계층적 분류구조의 퍼지시스템 설계 및 시계열 예측 응용

Design of Fuzzy System with Hierarchical Classifying Structures and its Application to Time Series Prediction

방영근* · 이철희**

Young-Keun Bang* and Chul-Heui Lee**

* 강원대학교 대학원 전기전자공학과

** 강원대학교 IT특성화학부대학 전기전자공학부

요 약

시스템의 동작특성을 표현하는 퍼지 규칙들은 퍼지 클러스터링 기법에 매우 의존적이다. 만약, 클러스터링 기법의 분류 능력이 개선된다면, 그들에 의해 생성되는 퍼지 규칙과 식별되는 파라미터들이 보다 정밀해 질 수 있으므로 시스템의 성능이 개선될 수 있다. 따라서 본 논문에서는 분류능력이 강화된 새로운 계층 구조 클러스터링 알고리즘을 제안한다. 제안된 클러스터링 기법은 데이터 사이의 통계적 특성과 상관성을 고려하여 보다 정확하게 데이터들을 분류할 수 있도록 2개의 클러스터의 구조를 갖는다. 또한, 본 논문은 차분 데이터를 이용하여 원형 데이터의 패턴이나 규칙들이 명확하게 반영될 수 있도록 하며, 각각의 차분 데이터들의 다양한 특성을 고려할 수 있도록 다중 퍼지 시스템을 구현한다. 마지막으로, 제안된 기법들의 유효성을 다양한 비선형 시계열 데이터들의 예측을 통해 검증한다.

키워드 : 분류, 계층적 클러스터링, 상관성, 차분데이터, 다중 퍼지 시스템

Abstract

Fuzzy rules, which represent the behavior of their system, are sensitive to fuzzy clustering techniques. If the classification abilities of such clustering techniques are improved, their systems can work for the purpose more accurately because the capabilities of the fuzzy rules and parameters are enhanced by the clustering techniques. Thus, this paper proposes a new hierarchically structured clustering algorithm that can enhance the classification abilities. The proposed clustering technique consists of two clusters based on correlation and statistical characteristics between data, which can perform classification more accurately. In addition, this paper uses difference data sets to reflect the patterns and regularities of the original data clearly, and constructs multiple fuzzy systems to consider various characteristics of the differences suitably. To verify effectiveness of the proposed techniques, this paper applies the constructed fuzzy systems to the field of time series prediction, and performs prediction for nonlinear time series examples.

Key Words : classification, hierarchical clustering, correlation, difference data, multiple fuzzy systems

1. 서 론

데이터 처리 기술들은 현대사회의 고도화에 따라 더욱 복잡한 구조를 가지게 되며, 이는 일반적으로 처리되어야 할 데이터들이 자연현상에 기인하는 강한 비선형적 특성을 보이기 때문이다. 따라서 이러한 데이터들의 비선형적 특성들을 효과적으로 분석할 수 있다면, 그들의 처리 기술들에 대한 구조적인 복잡성이나 많은 양의 데이터를 필요로 하지 않아도 될 것이다. 데이터 이면에 내재된 다양한 비선형적 특성들은 데이터가 가지는 일련의 패턴들이나 규칙성으로 대표될 수 있으며, 이러한 패턴들을 찾기 위해 많이 사용되는 기법이 퍼지 클러스터링 기법이다. 이러한 퍼지 클러스터링 기법들은 유사한 데이터들을 클러스터링 하고 또한,

데이터들이 클러스터에 소속되는 정도의 애매함을 퍼지 모델로 구현 가능하게 함으로써 비선형 데이터를 다루는 분야에 광범위하게 적용되어 왔다 [1-3]. 하지만, 강한 비선형성을 갖는 데이터들을 효과적으로 분류하고, 그들의 특성을 충분히 고려할 수 있는 시스템을 설계할 경우, 시스템의 구조적 복잡성이 초래되며, 이는 퍼지 시스템의 모델링 분야에 많은 제약점들을 야기시킨다. 따라서 본 논문에서는 비선형 데이터를 처리하는 퍼지 시스템의 모델링에 있어, 구조적 복잡성을 피하면서 효과적인 클러스터링 기법을 제안하며, 이에 따라 생성되는 퍼지 규칙과 추정되는 파라미터들의 적합성을 개선함으로써 우수한 성능의 퍼지 시스템을 모델링하는 기법을 제안하였다. 또한, 비선형 데이터를 보다 명확히 시스템에 반영시키기 위해 데이터의 원형이 아닌 그들의 차분 데이터를 활용하는 기법을 제안하였다. 제안된 퍼지 시스템은 먼저 통계적 특성이 원형 데이터 보단 안정된 차분 데이터들을 생성하고 [4-6] 상관 분석을 통해 원형 데이터의 특성을 잘 드러낼 수 있는 차분 데이터 후보군을 선별하게 된다. 또한, TSK 퍼지 모델을 이용하여 차분들의

접수일자 : 2009년 8월 5일

완료일자 : 2009년 10월 5일

본 과제(결과물)는 지식경제부의 지원으로 수행한 에너지 자원 인력 양성사업의 연구결과입니다.

