산업기술연구(강원대학교 산업기술연구소 논문집), 제29권 B호, 2009.
Journal of Industrial Technology, Kangwon Natl. Univ., Korea, No. 29 B, 2009.

계층 구조 클러스터링 알고리즘 설계 및 그 응용

Design of Hierarchically Structured Clustering Algorithm and its Application

방 영 근* 박 하 용** 이 철 희***
Bang, Young-Keun Park, Ha-Yong Lee, Chul-Heui

Abstract

In many cases, clustering algorithms have been used for extracting and discovering useful information from non-linear data. They have made a great effect on performances of the systems dealing with non-linear data. Thus, this paper presents a new approach called hierarchically structured clustering algorithm, and it is applied to the prediction system for non-linear time series data. The proposed hierarchically structured clustering algorithm (called HCKA: Hierarchical Cross-correlation and K-means clustering Algorithms) in which the cross-correlation and k-means clustering algorithm are combined can accept the correlationship of non-linear time series as well as statistical characteristics. First, the optimal differences of data are generated, which can suitably reveal the characteristics of non-linear time series. Second, the generated differences are classified into the upper clusters for their predictors by the cross-correlation clustering algorithm, and then each classified differences are classified again into the lower fuzzy sets by the k-means clustering algorithm. As a result, the proposed method can give an efficient classification and improve the performance.

Finally, we demonstrates the effectiveness of the proposed HCKA via typical time series examples.

키워드 : *계층구조 클러스터링, HCKA, 차분 데이터, 상관성, 비선형 시계열*
Keywords : *hierarchical clustering , HCKA, difference data, correlationship, non-linear time series*

1. 서론

현대 사회는 정보화와 고도화에 따라 좀 더 어려운 난제에 대한 연구들이 활발히 진행 중이다. 그러나 정보화와 고도화에 따라 처리되어야 할 데이터의 양도 증가 되었을 뿐만 아니라, 대부분의 데이터들은 자연현상에 기인하는 강한 비선형성을 내포함으로 많은 제약점들을 야기 시킨다. 더욱이, 비선형 데이터들을 퍼지 모델에 적용할 경우 구현되는 모델의 성능은 그들의 퍼지분할을 위한 클러스터링 기법에 매우 민감하게 반응한다. 이는 시스템의 성능과 밀접한 시스템의 퍼지 규칙들이 퍼지 클러스터링에 기반된 퍼지 집합에 의해 생성되기 때문이다. 일반적으로 퍼지 클러스터링 기법으로 가장 빈번하게 사용되는 k-means 클러스터링 기법이나 c-means 클러스터링 기법은 단지 데이터들이 가지는 통계적 특성만을 고려하기 때문에 데이터들의 이면에 내재된 다양한 특성들을 충분히 반영한다고 볼 수는 없다. 따라서 보다 우수한 성

* 강원대학교 대학원 전기전자공학과 박사과정
** 강원대학교 전기제어공학부 교수
*** 강원대학교 전기전자공학부 교수, 교신저자

능의 시스템을 구현하기 위해선 퍼지집합의 수를 증가해야 하며, 이는 결국 많은 수의 퍼지 규칙 생성에 따른 시스템의 복잡성을 초래하게 된다. [1-2] 이에 반해, 데이터의 이면에 내재된 패턴이나 경향 등 데이터의 다양한 특성들을 충분히 고려할 수 있도록 클러스터링이 된다면, 보다 적은 퍼지 규칙으로도 우수한 성능의 퍼지시스템을 구현 할 수 있을 것이다.

