustering 기법에서 데이터 패턴 x_k 는 출력공간상에서 정의된 언어적 변수에 의해 얻어진 부가적인 정보과 입력공간상에서의 데이터 패턴들의 분포를 고려하여 분류되어진다. 출력공간상에서 정의된 언어적 변수는 퍼지 집합 B로 표현되며 B와 같이 정의된다.

$$B : R \rightarrow [0, 1] \quad (3)$$

출력패턴 y_k 가 퍼지 집합 B에 소속되는 소속정도를 나타내는 f_k 는 (4)와

같이 정의된다. f_k 는 부가적 정보 입자의 멤버십을 의미한다.

$$f_k = B(y_k), \quad 1 \leq k \leq N \quad (4)$$

CFCM 기법을 위한 목적함수는 일반적인 FCM 알고리즘의 목적함수와 동일하며 (5)와 같이 정의된다.

$$J_l = \sum_{j=1}^R \sum_{k=1}^N (\mu_{jk}^l)^p \cdot \|x_k - v_j^l\|^2, \quad 1 \leq l \leq C \quad (5)$$

여기서 J_l 은 l번째 context (출력공간상에서 정의된 퍼지 집합)에서 정의된 목적함수를 의미하며, μ_{jk}^l 은 l번째 context에서 정의된 소속정도들의 행렬을 의미한다. v_j^l 은 동일 context에서 얻어진 각 클러스터의 prototype을 의미한다. 기존의 FCM과 동일한 목적함수를 사용하는 CFCM의 제한 조건은 FCM의 제한 조건과 다르다. FCM의 제한 조건은 (6)과 같이 정의되지만 CFCM의 제한 조건은 (7)과 같이 정의된다.

$$\sum_{j=1}^R \mu_{jk} = 1, \quad 1 \leq k \leq N \quad (6)$$

$$\sum_{j=1}^R \mu_{jk}^l = f_k, \quad 1 \leq k \leq N \quad (7-1)$$

$$f_k = B_l(y_k) \quad (7-2)$$

여기서, B_l 은 출력공간상에서 정의된 l번째 언어적 변수를 의미하며 퍼지 집합으로 정의된다. 제한 조건 (7-2)을 만족하며 목적함수 (7-1)을 최소화 시키는 최적화 문제는 (8)과 같이 정의한다.

$$\min_{U, v} J_l \text{ subject to } \sum_{j=1}^R \mu_{jk}^l = f_k \quad (8)$$

반복적 방법을 이용하여 최적화 문제 (8)을 최소화 시키는 partition matrix U와 각 클러스터의 prototype은 (9)와 (10)과 같다.

$$\mu_{ik}^l = \frac{f_k}{\sum_{j=1}^C \left(\frac{\|x_k - v_j^l\|}{\|x_k - v_j^l\|} \right)^{2/(p-1)}} \quad (9)$$

$$v_j^l = \frac{\sum_{k=1}^N (\mu_{jk}^l)^p \cdot x_k}{\sum_{k=1}^N (\mu_{jk}^l)^p} \quad (10)$$

각 클러스터의 partition matrix U를 구하기 위한 식(9)을 이용해서 새롭게 주어진 데이터 패턴 x_{new} 의 partition matrix를 구한다. 그러나 새롭게 주어진 데이터 패턴은 입력변수들의 값 x_{new} 만 알려져 있을 뿐 출력변수의 값 y_{new} 은 알려져 있지 않다. 그러므로 (7-2)를 통해 구해야 하는 f_{new}^l 을 구할 수 없게 된다. 그러므로 지금까지 CFCM을 이용하여 radial basis function의 위치와 모양을 정의하는 퍼지 모델의 경우에는 (9)를 그대로 사용하지 못하고 (11)과 같이 변경하여 사용하였다[12].

$$\mu_{ik}^l = \frac{1}{\sum_{j=1}^C \left(\frac{\|x_k - v_j^l\|}{\|x_k - v_j^l\|} \right)^{2/(p-1)}} \quad (11)$$

(11)은 기존의 FCM에 의해 정의된 partition matrix를 얻기 위한 식과 동일하다. 출력공간상에서 추출된 부가적 정보 B_l 은 각 클러스터의 prototype v_j^l 을 정의하기 위해서만 이용되고 RBF 신경회로망의 partition matrix를 구하기 위한 식 (11)에는 이용되어지지 않게 된다. 그러므로 출력공간에서 정의된 부가적인 정보 B_l 를 RBF 신경회로망의 partition matrix에도 반영하기 위해서는 새롭게 주어진 입력 패턴 x_{new} 에 대응하는 출력공간상에서 정의된 부가적 정보 f_{new}^l 를 추정할 필요성이 대두되었다.

