2008년도 대한전기학회 하계학술대회 논문집 2008. 7. 16 - 18

상관해석을 기반으로 한 데이터의 전처리와 오차 보정을 갖는 퍼지 시계열 예측

방영근, 이철희
강원대학교

Fuzzy Time Series Prediction with Data Preprocessing and Error Compensation Based on Correlation Analysis

Young-Keun Bang, Chul-Heui Lee
Kangwon National University

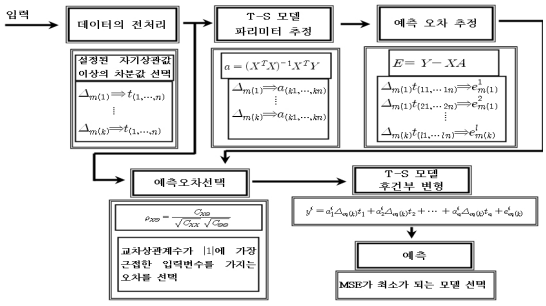
Abstract -유동적 비선형 특성을 보이는 혼돈 시계열에 대한 정확한 예측을 위해 예측 입력으로 차분 데이터를 사용하면 보다 나은 예측이 가능하다. 그러므로 본 논문에서는 상관 해석에 기반한 데이터의 전처리를 통해 적절한 최적 차분 간격 후보군을 선정하고 이들 각각에 대한 TS 퍼지 예측기로 다중 모델을 구성하여 성능 지수 평가에 의해 최적의 퍼지 예측기를 선택하여 예측을 수행하도록 하였으며, TS 퍼지 규칙 후건부에서 결정되는 예측 출력에 상관 해석에 기반한 오차 보정 메커니즘을 추가함으로써 예측 성능을 더욱 향상시킬 수 있도록 하였다.

1. 서 론

시계열 예측을 위한 시계열 분석방법에는 기존의 선형적 접근방법들의 문제를 개선하고자 최근에는 퍼지이론, 신경망, 유전 알고리즘 등과 같은 soft computing 기법들에 대한 연구가 많이 이루어지고 있으며, 그 중에서 특히 퍼지 이론은 실제 세계의 근사적이고 부정확한 성질을 표현하는데 보다 효과적이다.[1] 이러한 퍼지 모델을 구현하기 위해서, 비정상 시계열의 변화되는 통계량들을 잘 반영할 수 있도록 원시계열의 1차 차분 데이터를 사용하는 방법이 제안[2]되기도 했지만, 이러한 방법이 모든 비정상 시계열 데이터의 유동적 특성을 잘 표현해주는 것은 아니므로 그 효과와 활용이 제한적이다. 우리의 선행연구[3]에서는 이러한 문제에 대처할 수 있는 방법을 제안한 바 있으며, 본 논문에서는 [3]의 구조에 상관 해석에 기반하여 차분 간격 후보군이 적절히 선택될 수 있도록 하고, TS 퍼지 규칙 기반 생성에 k-means 알고리즘을 적용하여 전건부 입력 공간의 퍼지 분할이 데이터에 최적화될 수 있도록 하였다. 또한 예측 오차 보정 메커니즘을 추가함으로써 예측 성능을 더욱 향상시킬 수 있도록 하였다.

2. 제안된 퍼지 시계열 예측 시스템의 구조

그림 1은 제안된 알고리즘의 구조이며 다음과 같은 단계를 통하여 수행된다.



<그림 1> 제안된 알고리즘의 구조

1단계 : 데이터의 전처리에 의한 차분간격 후보군 선정
2단계 : TS 퍼지 규칙 생성 및 파라미터 식별
3단계 : 상관계수를 이용한 오차 선택 및 TS 모델 후건부 변형
4단계 : 차분간격 후보군에 대한 다중 퍼지 모델 중 MSE가 최소인 모델 선택 및 선택된 모델을 이용한 예측 수행

3. 데이터의 전처리와 모델선택

1단계의 데이터의 전처리 과정은 비정상시계열 뒤에 내재된 패턴이나 규칙성이 예측 시스템에 잘 표착될 수 있도록 하는 다수의 차분 간격 후보군을 선정하고 이를 이용하여 다중 퍼지 모델을 구현 하게 된다. 4단계에서는 이렇게 구현된 다중 퍼지 모델 중 규정된 평가 지수를 최소화하는 모델을 예측 시스템으로 선정하여 예측을 수행하게 된다.