따라서 본 논문에서는 이러한 문제에 대한 접근 방법으로 계층구조 클러스터링 기법을 제안한다. 제안된 방법은 k-means 클러스터링 기법과 상관 클러스터링 기법을 이용하여 데이터를 2단계로 클러스터링 함으로써, 데이터의 다양한 특성이 시스템에 잘 반영될 수 있도록 하였다. 또한 시스템의 클러스터링을 위해 사용될 비선형 데이터를 1차적으로 가공하여 평균적으로 보다 안정된 차분 데이터로 생성하고 [3], 생성된 차분 데이터들 중 원형 데이터의 특성을 잘 드러내는 최적 차분 후보군을 선별하여 사용함으로써 시스템이 데이터의 특성들을 충분히 반영할 수 있도록 하였다. 또한, 최적 차분 후보군에 상응하여 구현되는 다중 모델 퍼지 예측기들은 성능평가를 통해 그 성능이 가장 우수한 하나의 예측기만을 선택하고, 선택된 예측기가 이후의 모든 예측을 수행하게 함으로써 시스템의 구조적 복잡성을 최소화 하도록 하였다. 마지막으로, 본 논문은 비선형 시계열 데이터를 이용하여 제안된 시스템이 계층 구조 클러스터링 기법을 통해 적은 수의 퍼지 규칙만으로도 우수한 예측 성능을 보임을 증명하였다.

2. 제안된 시스템의 구조

차분 데이터를 이용한 시스템의 경우, 비선형 또는 혼돈데이터의 원형을 이용하는 것 보단 데이터의 이면에 내재된 패턴이나 법칙성, 경향 등을 보다 잘 드러낼 수 있으므로 좋은 성능을 보일 수 있음이 많은 연구를 통해 증명되어 왔으며, 또한 많은 양의 차분 데이터들 중 원형 데이터와 차분 데이터들 간의 유사성을 판별하여 시스템에 적용시킬 최적 차분 후보군을 선택하고 이를 통해 다중 모델 퍼지 예측기를 구현하는 방법들이 우리의 기존 논문에서 제시된 바 있다 [4-5]. 본 논문에선 이러한 차분 후보군의 선별 과정을 보다 명확히 개선하였으며, 또한 이를 사용하는 각각의 예측기의 클러스터링 방법을 계층구조로 구현함으로써 보다 우수한 분류능력 통해, 적은 수의 퍼지 규칙만으로도 우수한 예측 특성을 가지는 시스템이 구현될 수 있도록 하였다. 아래의 그림 1은 본 논문에 사용된 전체 시스템의 순서도 이다.

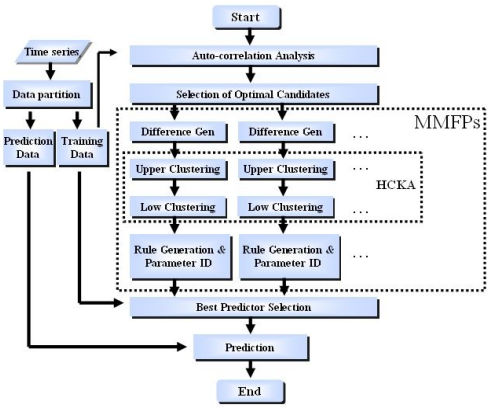


그림 1. 제안된 시스템의 전체 순서도

그림 1을 살펴보면, 제안된 시스템은 먼저 적절히 정의된 훈련데이터를 이용하여, 그 원형 데이터의 특징을 시스템에 잘 반영시키기 위한 1차 차분 처리 과정과, 그렇게 처리된 각각의 차분 데이터로 다중 모델 예측 시스템을 구현하게 된다. 다중 모델 예측 시스템에서는 그들의 입력으로 사용되는 1차 처리된 차분 데이터들을 이용하여 상관성 판별에 따른 상위 클러스터를 구현하게 되며, 분류된 상위 클러스터 내 데이터들은 그들의 통계적 특성에 따라 하위 퍼지집합으로 다시 분류되는 구조를 가진다. 또한, 다수의 예측기들은 모델선택 절차에서 성능이 우수한 하나의 예측기가 선택되도록 하였으며, 이를 통해 실제 예측에서의 구조적 복잡성을 최소화 하도록 구현 된다.

3. 비선형 데이터의 전처리

다중 모델 퍼지 예측기의 입력 및 파라미터 추정을 위해 사용될 최적 차분 후보군은 다음과 같은 과정을 통해 선별된다.

