# 클러스터링 기법과 유전자 알고리즘에 의한 다중 퍼지 모델의 동정

박병준*, 이수구*, 오성권*, 김현기**
원광대학교 전기전자공학부*, 수원대학교 전기전자정보통신공학부**

# The Identification of Multi-Fuzzy Model by means of HCM and Genetic Algorithms

Byoung-Jun Park*, Su-Gu Lee*, Sung-Kwun Oh*, Hyun-Ki Kim**,
*School of Electrical and Electronic Engineering, Wonkwang Univ., Iksan, KOREA
**School of Electrical Engineering, Suwon Univ. Whasung-gun, KOREA

**Abstract** - In this paper, we design a Multi-Fuzzy model by means of clustering method and genetic algorithms for a nonlinear system. In order to determine structure of the proposed Multi-Fuzzy model, HCM clustering method is used. The parameters of membership function of the Multi-Fuzzy are identified by genetic algorithms. We use simplified inference and linear inference as inference method of the proposed Multi-Fuzzy model and the standard least square method for estimating consequence parameters of the Multi-Fuzzy. Finally, we use some of numerical data to evaluate the proposed Multi-Fuzzy model and discuss about the usefulness.

## 1. 서 론

많은 연구자들은 자연 현상을 이해하고 그것을 이용하기 위한 연구가 이루어지고 있다. 그 중 1965년에 Zadeh[1]에 의해 창안된 "퍼지집합"이론은 비선형적이고 복잡한 실 시스템의 특성을 해석하는데 적용함으로써 수학적 모델보다 좋은 결과를 가져왔다. 설계할 시스템의 성능 및 기능의 요구조건에 따라 퍼지 모델은 애매모호한 언어적 변수를 수치적으로 표시할 수 있어서 융통성 있는 시스템 설계를 가능하게 하고 시스템의 기능을 향상시키며 설계를 간단하게 해주는 장점이 있다. 그러나 퍼지 규칙을 결정하는데 있어서 전문가와 시행착오(trial and error)에 의존해야 하는 어려움이 있어 동적으로 변화하는 환경에서 적응적으로 대처할 수 있는 추론 시스템을 구성하기가 힘들다. 문제점을 해결하기 위한 한가지 방법으로 다중 퍼지 모델을 제안한다. 다중 퍼지 모델은 HCM[2] 클러스터링 알고리즘을 이용하여 입출력 데이터들의 유사한 특성에 따라 분류하고 그 분류된 데이터군들을 가지고 유전자 알고리즘을 이용하여 전반부 파라미터를 최적으로 동정한다. HCM 클러스터링과 유전자 알고리즘에 의해 최적화된 다중 퍼지 모델은 다른 기존의 모델들보다 우수함을 보인다. 더 나아가 기존의 학습 데이터뿐만 아니라 테스트 데이터를 고려한 퍼지 모델 성능 즉 근사화 능력과 예측 능력 모두를 고려하여 실 공정 적용의 유용성에 그 방향을 맞추었다. 하중 값을 가진 목적 함수를 사용하여 설계자의 의도, 성능 결과의 상호균형, 시스템의 비선형 정도 등에 따른 최적 모델 동정을 시도하였다. 두 형태의 퍼지 모델 방법은 간략 추론과 선형 추론에 의해 시행되며, 멤버쉽 함수의 형태로는 삼각형 형태를 사용한다. 제안된 다중 퍼지 모델의 성능을 평가하기 위해서 Box[7]와 Jenkins[8]가 사용한 시계열 데이터와 가스 터빈 발전소의 데이터를 이용하였다.

## 2. 다중 퍼지 모델

### 2.1 다중 퍼지 모델 구조

기존의 퍼지 모델들에서 사용하였던 HCM 클러스터링은 시스템의 입력 공간의 퍼지 분할을 통하여 초기 멤버쉽 함수의 파라미터를 결정하는데 사용되었으나[10], 이는 데이터의 특성에 맞는 모델을 구축하지는 못하였다. 본 논문에서는 HCM 클러스터링을 이용하여 서로 유사한 특성을 가진 그룹으로 데이터를 분류하고 그 분류된 데이터의 특성에 맞는 각각의 퍼지 모델을 그림 1과 같이 구축한다. 여기에서 최적의 전반부 파라미터를 동정하기 위해 유전자 알고리즘을 이용한다. HCM 클러스터링에 의해 분류된 클러스터 수는 다중 퍼지 모델을 구성하기 위한 단일 모델의 수가되고, 분류된 데이터 집합들은 각각 기본 모델의 입출력이 되어 단일 퍼지 모델의 기능을 수행하게 된다.