특성을 최대한 고려할 수 있도록 다중 퍼지 예측 시스템을 구현하고, 선별된 차분 데이터들을 그들에 상응하는 다중 퍼지 시스템의 각각의 입력으로 사용하였다. 다중 퍼지 시스템 내에서 퍼지 규칙생성을 위한 퍼지 분할에는 계층 구조 클러스터링 알고리즘(HSCA : Hierarchically structured clustering algorithm)을 적용하여 시스템의 정밀성을 높일 수 있도록 하였다. 본 논문에서 제안된 계층 구조 클러스터링 알고리즘은 크게 상위 층의 클러스터와 그 클러스터들 내의 하위 층의 퍼지 집합으로 구성되며, 상위 층의 클러스터에 적용된 분류 기법으로는 데이터들 상호간의 상관성 (correlation) 을 기반으로 하는 교차상관 클러스터링 알고리즘(cross-correlation clustering algorithm)을, 하위 층의 퍼지집합에는 일반적으로 널리 쓰이는 K-평균 클러스터링 알고리즘(k-means clustering algorithm)을 적용하여 구조적 복잡성을 간소화 할 수 있도록 하였다. 따라서 퍼지 규칙 생성을 위한 퍼지 분할과 분할된 퍼지 집합에 분류된 데이터들은 그들의 상관성과 통계적 특성이 동시에 고려될 수 있기 때문에 규칙의 증가나 구조적 복잡성을 피하면서도 TSK 퍼지 모델의 전반부 규칙들과 후반부에 추정되는 파라미터들이 데이터들에 대하여 높은 적합성을 반영할 수 있도록 함으로써 시스템의 성능이 강화될 수 있도록 하였다. 마지막으로 제안된 퍼지 시스템의 성능을 검증하기 위해 호주의 전력생산량 데이터와 Mackey-Glass 시계열 데이터를 이용하여 예측분야에 적용하였으며, 시뮬레이션 결과를 통해 제안된 시스템의 효용성을 증명 하였다.

2. 제안된 TSK 다중 퍼지 시스템의 구조

원형 데이터로부터 생성된 차분 데이터들은 원형 데이터 보단 평균적으로 안정된 정보를 제공할 수 있으므로 비선형 시스템 모델링 분야에 자주 언급 된다. 하지만 생성된 모든 차분 데이터들이 원형 데이터의 특성을 잘 반영할 수 있는 것은 아니며, 모든 차분을 사용할 경우 시스템의 구조를 더욱 복잡하게 만들 것이다. 따라서 일부의 차분 데이터를 추출하여 사용하여야 하지만 추출된 차분데이터들이 다른 차분 데이터들 보다 원형 데이터의 특성을 잘 반영할 수 있는지는 경험에 기반 될 수밖에 없다. 따라서 이러한 문제의 해결 방법으로 본 논문에서는 데이터의 전처리 과정을 통해 일부의 최적 차분 데이터들만을 추출할 수 있도록 하여 사용하였으며, 각각의 차분 데이터들의 특성을 적절히 반영하기 위해 다중 퍼지 시스템을 구현하였다. 또한, 서론에서도 언급되었듯이 각각의 퍼지 시스템에는 계층구조 클러스터링 알고리즘을 적용하여 TSK 퍼지 모델의 전반부 규칙과 후반부 파라미터가 데이터에 최적화 될 수 있도록 하였다. 아래의 그림 1은 제안된 시스템의 전체 순서도를 보여 준다. 그림 1을 살펴보면, 제안된 시스템은 먼저 적절히 정의된 훈련데이터를 이용하여, 그 원형 데이터의 특징을 시스템에 잘 반영시키기 위한 1차 차분 처리 과정(단원 3:최적 차분 데이터의 생성)과, 그렇게 처리된 각각의 차분 데이터로 다중 모델 퍼지 시스템을 구현하는 과정(단원 4:다중 퍼지시스템 설계)을 거치게 된다. 다중 모델 퍼지 시스템의 구현에서는 그들의 입력으로 사용되는 1차 처리된 차분 데이터들을 상관성에 따른 분류기법 (상관 클러스터링)을 통하여 상위 층의 클러스터로 2차 처리 과정을 거치게 되고, 이렇게 처리된 데이터들은 또 다시 통계적 특성에 따른 분류 기법 (k-means 클러스터링)을 통해 하위 층의 퍼지 집합

로 다시 3차 처리 되는 구조(단원 4.1:계층구조 클러스터링)를 가지고 있다. 따라서 시스템의 동작 특성을 결정짓는 규칙의 생성을 위한 퍼지집합에 분류된 데이터들은 서로간의 연관성이 매우 클 것이며, 이는 보다 데이터의 특성에 적합한 규칙의 생성(단원 4.2:퍼지 규칙 생성)과 보다 정밀한 파라미터 추정(단원 4.3:파라미터 추정)을 가능하게 할 것이다. 또한, 마지막으로 훈련데이터를 이용하여 가장 성능이 우수한 하나의 시스템을 선택(단원 5:동작 시스템 선택)하여 동작하게 함으로써 효과적인 시스템의 운용을 가능케 할 수 있는 구조로 이루어져 있다.

그림 1. 제안된 시스템의 전체 순서도
Fig. 1. The total flow chart of the proposed systems

3. 최적 차분 데이터의 생성

시스템의 설계에 사용할 최적 차분 데이터를 선별하기 위하여 먼저, 주어진 시계열 데이터 중 예측 모델을 구성하는데 어려움이 없을 정도의 적정 길이의 데이터를 훈련 데이터로 선정한 후 다음의 과정을 통해 최적 차분 후보군을 선별한다.

1 단계) 차분 간격 값에 따라 원형 데이터와 차분 데이터들의 자기 상관 계수를 구한다.

$$r_j = \frac{\frac{1}{N-j} \sum_{i=1}^{N-j} (y(i) - \bar{y})(y(i+j) - \bar{y})}{\frac{1}{N} \sum_{i=1}^N (y(i) - \bar{y})^2} \quad (1)$$

여기서, N 은 훈련 데이터의 길이이고, j 는 차분 간격 값이다. 또한, $y(i)$ 는 i 번째 훈련 데이터이며, $\bar{y}$ 는 훈련 데이터의 평균이다. 또한, r 은 각각의 차분 간격에 따른 상관계수 값이다.

