

전처리과정을 갖는 시계열데이터의 퍼지예측

윤상훈, 이철희
강원대학교 전기공학과

A Fuzzy Time-Series Prediction with Preprocessing

Sang-Hun Yoon, Chul-Hee Lee
Department of Electrical Engineering, KangWon National University

Abstract - In this paper, a fuzzy prediction method is proposed for time series data having uncertainty and non-stationary characteristics.

Conventional methods, which use past data directly in prediction procedure, cannot properly handle non-stationary data whose long-term mean is floating. To cope with this problem, a data preprocessing technique utilizing the differences of original time series data is suggested.

The difference sets are established from data. And the optimal difference set is selected for input of fuzzy predictor. The proposed method based the Takagi-Sugeno-Kang(TSK or TS) fuzzy rule. Computer simulations show improved results for various time series.

1. 서 론

일반적으로 얻어지는 데이터들은 일정한 시간을 주기로 하여 얻어지는 시계열데이터가 대부분이고, 이러한 시계열데이터의 분석 및 예측은 많은 분야에서 필요로하게 된다. 주어진 시계열데이터의 생성과정에 대한 모형이 선형적인 수식으로 모델링 할 수 있다면 얻어진 초기 데이터로부터 쉽게 예측을 수행할 수 있다.

시계열데이터의 분석모형의 개발은 태양혹점의 연간변동을 예측하기 위하여 사용한 자기회귀(AR, Autoregressive)모형으로부터 시작되었으며, 그 후 반세기정도 Box-Jenkins의 ARIMA모형(1-2)에 의해 대표되어지는 선형시스템에 의한 시계열분석이 주류를 이루어왔다.

그러나 많은 경우에 데이터 생성과정에 대한 모형을 알 수 없는 uncertainty가 존재하고, 확률적으로 처리되지 않는 특성을 가지고 있으며, 비선형적 특성을 나타내는 경우가 많기 때문에 시계열데이터의 예측에 fuzzy나 neural network과 같은 soft computing 기법을 사용하는 것이 더 효과적이다. 따라서 Fuzzy AR(3), Fuzzy-Neural Approach[4-5], Fuzzy Clustering[6], Genetic Fuzzy[7], Trend Analysis[8]등과 같이 soft computing을 응용한 많은 기법들이 제안되어왔다. 그 중에서도 fuzzy system이 다른 알고리즘들보다도 자주 사용되는 이유는 훨씬 복잡한 물리적 시스템의 동작에 나타나는 uncertainty를 좀 더 쉽게 표현할 수 있기 때문이다.

본 논문에서는 TS 퍼지규칙을 기반으로 한 시계열데이터의 퍼지예측기법을 제안하였다. 제안된 방법에서는 데이터 발생 시스템에 대한 사전정보가 부족하거나, 시변특성을 갖는 경우에도 잘 동작할 수 있도록 원데이터를 전처리과정을 통하여 데이터의 특성과 유사성이 잘 드러나게 하여, 데이터에 대한 예측확도를 향상시킬 수 있도록 하였다.

예측기법에 과거의 데이터를 직접적으로 사용하는 기존의 방법들은 비정상적인 데이터를 다루기에는 적합하

지 않다. 본 논문에서는 원데이터의 차분값을 사용하는 전처리과정을 사용함으로써 이러한 문제점을 처리할 수 있도록 하였다. 데이터의 전처리 과정에서는 차분 간격이 다른 각각의 차분값들의 집합을 설정한다. 이를 중에서 가장 최적화된 차분값을 선택하여 예측을 하는 일종의 다중모델기법을 사용하였다. 퍼지규칙의 전진부에는 입력공간([차분값의 최소, 차분값의 최대])을 퍼지영역으로 분할하였고, 후진부에는 선택된 입력데이터로부터 파라미터 추정을 거친 AR 모델을 사용하는 TS 퍼지규칙을 기반으로 하였다. TS 퍼지모델은 복잡한 시스템을 간단하게 표현하기에 용이하므로, uncertainty가 존재하는 시스템의 모델을 예측하기에 적합하다.