2.2 Fuzzy K-nearest neighbors 기법을 이용한 부가적 정보 입자의 추정

Conditional fuzzy C-means clustering의 단점은 부가적 정보 입자인 f_k 를 새롭게 주어지는 데이터 패턴에 대해서는 알 수 없다는 것이다. 이러한 단점을 극복하여 CFCM을 이용하여 Radial Basis Function Neural Networks(RBFNN)의 Radial Basis Function (RBF)를 정의하는 모델의 성능개선을 이루기 위하여 classification 기법의 일종인 Fuzzy K-Nearest Neighbors(FKNN) 기법을 적용한다. FKNN 기법을 적용하여 새롭게 주어진 데이터에 대응하는 부가적 정보 입자 f_k 를 추정하고자 한다.

2.2.1 Fuzzy K-Nearest Neighbors 기법

Fuzzy K-Nearest Neighbors 기법은 기존의 Crisp K-Nearest Neighbors기법에

퍼지 집합의 개념을 적용한 기법으로 주어진 query instance와 학습 데이터 패턴들간의 유사성(similarity)의 정의에 기반을 두고 있다. 데이터 패턴들간의 유사성은 (19)와 같이 정의한다.

$$S(x_{L(i)}, q) = \begin{cases} 1, & \text{if } x_{L(i)} = q \\ \left(\frac{1}{\|x_{L(i)} - q\|} \right)^{\frac{2}{p-1}}, & \text{if } x_{L(i)} \neq q \end{cases} \quad (19)$$

여기서, S는 데이터 $x_{L(i)}$ 와 query instance q 와의 유사성을 의미하며, K는 Nearest neighbor들의 수, p는 퍼지화 계수 (fuzzification coefficient)를 의미한다. (19)를 통해 얻어진 유사성 $S(x_{L(i)}, q)$ 는 다음 조건을 만족한다.

$$0 < S(x_{L(i)}, q) \leq 1 \quad (20)$$

근접 위치에 있는 K개의 학습 데이터들의 출력 label을 이용하여 fuzzy K-Nearest Neighbor 기법을 적용하여 주어진 query instance의 출력 label은 (21)과 같이 구한다.

$$\hat{y}(q) = \sum_{i=1}^K S(x_{L(i)}, q) \cdot y(x_{L(i)}) = \sum_{i=1}^K \left(\frac{y(x_{L(i)})}{\sum_{j=1}^K \left(\frac{1}{\|x_{L(j)} - q\|} \right)^{\frac{2}{p-1}}} \right) \quad (21)$$

여기서, $\hat{y}(q)$ 는 query instance와 연관된 fuzzy K-Nearest Neighbor 기법을 이용하여 추정된 출력 label을 의미하며, $y(x)$ 는 데이터 패턴 x 와 연관된 알려진 출력 label을 의미한다.

2.2.2 Fuzzy K-Nearest Neighbors 기법을 이용한 부가적 정보 입자의 추정
fuzzy K-Nearest Neighbors 기법을 이용하여 테스트 데이터 집합의 부가적 정보를 추정한다. 테스트 집합의 부가적 정보 입자를 추출하기 위한 과정은 아래와 같다.

Step 1. 학습 데이터 집합에 속한 데이터 패턴들과 관련된 출력공간상에서 정의된 부가적 정보 입자(f_i^j)를 (13)을 이용하여 추출한다. (여기서 f_i^j 는 k 번째 데이터 패턴과 관련된 1번째 부가적 정보입자(퍼지 집합)의 적합도를 의미한다.)

Step 2. 학습 데이터 집합에 정의된 부가적 정보입자들의 적합도와 fuzzy K-Nearest Neighbor 기법을 이용하여 테스트 데이터 집합의 부가적 정보입자들의 적합도를 추정한다.

Sub-Step 2.1 테스트 데이터 집합의 j번째 데이터를 query instance로 정의한다.

