2003년도 대한전기학회 하계학술대회 논문집 2003.7.21-23

개선된 PSFS를 이용한 태양흑점 시계열 데이터의 예측

김민수, 유치형, 이해수, 정찬수
숭실대 전기공학과

Prediction of Sun Spots Time Series using the Improved Parallel-Structure Fuzzy Systems

Min-Soo Kim, Chi-Hyoung You, Hae-Soo Lee, and Chan-Soo Chung
Dept. of Electrical Eng. Soongsil Univ.

Abstract - 흑점은 태양 표면에 검은 구멍처럼 보이는 것으로 흑점이 나타나면 태양활동이 활발함을 의미한다. 이러한 태양활동은 플레어나 홍염 등의 형태로 표출되어 지구의 자기장을 변동시킴으로써 전력, 통신 시스템의 장애를 유발하게 된다. 따라서 이러한 흑점 데이터를 예측함으로써 사전에 대비할 수 있도록 할 필요가 있다.
흑점 시계열 데이터의 예측에 사용된 시스템은 병렬구조를 갖는 퍼지시스템(PSFS)으로 각 퍼지시스템의 규칙은 주어진 입출력 데이터를 클러스터링하여 생성하였다. 특히, 흑점 시계열 데이터와 같이 주기성향을 갖는 데이터의 경우에도 적용가능하도록 유연한 구조를 갖는 개선된 PSFS를 제안하여 그 성능을 검증하였다.

1. 서 론

시스템에 대한 정보나 과거 특성이 주어졌을 때, 과학의 핵심적인 역할 중 하나는 미래에 대한 예측에 있다. 예측은 모델에 근거한 접근과 통계적 특성에 기초한 접근으로 나눌 수 있는데, 통계적 특성에 기초한 접근 방법은 주어진 데이터의 해석을 통해 예측하는 방법이 시계열 데이터의 예측에 이용하고 있다.
카오스 신호는 자연현상과 비슷한 특성으로 초기조건에 민감하며 장기 예측이 어려운 신호로서 카오스 신호를 정확히 예측한다는 것은 불가능하나 통계적 특성에 기초한 접근 방법에 기초하여 장기 예측 보다는 단기 예측이나 경향 분석에 중점을 두고 예측이 이루어지고 있다.
퍼지 시스템은 전문가의 경험이나 지식을 IF-THEN 형태의 언어 규칙을 이용하여 적절한 동작을 취하게 하는 시스템으로 퍼지규칙은 전문가의 경험이나 공학적 지식을 토대로 구할 수 있다.
본 논문에서는 카오스 시계열 데이터와 같은 동특성을 보여주는 태양흑점 시계열 데이터를 예측하기 위한 개선된 병렬구조 퍼지시스템 (PSFS; Parallel-Structure Fuzzy System)을 제안하였다. 각 퍼지시스템은 수치정보로 주어진 경우 주로 사용되는 Takagi와 Sugeno가 제안한 혼합추론방법을 사용하였으며, 퍼지규칙을 생성하기 위한 방법으로 데이터량이 많은 경우 그 특징 추출이 용이한 데이터 클러스터링 방법[1]을 사용하였다. 퍼지시스템의 입력은 지연시간(Time delay) τ에 따른 임베딩차원(Embedding dimension) m으로 결정되며, m은 각 τ에 따라 m을 변경하면서 구한 한단계 이후 예측의 평균제곱오차(MSE)와 최대절대오차(MAE)가 작은 상수값들로 결정하였다. 이렇게 구성된 퍼지시스템은 PSFS을 구성하게 되며 PSFS의 최종 출력은 각 퍼지시스템의 출력들의 평균값으로 다음 예측을 위한 입력으로 사용된다.

2. 시계열 데이터의 예측을 위한 병렬구조 퍼지시스템

시계열 데이터의 예측 방법은 입력 값이 예측한 값에 이용되는가에 따라 한 단계 이후 예측(One step ahead prediction)과 단기예측(Short-term prediction) 또는 장기예측(Long-term prediction)으로 구분되어진다. 한 단계 이후 예측은 식 (1)과 같이 k번째 단계에서 τ 이후 값만 예측하는 것을 의미하며 단기예측이나 장기예측은 예측된 값이 다시 입력으로 사용된다.

$$\hat{x}(k+\tau)=f[x(k),x(k-\tau),x(k-2\tau),\cdots,x(k-(m-1)\tau)] \quad (1)$$

여기에서 $f()$는 예측을 위한 시스템을, $\hat{x}()$은 예측된 값을 각각 의미한다.
$(m\times\tau)$ 개의 초기 데이터가 주어졌을 때, 첫번째 단계에서 $\hat{x}(1)$이 예측되며, 이 값은 두번째 값 $\hat{x}(2)$를 예측하기 위한 입력으로 사용된다. 초기 데이터의 수는 각 퍼지시스템이 가지는 τ와 m의 곱 중에서 최대값에 의해 결정된다.

