2009년도 대한전기학회 하계학술대회 논문집 2009. 7. 14 - 17

HCKA 기반 다중 모델 퍼지 예측 시스템의 구현

방영근, 심재선, 박하용, 이철희
강원대학교

Design of Multiple Model Fuzzy Prediction Systems Based on HCKA

Young-Keun Bang, Jae-Son Shim, Ha-Yong Park, Chul-Heui Lee
Kangwon National University

Abstract - 일반적으로, 퍼지 예측 시스템의 성능은 데이터의 특성과 퍼지 집합을 생성하기 위한 클러스터일 기법에 매우 의존적이다. 하지만, 예측을 위한 시계열 데이터들은 자연현상에 기인하는 강한 비선형적 특성을 가지고 있으므로 적합한 시스템을 구현하는 것에 많은 제약이 따른다. 따라서 본 논문에서는 시계열의 비선형적 특성을 적절히 취급하기 위하여, 그들로부터 생성 가능한 차분 데이터 중, 유효한 차분데이터를 이용하여 다중 모델 퍼지 예측 시스템을 구현함으로써, 보다 우수한 예측이 가능하도록 하였으며, 퍼지 시스템의 모델링에는 교차 상관분석 기법에 따른 계층적 구조의 클러스터링 기법 (Hierarchical Cross-correlation and K-means Clustering Algorithms: HCKA)을 적용하여, 시스템을 위한 규칙기반의 적합성을 높일 수 있도록 하였다.

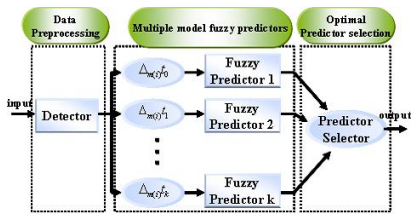
1. 서 론

시계열 분석 및 예측 기법들은 대부분 이들을 위해 사용되는 데이터들이 가지는 강한 비선형적 특성으로 인해 많은 제약점들이 따른다. 일반적으로 퍼지 모델은 비선형 데이터를 적절히 취급할 수 있는 좋은 수단이 될 수 있으나, 이들의 성능은 비선형 데이터의 취급에 있어 종종 잘못된 결과를 도출하기도 한다. 따라서, 이러한 퍼지 모델의 한계를 극복하기 위해 신경망이나 유전자 알고리즘과 같은 다양한 soft computing 기법들이 사용되며, 이들은 퍼지 모델이 갖는 한계점에 있어 유연하게 대처해 왔다 [1-2]. 하지만, 이런 Hybrid형 시스템의 경우, 그들의 성능 개선을 위해선 구조적인 복잡성이 초래된다. 따라서, 본 논문에서는 시스템의 구조적 복잡성을 피하면서도 비선형 데이터들을 적절히 취급할 수 있는 퍼지 예측 시스템을 제안한다. 먼저, 시스템 구현의 용이성을 위해, 비선형 데이터의 원형 보단 통계적 특성이 안정되어 있는 그들의 차분데이터를 이용하며, 생성 가능한 차분 데이터들 중 상관분석 기법을 이용하여 원형 데이터의 특성을 잘 드러내는 최적의 후보군을 선별한다. 그 후, 선별된 후보군들을 이용하여 다중 모델 퍼지 예측 시스템을 구현함으로써, 원형 데이터의 다양한 특성들을 고려할 수 있게 하였다. 또한, 시스템의 성능과 밀접한 클러스터링 기법에는 계층구조형태의 클러스터링 기법을 적용하여, 그들로부터 생성되는 시스템의 규칙기반의 적합성을 높일 수 있도록 하였다. 일차적으로 처리된 차분데이터들은 시스템에서 상관 클러스터링 기법에 의해 각각의 시스템의 상위 클러스터에 분류가 되며, 분류된 데이터들은 또다시 k-means클러스터링 기법에 의해 하위 퍼지 집합을 구현하게 된다. 이러한 방법에 의해 구현된 다중 모델 퍼지 예측시스템은 성능 평가를 통해 최종 예측기로 선택된 하나의 예측기를 통해 실제 예측을 수행함으로써 최상의 예측이 수행될 수 있도록 하였다.