2단계) 구하여진 자기 상관 계수 값을 큰 것부터 작은 순서대로 나열한다.

3단계) 상관계수 값이 0이하인 차분 간격 값들을 삭제한다.

4단계) 상관계수 값이 가장 높은 순으로 5개에 상응하는 차분 간격 값을 1차적으로 최적 차분 간격 값으로 선택한다.

이는 원형 데이터의 특성을 시스템이 최소한 고려할 수 있도록 하기 위한 방법이다.

5단계) 선택된 상관값들을 제외한 나머지 상관 계수 값들에 대하여 인접한 두 상관 계수 값의 차를 계산한다.

$$CH_s = (cof_s - cof_{s+1}), \text{ where } s = [1, N-7] \quad (2)$$

여기서, cof 는 (1)에 의해 계산된 상관계수 값들이고 s 는 차 연산을 위한 총 수행 길이를 의미한다.

6단계) 차 연산 값이 가장 크게 나타는 구간의 상관 계수 값을 임계치로 정의 하고, 임계치 보다 큰 상관 계수 값에 상응하는 차분 간격 값을 2차적으로 최적 차분 간격 값으로 선택한다. 따라서 최적 차분 간격 값의 개수는 적어도 6개 이상이 될 것이며, 생성되는 차분 데이터들과 구현되는 퍼지 시스템의 수 또한 6개 이상이 될 것이다. 이러한 방법은 원형 데이터의 특성을 충분히 고려 할 수 있도록 하기 위한 것이며, 각각의 차분 데이터들의 특성들 또한 충분히 고려 하기 위한 것이다

7단계) 이렇게 선별된 차분 간격 값에 따라 각각의 차분 데이터를 다음에 의해 생성한다.

$$\begin{aligned} d_{m(i)}t_1 &= y(N) - y(N-m(i)) \\ d_{m(i)}t_2 &= y(N-1) - y(N-m(i)-1) \\ &\vdots \\ d_{m(i)}t_n &= y(N-n-1) - y(N-m(i)-n-1) \\ &\vdots \\ d_{m(i)}t_{N-m(i)} &= y(m(i)+1) - y(1) \end{aligned} \quad (3)$$

여기서, $m(i)$ 는 각각의 선택된 차분 간격 값들의 i 번째 값이며, $d_{m(i)}t_n$ 은 생성된 n 번째 차분 데이터이다. (3)에 의해 생성된 차분 데이터들은 다중 퍼지 시스템의 구현을 위해 각각 사용된다.

4. 계층 구조를 이용한 TSK 다중 퍼지 시스템 설계

본 논문에서는 비교적 적은 입력과 퍼지 분할로 적합한 언어적 규칙 기반을 구현할 수 있으며, 후반부 선행식을 통해 고전 선형 회귀 모델의 이점도 취할 수 있는 TSK 퍼지 모델을 이용하여 다중 퍼지 시스템을 설계 하였다. TSK 퍼지 모델의 언어적 규칙의 일반식은 다음과 같이 정의 된다.

$$\begin{aligned} R: & \text{ If } x_1 \text{ is } A_1 \text{ and } x_2 \text{ is } A_2 \text{ and } \dots \text{ and } x_n \text{ is } A_n \\ & \text{ Then } y = p_0 + p_1x_1 + p_2x_2 + \dots + p_nx_n \end{aligned} \quad (4)$$

TSK 퍼지 모델은 언어적 규칙을 표현하는 전반부와 규칙에 따른 동작을 제어하는 후반부로 구성되어 있다. 전반부의 규칙을 생성하기 위해서는 입력공간의 퍼지 분할이 필요하며, 규칙의 출력을 위해서는 파라미터 식별이 필요하다. 본 논문은 입력공간의 퍼지 분할을 위해 계층구조 클러스터링 기법을 적용하였으며, 파라미터 추정에는 구조가 간단하면서도 효율적인 최소 자승법 (least square method)을 사용하였다. 또한, 전반부의 언어적 규칙의 표현과 후반부의 파라미터 추정을 위한 입력데이터 쌍은 기존의 논문[5-6]에서 언급된 것처럼 4개의 연속된 차분 데이터를 하나의 입력

데이터 쌍으로 사용하였으며, 따라서 각각의 차분데이터들에 따라 생성될 수 있는 입력 데이터 쌍은 아래와 같다.

$$\begin{aligned} Sets &= \{d_{m(i)}t_{k+1}, d_{m(i)}t_{k+2}, d_{m(i)}t_{k+3}, d_{m(i)}t_k\} \\ \text{where, } k &= [1, N-m(i)-3] \end{aligned} \quad (5)$$

여기서, $N-m(i)-3$ 은 총 생성될 수 있는 입력 데이터 쌍의 개수이다. 또한 수식(5)의 $d_{m(i)}t_k$ 는 파라미터 추정을 위해 사용되는 출력 값을 의미한다. 따라서 차분 간격 $m(i)$ 에 대한 TSK 퍼지 모델의 j 번째 규칙 R_j 는 수식 (6)과 같이 다시 정의된다.

$$\begin{aligned} \text{Rule } (R_j) \\ \text{If } d_{m(i)}t_{k+1} \text{ is } A_j \text{ and } d_{m(i)}t_{k+2} \text{ is } B_j \text{ and } d_{m(i)}t_{k+3} \text{ is } C_j \\ \text{Then } d_{m(i)}t_k = p_0^j + p_1^j d_{m(i)}t_{k+1} + p_2^j d_{m(i)}t_{k+2} + p_3^j d_{m(i)}t_{k+3} \end{aligned} \quad (6)$$