최적 차분 간격 후보군 판별 과정

Step1) 수식 1을 통한 상관 계수 계산
Step2) 높은 순으로 상관 계수 나열
Step3) 상위 5개의 상관값에 해당하는 간격 선택
Step4) 수식 3을 이용한 남은 계수의 차연산
Step5) 단계4의 가장 큰 값 이상의 상관 값 선택
Step6) 선택된 상관값에 상응하는 간격값을 선택
Step7) 선택된 간격 값들을 최적 차분 간격 후보군으로 판별.

아래의 수식은 원형 데이터와 차분 데이터간의 유사성을 판별하기 위한 자기 상관 함수 이다.

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

여기서, N는 사전 정의된 훈련데이터의 길이이고, j는 차분 간격 값이며, $y(i)$는 i번째 훈련데이터이고, $\bar{y}$는 훈련데이터의 평균이다. 또한 Step3에서 차분 후보군을 위하여 1차적으로 5개의 차분 간격 값을 선별하는 것은, 유사성이 높다고 해서 시스템에 대한 적합성까지 높다고 판별될 수 없기 때문에 차분 후보군의 최소 개수를 정의하기 위한 것이다. 그리고 아래의 수식은 Step 4의 차 연산을 위한 수식이다.

$$CH_s = (cof_s - cof_{s+1}) \quad (2)$$
$$where,\ s = [1, N-1-5(step3)-1(subtra)]$$

여기서, cof는 (1)에 의해 계산된 상관계수 값들이고 s는 차 연산을 위한 총 수행 길이를 의미한다. 또한 $5(step3)$은 step3에 의해 먼저 선택된 차분 간격 값들의 개수를 의미하며, $1(subtra)$는 구해준 계수 값들의 차를 연산 하는 것이므로 1적은 개수만큼 연산이 수행됨을 의미한다. 따라서 후보군의 개수는 적어도 6이상이 될 것이며, 이는 또한 원형 데이터의 특성을 충분히 고려 할 수 있을 것이다. 이렇게 선별된 차분 간격 후보군의 개수가 만약 m개라면, 이들에 상응하는 차분 데이터들은 (3)에 의해 생성된다.

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

(3)에 의해 생성된 차분 데이터들은 그들을 입력으로 사용하는 시스템의 규칙기반 및 파라미터 식별을 위해 사용된다.

4. 다중 예측 시스템의 구현

T-S퍼지 모델은 언어적 규칙 기반의 구현과 선형식을 이용한 수학적 판별의 용이성을 꽤할 수 있기 때문에 본 논문에서는 이를 사용하며, T-S 퍼지 모델의 일반식은 다음과 같이 정의된다.

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

일반적으로, T-S퍼지 모델은 (4)와 같이 조건부의 언어적 규칙생성을 위한 입력공간의 퍼지 분할과 규칙의 출력을 위한 파라미터 식별이 요구된다. 본 논문에서는 입력공간의 퍼지 분할을 위해 계층구조 클러스터링 기법 (HCKA)을 적용하였다. 또한 조건부와 결론부의 적절한 규칙 생성과 파라미터 추정을 위해 본 논문에서는 3개의 연속된 차분 데이터를 하나의 입력 데이터 쌍으로 사용하였다. 따라서 각각의 차분 간격값들에 상응하여 생성 가능한 입,출력 데이터 쌍은 다음과 같다.

$$Sets = \{d_{m(i)}t_{k+1},\ d_{m(i)}t_{k+2},\ d_{m(i)}t_{k+3}, \nabla_k\} \quad (5)$$
$$where,\ k = [1, N-m(i)-3]$$

여기서, $N-m(i)-3$은 총 생성될 수 있는 입출력 데이터 쌍의 개수 이다. 또한 수식(5)의 ∇_k는 $d_{m(i)}t_k$을 의미하며 파라미터 추정을 위한 출력 값을 의미한다. 따라서 차분 간격 $m(i)$의 퍼지 예측기에서, 하나의 입력 쌍에 대한 j번째 퍼지 규칙 R_j는 수식 (6)과 같이 다시 정의 된다.