2. 다중 모델 퍼지 예측 시스템

2.1 데이터의 전처리

아래의 그림 1은 제안된 다중 모델 퍼지 예측 시스템의 구조를 보여준다. 다중 모델 퍼지 예측 시스템을 구성하는 각각의 예측기 구현을 위해, 원형 데이터는 자기상관 분석 기법을 통해 높은 상관성을 지니는 다수의 최적 차분 후보군으로 선별되는 차분데이터들로 가공된다.



〈그림 1〉 다중 모델 퍼지 예측 시스템의 구조

수식 (1)은 최적 차분 후보군을 선별하기 위한 자기상관 함수를 보여준다.

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

여기서, N은 훈련데이터의 길이이고, j는 차분 간격 값을 의미한다, 또한, $y(i)$는 i번째 훈련데이터이며, $\bar{y}$는 훈련데이터의 평균이다. 최적 차분 후보군의 선별을 위해, 구하여진 상관 계수들은 그들의 순위에 따라 재배치되며, 순위 따라 처음 5개의 상관계수에 상응하는 차분 간격 j가 먼저 최적 차분 후보군으로 선별된다. 그 후, 남겨진 계수들 사이의 차 연산을 통해 가장 큰 폭의 변화를 보인 계수 값에서 절단하고, 절단된 계수 값의 순위 이상의 순위에 해당되는 차분 간격들이 다시 최적 차분 간격 후보군에 포함된다. 이렇게 선별된 최적 차분 후보군들이 $m(i)$개라면, 그들의 차분 데이터들은 다음과 같은 방법으로 생성된다.

$$\begin{aligned}
d_{m(i)}t_1 &= y(N)-y(N-m(i)) \\
d_{m(i)}t_2 &= y(N-1)-y(N-m(i)-1) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \\
d_{m(i)}t_n &= y(N-n-1)-y(N-m(i)-n-1) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \\
d_{m(i)}t_{N-m(i)} &= y(m(i)+1)-y(1)
\end{aligned} \tag{2}$$

2.2 퍼지 규칙 기반

아래의 그림 1은 제안된 다중 시스템을 위한 모델로 TSK퍼지 모델을 사용하였으며, TSK 퍼지모델의 일반식은 다음과 같이 언어적 규칙기반을 표현하는 조건부와 그에 따른 출력을 위한 결론부로 구성된다.

$$\begin{aligned}
R\text{: } &If \quad x_1 \text{ is } A_1 \text{ and } x_2 \text{ is } A_2 \text{ and} \cdots \text{ and } x_n \text{ is } A_n \\
&Then\ y = p_0 + p_1x_1 + p_2x_2 + \cdots + p_nx_n
\end{aligned} \tag{3}$$

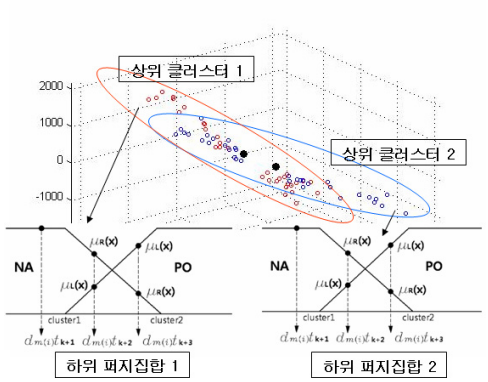
제안된 논문에서는 퍼치 규칙생성의 효율성을 위하여 생선된 차분 데이터들의 3개의 연속된 값을 하나의 입력 집합으로 사용한다. 또한, 계층적 구조 클러스터링 기법을 적용하였기 때문에 수식 (3)은 다음과 같이 수정된다.