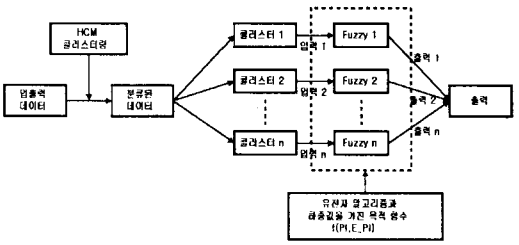



그림 1. 다중 퍼지 모델 구조

### 2.1.1 전반부 동정

퍼지 모델링에서 전반부 동정, 즉 구조 동정 및 파라미터 동정은 비선형 시스템을 표현하는데 있어서 매우 중요하다. 왜냐하면 전반부 입력 변수의 선택과 선택된 입력 변수의 공간 퍼지 분할 그리고 파라미터 동정은 비선형 시스템의 성능을 결정하는데 많은 영향을 미치기 때문이다. 전반부에서 멤버쉽 함수의 형태로는 삼각형 형태를 사용한다. 기존의 방법은 멤버쉽 함수를 그림 2처럼 입력 변수의 최소값과 최대값 사이를 임의의 개수로 등분해서 일률적으로 정의하였으나 이는 데이터들이 가지고 있는 특성을 제대로 반영하지 못하는 단점이 있다. 그래서 전반부 파라미터 동정을 위해 유전자 알고리즘을 이용한다. 유전자 알고리즘을 이용하면 위에서 언급한 문제점을 해결할 수 있고 멤버쉽 함수의 정점과 같은 파라미터들을 최적으로 동정할 수 있다.

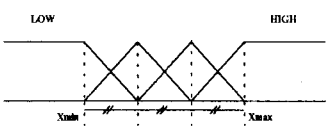



그림 2. Max-Min에 의한 멤버쉽 함수 정의

### 2.1.2 후반부 동정

퍼지 모델의 후반부 동정도 전반부와 마찬가지로 구조 동정과 파라미터 동정으로 나뉘어진다. 후반부 구조로는 퍼지추론에 의해 구별되는 구조 1, 2를 사용하며, 각 구조에 대한 후반부 파라미터 동정은 다음과 같다.

구조 1(후반부 : 상수)

후반부가 단일 상수항만을 가지는 것으로, 이와 같은 추론법을 간략추론법이라 한다. 이 퍼지 모델은 식 (1)과 같은 형태를 가지는 구현 규칙들로 구성되며, 퍼지 추론에 의해 추론된 값 $y^*$는 식 (2)와 같다.

$$R^n: IF\ x_1\ is\ A_{n1}\ \text{and}\ \cdots\ \text{and}\ x_k\ is\ A_{nk}\ Then\ a_{no} \quad (1)$$

$$y^* = \frac{\sum_{j=1}^{n} w_{ji}\, y_j}{\sum_{j=1}^{n} w_{ji}} = \frac{\sum_{j=1}^{n} w_{ji}\, a_{j0}}{\sum_{j=1}^{n} w_{ji}} \quad (2)$$

후반부 파라미터는 $a_{j0}$로써 입출력 데이터가 주어졌을 때 최소 자승법에 의해 구해진다. 최소 자승법에 의한 후반부 파라미터의 동정은 식 (3)에 의해 구해진다.

$$\hat{a} = (X^T X)^{-1} X^T Y \quad (3)$$

구조 2(후반부 : 일차 선형식)

후반부가 일차 선형식을 가지는 것으로, 이와 같은 추론법을 선형추론법이라 한다. 이 퍼지 모델은 식 (4)의 형태를 가지는 구현 규칙들로 구성된다.

$$R^n\ :\ IF\ \ x_1\ is\ A_{n1}\ \text{and}\ \cdots\ \text{and}\ x_k\ is\ A_{nk}$$
$$Then\ a_{10} + a_{n1}x_1 + \cdots + a_{nk}x_k \quad (4)$$

선형 추론법에 의해 추론된 값 $y^*$는 다음과 같다.

$$y^* = \frac{\sum_{j=1}^{n} w_{ji}\, y_j}{\sum_{j=1}^{n} w_{ji}} = \frac{\sum_{j=1}^{n} w_{ji}\, (a_{j0} + a_{j1}x_{1i} + \cdots + a_{jk}x_{ki})}{\sum_{j=1}^{n} w_{ji}} \quad (5)$$

최소 자승법에 의한 후반부 파라미터 동정은 구조 1과 같이 식 (3)에 의해 구해진다.