3.1 데이터의 전처리

적당한 길이로 선정된 훈련 데이터로부터 다음과 같이 정의되는 자기상관 계수를 구한다.

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

여기서, N은 훈련 데이터의 길이이고, j는 차분 간격 값이다. 또한, $y(i)$는 i번째 훈련 데이터이며, $\bar{y}$는 훈련 데이터의 평균이다. 최적 차분 간격 후보군은 자기 상관 계수 값이 큰 것부터 작은 순서대로 나열한 뒤 인접한 두 상관 계수의 값의 차가 가장 크게 나는 상관값 이상의 데이터들을 선정한다.

아래의 식은 결정된 차분 간격의 후보군 $\{m(i)\}$에 대한 차분 값들을 구하는 방법이다.

$$\begin{aligned} \Delta_{m(i)}t_0 &= y(t)-y(t-m(i)) \\ \Delta_{m(i)}t_1 &= y(t-1)-y(t-1-m(i)) \\ &\vdots \\ \Delta_{m(i)}t_{t-m(i)-1} &= y(m(i)+1)-y(1) \end{aligned} \tag{2}$$

이렇게 생성된 각각의 차분 간격 값을 갖는 데이터들은 4장의 퍼지 예측기 설계를 위한 입력 데이터로 사용된다.

3.2 모델 선택

3.1절에 의해 구현된 다중 퍼지 모델들 중에서 훈련구간의 예측을 수행한 뒤 최소의 자승 오차 평균(Mean Square Error : MSE)을 가지는 모델을 선택하여 예측을 수행하게 된다.

$$MSE = \frac{1}{N-m(i)} \sum_{n=m(i)+1}^{N} (y(n)-\hat{y}(n))^2 \tag{3}$$

여기서, $\hat{y}(n)$은 퍼지 예측기에서 출력한 $y(n)$의 예측 값이며, 차분 간격 $m(i)$인 퍼지 예측기 모델에서는 예측에 사용되는 차분 데이터가 $N-m(i)$개이므로 위와 같이 평균이 구해진다.

4. TS 퍼지 예측기 설계

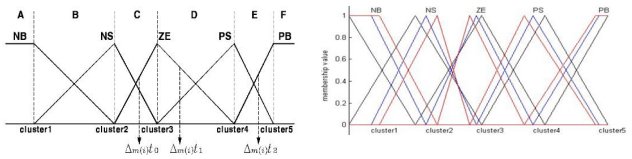
차분 간격 $m(i)$에 대해 예측이 수행되는 시간 t에서 가장 최근의 차분 데이터 $\Delta_{m(i)}t_0$, $\Delta_{m(i)}t_1$, $\Delta_{m(i)}t_2$의 3개를 입력 변수으로 하는 i번째 TS 퍼지 예측 모델의 언어적 규칙은 다음과 같은 형태가 된다.

$$\begin{aligned} R_j : \text{ if } & \Delta_{m(i)}t_0 \text{ is } A_j \text{ and } \Delta_{m(i)}t_1 \text{ is } B_j \text{ and } \Delta_{m(i)}t_2 \text{ is } C_j \\ & \text{then } \widehat{\nabla_t^j} = a_0 + a_1\Delta_{m(i)}t_0 + a_2\Delta_{m(i)}t_1 + a_3\Delta_{m(i)}t_2 \end{aligned} \tag{4}$$

여기서 후건부의 출력은 $\widehat{\nabla_t^j} = \hat{y^j}(t+p) - y(t)$로서 현재 순간의 데이터 $y(t)$와 j번째 규칙에 의해 결정되는 예측하고자 하는 p스텝 앞의 예측 값 $\hat{y^j}(t+p)$와의 차분 값이다.

4.1 입력 공간 퍼지분할 및 규칙 생성

퍼지분할을 위해, 먼저 N개의 훈련 데이터로부터 차분 간격 $m(i)$에 대해 생성된 차분 값 $\Delta_{m(i)}N_0$, …, $\Delta_{m(i)}N_{N-m(i)-1}$의 최소값과 최대값 사이를 퍼지 분할의 전체 영역(universe of discourse)으로 하고, 입력 변수 데이터에 대해 k-means 클러스터링 알고리즘을 이용하여 그림 3과 같이 NB, NS, ZE, PS, PB의 5개 퍼지 집합으로 분할하였다.