4.1 계층 구조 클러스터링에 의한 퍼지 집합의 생성

기존의 다중모델을 이용한 논문[5-6]들은 우수한 예측 성능을 보였지만, 3개의 입력데이터와 5개의 퍼지집합으로 인한 많은 수의 퍼지 규칙이 생성되었고, 이로 인해 규칙의 생성이나 파라미터 추정과정의 복잡성이 야기 되었다. 본 논문에서는 최소의 퍼지 규칙을 생성하면서도 효율적으로 시스템을 설계하기 위해 계층 구조 클러스터링 알고리즘을 제시하며 이를 통해 적은 수의 퍼지규칙 생성만으로 우수한 예측이 가능할 수 있도록 하였다. 본 논문에 제시된 클러스터링 기법은 상위 층의 클러스터들과 하위 층의 퍼지 집합으로 구성된다. 각각의 상위 클러스터들은 crisp한 집합으로 구분하며, 이들 각각은 그들의 퍼지집합으로 구성된다. 상위 클러스터에 분류되는 데이터들은 각각의 클러스터 중심과 입력 데이터들의 상관성을 분석하여 상관성이 높은 클러스터 쪽으로 데이터들이 분류되도록 하였다. 따라서 각각의 상위 클러스터에 분류된 데이터들은 서로간의 상관성이 높을 것이며, 이는 상위 클러스터내의 하위 퍼지집합의 수를 최소한으로 구성하여도 적합한 규칙의 생성이 가능하도록 할 것이다. 또한, 상위 클러스터의 수를 증가하면 클러스터에 분류되는 데이터의 양이 그만큼 적어지게 되므로 이는 하위 퍼지집합의 구성에 필요한 충분한 데이터의 양을 보장하지 못하게 될 수도 있다. 따라서 본 논문은 상위 층의 클러스터의 수를 2개로 제한하며, 또한 각각의 클러스터내의 퍼지 집합의 수도 2개로 제한한다. 따라서 하나의 상위 층 클러스터로부터 생성될 수 있는 최대 생성 규칙 수는 8개 이하가 될 것이며, 하나의 퍼지 시스템에 2개의 상위 클러스터가 존재하므로 하나의 시스템은 16개 이하의 퍼지 규칙을 가지게 된다. 비록 이러한 방법에 의해 생성되는 퍼지 규칙 수는 적을 수 있으나 데이터의 2중분류 구조를 통해 하위 퍼지집합에 분류된 데이터들은 그들 간의 상관성과 통계적 특성이 모두 고려될 수 있으므로 적합한 퍼지규칙의 생성과 보다 정밀한 파라미터 추정이 가능할 것이며, 이를 통해 적은 규칙으로도 시스템의 효율적인 운영이 가능할 것이다. 아래의 그림 2는 제안된 계층 구조 클러스터링 알고리즘의 구조를 보여 준다.

상위 층의 클러스터로의 데이터 분류를 위해선 교차상관 클러스터링 알고리즘을 적용하였으며 다음과 같이 수행되었다. 먼저, 2개의 임의의 상위 클러스터 중심 $V_{m(i)}^{upper}$ 가 $[v1_{m(i)}^{upper}, v2_{m(i)}^{upper}, v3_{m(i)}^{upper}]$ 이라면, 두 개의 클러스터에 분류되는 데이터에 대한 적합도는 다음과 같이 교차상관 함수에

의해 판별된다.

그림 2. 계층 구조 클러스터링 알고리즘의 구조
Fig. 2. Structure of hierarchical clustering algorithm

$$\rho_{XV}^{upper} = \frac{C_{XV}^{upper}}{\sqrt{C_{XX}} \sqrt{C_{VV}^{upper}}} \quad (7)$$

여기서, C_{XX} 는 각각의 입력데이터 쌍의 공분산이고, C_{VV}^{upper} 는 각각의 상위 클러스터 중심 V^{upper} 의 공분산을 의미한다. 또한, $upper$ 는 상위 층 클러스터의 번지를 의미하며 C_{XV} 는 클러스터 중심과 입력쌍들 간의 교차 공분산을 의미하며, 각각의 공분산들은 아래와 같이 정의된다.

$$C_{XX} = \sum_{l=1}^3 (d_{m(i)t_l} - \bar{X}_k)^2 \quad (8)$$

$$C_{VV}^{upper} = \sum_{l=1}^3 (V_l^{upper} - \bar{V}^{upper})^2 \quad (9)$$

$$C_{XV}^{upper} = \sum_{l=1}^3 (d_{m(i)t_l} - \bar{X}_k)(V_l^{upper} - \bar{V}^{upper}) \quad (10)$$

여기서, $\bar{X}_k$ 는 입력데이터 쌍의 평균을 의미하며, $\bar{V}^{upper}$ 는 상위 클러스터 중심 값들의 평균을 의미한다. 따라서 입력 데이터쌍은 더 높은 상관성을 나타내는 클러스터 쪽으로 분류되고, 이렇게 분류된 상위 클러스터의 중심은 다음과 같은 방법에 의해 갱신된다.

$$V_{m(i)}^{upper} = \frac{1}{D^{upper}} \sum_{n=1}^{D^{upper}} V^{upper}(D) \quad (11)$$

여기서, $V^{upper}(D)$ 는 V^{upper} 에 분류된 입력 데이터 쌍들이고, D^{upper} 는 데이터의 개수이다. 이러한 중심값의 갱신은 다음과 같이 정의되는 왜곡을 만족할 때까지 반복된다.

