빅 데이터 처리를 위한 증분형 FCM 기반 RBF Neural Networks 패턴 분류기 설계

이승철*, 오성권*, 노석범**
수원대학교 전기공학과*, 원광대학교 전자융합공학과**

Design of Incremental FCM-based RBF Neural Networks Pattern Classifier for Processing Big Data

Seung-Cheol Lee*, Sung-Kwun Oh* and Seok-Beom Roh**
Department of Electrical Engineering, The University of Suwon*
Department of Electronic Convergence Engineering, The University of Wonkwang**

Abstract - 본 연구에서는 증분형 FCM(Incremental Fuzzy C-Means: Incremental FCM) 클러스터링 알고리즘을 기반으로 방사형 기저함수 신경회로망(Radial Basis Function Neural Networks: RBFNN) 패턴 분류기를 설계한다. 방사형 기저함수 신경회로망은 조건부에서 가우시안 함수 또는 FCM을 사용하여 적합도를 구하였지만, 제안된 분류기에서는 빅 데이터간의 적합도를 구하기 위해 증분형 FCM을 사용한다. 또한, 빅 데이터를 학습하기 위해 결론부에서 재귀최소자승법(Recursive Least Square Estimation: RLSE)을 사용하여 다항식 계수를 추정한다. 마지막으로 추론부에서는 증분형 FCM에서 구한 적합도와 재귀최소자승법으로 구한 다항식을 이용하여 최종 출력을 구한다.

1. 서 론

컴퓨팅 기술의 발달과 디지털 경제의 확산으로 인해 최근 세계적으로 빅 데이터를 주목하고 있고, 우리나라 IT 10대 핵심기술 중 하나로 빅 데이터를 선정하여 빅 데이터의 중요성이 더욱 부각되고 있다. 또한 세계 경제 포럼은 2012년 떠오르는 10대 기술 중 하나로 빅 데이터 처리 기술을 선정하여 실시간으로 수집 가능한 데이터를 처리하기 위한 기술들의 중요성이 부각되고 있다. 특히, 방대한 양의 데이터를 어떻게 학습시킬 것인가에 대한 연구는 인공지능 및 기계학습 분야에서 활발히 진행 되고 있다. 본 논문에서는 방대한 양의 데이터를 학습하기 위한 방사형 기저함수 신경회로망을 설계한다. 증분형 Fuzzy C-Means 클러스터링 알고리즘과 재귀최소자승법을 사용하여 방대한 양의 데이터를 한번에 학습하는 것이 아닌 데이터를 순환적으로 학습하는 방법으로 설계한다.

2. 본 론

2.1 방사형 기저함수 신경회로망

제안된 패턴 분류기는 그림 1과 같이 조건부, 결론부 그리고 추론부로 구성된다. 조건부에서 일반적으로 사용되는 가우시안 함수 대신에 증분형 FCM 클러스터링 알고리즘을 사용하여 하나의 클러스터에 속해져 있는 각각의 데이터 점의 소속정도를 퍼지 집합으로 출력하여 활성함수로 이용한다. 그리고 결론부의 연결가중치는 식(1)~(4)와 같이 다항식 형태로 구성되고, 재귀최소자승법을 이용하여 계수를 추정한다. 증분형 FCM 클러스터링 알고리즘과 재귀최소자승법을 사용함으로써 방대한 양의 데이터를 순환적으로 처리한다.

Type 1 : 상수항(Constant)

$$f_j(x_1,\cdots,x_k) = a_{j0} \tag{1}$$

Type 2 : 1차 선형식(Linear)

$$f_j(x_1,\cdots,x_k) = a_{j0} + \sum_{i=1}^{k} a_{ji}x_i \tag{2}$$

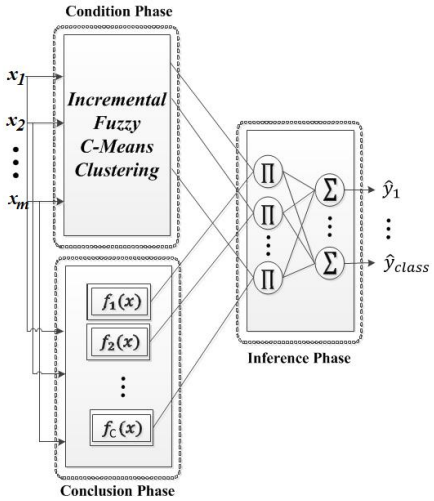
Type 3 : 2차 선형식(Quadratic)

$k=2$:

$$f_j(x_1,\cdots,x_k) = a_{j0} + \sum_{i=1}^{k} a_{ji}x_i + \sum_{i=1}^{k} a_{j(k+i)}x_i^2 + a_{(2k+1)}x_1x_2 \tag{3}$$

$k \geq 3$:

$$f_j(x_1,\cdots,x_k) = a_{j0} + \sum_{i=1}^{k} a_{ji}x_i + \sum_{i=1}^{k} a_{j(k+i)}x_i^2 + a_{(2k+1)}x_1x_2 + \cdots + a_{(k(k+3)/2)}x_{(k-1)}x_k \tag{4}$$