### 2.2 HCM 클러스터링

클러스터링 알고리즘이란 데이터의 분류를 위해서 사용되는 것으로 데이터의 내부가 비슷한 패턴, 속성, 형태 등의 기준을 통해 데이터를 분류하여 내부의 구조를 찾아내는 것이다. 본 논문에서는 클러스터링 알고리즘 중에서 데이터들간의 거리를 기준으로 근접한 정도를 측정하고, 이를 바탕으로 데이터를 특성별로 분류하는 HCM 클러스터링을 이용하여 데이터들의 특성을 파악한다.

주어진 데이터 분류는 먼저 학습 데이터를 분류하고, 분류된 학습 데이터의 중심에 의해 테스트 데이터를 분류한다. HCM 클러스터링에 의한 학습 데이터 분류는 다음과 같다.

〔단계 1〕 클러스터의 개수 ($2 \le c \le n$)를 결정하고, 소속 행렬 $U$를 $U^{(0)} \in M_c$으로 초기화

$$M_c = \left\{ U \mid u_{ij} \in \{0,1\},\ \sum_{i=1}^{c} u_{ik} = 1,\ 0 < \sum_{k=1}^{n} u_{ik} < n \right\} \quad (6)$$

〔단계 2〕 클러스터의 중심벡터를 구한다.

$$v_{ij} = \frac{\sum_{k=1}^{n} u_{ik} \cdot x_{kj}}{\sum_{k=1}^{n} u_{ik}} \quad (7)$$

〔단계 3〕 각각의 클러스터 중심과 데이터와의 거리를 계산하여 새로운 소속행렬 $U^{(r)}$을 생성

$$d_{ik} = d(x_k - V_i) = \| x_k - V_i \| = \left[ \sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right]^{1/2} \quad (8)$$

$$u_{ik}^{(r+1)} = \begin{cases} 1 & d_{ik}^{(r)} = \min\{d_{jk}^{(r)}\}\ \text{for all } j \in c \\ 0 & otherwise \end{cases} \quad (9)$$

〔단계 4〕 만일 식 (7)을 만족하면 종료하고, 그렇지 않으면 $r = r+1$로 놓고 〔단계 2〕로 간다.

$$\| U^{(r+1)} - U^{(r)} \| \le \varepsilon\ (\text{tolerance level}) \quad (10)$$

### 2.3 유전자 알고리즘

유전자 알고리즘[3,4]은 자연 선택과 유전학에 기반으로 하는 확률적인 탐색방법으로써 탐색과 해의 가능영역들을 균형있게 이용하기 위하여 생산, 교차, 돌연변이의 과정을 수행하는 일반성 있는 탐색법으로 비선형 최적화 이론에 탁월한 성능을 발휘하고 있다. 기존의 다른 탐색방법들은 탐색공간에서 최적값을 찾기도 전에 지역극소(local minimum)에 빠질 위험이 있지만 유전자 알고리즘은 해가 될 가능성이 있는 개체집단을 유지하면서 그들 모두가 동시에 최적값을 찾아나가기 때문에 지역 극소에 빠질 위험을 어느 정도 해결할 수 있다는 것이다. 또한 모델의 성능지수가 최소가 되는 전역 극소 영역을 찾는 능력을 갖고 있으며, 기존의 방법들과는 달리 선형, 연속, 미분 가능 등의 제한이 없기 때문에 다양한 분야에 별다른 제한 없이 적용할 수 있다는 장점을 가진다는 것이 중요한 특징이다. 이러한 장점들을 이용하여 전반부 파라미터를 최적으로 동정하는데 유전자 알고리즘을 사용한다.

### 2.4 하중 값을 가지는 목적 함수

학습 데이터와 테스트 데이터를 고려한 퍼지 모델 성능 다시 말해서 근사화 능력과 예측 능력 모두를 고려하기 의해 목적 함수[5,6]를 다음과 같이 정의한다.

$$f(PI, E\_PI) = \theta \times PI + (1-\theta) \times E\_PI \quad (11)$$

$\theta$는 $PI$와 $E_PI$에 대한 하중 값이다. $PI$는 학습 데이터에 대한 성능지수를, $E_PI$는 테스트 데이터에 대한 성능지수이다. 이 목적 함수는 $\theta$의 선택에 따라 퍼지 모델의 근사화와 일반화 사이에서 최적화에 대한 방향을 설정할 수 있는 특징을 가진다.