$$V_D = \left| \frac{V_{pre} - V_{curr}}{V_{pre}} \right| < 10^{-4} \quad (12)$$

여기서, V_{pre} 는 이전의 중심 값들이며, V_{curr} 는 현재의 갱신된 중심값을 의미한다. 이렇게 상관 클러스터링에 의해 각각의 상위 클러스터에 데이터가 분류되면, 분류된 데이터 쌍들은 다시 k-means 클러스터링에 의해 하위 퍼지 집합으로 분할된다. 각각의 상위 클러스터 내의 데이터 중 최소값과 최대값 사이를 퍼지분할의 전체 영역으로 정의하고 k-means 클러스터링 방법을 적용하여 그림 2와 같이 NA, PO로 분할하였으며, 멤버십 함수로는 사다리꼴 멤버십 함수를

사용하였다. 또한, 그림 2와 같이 각각의 입력데이터 $d_{m(i)t}$ 가 퍼지 집합에 소속되는 정도는 다음과 같이 정의 된다.

$$\begin{aligned} & \text{If } cluster(1) \geq d_{m(i)t} \text{ or } cluster(2) \leq d_{m(i)t} \\ & \quad \mu_L(d_{m(i)t}) \text{ or } \mu_R(d_{m(i)t}) = 1 \\ & \text{Else} \\ & \quad \mu_L(d) = \frac{d_{m(i)t} - cluster(c-1)}{cluster(c) - cluster(c-1)} \\ & \quad \mu_R(d) = \frac{cluster(c+1) - d_{m(i)t}}{cluster(c+1) - cluster(c)} \end{aligned} \quad (13)$$

여기서, c 는 입력 데이터가 만족하는 퍼지 집합의 중심값을 의미하며, μ_L 은 그 퍼지 집합의 중심으로부터 왼쪽의 소속함수 값이며, μ_R 은 오른쪽의 소속함수를 의미한다. 다시 표현하면, 소속정도가 1인 경우는 데이터들이 퍼지집합들의 수평면상에 존재할 경우이며, ELSE 이하의 소속 값들은 각각의 데이터들이 퍼지집합의 경사면에 위치할 경우에 해당되는 것이다. 또한, 각각의 퍼지 시스템 마다 사용되는 입력 차분 데이터가 다르기 때문에, 이러한 과정은 각각의 시스템에 독립적으로 수행된다.

4.2 TSK 퍼지 시스템 규칙 기반 생성

각각의 퍼지 시스템의 규칙기반을 위한 퍼지 규칙은 상위 클러스터를 만족한 데이터들을 이용하여 하위 퍼지 집합에 의해 생성될 수 있다. 따라서 그림 2와 같이 입력데이터 쌍이 $upper$ 번째 상위 클러스터를 만족하였다면, 이 입력 데이터쌍은 다음과 같이 퍼지 규칙을 생성하게 된다.

$$\begin{aligned} R_1^{upper}: & d_{m(i)t_{k+1}} \text{ is NA and } d_{m(i)t_{k+2}} \text{ is NA and } d_{m(i)t_{k+3}} \text{ is NA} \\ R_2^{upper}: & d_{m(i)t_{k+1}} \text{ is NA and } d_{m(i)t_{k+2}} \text{ is NA and } d_{m(i)t_{k+3}} \text{ is PO} \\ R_3^{upper}: & d_{m(i)t_{k+1}} \text{ is NA and } d_{m(i)t_{k+2}} \text{ is PO and } d_{m(i)t_{k+3}} \text{ is NA} \\ R_4^{upper}: & d_{m(i)t_{k+1}} \text{ is NA and } d_{m(i)t_{k+2}} \text{ is PO and } d_{m(i)t_{k+3}} \text{ is PO} \end{aligned} \quad (14)$$

본 논문에서는 입력데이터쌍이 만족하는 규칙만을 생성하고, 규칙 생성과정에서 중복되는 규칙은 삭제하면서 퍼지 규칙을 생성하므로 하나의 퍼지 시스템의 규칙기반을 이루는 퍼지 규칙의 수는 16개 이하가 된다.

4.3 TSK 퍼지 시스템의 파라미터 식별

만약 차분간격 $m(i)$ 에 대한 i 번째 TSK 퍼지 시스템의 $upper$ 번째 상위 클러스터의 j 번째 퍼지규칙 R_j^{upper} 을 만족하는 입력 데이터쌍의 수가 n 개 이면, 이들에 의한 규칙 후반부의 선형 식은 다음과 같이 n 개의 연립 방정식으로 표현된다.

$$\begin{aligned} d_{m(i)t_k}^{(1)} &= p_0^j + p_1^j d_{m(i)t_{k+1}}^{(1)} + p_2^j d_{m(i)t_{k+2}}^{(1)} + p_3^j d_{m(i)t_{k+3}}^{(1)} \\ d_{m(i)t_k}^{(2)} &= p_0^j + p_1^j d_{m(i)t_{k+1}}^{(2)} + p_2^j d_{m(i)t_{k+2}}^{(2)} + p_3^j d_{m(i)t_{k+3}}^{(2)} \\ &\vdots \\ d_{m(i)t_k}^{(n)} &= p_0^j + p_1^j d_{m(i)t_{k+1}}^{(n)} + p_2^j d_{m(i)t_{k+2}}^{(n)} + p_3^j d_{m(i)t_{k+3}}^{(n)} \end{aligned} \quad (15)$$

이 연립방정식은 최소자승법에 의한 파라미터 추정을 위해 다음과 같이 변형 될 수 있다.