$$\begin{aligned} &Rule(R_j) \\ &If\ d^j_{m(i)}t_{k+1}\ is\ A_j \text{ and } d^j_{m(i)}t_{k+2}\ is\ B_j \text{ and } d^j_{m(i)}t_{k+3}\ is\ C_j \\ &Then\ \nabla^j_k = p^j_0 + p^j_1 d^j_{m(i)}t_{k+1} + p^j_2 d^j_{m(i)}t_{k+2} + p^j_3 d^j_{m(i)}t_{k+3} \end{aligned} \quad (6)$$

4.1 계층구조 클러스터링(HCKA).

본 논문에 제안된 계층구조 클러스터링 기법은 크게 상위 클러스터로 데이터들을 1차 분류하고, 수행된 결과를 이용하여 다시 하부 퍼지집합을 생성하는 구조를 가진다. 시스템의 상부 클러스터는 교차상관 함수에 기반된 상관클러스터링 기법 (cross-correlation clustering algorithm)을 통해 생성된다. 만약 임의의 상위 클러스터 중심 $V^{upper}_{m(i)}$가 $[v1^{upper}_{m(i)}, v2^{upper}_{m(i)}, v3^{upper}_{m(i)}]$이라면, 상위 클러스터에 분류되는 데이터에 대한 적합도는 다음과 같이 교차상관 함수에 의해 판별된다.

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

여기서, C_{XX}는 각각의 입력데이터 쌍의 공분산이고, C^{upper}_{VV}는 각각의 상위 클러스터 중심 V^{upper}의 공분산을 의미한다. 그리고 C_{XV}는 중심과 입력 쌍들 간의 교차 공분산을 의미하며, 각각의 공분산들은 아래와 같이 정의된다.

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

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

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

여기서, $\overline{X_k}$는 입력데이터 쌍의 평균을 의미하며, $\overline{V_{upper}}$는 상위 클러스터 중심 값들의 평균을 의미한다. 또한 $upper$는 상위 클러스터의 위치를 의미한다. 따라서 입력 데이터 쌍은 더 높은 상관성을 나타내는 클러스터 쪽으로 분류되고, 이렇게 분류된 상부 클러스터의 중심은 다음과 같이 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 쌍들이고, D^{upper}는 데이터의 개수이다. 이러한 중심값의 갱신은 다음과 같이 정의되는 왜곡을 만족할 때 까지 반복된다.

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

여기서, V_{pre}는 이전의 중심 값들이며, V_{curr}는 현재의 갱신된 중심값을 의미한다. 아래의 그림 2는 상관 클러스터링을 통한 상부 클러스터의 한 예를 보여 준다.

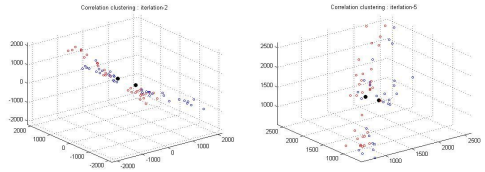


그림 2. 상부 클러스터링(cross-correlation)의 예

이렇게 상관 클러스터링에 의해 각각의 상위 클러스터에 데이터가 분류되면, 분류된 데이터 쌍들은 다시 k-means 클러스터링에 의해 하부 퍼지 집합으로 분류된다. 각각의 상위 클러스터 내의 데이터 중 최소값과 최대값 사이를 퍼지분할의 전체 영역으로 정의하고 k-means 클러스터링 방법을 적용하여 퍼지 분할하게 된다. 본 논문에서는 최소의 퍼지 규칙을 생성하는 것을 목적으로 함으로 상위 클러스터와 하위 퍼지집합의 수를 2개로 제한하였다. 따라서 하위 퍼지집합은 그림 3과 같이 NA ,PO로 정의되며, 입력 데이터의 소속정도는 는 (13)과 같이 사다리꼴 함수를 통해 얻어지게 된다.