$$\begin{aligned}
R_j^{up}\text{: } &If \ \ d_{m(i)}^j t_{k+1}^{up} \text{ is } A_1 \text{ and } d_{m(i)}^j t_{k+2}^{up} \text{ is } A_2 \text{ and } d_{m(i)}^j t_{k+3}^{up} \text{ is } A_3 \\
&Then \ \ y_j^{up} = p_o^{up} + p_1^{up} d_{m(i)}^j t_{k+1}^{up} + p_2^{up} t_{k+2}^{up} + p_3^{up} d_{m(i)}^j t_{k+3}^{up}
\end{aligned} \tag{4}$$

여기서, up은 입력데이터쌍이 포함되는 상위 클러스터(본 논문에서 2로 정의 됨)를 의미하며, R_j^{up}는 up번째 상위 클러스터에서 생성된 j번째 퍼지 규칙을 의미한다. 또한, A는 각각의 상위 클러스터에서 생성되는 퍼지집합에 대하여, 입력 데이터들 가지는 맴버쉽 값들로 삼각형 소속 함수에 의해 정의 된다.

2.3 계층 구조 클러스터링

아래의 그림 2는 본 논문에서 제안된 계층구조 클러스터링 (HCKA : hierarchical cross-correlation and k-means clustering algorithms)의 구조를 보여 준다.



〈그림 2〉 HCKA(Hierarchical cross-correlation and k-means clustering algorithms)의 구조

먼저 입력 데이터 쌍들은 임의의 중심벡터 $V_{m(i)}^{up}$와의 상관성에 의해 보다 높은 상관성을 보이는 상위 클러스터로 분류되며, 이러한 상관성의 판별을 위해 다음과 같은 교차 상관함수를 이용한다.

$$\rho^{up} = \frac{C_{XV}^{up}}{\sqrt{C_X^{up}}\sqrt{C_V^{up}}} \tag{5}$$

여기서, C_X^{up}는 각각의 상위클러스터에 분류된 입력데이터 쌍의 공분산이며, C_V^{up}는 각각의 상위클러스터 중심의 공분산을 의미한다. 또한, C_{XV}^{up} 는 이들의 교차 공분산이다. 상위 클러스터의 중심 값은 분류되는 데이터들과 중심 값과의 상관성의 크기가 더 이상 변화지 않을 때 까지 반복하여 수행된다. 또한, 본 논문에선, 상위 클러스터를 crisp 집합으로 간주한다. 따라서, 각각의 상위 클러스터에 분류된 데이터들은 서로간의 상관성이 높을 것이며, 이들에 의해 생성되는 퍼지집합 또한 더욱 명료한 형태에서 정의될 수 있을 것이다. 퍼지 집합의 생성은 k-means 클러스터링 기법을 적용하며, 퍼지집합의 수는 상위 클러스터 당 2개로 정의된다. 위와 같은 방법에 의해 각각의 예측기에서 생성되는 퍼지규칙의 수는 16개 이하로 제한될 것이다.

2.4 규칙 파라미터 추정 및 최종 예측기 선택

up번째 상위 클러스터에서 생성된 j번째 퍼지 규칙 R_j^{up}를 만족하는 입력데이터 쌍이 n개라면, 규칙의 파라미터는 다음과 같이 최소자승법(LSM : least square method)에 의해 판별될 수 있다.

$$\widehat{p_j^{up}} = (D_j^{up\,T} D_j^{up})^{-1} D_j^{up\,T} Y_j^{up} \tag{6}$$

여기서 $\widehat{p_j^{up}}$은 추정되는 파라미터 벡터를 의미하고, D_j^{up}은 규칙 R_j^{up}를 만족하는 n개의 입력데이터들의 벡터이며, Y_j^{up}은 파라미터 추정을 위한 출력 값들의 벡터로 $d_{m(i)}^{j} t_k^{up}$을 의미한다. 따라서, 예측을 위한 입력 데이터쌍이 up번째 클러스터에 분류되고, 이들이 q개의 퍼지 규칙을 만족한다면, 출력은 다음과 같이 정의된다.