$$\begin{bmatrix} d_{m(i)}^j t_k^{(1)} \\ d_{m(i)}^j t_k^{(2)} \\ \vdots \\ d_{m(i)}^j t_k^{(n)} \end{bmatrix} = \begin{bmatrix} 1 & d_{m(i)}^j t_{k+1}^{(1)} & d_{m(i)}^j t_{k+2}^{(1)} & d_{m(i)}^j t_{k+3}^{(1)} \\ 1 & d_{m(i)}^j t_{k+1}^{(2)} & d_{m(i)}^j t_{k+2}^{(2)} & d_{m(i)}^j t_{k+3}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & d_{m(i)}^j t_{k+1}^{(n)} & d_{m(i)}^j t_{k+2}^{(n)} & d_{m(i)}^j t_{k+3}^{(n)} \end{bmatrix} \begin{bmatrix} p_0^j \\ p_1^j \\ p_2^j \\ p_3^j \end{bmatrix} \quad (16)$$

$$Y_j = D_j P_j \quad (17)$$

여기서, Y_j 는 출력벡터, D_j 는 입력벡터, P_j 는 파라미터 벡터를 의미한다. 따라서 파라미터들은 식(18)의 최소자승법을 통해 추정되게 된다.

$$\hat{P}_j = (D_j^T D_j)^{-1} D_j^T Y_j \quad (18)$$

식(18)에 의해 추정된 계수들은 식(19)과 같은 오차 파워의 합을 최소로 하는 최적해가 될 것이다.

$$E_j = (Y_j - D_j \hat{P}_j)^T (Y_j - D_j \hat{P}_j) \quad (19)$$

하나의 입력쌍은 여러개의 규칙을 만족 할 수 있으며, 따라서 하나의 입력쌍이 총 q 개의 퍼지 규칙을 만족한다면, 시스템의 전체 출력 $\hat{d}_{m(i)} y_k$ 는 입력쌍이 만족한 각각의 규칙 R_j^{upper} 의 전반부에서 결정되는 μ_j 와 각각의 출력 값 $\hat{d}_{m(i)}^j t_k$ 로부터 식(20)과 같이 가중 합으로 구할 수 있다.

$$\hat{d}_{m(i)} y_k = \frac{\sum_{j=1}^q \mu_j \hat{d}_{m(i)}^j t_k}{\sum_{j=1}^q \mu_j} \quad (20)$$

또한 식(20)의 출력값 $\hat{d}_{m(i)} y_k$ 은 원하는 예측값과 현재의 입력값 사이의 증가분을 의미하므로 최종 출력값은 다음과 같이 구해진다.

$$\hat{y}(k+p) = y(k) + \hat{d}_{m(i)} y_k \quad (21)$$

본 논문에서 p 는 1로 정의 되며, 따라서 시스템의 출력값은 현재 값에 대한 한 스텝 전방 예측값이 된다.

5. 동작 시스템 선택

구현된 다중 퍼지 시스템이 모델링 과정에서 뿐만 아니라 실제 동작 과정에서 전부 동작된다면 연산의 과정에 있어 상당한 부담이 생길 것이다. 이러한 부담을 줄이기 위해 본 논문에서는 다중 퍼지 시스템들을 훈련데이터를 통해 동작한 후, 성능평가 지수를 가장 잘 만족하는 하나의 시스템만을 실제 동작에 적용함으로써 문제를 해결할 수 있도록 하였다. 이는 훈련 구간의 평균 동작이 가장 우수하였으므로 선택된 퍼지 시스템이 가장 우수한 성능을 나타내는 것으로 간주할 수 있기 때문이다. 아래의 수식은 논문에 사용된 성능평가 지수로서 이를 최소화 하는 시스템이 선택되게 된다.

$$MSE = \frac{1}{N - m(i) - 4} \sum_{n=m(i)+4}^N (y(n) - \hat{y}(n))^2 \quad (22)$$

수식 (22)에서, 차분 간격 $m(i)$ 의 3입력 데이터에 대하여 실제로 예측이 수행된 결과 값들은 4번째 데이터부터이므로 식(22)와 같이 예측된 값들만의 평균을 이용함으로써

정확한 성능평가를 할 수 있다.

6. 컴퓨터 시뮬레이션 및 검토

제안된 시스템의 성능을 검증하기 위하여 2개의 다른 특성을 보이는 데이터를 이용하였으며, 본문에 언급되었듯이, 상위 클러스터의 수를 2개, 하위 퍼지집합의 수도 2개로 하여 최대 생성 규칙이 16개 이하가 되도록 하여, 보다 적은 규칙에서도 좋은 성능의 예측이 가능함을 보였다. 또한, 성능 평가는 다른 논문들에서 제안된 방법들과의 비교를 위해 같은 길이의 훈련구간과 예측구간을 정의하여 시뮬레이션 하였다.

6.1 시계열 데이터의 시뮬레이션

첫 번째 시뮬레이션 데이터는 호주의 분기별 전력생산량 데이터로 총 155개의 데이터 중 70개를 훈련데이터로 사용하였으며, 나머지 데이터를 성능검증을 위한 예측 데이터로 사용하였다. 아래의 그림 3은 다중 퍼지 시스템들 중 성능이 우수한 3개의 시스템의 훈련구간에서의 상위 클러스터링 결과를 보여 준다. 제안된 논문에서의 최종 출력은 하나의 예측기를 통해 이루어지지만, 아래의 결과들은 상위 클러스터의 결과들이 예측 출력에 미치는 영향을 보여주기 위한 것이다. 실제로 본 논문의 최종 결과는 훈련구간에서 선택된 차분 간격 2에 상응하는 예측기의 출력 값이다.

a) 차분간격 8의 상위 클러스터링 결과

a) Results of the upper clusters with interval 8

b) 차분간격 4의 상위 클러스터링 결과

b) Results of the upper clusters with interval 4

c) 차분간격 2의 상위 클러스터링 결과

c) Results of the upper clusters with interval 2

그림 3. 차분 간격에 따른 예측기의 상위 클러스터링 결과
Fig. 3. Results corresponding to the upper clusters of each interval