2. 본 론

2.1 퍼지 예측시스템의 구조

일반적으로 TS fuzzy 모델은 다음과 같은 퍼지규칙의 형식을 갖는다.

$$\begin{aligned} R^i : & \text{ IF } x_1 \text{ is } A_1^i \text{ and } \cdots \text{ and } x_n \text{ is } A_n^i \\ & \text{ THEN } y^i = a_0^i + a_1^i x_1 + a_2^i x_2 + \cdots + a_n^i x_n \end{aligned} \quad (1)$$

여기서, R^i 는 i 번째 퍼지규칙을 의미하고, x_n 은 입력값을 나타내며, A_n^i 는 입력값을 퍼지영역으로 분할하여 할당된 언어적 퍼지변수이다. 후진부의 y^i 는 입력값들의 선형결합으로 구해지는 출력값이고, a_n^i 는 선형식의 파라미터값을 나타낸다.

본 논문에서, 입력값은 원데이터의 차분값을 사용하였고, 차분값에 대한 최소값과 최대값 사이를 퍼지분할영역의 전체영역(universe of discourse)으로 하고, 다음 그림에서 보여지는 바와 같이 NL , NS , ZE , BS , BL 의 5구간으로 퍼지분할을 하였다.

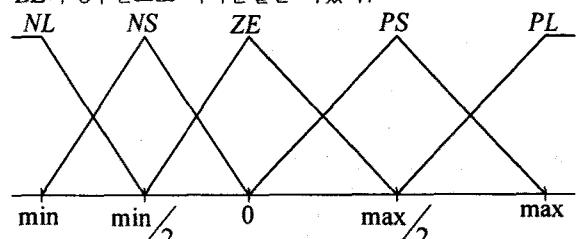


그림 1. 입력데이터의 퍼지분할

위의 그림에서 NL , NS , ZE , BS , BL 는 퍼지집합에 할당된 언어적인 값이다. 입력된 데이터는 위의 퍼지집합 중 하나 이상의 퍼지집합에 속하게 된다.

퍼지시스템의 입력으로 원데이터의 차분값을 사용하기 위하여 다음과 같이 값을 구한다.

$$\Delta_1 t_1 = x_t - x_{t-1}, \Delta_1 t_2 = x_{t-1} - x_{t-2}, \dots, \Delta_1 t_n = x_{t-n+1} - x_{t-n} \quad (2)$$

여기서, Δ_1 은 데이터의 차분을 나타내며, 데이터간의

차분 간격이 1이고, n 은 입력데이터의 개수를 나타낸다.

퍼지 예측시스템의 구조를 살펴보면 다음과 같다.

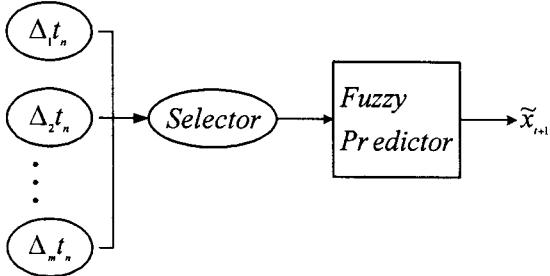


그림 2. 퍼지 예측시스템의 구조

위 그림에서 $\Delta_1 t_n, \Delta_2 t_n, \dots, \Delta_m t_n$ 은 차분값들의 집합이고, Selector는 차분 간격값이 다른 차분값의 집합 ($\Delta_1 t_n, \Delta_2 t_n, \dots, \Delta_m t_n$) 중에서 퍼지시스템의 입력으로 적합한 차분값을 선택하는 기능을 한다. \tilde{x}_{t+1} 은 한단계 미래의 예측값이다.

2.2 데이터의 전처리과정

차분값들의 집합을 설정하기 위하여 차분 간격값을 서로 다르게 만든 값들은 (2)의 식을 좀더 확장하여 생각하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \Delta_2 t_1 &= x_t - x_{t-2}, \quad \Delta_2 t_2 = x_{t-1} - x_{t-3}, \dots, \\ \Delta_2 t_n &= x_{t-n+1} - x_{t-n-1} \\ \Delta_3 t_1 &= x_t - x_{t-3}, \quad \Delta_3 t_2 = x_{t-1} - x_{t-4}, \dots, \\ \Delta_3 t_n &= x_{t-n+1} - x_{t-n-2} \\ &\vdots \quad \vdots \quad \vdots \\ \Delta_m t_1 &= x_t - x_{t-m}, \quad \Delta_m t_2 = x_{t-1} - x_{t-m-1}, \dots, \\ \Delta_m t_n &= x_{t-n+1} - x_{t-n-m+1} \end{aligned} \quad (3)$$

여기서, m 은 차분 간격이고, n 은 입력데이터의 개수를 나타낸다.