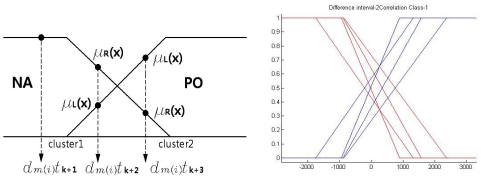


그림 3. 하위 퍼지집합(k-means)의 예

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

여기서, c는 입력 데이터가 만족하는 클러스터의 중심값을 의미하며, μ_L은 그 클러스터의 중심으로부터 왼쪽의 소속함수 값이며, μ_R은 오른쪽의 소속함수를 의미한다. 또한, 각각의 퍼지 예측기마다 사용되는 입력 차분 데이터가 다르기 때문에, 이러한 과정은 각각의 예측기에 독립적으로 수행된다.

4.2 퍼지 규칙의 생성.

각각의 퍼지 예측기의 규칙기반을 위한 퍼지 규칙은 상위 클러스터를 만족한 데이터들을 이용하여 하부 퍼지 집합에 의해 생성 될 수 있으며, 본 논문에서는 입력 데이터쌍이 만족하는 퍼지 규칙만을 생성한다. 따라서 그림 3의 좌측과 같은 입력 데이터 쌍은 (14)와 같이 규칙을 생성할 수 있으며, 하나의 입력 쌍 최대 8개의 퍼지 규칙까지 생성 가능 할 것이다.

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

정의된 상위 클러스터의 수가 2개이고, 각각에 대하여 2개의 퍼지집합이 존재하므로 3입력에 대하여 최대 생성될 수 있는 규칙의 수는 16개가 될 것이고, 입력데이터쌍이 만족하는 규칙만을 생성하므로 규칙의 수가 16개 보다 적을 수 있다. 이것은 최소 규칙만을 생성하면서도, 생성된 규칙의 적합

성을 높여 우수한 예측이 가능하도록 하기 위한 것이다.

4.3 퍼지 규칙의 파라미터 식별.

퍼지 규칙의 결정부의 선형 수식 파라미터 식별에는 그 규칙의 조건부를 만족하는 모든 입력 데이터 쌍을 이용하여 LSM(least square method)에 의해 식별된다. 따라서 차분간격 $m(i)$에 대한 i번째 T-S퍼지 예측기의 $upper$번째 상부 클러스터의 j번째 퍼지규칙 R_j^{upper}를 만족하는 입력 데이터쌍의 수가 n개 이면, 이들에 의해 그 규칙의 결정부 선형 식은 (15)와 같이 n개의 연립 방정식이 될 것이다.

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

이를 다시 벡터-행렬식으로 표현하면

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

$$Y_j = D_j P_j \tag{17}$$

여기서, Y_j는 출력벡터, D_j는 입력벡터, P_j는 파라미터 벡터를 의미한다. 따라서 파라미터들은 식(18)의 LSM을 이용하여 추정될 수 있다.

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

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

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

만약, 하나의 입력쌍이 총 q개의 퍼지 규칙을 만족한다면, 출력 $\widehat{\nabla}(t)$는 입력쌍이 만족한 각각의 규칙 R_j^{upper}의 조건부에서 결정되는 μ_j와 각각의 출력값 $\widehat{\nabla_k^j}$로부터 식(20)과 같이 가중 합으로 구할 수 있다.

$$\widehat{\nabla}(t) = \frac{\sum_{i=1}^{q} \mu_i \widehat{\nabla}_k^i}{\sum_{i=1}^{q} \mu_i} \tag{20}$$

또한 식(20)의 출력값은 실제 예측을 원하는 미래값과 현재의 입력값 사이의 증가분을 의미하므로 원하는 최종 출력값은 식(21)과 같을 것이다.