아래의 표 1은 상위 3개의 시스템의 성능을 비교한 것으로 3개의 시스템의 성능이 적은 수의 퍼지 규칙을 사용하여

서도 비교적 모두 우수한 것으로 나타났으며, 성능비교를 위한 성능 지표로는 식 (23)으로 정의되는 MRE(mean relative error)를 사용하였다.

$$MRE = \left(\frac{1}{N} \sum_{i=1}^N \frac{|y(i) - \hat{y}(i)|}{|y(i)|} \right) \times 100\% \quad (23)$$

표 1. 상위 3개의 퍼지 시스템의 성능비교

Table 1. Comparative results of the three superior fuzzy systems

평가 순위	차분 간격	상부 클러스터	하부 퍼지집합	전체 규칙	결과 MRE
1	2(selected)	2 clusters	2 sets per cluster	12규칙	1.5315
2	4			16규칙	1.5797
3	8			16규칙	1.8579

표 1을 그림 3과 비교해 볼 때, 상위 클러스터링이 가장 잘 된 그림 3의 차분 간격 2에 해당되는 퍼지 시스템의 성능이 표 1에서 가장 우수했음을 알 수 있다.

아래의 그림 4는 시스템 선택과정에서 선택된 차분간격 2에 상응하는 퍼지 시스템의 출력 값과 예측 오차를 보여준다. 그림 4의 예측 결과에서 파란색은 예측 값, 검은색은 원 데이터의 값을 의미한다.

a) 제안된 시스템의 예측 결과

a) Prediction result of the proposed system

b) 제안된 시스템의 예측 오차

b) Prediction error of the proposed system

그림 4. 호주 전력 생산량 데이터의 예측 결과

Fig. 4. Prediction results of Australian Power Production data

그림 4의 a를 살펴보면, 실제 데이터의 값과 예측 데이터의 값이 거의 중복되어 있으며, 실제 데이터와 시스템 출력 사이의 오차는 그림 b를 통해 확인 할 수 있다.

아래의 표2는 다른 논문[2][4][7][8]들과 제안된 시스템의 성능 비교표이다.

표 2. 다른 방식들과의 성능 비교

Table 2. Comparative results between our method and other methods

방법 지표	Mamdani 퍼지모델	다중 퍼지모델	Fuzzy-AR	GA-RS 방식	제안된 방식
MRE	7.8123	2.7125	3.1254	1.8100	1.5315

표2를 살펴보면, 제안된 방식이 적은 수의 퍼지규칙으로도 다른 방법들 보다 좋은 예측성능을 보임을 알 수 있다.

6.2 비선형 데이터의 시뮬레이션

두 번째 시뮬레이션 데이터는 혼돈 비선형 시계열 예측에 자주 이용되는 Mackey-Glass 시계열 데이터로 다음과 같이 정의되는 수식으로부터 발생된다.

$$\frac{dx(t)}{dt} = \frac{0.2x(t-17)}{1+x^{10}(t-17)} - 0.1x(t) \quad (24)$$

시뮬레이션에 사용된 데이터는 x(124)부터 x(1123)까지 1000개의 데이터를 추출하여 사용하였으며, 그 중에 500개의 데이터를 훈련데이터로, 나머지데이터를 성능비교를 위한 예측 데이터로 사용하였다. 아래의 그림 5는 상위 3개의 우수한 예측 성능을 보이는 퍼지 시스템에 대한 2개의 상위 클러스터링 결과 이다.

a) 차분간격 1의 상위 클러스터링 결과

a) Results of the upper clusters with interval 1

b) 차분간격 2의 상위 클러스터링 결과

b) Results of the upper clusters with interval 2

c) 차분간격 3의 상위 클러스터링 결과

c) Results of the upper clusters with interval 3

그림 5. 차분 간격에 따른 예측기의 상위 클러스터링 결과
Fig. 5. Results corresponding to the upper clusters of each interval

그림 5를 살펴보면, 상위 클러스터에 분류되는 데이터의 분포 패턴이 호주 시계열 데이터보다 좀 더 뚜렷한 경계를 기준으로 분포되어 있음을 보여주며, 이러한 결과는 상위 클러스터내의 퍼지집합들이 좀 더 데이터에 적합하게 분할 될 수 있으며, 따라서 생성되는 규칙들이나 추정되는 파라미터들의 적합성이 우수한 것이다. 이러한 결과들은 아래의 표 3에 나타나 있으며, 성능 평가 지표로는 RMSE를 사용하였다.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y(i) - \hat{y}(i))^2} \quad (25)$$

표 3. 상위 3개의 퍼지 시스템의 성능비교

Table 3. Comparative results of the three superior fuzzy systems

평가 순위	차분 간격	상부 클러스터	하부 퍼지집합	전체 규칙	결과 RMSE
1	1(selected)	2 clusters	2 sets per cluster	16규칙	0.0009
2	2			16규칙	0.0017
3	3			16규칙	0.0023

표 3의 예측 결과를 보면 3개의 퍼지 시스템이 적은 수의 규칙으로도 상당히 우수한 성능을 나타냄을 알 수 있다.

아래의 그림 6은 제안된 시스템의 최종 출력으로 예측 결과와 그 오차를 보여주며, 파란색은 예측값, 검은 색은 원시계열을 의미한다.

a) 제안된 시스템의 예측 결과

a) Prediction result of the proposed system

b) 제안된 시스템의 예측 오차

b) Prediction error of the proposed system

그림 6. Mackey-Glass 시계열 데이터의 예측 결과

Fig. 6. Prediction results of Mackey-Glass Time Series data

위의 그림 6을 살펴보면 실제값과 예측값이 앞선 시뮬레이션의 결과와 같이 거의 중복되었음을 알 수 있고, 아래의 오차 그림을 살펴봐도 오차가 거의 10e-3단위로 나타나므로 매우 정확한 예측이 이루어짐을 알 수 있다. 이러한 결과는 상관 클러스터링을 통한 1차 데이터 분류 결과가 그림 5에서 보듯이 비교적 정확히 이루어짐으로써 하위 퍼지집합으로부터 생성된 규칙기반의 적합도가 데이터에 대하여 매우 적합한 형태로 생성된 결과로 볼 수 있다. 아래의 표 4는 제안된 시스템의 성능과 다른 논문의 방식[9-10]들과의 성능비교표이다.