일반적으로 퍼지시스템의 퍼지규칙은 전문가의 지식이나 경험에 의해서 얻어진다. 그러나 시계열데이터에서 퍼지규칙을 생성할 때는 이전에 얻어진 데이터를 훈련데이터로 사용하여 퍼지규칙을 얻어낸다. 또한 TS 퍼지규칙의 경우에는 후건부의 파라미터값도 훈련데이터를 사용하여 얻게된다.

n 개의 입력값 ($\Delta_m t_1, \Delta_m t_2, \dots, \Delta_m t_n$)을 갖는 TS 모델은 다음과 같은 퍼지규칙을 갖는다.

$$\begin{aligned} R'_m : & \text{ IF } \Delta_m t_1 \text{ is } A_{m1}^i \text{ and } \dots \text{ and } \Delta_m t_n \text{ is } A_{mn}^i \\ & \text{ THEN } y^i[x_{t+1} - x_t] = a_{m0}^i + a_{m1}^i \Delta_m t_1^i + \dots + a_{mn}^i \Delta_m t_n^i \end{aligned} \quad (4)$$

여기서, $y^i[x_{t+1} - x_t]$ 는 i 번째 퍼지규칙에 의하여 예측된 증가분의 예측값이다. m 은 선택된 차분 간격을 나타내고, a_{mn}^i 는 i 번째 퍼지규칙의 후건부 선형식의 파라미터를 나타낸다. 파라미터를 추정하는 방법은 다음 단락에서 설명하게 될 것이다.

위의 그림 2에서 Selector의 기능에 대한 알고리즘을 살펴보면 다음과 같다.

- ① 훈련데이터와 비교데이터를 설정한다.(비교데이터는 훈련데이터 바로 직후의 데이터로 설정)
- ② 차분 간격 $m = 1$ 과 허용오차 ϵ 을 설정한다.
- ③ 훈련데이터에서 차분간격 m 인 차분값들 $\Delta_m t_n$ 으로 퍼지규칙을 생성한다.
- ④ $\Delta_m t_n$ 의 데이터로 비교데이터를 예측한다.
- ⑤ ④에서 예측한 값과 비교데이터와의 최소자승오차

(MSE)를 구한다.

- ⑥ $\text{MSE} > \epsilon$ 이면, m 을 1증가시킨 후 ②로 돌아가고, $\text{MSE} \leq \epsilon$ 이면, m 을 선택하고, ②로 돌아간다.
- ⑦ 선택된 m 값들 중에서 가장 작은 값을 선택한다.

위의 방법으로 선택된 m 차분의 차분값을 퍼지시스템에 입력하여 한단계 미래의 값을 예측하여 출력한다.

퍼지시스템의 입력값으로 원데이터를 직접 사용하는 것보다는 입력값들의 차분값을 사용하는 것이 더 좋은 결과를 보여준다는 것은 이전에 발표된 논문들[9]에서 연구되었다.

주어진 데이터가 주기나 증가추세와 같은 일련의 경향을 보이는 시계열데이터라면 원데이터를 그대로 퍼지예측의 입력으로 사용하는 것보다는 위와 같이 데이터들의 차분값을 사용하는 것이 시계열예측에 적합하다. 이런 이유는 데이터들의 차분값이 시스템의 동적특성을 표현하는데 있어 매우 유용하기 때문이다. 또한 위에서와 같이 차분값을 선택적으로 사용함으로써 고정된 하나의 차분값만을 사용하는 시스템에 비하여 더 좋은 결과를 나타내었다.

2.3 파라미터 식별

퍼지규칙의 후건부에 있는 선형식의 파라미터값을 구하기 위하여 최소자승법을 사용하였다. 먼저 후건부의 선형식을 살펴보자.

$$y^i = a_{m0}^i + a_{m1}^i \Delta_m t_1^i + a_{m2}^i \Delta_m t_2^i + \dots + a_{mn}^i \Delta_m t_n^i \quad (5)$$

위 (5)식에서 $\Delta_m t_n^i$ 의 항과 상수항을 분리하고, 행렬을 사용하여 간략화하면 다음과 같은 식으로 나타낼 수 있다.

$$y^i = \mathbf{A}x^i + b \quad (6)$$

여기서 \mathbf{A} 는 $[a_{m1}^i, a_{m2}^i, \dots, a_{mn}^i]$, x^i 는 $[\Delta_m t_1^i, \Delta_m t_2^i, \dots, \Delta_m t_n^i]$ 이고, $b = a_{m0}^i$ 를 나타낸다. 위 (6)식에서 파라미터 \mathbf{A} 와 b 를 구하는 문제는 이미 주어진 시계열데이터로부터 다음 식의 최소값을 구하는 문제와 같다.

$$\sum_{n=1}^k [y_n^i - (\mathbf{A}x_n^i + b)]^2 \quad (7)$$

여기서, k 는 시계열데이터에서 i 번째 퍼지규칙에 해당하는 데이터들의 개수를 나타낸다.