$$\hat{y}(t+p) = y(t) + \widehat{\nabla}(t) \tag{21}$$

5. 최적 예측기 선택

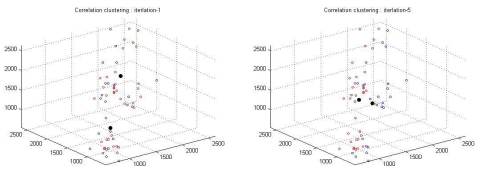
구현된 다중 퍼지 예측 시스템은 최적 차분 간격 후보군의 개수만큼의 예측기가 구현되고, 이러한 예측기 중 하나의 예측기가 선택되어 최종 예측을 수행하게 된다. 최종 예측을 수행할 예측기를 선택하는 과정에서 본 논문에서는 훈련데이터를 이용하여 식 (22)로 정의되는 MSE(mean squared error)를 평가한 후, 이를 최소화하는 예측기를 선택하여 최종 예측을 수행할 예측기로 사용하였다.

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

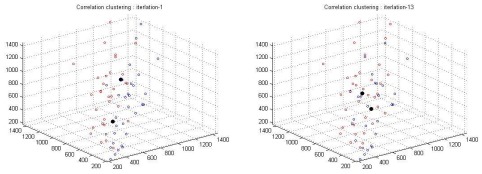
차분 간격 $m(i)$의 3입력 데이터에 대하여 실제로 예측이 수행된 결과 값들은 4번째 데이터부터 이므로 식(22)와 같이 예측된 값들만의 평균을 이용함으로써 정확한 성능평가를 할 수 있다.

6. 시뮬레이션

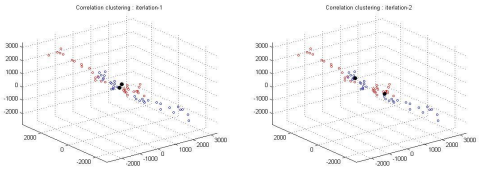
제안된 방법의 성능을 검증하기 위하여 호주 시계열 데이터를 이용하여 성능을 검증하였다. 본문에 언급되었듯이, 상부클러스터의 수를 2개, 하부 퍼지집합의 수도 2개로 하여 최대 생성 규칙이 16개 이하가 되도록 하였다. 총 155개의 데이터 중 70개를 훈련데이터로 사용하였으며, 나머지 데이터를 성능검증을 위한 예측 데이터로 사용하였으며, 아래의 그림 4는 다중 퍼지 예측기들 중 성능이 우수한 3개의 예측기의 상부 클러스터링 결과를 보여 준다.



a). 차분 간격 8의 상부 클러스터링 결과

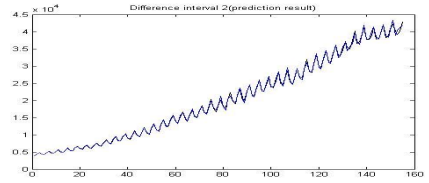


b). 차분 간격 4의 상부 클러스터링 결과

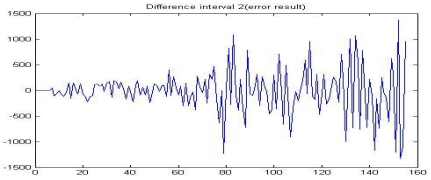


c). 차분 간격 2의 상부 클러스터링 결과
그림 4. 차분 간격에 따른 예측기의 상부 클러스터링 결과

아래의 그림 5는 예측 모델 선택에 의해 최종 선택된 차분간격 2의 예측기의 출력값과 예측 오차를 보여준다. 그림 5의 예측 결과에서 파란색은 예측 값, 검은색은 원시계열의 값을 의미한다.



a). 차분 간격 2의 예측 결과



b). 차분 간격 2의 예측 오차
그림 5. 호주 전력생산량 데이터의 예측결과

아래의 표 1은 상위 3개의 예측기의 성능을 비교한 것으로 3개의 예측기의 성능이 적은 수의 퍼지 규칙을 사용하여서도 비교적 모두 우수한 것으로 나타났다. 성능비교를 위한 성능 지표로는 식(23)으로 정의되는 MRE(mean relative error)를 사용하였다.