표 4. 다른 방식들과의 성능 비교

Table 4. Comparative results between our method and other methods

방법 지표	WANG	MCM-1	MCM-2	제안된 방식
RMSE	0.0108	0.0108	0.0105	0.0009

표 3과 표 4는 제안된 방식이 다른 방법들의 비해 비교적 적은 수의 규칙으로도 매우 우수한 성능으로 비선형 시계열 데이터를 예측하였음을 보여주고 있다.

7. 결 론

본 논문에서는 계층적 분류구조를 가지는 다중 퍼지 시스템의 설계 방법을 다루었으며, 설계된 시스템을 시계열 예측 분야에 적용하였다. 1차적으로 원형 데이터의 특성을 충분히 고려할 수 있는 시스템의 설계를 위해 원형데이터의 차분 데이터를 사용하였고 이를 이용하여 다중 퍼지 시스템을 설계하여 다양한 패턴이나 규칙성을 최대한 고려할 수 있도록 하였다. 또한, 구현되는 시스템들이 보다 데이터의 특성을 잘 반영하며 동작할 수 있도록 하기 위해 계층 구조 클러스터링 알고리즘을 적용하였다. 계층 구조 클러스터링 알고리즘은 각각의 퍼지 시스템의 규칙기반의 형태를 보다 데이터의 특성에 적합하게 생성되도록 하여 시스템의 성능에 중요한 요소인 퍼지 규칙의 수를 최소화 하면서도 우수한 성능을 보이는 시스템 설계를 가능하게 하였으며, 시계열 예측에 대한 시뮬레이션 결과 비선형 데이터의 이면에 내재된 다양한 특성들을 잘 반영할 수 있어 충분히 좋은 결과를 나타내었다. 따라서 본 논문에 제안된 방법들은 좀 더 복잡한 특성을 나타내는 다양한 비선형 데이터들을 다루는 여러 분야에서 정보의 추출이나, 비선형 시스템의 제어 등에 응용될 수 있을 것이다. 향후에는 다중 시스템의 운영을 적절히 할 수 있는 적응 시스템 선택방법이 연구되어야 할 것으로 생각된다.

참 고 문 헌

- [1] Stephen J. Redmond, Conor Heneghan, "A method for initialising the K-means clustering algorithm using kd-trees", *pattern recognition letters*, vol. 28, pp. 965-973, 2007
- [2] K.Ozawa, T.Nimura, "Fuzzy Time-Series Model of Electric Power Consumption", *IEEE Canadian conference on Electrical and Computer Engineering*, pp.1195-1198, 1999
- [3] Juhong Nie, "Nonlinear Time-Series Forecasting: A Fuzzy Neural Approach", *Neuro computing*, vol.16, pp.66-76, MacMaster University, 1997
- [4] Inteak Kim, Song-Rock Lee, "A Fuzzy Time Series Prediction Method based on Consecutive Values", *1999IEEE International Fuzzy Systems conference proceedings*, vol.2, pp.703-707, 1999
- [5] Chul-Heui Lee, Sang-Hun Yoon, "Fuzzy Nonlinear Time Series Forecasting with Data Preprocessing and Model Selection", *Journal of Telecommunications and Information*, vol.5, pp.232-238, 2001
- [6] Young-Keun Bang, Chul-Heui Lee "Fuzzy Time Series prediction with Data Preprocessing and Error Compensation Based on Correlation Analysis", *International Conference on*

Convergence and Hybrid Information Technology,
vol.2, pp.714-721, 2008

- [7] Daijin Kim, Chulhyun Kim, "Forecasting Time Series with Genetic Fuzzy Predictor Ensemble". *IEEE Trans. on Fuzzy Systems*, vol. 5, pp.523-535, 1997
- [8] 주용석, 유전알고리즘과 러프집합을 이용한 퍼지 시스템 모델링, 강원대학교 석사학위논문, 2003
- [9] L. X. Wang, J. M. Mendel, "Generating fuzzy rules from numerical data, with applications", *IEEE Trans. on Systems, Man, and Cybern*, 22 No.6, pp1414-1427, 1992
- [10] 김인택, 공창욱, "시계열 예측을 위한 퍼지 학습 알고리즘", *한국 퍼지 지능시스템 학회*, vol.7, No.3, pp. 34-42, 1997

저 자 소 개

방영근(Young-Keun Bang)

2000년 : 강원대학교(삼척) 공학사.
2003년 : 강원대학교(삼척) 공학석사
2004년~현재 : 강원대학교(춘천) 전기전자 공학과 박사과정

관심분야 : 지능제어, 최적화, 데이터마이닝
E-mail : b2y2c1@hanmail.net

이철희(Chul-Heui Lee)

1983년 : 서울대학교 공학사.
1985년 : 서울대학교 공학석사
1989년 : 서울대학교 전기공학과(공학박사)
1994년~1994년 : 미국 IONA대학 방문교수
1990년~현재 : 강원대학교 전기전자공학부 교수

관심분야 : Soft computing & Computational intelligence
(신경망, 퍼지 시스템, 유전자 알고리즘), 지능제어 및 신호처리
E-mail : chlee@kangwon.ac.kr