위 (7)식에서 파라미터 \mathbf{A} 와 b 에 대해서 최소화하기 위하여, 다음과 같은 식을 유도해 볼 수 있다.

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mathbf{A}} \sum_{n=1}^k [y_n^i - (\mathbf{A}x_n^i + b)]^2 \\ &= 2 \sum_{n=1}^k (y_n^i - \mathbf{A}x_n^i - b)(-x_n^i) \end{aligned} \quad (8)$$

$$\begin{aligned} 0 &= \frac{\partial}{\partial b} \sum_{n=1}^k [y_n^i - (\mathbf{A}x_n^i + b)]^2 \\ &= 2 \sum_{n=1}^k (y_n^i - \mathbf{A}x_n^i - b)(-1) \end{aligned} \quad (9)$$

위 (8), (9)의 두 식을 간단히 해보면, 다음과 같은 정규식으로 나타낼 수 있다.

$$\mathbf{A} \sum_{n=1}^k (x_n^i)^2 + b \sum_{n=1}^k x_n^i = \sum_{n=1}^k x_n^i y_n^i \quad (10)$$

$$\mathbf{A} \sum_{n=1}^k x_n^i + kb = \sum_{n=1}^k y_n^i \quad (11)$$

최종적으로 식 (10), (11)를 풀어보면, \mathbf{A} 와 b 에 대하여 다음과 같은 해를 얻을 수 있다.

$$\mathbf{A} = \frac{k \left(\sum_{n=1}^k x_n^i y_n^i \right) - \left(\sum_{n=1}^k x_n^i \right) \left(\sum_{n=1}^k y_n^i \right)}{k \left(\sum_{n=1}^k (x_n^i)^2 \right) - \left(\sum_{n=1}^k x_n^i \right)^2} \quad (12)$$

$$b = \frac{\left(\sum_{n=1}^k (x_n^i)^2 \right) \left(\sum_{n=1}^k y_n^i \right) - \left(\sum_{n=1}^k x_n^i y_n^i \right) \left(\sum_{n=1}^k x_n^i \right)}{k \left(\sum_{n=1}^k (x_n^i)^2 \right) - \left(\sum_{n=1}^k x_n^i \right)^2} \quad (13)$$

2.4 Simulation

이번 단락에서는 위에서 서술된 방법에 의하여 직접 시뮬레이션한 결과를 설명하고자 한다. 먼저, 비정상(non-stationary)적인 특성을 갖는 데이터와 주기적인 특성을 갖는 데이터를 가지고 시뮬레이션 해 보았다.

시뮬레이션을 위하여 사용된 위 두 가지 데이터의 형태는 다음과 같다.

비정상 데이터 : $y = \sin(ax) + \sin(bx) + cx$ (14)

주기적 데이터 : $y = \sin(ax) + \cos(bx)$ (15)

여기서 a , b , c 는 양의 실수이다. 위의 두 가지 데이터 모두 총 데이터 수는 500개이고, 최초의 100개의 데이터를 훈련데이터로써 사용하였고, 직후의 50개의 데이터를 비교데이터로 사용하였다. 이러한 설정이 끝나고 예측과정에서는 선택된 m 차분 초기 n 개의 데이터를 입력값으로 사용하여 전체적인 구간을 예측하였다.

시뮬레이션 결과는 아래에 나오는 그림 3, 4와 같다. 그림에서 보는 바와 같이 비정상적인 특성을 갖는 데이터와 주기적인 특성을 갖는 데이터에 대해서 정확하게 데이터를 예측해 나아가고 있음을 알 수 있다. 아래 그림에서 실선으로 나타난 것이 실제 데이터이고, 'o'으로 표시된 것이 예측된 데이터이다.