<그림 1> 제안된 방사형 기저함수 신경회로망의 구조

2.2 증분형 Fuzzy C-Means 클러스터링 알고리즘

제안된 패턴 분류기의 조건부에서 가우시안 함수 대신 사용하는 증분형 Fuzzy C-Means(FCM) 클러스터링 알고리즘을 사용한다. 먼저 일반적인 FCM 클러스터링 알고리즘은 n개의 입력변수 집합을 c개의 퍼지 그룹들로 분할하고 목적함수가 최소가 되도록 각 클러스터의 중심점을 데이터 전체를 이용하여 찾는 알고리즘이고, 증분형 FCM은 데이터의 일부를 이용하여 중심점을 찾고, 추가적으로 들어오는 데이터에 따라 중심점을 변경하는 알고리즘이다. 증분형 FCM 클러스터링 알고리즘의 단계는 다음과 같다.

[단계 1] 클러스터의 개수 $c(2 \leq c \leq h)$를 정하고 퍼지화 계수 $m(1<m<\infty)$을 선택한다. 그리고 초기 소속행렬 $U^{(r)}$을 초기화 한다.

$$U^{(r)} = \left\{ u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1 \forall l, 0 < \sum_{k=1}^{N} u_{ik} < n \forall j \right\} \tag{5}$$

[단계 2] 클러스터의 중심 v를 계산한다. 식 (6)은 알고리즘의 목적함수이다.

$$J(u_{ik}, v_i) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m \| x_k - v_i \|^2 \tag{6}$$

여기서, u_{ik}은 0과 1사이의 소속정도를 나타내는 값으로 $i(i=1,...,c)$번째 클러스터에 속해져 있는 x_k의 $k(k=1,...,n)$번째 데이터의 소속정도를 나타낸다. v는 $i(i=1,...,c)$번째 클러스터 중심 벡터이다. m은 퍼지화 계수를 나타내며 $m \in [1,\infty]$와 같은 범위를 가지고 있다.

식 (6)의 목적함수를 최소화하기 위해서 다음과 같이 목적함수를 세분

화 시켜야 한다.

$$v_{il}^{(r)} = \frac{\sum_{k=1}^{N} (u_{ik})^m \cdot x_{kl}}{\sum_{k=1}^{N} (u_{ik})^m} \tag{7}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|} \right)^{2/(m-1)}} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)}} \tag{8}$$

[단계 3] 계산된 중심 v를 이용하여 새로운 소속행렬 $U^{(r+1)}$을 계산한다.

$$u_{ik}^{(r+1)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{2/(m-1)}} \tag{9}$$

[단계 4] 식 (10)을 계산하고, 만약 $\Delta > \varepsilon$이면 $r = r+1$로 정하고 단계 2로 돌아간다. $\Delta \leq \varepsilon$이면 다음 단계로 넘어간다.

$$\Delta = \| U^{(r+1)} - U^{(r)} \| = \max_{i,k} \left| u_{ik}^{(r+1)} - u_{ik}^{(r)} \right| \tag{10}$$

[단계 5] N개의 데이터를 이용하여 단계 1~단계 4를 거쳐 소속행렬을 계산한 후, 순차적으로 들어오는 데이터를 이용하여 새로운 소속행렬 $u_{i,N+1}$을 계산한다.

$$u_{i,N+1} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| x_{N+1} - v_j \|}{\| x_{N+1} - v_i \|} \right)^{2/(m-1)}} \tag{11}$$

[단계 6] 새로운 소속행렬 $u_{i,N+1}$을 이용하여 데이터의 새로운 중심점을 계산한다. 그리고 단계 3으로 돌아가서 알고리즘을 반복하고, 새로운 데이터가 없을시 알고리즘을 종료한다. 단계 3으로 되돌아가면 그 이후로 단계 4는 무시한다.

$$v_i^{N+1} = \frac{\sum_{k=1}^{N+1} (u_{ik})^m \cdot x_{kl}}{\sum_{k=1}^{N+1} (u_{ik})^m} = \frac{\sum_{k=1}^{N} (u_{ik})^m \cdot x_{kl} + u_{i,N+1} \cdot x_{kl}}{\sum_{k=1}^{N} (u_{ik})^m + u_{i,N+1}} \tag{12}$$

$$= \frac{v_{il} + \frac{u_{i,N+1}}{\sum_{k=1}^{N} u_{ik}} \cdot x_{kl}}{1 + \frac{u_{i,N+1}}{\sum_{k=1}^{N} u_{ik}}}$$

2.3 재귀최소자승법

결론부의 연결가중치는 방대한 양의 데이터를 모두 사용하여 연결가중치를 추정하는 최소자승법(LSE)으로는 컴퓨터 메모리 부족과 같은 문제를 발생시키기 때문에 일부 데이터까지만 최소자승법으로 추정하고, 그 이후로는 재귀최소자승법(RLSE)을 이용하여 추정한다. 재귀최소자승법은 데이터가 순환적으로 추가됨에 따라 연결가중치를 업데이트 한다. 먼저 최소자승법은 식 (13)으로 표현될 수 있다. θ_k는 k번째 데이터까지의 연결가중치를 의미하고, θ_{k+1}은 식 (14)로 표현할 수 있다.