$$d_{m(i)}\hat{y}(t) = \frac{\sum_{i=1}^{q} \mu_i^{up}\ \hat{y}_i^{up}}{\sum_{i=1}^{q} \mu_i^{up}} \tag{7}$$

여기서, 본 논문에서는 입력데이터 쌍으로 원시계열의 차분을 이용하였음으로 $d_{m(i)}\hat{y}(t)$는 현재와 예측하고자 하는 미래 값사이의 증가분을 의미하며, 따라서 최종 예측 값은 다음과 같이 정의된다.

$$\hat{y}(t+p) = y(t) + d_{m(i)}\hat{y}(t) \tag{8}$$

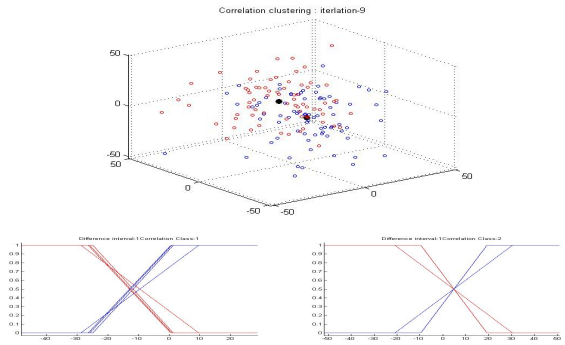
여기서, p는 예측 스텝 값으로 본 논문에서는 1로 정의된다. 마지막으로 구현된 다중 예측기들 중 다음과 같이 정의된 평균 자승 오차를 이용하여, 이를 최소화하는 예측기를 통해 최종 예측을 수행한다.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y(i) - \hat{y}(i))^2 \tag{9}$$

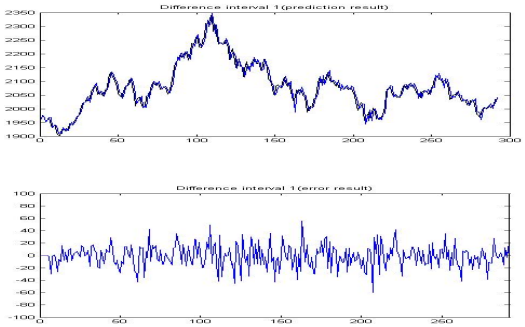
여기서, N은 성능을 검증하기 위한 샘플 값들의 길이이다.

3. 시뮬레이션 및 결과 고찰

시뮬레이션을 위하여 호주의 다우지수 데이터를 이용하였다. 총 292개의 데이터 중 150개를 훈련데이터로 이용하고, 나머지를 성능 평가를 위한 예측 데이터로 이용하여 시뮬레이션을 수행하였다. 그림 3은 상위 클러스터와 그에 상응하는 하위 퍼지집합의 모양을 보여 주고, 그림 4는 최종 예측 결과와 오차를 보여준다. 또한, 표1은 제안된 시스템의 예측 성능을 보여준다.



〈그림 3〉 다우 지수 데이터의 상위 클러스터와 하위 퍼지집합의 모양



〈그림 4〉 다우 지수 데이터의 상위 클러스터와 하위 퍼지집합의 모양

〈표1〉 제안된 퍼지 예측 시스템의 성능

지표 / Data	상위 클러스터 수	하위 퍼지집합 수	퍼지 규칙 수	MRE
다우지수	2	클러스터 당 2집합	16 규칙	0.7111

그림 4를 살펴보면 검은 색의 원시계열 값과 파란색의 예측 값들이 거의 중복되어 표시됨을 알 수 있고, 오차 또한 비교적 적은 모습을 보여준다. 또한, 표1을 살펴보면 제안된 퍼지 예측 시스템은 16개의 규칙을 사용하여 평균상대오차가 0.7111로 비교적 우수한 예측을 수행하였음을 알 수 있다.

감사의 글

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

[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 Prototypes-Embedded Genetic K-means Algorithm,” ICPR. 18th International Conference on Pattern Recognition, Vol. 2, pp. 724-727, 2006