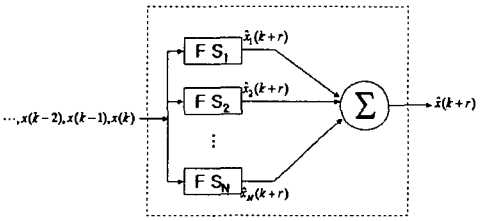

그림 1: PSFS의 구조

병렬구조 퍼지시스템이 N개의 퍼지시스템으로 구성되고 k번째까지 예측 되었을 때, $(k+1)$번째 단계에서의 입력은 $\{\hat{x}(k-(m\cdot\tau-1)),\ \hat{x}(k-(m\cdot\tau-2)),\ \cdots,\ \hat{x}(k-2),\ \hat{x}(k-1),\ \hat{x}(k)\}$이 되고 각 퍼지시스템의 출력은 $\hat{x}_1(k+1),\hat{x}_2(k+1),\cdots,\hat{x}_N(k+1)$이 된다. 그리고 최종 출력인 예측값은 $\hat{x}^*(k+1)$이 된다.
$(k+r)$번째 단계에서의 병렬구조 퍼지시스템의 구성도를 그림 1에 나타내었다. 이 때 N개의 퍼지시스템은 서로 다른 (τ,m)을 각각 가지며, 각 퍼지시스템의 입력 데이터는 τ및 m에 의해 결정된다.
예를 들면, (τ, m)이 (1,3),(2,4),(3,3)이면 퍼지시스템은 3개로 구성이 되어 있으며, 각 퍼지시스템의 입력은, 퍼지시스템1은 $\{\hat{x}(k-2),\hat{x}(k-1),\hat{x}(k)\}$을, 퍼지시스템2는 $\{\hat{x}(k-7),\hat{x}(k-5),\hat{x}(k-3),\hat{x}(k-1)\}$을, 그리고 퍼지시스템3은 $\{\hat{x}(k-8),\hat{x}(k-5),\hat{x}(k-2)\}$을 각각 입력으로 사용하게 된다. 각 퍼지시스템의 출력은 각각 $\hat{x}_1(k+1),\hat{x}_2(k+1),\hat{x}_3(k+1)$이 되어 최종 예측값을 결정하게 된다.
τ 에 따른 퍼지시스템의 m결정은 τ 가 1부터 N까지 변할 때, 각 τ 마다 m을 1부터 8까지 변화시키면서 성능지표를 비교하는 방법으로 결정하였다. 성능지표는 한 단계 이후 예측 결과값들의 MSE와 MAE를 사용하였는

데 성능지표가 작은 상수 값들을 r 에서의 최적 m으로 결정하였다. 각 r에서 결정될 m은 중복을 허용하였다.

(k+1)단계에서 PSFS의 최종 예측 값 $\hat{x}^*(k+1)$은 식 (2)와 같이 각 퍼지시스템 출력의 평균으로 결정된다.

$$\hat{x}^*(k+1)=\frac{1}{N}[\sum_{i=1}^{N}\hat{x}_i(k+1)] \quad (2)$$

단기예측이나 장기예측의 경우에는 예측된 값은 다음 단계에서 입력으로 사용되기 때문에 작은 오차도 누적되어 몇 단계를 지나면 예측이 어려워진다. 이러한 문제로 인해 여러 시스템 출력들의 평균값을 최종 출력으로 사용하였다.