## 3. 시뮬레이션 및 결과 고찰

### 3.1 가스로 공정

제안된 다중 퍼지 모델을 Box와 Jenkins가 사용한 가스로 시계열 데이터[7]를 이용하여, 입출력 데이터인 가스 흐름율과 연소된 이산화탄소 농도의 가스로 공정을 퍼지 모델링한다.

본 논문에서는 수치적 실험에 이용되는 성능지수로 식 (7)을 사용한다.

$$PI = \frac{1}{N} \sum_{i=1}^{N} (y_i - y_i^*)^2 \quad (12)$$

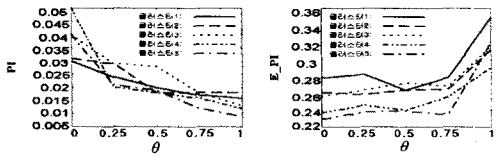
표 1은 단일 퍼지 모델과 본 논문에서 제안한 다중 퍼지 모델의 성능지수를 비교하여 나타내고 있다.

그림 2는 표 1에서 나타난 단일 퍼지와 다중 퍼지의 각 클러스터 개수와 하중 계수 $\theta$에 따른 학습 및 테스트 입출력 데이터군에 의한 각 모델의 성능지수를 그래프로 표시한 것이다.

표 1. 클러스터 개수에 따른 성능지수(구조 2)

| $\theta$ | 단일 Fuzzy | | 다중 Fuzzy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 클러스터 2 | | 클러스터 3 | | 클러스터 4 | | 클러스터 5 | | | |
| | PI | E_PI | PI | E_PI | PI | E_PI | PI | E_PI | PI | E_PI |
| 0.0 | 0.031 | 0.278 | 0.030 | 0.261 | 0.040 | 0.256 | 0.050 | 0.237 | 0.040 | 0.229 |
| 0.25 | 0.029 | 0.283 | 0.024 | 0.259 | 0.029 | 0.264 | 0.020 | 0.246 | 0.021 | 0.238 |
| 0.5 | 0.018 | 0.264 | 0.020 | 0.264 | 0.028 | 0.272 | 0.018 | 0.240 | 0.019 | 0.239 |
| 0.75 | 0.018 | 0.279 | 0.017 | 0.265 | 0.018 | 0.269 | 0.016 | 0.256 | 0.012 | 0.235 |
| 1.0 | 0.018 | 0.350 | 0.016 | 0.312 | 0.013 | 0.316 | 0.012 | 0.291 | 0.009 | 0.320 |

그림에서 볼 수 있듯이 학습 데이터의 경우(PI)와 테스트 데이터의 경우(E_PI) 모두 클러스터 5의 성능지수가 가장 좋음을 알 수 있다.




(a) PI (b) E_PI

그림 3. 하중 계수 $\theta$에 따른 다중 퍼지의 성능지수

표 2에서는 기존의 모델과 성능지수를 비교하였다. 표에 나타난 바와 같이 다중 퍼지 모델(구조 2의 클러스터 5)이 기존의 다른 모델들보다 성능이 우수함이 입증된다.

표 2. 기존 모델과 성능지수 비교

| 모델 | 추론법 | 규칙 | PI | E_PI |
|---|---|---|---|---|
| Oh's model [5,6] | 간략 | 4 | 0.024 | 0.328 |
| | 선형 | 4 | 0.022 | 0.326 |
| | | 6 | 0.021 | 0.364 |
| HCM+Hybrid[12] | 간략 | 6 | 0.022 | 0.332 |
| | 선형 | 6 | 0.017 | 0.269 |
| Max-Min[11] | 간략 | 6 | 0.022 | 0.336 |
| | 선형 | 6 | 0.021 | 0.362 |
| Our model | 간략 | 4 | 0.017 | 0.307 |
| | 선형 | 4 | 0.009 | 0.320 |

### 3.2 가스 터빈 발전소에서의 NOx 배출 공정

제안한 다중 퍼지 모델의 성능을 평가하기 위해서 가스 터빈 발전소의 데이터[8]를 이용한다. 모델의 입력은 Tamb, COM, LPT, Pcd, Texh이고 출력은 NOx이다. NOx의 성능 평가를 위해서 식 (12)을 이용한다.

표 3은 후반부 구조 1의 성능지수를 나타낸다. 또한 단일 퍼지 모델과 다중 퍼지 모델을 비교하고 있다. 다중 Fuzzy에서 클러스터 4의 성능이 우수함을 알 수 있다.