a)입력 변수에 대한 퍼지집합 b) k-means 알고리즘 동작

<그림 2> 입력 공간의 퍼지 분할과 소속함수 설정

위희 퍼지 분할과 소속함수는 다중 모델을 구성하는 k개의 퍼지 예측기 각각에 대해 독립적으로 수행된다. 그림 3과 같은 소속 함수에 대해서 입력 값 x에 대한 소속함수 값은 다음과 같이 구한다.

$$\begin{aligned} &\text{A, F 구간 : } \mu_{NB}(x) \text{ or } \mu_{PB}(x) = 1 \\ &\text{B, C, D, E 구간 : } \mu_L(x) = \frac{C_R - x}{C_R - C_L} \\ &\qquad\qquad\qquad \mu_R(x) = \frac{x - C_L}{C_R - C_L} \end{aligned} \tag{5}$$

여기서 $\mu_L(x)$는 x가 속한 구간의 좌측 클러스터 중심값 C_L을 중심값으로 하는 퍼지 집합에 대한 소속 함수 값을 나타내며, $\mu_R(x)$는 우측 클러스터 중심값 C_R을 중심값으로 하는 퍼지 집합에 대한 소속 함수 값을 나타낸다. 또한, 그림 3에 표시한 것과 같은 입력 쌍에 대해서는 다음과 같은 형태로 8개의 퍼지 규칙이 생성된다.

R_1: if $\Delta_{m(i)}t_0$ *is* *NS* and $\Delta_{m(i)}t_1$ *is* *ZE* and $\Delta_{m(i)}t_2$ *is* *PS* *then* ~
⋮ ⋮ ⋮ ⋮

모든 입력 쌍에 대해 이러한 방법으로 전건부를 생성하여 중복되는 규칙들은 제거하면 퍼지 규칙들이 생성되며, 이러한 규칙 생성 작업은 다중 퍼지 모델 각각에서 독립적으로 수행된다.

4.2 퍼지규칙 파라미터 식별

TS 퍼지모델 후건부의 선형 수식의 파라미터 식별에는 그 규칙의 전건부를 생성하는 모든 입력 쌍들을 데이터로 사용하게 되며, 최소자승법을 적용하여 파라미터를 식별한다. 차분 간격 $m(i)$에 대한 i번째 TS 퍼지 예측기의 j번째 퍼지 규칙 R_j의 생성에 기여한 n개의 입력 쌍 데이터에 대한 후건부 선형식은 다음과 같다.

$$\begin{aligned} \widehat{\nabla_{t1}^j} &= a_0^j \Delta_{m(i)} t1_0^j + a_1^j \Delta_{m(i)} t1_1^j + a_2^j \Delta_{m(i)} t1_2^j \\ \widehat{\nabla_{t2}^j} &= a_0^j \Delta_{m(i)} t2_0^j + a_1^j \Delta_{m(i)} t2_1^j + a_2^j \Delta_{m(i)} t2_2^j \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \widehat{\nabla_{tn}^j} &= a_0^j \Delta_{m(i)} tn_0^j + a_1^j \Delta_{m(i)} tn_1^j + a_2^j \Delta_{m(i)} tn_2^j \end{aligned} \tag{6}$$

이를 벡터-행렬식으로 표현하면

$$\begin{bmatrix} \widehat{\nabla_p^j}(1) \\ \widehat{\nabla_p^j}(2) \\ \vdots \\ \widehat{\nabla_p^j}(n) \end{bmatrix} = \begin{bmatrix} \Delta_{m(i)} t_0^j(1) & \Delta_{m(i)} t_1^j(1) & \Delta_{m(i)} t_2^j(1) \\ \Delta_{m(i)} t_0^j(2) & \Delta_{m(i)} t_1^j(2) & \Delta_{m(i)} t_2^j(2) \\ \vdots & \vdots & \vdots \\ \Delta_{m(i)} t_0^j(n) & \Delta_{m(i)} t_1^j(n) & \Delta_{m(i)} t_2^j(n) \end{bmatrix} \begin{bmatrix} a_0^j \\ a_1^j \\ a_2^j \end{bmatrix} \tag{7}$$

$$Y_j = X_j \Theta_j \tag{8}$$

여기서 Y_j는 출력 벡터, X_j는 입력 데이터 행렬, Θ_j는 계수 벡터이며 Θ_j는 최소 자승법을 이용하여 다음과 같이 구할 수 있다.

$$\hat{\Theta}_j = (X_j^T X_j)^{-1} X_j^T Y_j \tag{9}$$

4.3 오차 보정

입력 쌍과 규칙 R_j의 생성에 기여한 훈련 데이터들의 유사성을 판별하는 방법으로는 교차 상관 계수를 이용한 상관성 분석을 사용한다.