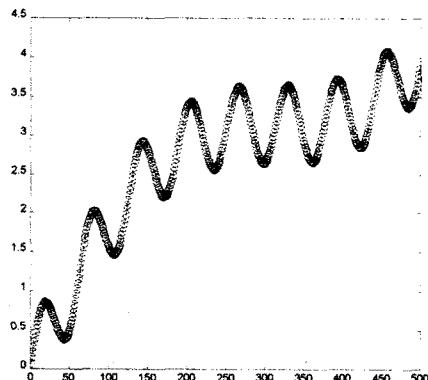


그림 3. 비정상적인 특성을 갖는 데이터의 예측

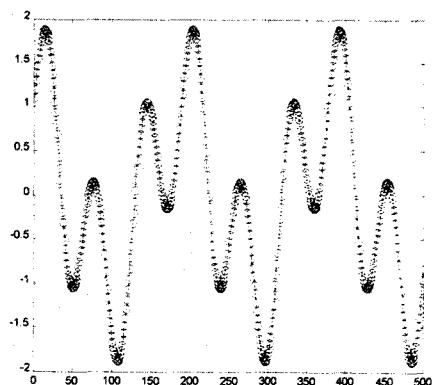


그림 4. 주기적인 특성을 갖는 데이터의 예측

그림 3의 비정상적인 특성을 갖는 데이터의 경우에는

원데이터의 차분이 6차($\Delta_6 t_n$)인 경우에 평균에러가 가장 낮은 값을 보였고, 그럼 4의 주기적 데이터의 경우에는 원데이터의 차분이 8차($\Delta_8 t_n$)인 경우에 평균에러가 가장 낮은 값을 보였다. 또한, 훈련데이터의 개수를 줄여가면서 테스트를 해보았는데, 훈련데이터의 개수가 불충분한 경우에서도 훌륭한 예측성능을 보여주었다.

위의 결과에서 보듯이 본 논문에서 제안된 선택적 퍼지시스템이 비정상적인 특성과 주기적인 특성을 갖는 데이터들을 예측하는데 있어서 적합한 시스템임을 확인할 수 있었다.

3. 결 론

위에서 살펴본 바와 같이 본 논문에서 시계열데이터의 예측에 적합한 입력에 사용할 차분 간격을 선택적으로 사용하는 퍼지시스템을 구현하고, 시뮬레이션을 해 보았다. 일반적으로 시계열데이터를 정확히 예측하는 것은 매우 어려운 작업이다. 비선행적인 특성을 갖는 물리계에서 얻어지는 데이터들이 uncertainty를 가지고 있기 때문이다.

본 논문에서처럼 차분값을 시계열데이터의 분석 및 예측에 사용하는 것은 매우 유용한 일이다. 위에서 살펴본 바와 같이 원데이터로부터 얻어진 차분값으로 입력값을 선택하는 퍼지시스템을 구현하여 적용함으로써 비선행적 물리계의 시계열데이터에 대하여 다소 정확하게 예측함을 살펴보았다.

앞으로의 연구과제는 현재의 퍼지시스템에서 퍼지규칙과 파라미터값이 훈련데이터에 의해 고정된 방식이 아니라, 시간이 흘러감에 따라서 새로이 들어온 값에 의해 퍼지규칙과 파라미터값이 생성되도록 하여 실시간적으로 값들을 보상하고, 또한 예측값을 궤환하여 차분간격값을 재선택하는 알고리즘을 개발하는 것이다.

(참 고 문 헌)

- [1] George E.P.Box, Gwilym M.Jenkins, "Time Series Analysis: forecasting and control", Revised Ed., Prentice Hall, 1976
- [2] G.Janacek, L.Swift, "Time Series forecasting, simulation, applications", Ellis Horwood, 1993
- [3] M.Kanke, K.Ozawa, T.Niimura, T.Watanabe, "Fuzzy Model-Based AR and Its Application", Proceedings of Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications, pp.374-377, 1998
- [4] Junhong Nie, "Nonlinear Time-Series Forecasting: A Fuzzy-Neural Approach", Neurocomputing vol.16 pp.63-76, MacMaster University, 1997
- [5] S.Papadakis, J.B.Theocharis, S.J.Kiartzis, A.G. Bakirtzis, "A Novel Approach to Short-term Load Forecasting Using Fuzzy Neural Networks", IEEE Trans. on Power Systems, vol.13 pp.480-492, 1999
- [6] A.B.Gev, "Non-Stationary Time-Series Prediction Using Fuzzy Clustering", North American Fuzzy Information, pp.413-417, 1999
- [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] J.F.Baldwin, T.P.Martin, J.M.Rossiter, "Time Series Modelling Using Fuzzy Trend Information", Proceedings of IIZUKA '98, pp.499-502, 1998
- [9] K.Ozawa, T.Niimura, "Fuzzy Time-Series Model of Electric Power Consumption", Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, pp.1195-1198, 1999'