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

표 1 상위 3개의 예측기의 성능 비교

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

아래의 표2는 다른 논문[1][3][6][7]들과 제안된 논문의 방식과의 성능 비교표이다.

표 2 다른 방식들과의 성능 비교표

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, mamdani 퍼지 모델과 같은 경우 입출력 공간에 대한 퍼지 분할로 인해 규칙의 수가 상당히 증가하게 되며, Fuzzy-AR 방식이나 GA-RS방식은 Hybrid soft computing 기법을 적용하였으므로 제안된 기법보다 그 구조가 복잡하다. 반면에 제안된 방법은 계층 구조 클러스터링 기법을 적용하여 구조를 단순화 하였으며, 또한 적은 수의 규칙을 이용하였음에도 성능이 우수함을 알 수 있다.

7. 결론

본 논문에서는 계층구조 클러스터링 기법을 적용하여 적은 퍼지 규칙을 생성하면서도, 생성되는 규칙들의 적합성을 높여, 시스템의 복잡성을 피하면서도 효과적으로 예측을 수행할 수 있는 방법을 제안 하였다. 본 논문에 사용된 상관 클러스터링과 k-means 클러스터링이 결합된 HCKA기법은 비선형 시계열의 상관성과 통계적 특성을 계층구조 형태로 동시에 고려함으로써 원 시계열에 대한 시스템의 적합성뿐만 아니라 보다 적은 규칙으로도 우수한 예측이 가능하도록 하였다. 시뮬레이션 결과를 살펴보면, 본 논문에 사용된 HCKA가 비선형 시계열의 특성 분류에 매우 유연하게 대처할 수 있을 뿐만 아니라 이를 통해 생성된 규칙이나 추정된 파라미터들이 시계열에 내재된 특성들을 잘 반영할 수 있음을 보여 준다. 따라서 제안된 방식은 비선형 데이터에 대한 예측 분야뿐만 아니라 시스템 제어, 데이터 마이닝 등 다양한 분야로의 접근에 있어 매우 효과적일 것으로 생각된다.

참 고 문 헌

[1] K.Ozawa, T.Niimura, “Fuzzy Time-Series Model of Electric Power Consumption”, *IEEE Canadian Conference on Electrical*

and Computer Engineering, Vol. 2, pp.1195-1198, 1999.

[2] S. S. Cheng, Y. H. Chao, H. M. Wang, H. C. Fu, "A Protypes-Embedded Genetic K-means Algorithm," *ICPR. 2006. 18th international Conference on Pattern Recognition*, Vol. 2, pp. 724-727, 2006.

[3] Inteak Kim, Song-Rock Lee, "A Fuzzy Time Series Prediction Method Based on Consecutive Values", *IEEE International Conerence on Fuzzy Systems*, Vol.2, pp.703 -707, 1999.

[4] Chul-Heui Lee, Sang-Hun Yoon, "Fuzzy Nonlinear Time Series Forecasting with Data Preprocessing and Model Selection", *Joural of Telecommunications and Information*, Vol.5, pp.232-238, 2001.

[5] Y. K. Bang, C. H. 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.

[6] Dai-jin Kim, Chul-hyun Kim, "Forecasting Time Series with Genetic Fuzzy Predictor Ensemble". *IEEE Trans. on Fuzzy Systems*, Vol.5, pp.523-535, 1997.

[7] Y. S. Joo, "Fuzzy System Modeling Using Genetic Algorithm and Rough Set Theory." *M. S. thesis, Dept. Electrical and Electronics. Eng, Kangwon Univ, Chunchon, Korea,* 2003.

[8] L. X. Wang, J. M. Mendel, "Generating Fuzzy Rules from Numerical Data, with Applications", *IEEE Trans. on Systems, Man, and Cybernetics,* Vol. 22 No.6, pp1414-1427, 1992.

[9] I. T. Kim, C. W. Kong, "Fuzzy Learning Algorithms for Time Series Prediction," *Journal of Intelligence and Information systems*, Vol. 7, No. 3, pp. 34-42, 1997.