예측을 위한 입력 쌍 $X_t = [\Delta_{m(i)}t_0,\ \Delta_{m(i)}t_1,\ \Delta_{m(i)}t_2]$와 훈련 데이터 $T_n = [\Delta_{m(i)}n_0,\ \Delta_{m(i)}n_1,\ \Delta_{m(i)}n_2]$의 교차 상관 계수 ρ_{XT}는 다음과 같이 정의된다.

$$\rho_{XT} = \frac{C_{XT}}{\sqrt{C_{XX}}\sqrt{C_{TT}}} \tag{10}$$

여기서 C_{XT}는 X_t와 T_n의 교차 공분산, C_{XX}, C_{TT}는 각각의 공분산으로서 X_t의 평균을 $\overline{X_t}$, T_n의 평균을 $\overline{T_n}$이라고 하면 다음과 같이 계산된다.

$$C_{XX} = \sum_{l=0}^{2} (\Delta_{m(i)} t_l - \overline{x_t})^2 \qquad C_{TT} = \sum_{l=0}^{2} (\Delta_{m(i)} n_l - \overline{T_n})^2 \tag{11}$$

$$C_{XT} = \sum_{l=0}^{2} (\Delta_{m(i)} t_l - \overline{x_t})(\Delta_{m(i)} n_l - \overline{T_n}) \tag{12}$$

식(10)의 교차 상관 계수는 데이터간 상관성의 척도로 사용할 수 있으므로 가장 큰 상관값의 오차를 오차 보정 값 $\hat{e_t^j}$로 사용하여 X_t에 대한 퍼지 규칙 R_j의 후건부 출력을 다음과 같이 계산한다.

$$\widehat{\nabla_t^j} = \hat{a_0^j} \Delta_{m(i)} t_0 + \hat{a_1^j} \Delta_{m(i)} t_1 + \hat{a_2^j} \Delta_{m(i)} t_2 + \hat{e_t^j} \tag{13}$$

총 q개의 퍼지 규칙을 갖는 입력 쌍 X_t에 대한 출력 $\widehat{\nabla_t}$은 소속함수 값과 위 (16)식 이용하여 구할 수 있다.

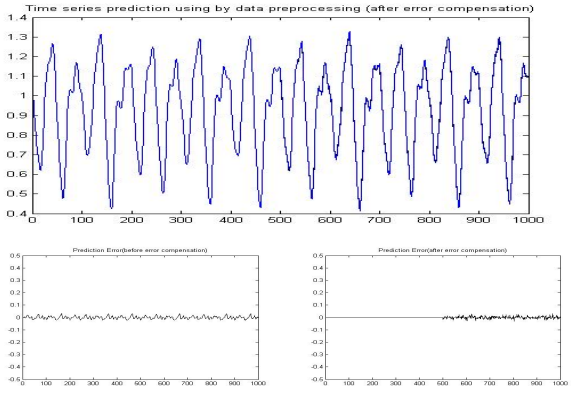
$$\widehat{\nabla}(t) = \frac{\sum_{i=1}^{q} \mu_i \widehat{\nabla_t^i}}{\sum_{i=1}^{q} \mu_i} \tag{14}$$

$\widehat{\nabla_t}$은 현재와 미래 값의 증가분이므로 최종 예측 값 $\hat{y}(t+p)$는 다음과 같이 구한다.

$$\hat{y}(t+p) = y(t) + \widehat{\nabla}(t) \tag{15}$$

5. 시뮬레이션 및 결론

시뮬레이션을 위해 mackey-glass time series를 이용하였으며 아래의 그림은 예측 결과를 나타낸다. 검은색은 원시계열 값이고 파란색은 예측된 값이다.



<그림 4> 예측 결과 및 오차(좌:원시계열, 우:제안된 방식)

예측된 결과를 살펴보면 원시계열 값과 예측값이 거의 중첩이 되었고, 또한 원시계열을 이용한 예측의 오차(좌)에 비해 제안된 방법의 오차(우)가 상당히 감소함을 알 수 있다. 이는 제안된 방법이 비정상 시계열의 예측에 우수한 성능을 나타냄을 알 수 있으며, 제안된 방법이 예측모델 구현에 유용한 방법이 될 수 있을 것으로 생각된다.

[참 고 문 헌]

[1] George E. P. Box and Gwilym M. Jenkins, Time series analysis : Forecasting and Control, Holden-Day, 1970.

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

[3] 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