$$\theta_k = (A^T A)^{-1} A^T Y \tag{13}$$

$$\theta_{k+1} = \left(\begin{bmatrix} A \\ a^T \end{bmatrix}^T \begin{bmatrix} A \\ a^T \end{bmatrix} \right)^{-1} \begin{bmatrix} A \\ a^T \end{bmatrix}^T \begin{bmatrix} Y \\ y \end{bmatrix} \tag{14}$$

여기서, P_k를 식 (15)라 가정하면, θ_{k+1}과 P_{k+1}은 식 (16), 식 (17)로 정의 될 수 있다.

$$P_k = (A^T A)^{-1} \tag{15}$$

$$P_{k+1} = P_k - \frac{P_k a_{k+1} a_{k+1}^T P_k}{1 + a_{k+1}^T P_k a_{k+1}} \tag{16}$$

$$\theta_{k+1} = \theta_k + P_{k+1} a_{k+1} (y_{k+1} - a_{k+1}^T \theta_k) \tag{17}$$

2.4 실험 및 결과 고찰

본 논문에서 제안한 패턴분류기의 성능을 평가하기위해 Machine Learning 데이터인 Shuttle 데이터를 이용하였다. Shuttle 데이터는 총 58,000개의 데이터로 구성되어 있고, 입력은 9개로 구성되어 있다. 그리고 제안된 패턴 분류기의 성능을 객관적으로 평가하기 위해 5-fold cross validation을 이용하였다. 표 1은 제안된 패턴 분류기의 파라미터를 나타낸다.

<표 1> 제안된 패턴 분류기의 파라미터

Parameters	Value
No. of Inputs	9
No. of Classes	7
Fuzzification Coefficient	2.0
Polynomial Type	Linear

표 2는 데이터의 일부를 학습시킨 후, 추가적으로 들어오는 데이터를 순차적으로 학습시키는 제안된 분류기(Classifier 1)와 데이터 전체를 한 번에 학습하는 분류기(Classifier 2)와의 성능 비교를 나타낸다. 실질적으로 Classifier 1과 Classifier 2의 성능차이는 크게 나타나지 않는다. 그 이유로 제안된 분류기는 방대한 양의 데이터를 학습하기 위한 것으로써, 성능 개선을 위한 분류기가 아니기 때문이다. 실험에 사용된 데이터는 빅 데이터로 보기에는 데이터 수가 다소 적기 때문에 제안된 분류기의 효율성이 부각되지 않는다. 하지만 방대한 양의 데이터를 이용한다면 제안된 분류기의 효율성이 부각될 것이라 판단된다.

<표 2> 제안된 패턴 분류기의 성능평가

	Classifier 1		Classifier 2	
No. of Rules	TR	TE	TR	TE
4	95.68±0.06	94.71±0.23	95.91±0.06	94.94±0.22
8	97.32±0.04	97.30±0.16	97.49±0.05	97.43±0.21
10	97.46±0.07	97.44±0.15	97.74±0.04	97.70±0.18
15	98.26±0.09	98.23±0.36	98.22±0.14	98.20±0.13
20	98.68±0.11	98.65±0.27	98.42±0.20	98.40±0.17

3. 결 론

본 논문에서는 빅 데이터 처리를 위한 증분형 FCM 기반 RBF Neural Networks를 설계하였다. 제안된 패턴 분류기는 전체 데이터 중 일부를 학습한 후, 증분형 FCM 클러스터링 알고리즘과 재귀최소자승법을 이용하여 추가적으로 들어오는 데이터를 하나씩 순환적으로 학습할 수 있다. 이 방법을 통해 방대한 양의 빅 데이터를 학습시킬 때 메모리 용량 부족으로 학습이 어려웠던 것을 가능하게 해준다. 본 실험에서 사용된 데이터를 빅 데이터로 보기는 어렵기 때문에 제안된 패턴 분류기의 우수성을 확인하기 어려웠지만, 분류기의 가능성은 충분히 확인하였다. 향후 연구 계획으로는 방대한 기상 레이더 데이터를 이용하여 실험을 진행 할 예정이다. 기상 레이더 데이터는 빅 데이터로써 제안된 패턴 분류기에 적합한 데이터로 생각되며, 제안된 패턴 분류기의 우수성 또한 확인 할 수 있을 것이라 생각된다.

감사의 글

본 연구는 경기도의 경기도지역협력연구센터사업의 일환으로 수행하였음[GRRC 수원2015-B2, U-city 보안감시 기술협력센터] 그리고 한국산업단지공단의 10차년도 산업집적지경쟁력강화산업계획의 생산기술사업화 지원사업으로 연구를 수행하였음(NTIS-1415136442)

[참 고 문 헌]

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