Sub-Step 2.2 주어진 query instance와 가장 가까운 K개의 학습 데이터 패턴을 찾아 K개의 nearest neighbor로 정의한다.

Sub-Step 2.3 sub-step 2.2에서 정의된 K nearest neighbors와 (22)를 이용하여 query instance의 1번째 부가적 정보입자의 적합도를 추정한다.

$$\hat{f}_i^j(q) = \sum_{i=1}^K S(x_{L(i)}, q) \cdot f_i^j = \sum_{i=1}^K \left(\frac{f_i^j}{\sum_{j=1}^K \left(\frac{1}{\|x_{L(j)} - q\|} \right)^{\frac{2}{p-1}}} \right) \quad (22)$$

여기서, $\hat{f}_i^j(q)$ 는 query instance q 와 관련된 1번째 정보 입자의 적합도를 의미한다.

Sub-Step 2.4 sub-step 2.1에서 sub-step 2.3까지의 과정을 모든 테스트 데이터 집합의 데이터에 관하여 수행한다.

Step 3. Step 1과 Step 2를 통해 얻어진 학습 데이터와 테스트 데이터 집합들의 부가적 정보입자의 적합도와 (9)를 이용하여 학습 데이터와 테스트 데이터 집합의 partition matrix를 구한다.

Step 4. step 3에서 얻어진 학습 데이터의 partition matrix U 를 기반으로 Least Square Estimation을 이용하여 RBFNN의 후반부 계수를 추정한다.

CFCM을 이용한 RBFNN의 최종 출력은 (23)과 같다.

$$\hat{Y} = UA \quad (23)$$

$$\text{여기서, } U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} & u_{11}^2 & u_{12}^2 & \dots & u_{1n}^2 & \dots & u_{11}^C & u_{12}^C & \dots & u_{1n}^C \\ u_{21} & u_{22} & \dots & u_{2n} & u_{21}^2 & u_{22}^2 & \dots & u_{2n}^2 & \dots & u_{21}^C & u_{22}^C & \dots & u_{2n}^C \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nn} & u_{n1}^2 & u_{n2}^2 & \dots & u_{nn}^2 & \dots & u_{n1}^C & u_{n2}^C & \dots & u_{nn}^C \end{bmatrix}$$

$$A^T = [a_1^1 a_2^1 \dots a_n^1 a_1^2 a_2^2 \dots a_n^2 \dots a_1^C a_2^C \dots a_n^C]$$

C는 출력공간에 정의된 정보입자의 수, R은 출력공간에 정의된 각각의 정보입자와 관련된 입력공간상의 퍼지 규칙의 수를 의미한다.

RBFNN의 후반부 계수 행렬 A는 (24)와 같이 추정한다.

$$A = (U^T U)^{-1} (U^T Y) \quad (24)$$

3. 실험 연구 및 결과 고찰

Radial Basis Function Neural Networks의 성능을 개선하기 위한 방법으로서 conditional fuzzy C-means clustering을 이용한 RBFNN의 receptive field의 위치를 설정하는 방법이 Pedrycz에 의해 제안되었다. Pedrycz에 의해 제안된 방법은 conditional fuzzy c-means clustering의 장점인 출력공간상의 정보입자들의 정보의 활용이라는 특징을 최대한 이용하지 못하고 있다. 이러한 단점을 극복하기 위하여 본 논문에서는 fuzzy K-Nearest Neighbors 기법을 이용하여 테스트 집합에서도 출력 공간상에 정의된 부가적 정보 입자들의 정보를 활용할 수 있도록 개선하였다.

제안된 퍼지 KNN과 CFCM을 이용하여 설계된 RBFNN의 성능 평가를 위하여 본 논문에서는 두가지 종류의 machine learning 데이터 집합 (1: Automobile MPG (Miles Per Gallon) Data, 2: Boston Housing Data)에 제안된 알고리즘을 적용하여 성능을 평가하였다. 성능 평가를 위한 평가 지수는

Root Mean Square Error(RMSE)를 사용하였다.

$$SE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (25)$$

제안된 알고리즘의 성능 평가를 위한 실험 조건은 표 1과 같다.