표 3. 클러스터 개수에 따른 성능지수(구조 1)

| 모델 | 클러스터 | PI |
|---|---|---|
| 단일 Fuzzy | 1 | 0.8284 |
| 다중 Fuzzy | 2 | 0.2984 |
| | 3 | 0.0765 |
| | 4 | 0.0043 |

표 4는 기존의 모델과 여기서 제안한 다중 퍼지 모델(구조 1의 클러스터 4)과의 성능지수가 비교되고 있다. HCM 클러스터링 알고리즘과 유전자 알고리즘을 사용한 다중 퍼지 모델이 기존 모델보다 월등히 우수함을 알 수 있다.

표 4.기존 모델[9]과 성능지수 비교

| 모델 | 신경망 | FNNs | GMDH | 퍼지 추론 | Our model |
|---|---|---|---|---|---|
| PI | 1773.3453 | 0.0520 | 0.2484 | 2.2172 | 0.0043 |

## 4.결 론

본 논문에서는 HCM 클러스터링 알고리즘과 유전자 알고리즘을 이용하여 보다 효과적으로 최적의 다중 퍼지 모델을 동정하였다. HCM 클러스터링을 이용하여 입출력 데이터의 특징을 찾아 모델 설계에 반영하였고, 유전자 알고리즘이 가지고 있는 우수한 성질을 이용하여 전반부 멤버쉽 함수의 파라미터 값을 최적으로 동정하였다. 시뮬레이션에서 볼 수 있듯이 제안된 다중 퍼지 모델은 복잡하고 비선형성이 강한 공정에서 기존의 모델들보다 성능이 향상된 모델을 설계할 수 있었다. 전문가와 시행착오에 의존하지 않고 체계적인 방법에 의해 객관적인 모델을 설계할 수 있음을 보였다. 또한 하중 값을 가지는 목적 함수에 의해 다중 퍼지 모델의 근사화와 일반화 사이에서 최적화에 대한 방안을 제시함으로써 모델의 근사화와 일반화 사이에 상호 연계를 통해 최적화 향상을 위한 방향을 제시하였다.


**감사의 글**

이 논문은 1998년도 한국학술진흥재단의 연구비에 의하여 지원되었음(KRF-98-001-01048).


**[참 고 문 헌]**


[1] L. A Zadeh, "Fuzzy sets", Inf. Control 8, 338-353, 1965.

[2] M.A. Ismail, "Soft Clustering Algorithm and Validity of Solutions", Fuzzy Computing Theory, Hardware and Applications, edited by M.M. Gupta, North Holland, pp.445-471, 1988.

[3] David E. Goldberg, "Genetic Algorithms in search, Optimization & Machine Learning", Addison-wesley.

[4] Zbigniwe Michalewicz, "Genetic Algorithms + Data Structure = Evolution Programs", Springer-Verlag.

[5] S.K. Oh and W. Pedrycz, "Identification of Fuzzy Systems by means of an Auto-tuning Algorithm and Its Application to Nonlinear Systems", Fuzzy Sets and Syst., Vol. 115, issue 2, pp.205-230, Jul 2000.

[6] Sung Kwun Oh, "Fuzzy Identification by Means of an Auto-tuning Algorithm and a Weighted Performance Index", Journal of Fuzzy Logic and Intelligent Systems, Vol. 8, No. 6, pp.106-118, 1998

[7] Box and Jenkins, "Time Series Analysis, Forcasting and Control", Holden Day, SanFrancisco, CA.

[8] G. Vachtsevanos, V. Ramani and T. W. Hwang, "Prediction of Gas Turbine NOx Emissions using Polynomial Neural Network" Technical Report, Georgia Institute of Technology, Atlanta, 1995.

[9] 안태천, 오성권, "발전소의 대기오염물질 배출패턴 모델정립(최종보고서)", 기초전력공학 공동연구소, 과제관리번호 : 96-지-08, 1997년 8월.

[10] 오성권, 윤기찬, 김현기, "유전자 알고리즘과 합성 성능지수에 의한 퍼지-뉴럴 네트워크 구조의 최적 설계", 제어·자동화·시스템공학회, 제 6권, 제 3호, pp. 273-283, 2000년 3월.

[11] 박병준, 오성권, 안태천, 김현기, "유전자 알고리즘과 하중 값을 이용한 퍼지시스템의 최적화", 대한전기학회 논문지, 제 48권, 제 6호, pp.789-799, 1999년 6월.

[12] 박병준, 윤기찬, 오성권, 장성환, "클러스터링 및 하이브리드 알고리즘을 이용한 퍼지모델의 최적화", 대한전기학회 논문지, 제 G권, pp2908-2910, 1999년 7월.