표 1 제안된 모델의 실험 파라미터

Parameter	Value
Polynomial Order (O)	1 (linear)
No. of Nearest Neighbors (K)	2 ~ 7
Fuzzification Coefficient (p)	1.2 ~ 4.0 (0.2 step)
No. of Contexts (C)	2 ~ 6
No. of Rules per each context	2 ~ 6

3.1 Automobile MPG (Miles Per Gallon) Data

Automobile MPG 데이터 집합은 각각의 자동차 종류에 따라 1갈론(gallon)의 연료로 자동차가 달릴 수 있는 거리를 표현하는 데이터들의 집합이다. 표 3은 제안된 모델과 기존의 다른 모델과의 모델링 성능 비교를 보인다.

표 3 모델의 성능 비교

Model	No. of Clusters	PI	EPI	
RBFNN [4]	36	3.24±0.24	3.62±0.31	
RBFNN with FCM [4]	36	3.21±0.21	3.51±0.27	
Proposed Model	Evenly Partition	K=6, C=5, R=4, p=2.2	0±0	1.74±0.20
	K-Means Clustering	K=6, C=4, R=4, p=2.2	0±0	2.73±0.27

3.2 Boston Housing Data

Boston Housing Data 집합은 보스턴 지역의 실제 부동산에 대한 정보를 취급하는 데이터들의 집합이다. 506개의 데이터 집합을 60%와 40%의 데이터 집합으로 랜덤하게 분할한다. 10번 반복하여 랜덤 분할하여 제안된 모델의 성능을 평가한다.

표 4 모델의 성능 비교

Model	No. of Clusters	PI	EPI	
RBFNN [4]	25	6.36±0.24	6.94±0.31	
RBFNN with FCM[4]	25	5.52±0.25	6.91±0.45	
Proposed Model	Evenly Partition	K=4, C=2, R=4, p=1.5	0±0	1.74±0.20
	K-Means Clustering	K=6, C=4, R=5, p=1.4	0±0	3.73±0.27

3. 결 론

본 논문에서는 RBFNN의 receptive field들의 위치 설정을 위한 알고리즘으로 많이 연구되어지고 있는 conditional fuzzy c-means clustering의 단점인 출력 공간상에서의 부가적 정보입자들의 추정을 위하여 fuzzy K-nearest neighbors 기법을 적용하였다. Conditional fuzzy c-means clustering 기법은 출력 공간상에서 출력변수들의 분포를 분석하여 입력공간에 receptive field의 위치를 설정하여, RBFNN의 모델링 성능을 개선시킨다. 그러나 출력변수의 값이 알려져 있지 않은 테스트 데이터의 경우에는 출력공간상에서의 부가적 정보 입자를 얻을 수 없기 때문에 conditional fuzzy c-means clustering을 적용할 수 없게 된다. 이를 극복하기 위하여 fuzzy k-nearest neighbor 기법을 적용하여 출력변수가 알려져 있지 않은 테스트 데이터 집합에도 conditional fuzzy c-means clustering을 적용할 수 있도록 한다. 제안된 알고리즘에 의해 설계된 RBFNN은 기존 모델에 비해 학습 데이터의 모델링 성능 지수가 매우 우수한 것을 알 수 있었다. 더불어 제안된 모델의 일반화 능력 (테스트 데이터에 대한 모델링 성능) 또한 기존의 다른 모델에 비해 우수한 것을 보인다.

감사의 글

이 논문은 2008년도 한국전자통신연구원의 재원으로 지원을 받아 수행된 연구임 (과제번호:2008-35MIC). 그리고 경기도의 경기도 지역협력연구센터사업의 일환으로 수행하였음 [GGA0801-45700, U-city 보안감시 기술협력센터].

[참고 문헌]

- [1] R. M. Tong, "Synthesis of fuzzy models for industrial processes," *Int. J. Gen. Syst.*, vol 4, pp. 143-162, 1978.
- [2] Bezdek, J. C., "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum Press, NY, 1981
- [3] W. Pedrycz, "Conditional Fuzzy C-Means," *Pattern Recognition Lett.* 17 (6) (1996) 625-632.
- [4] W. Pedrycz and K. C. Kwak, "Linguistic Models as a Framework of User-Centric System Modeling", *IEEE Trans. on Systems, Man and Cybernetics-A*, Vol. 36, No. 4, pp. 727-745, 2006