3. 퍼지시스템 모델링

Takagi와 Sugeno가 제안한 혼합 퍼지추론 방법에서 i번째 퍼지규칙은 식 (3)과 같이 표현된다.

$$R_i : \text{IF } x_1 \text{ is } A_{1i} \text{ and .... and } x_n \text{ is } A_{ni}$$
$$\text{THEN } y_i=f_i(x_1,\cdots,x_n) \quad (3)$$

여기에서 $x_1, x_2,\cdots,x_n$은 n개의 입력을, A_{ji}는 퍼지 입력 변수를, y_i는 i번째 규칙의 출력을 나타내며 $f_i(\cdot)$는 입력 변수들과 상수들의 선형 함수로서 식 (4)와 (5)처럼 표현된다.

$$A_{ji}(x_j)=\exp\left(-\frac{1}{2}\cdot\left(\frac{x_j-c_{ji}}{w_{ji}}\right)^2\right) \quad (4)$$

$$f_i(x_1,\cdots,x_n)=a_{0i}+a_{1i}x_1+\cdots+a_{ni}x_n \quad (5)$$

이때 c_{ji}는 멤버쉽 함수의 중심을, w_{ji}는 멤버쉽 함수의 폭을 나타내는 변수이며 $a_{0i}, a_{1i},\cdots, a_{ni}$는 데이터에 의해 결정되어야 하는 상수이다.

n개의 입력과 M개의 규칙, 그리고 max-product 추론 방법을 사용하였을 때, i번째 전건부 규칙의 정규화한 적합도 μ_i는 식(6)과 같이 표현되며 퍼지시스템의 최종 추론값 y^*는 식 (7)과 같이 표현된다.

$$\mu_i=\frac{\prod_{j=1}^{n}A_{ji}(x_j)}{\sum_{k=1}^{M}\left(\prod_{j=1}^{n}A_{jk}(x_j)\right)} \quad (6)$$

$$y^*=\sum_{i=1}^{M}\mu_i\cdot f_i(x_1,\cdots,x_n) \quad (7)$$

퍼지규칙 생성 방법인 데이터 클러스터링 알고리즘은 Mountain 클러스터링 방법의 변형된 형태인 Subtractive 클러스터링 알고리즘을 사용하였다.

Subtractive 클러스터링 알고리즘은 데이터간의 거리의 함수로 주어지는 포텐셜 값이 최대인 점을 첫번째 클러스터 중심으로 선정하고, 첫번째 클러스터 중심의 영향을 제거한 상태에서 최대 포텐셜 값을 갖는 데이터가 다음 클러스터 중심이 된다. 입력이 n차원이고 출력이 1차원 공간을 이루고 있는 $(n+1)$차원 입출력 공간상에서 N개의 데이터 $\{X_1, X_2, X_3,\cdots, X_N\}$가 주어졌을 때 데이터 클러스터링은 다음과 같은 순서로 이루어진다.

첫째, 주어진 데이터를 [0,1]로 정규화 한다.

둘째, 데이터 간의 거리를 구한다. i번째 데이터에서 각 데이터 간의 거리는 식 (10)과 같이 함수 P로 표현되며 P_i를 포텐셜 값이라 한다.

$$P_i=\sum_{j=1}^{N}\exp\left(-\alpha\cdot\|X_i-X_j\|^2\right) \quad (8)$$

여기에서 α는 $4/r_a^2$로 주어지며, r_a는 양의 상수로서 r_a밖의 데이터는 포텐셜값에 영향을 거의 주지 못한다.

셋째, 첫번째 클러스터 중심을 구한다. N개의 포텐셜 값 중 가장 높은 값을 P_1^*라 놓고 이 때의 데이터가 첫번째 클러스터 중심 X_1^*이 된다.

넷째, 첫번째 클러스터 중심의 영향을 제거한다. 첫번째 클러스터 중심 근처에는 많은 데이터가 존재하기 때문에 그 영향을 제거하지 않으며, 두 번째 클러스터 중심 또한 첫번째 클러스터 중심 근처에 존재할 가능성이 높기 때문에 식 (9)과 같이 첫번째 클러스터 중심의 영향을 제거한 포텐셜 값을 구한다.

$$P_i'=P_i-P_1^*\cdot\exp\left(-\beta\cdot\|X_i-X_1^*\|^2\right) \quad (9)$$

여기에서 β는 $4/r_b^2$로 주어지며, r_b는 양의 상수로서 r_a보다 큰 값을 취해 클러스터 중심 근처에 다음 클러스터 중심이 나타나지 않도록 한다.

다섯째, 두 번째 클러스터 중심을 구한다. 첫번째 클러스터 중심의 영향을 제거한 N개의 포텐셜 값 P_i'중 가장 높은 값을 P_2^*라 놓고 이 때의 데이터가 두 번째 클러스터 중심 X_2^*가 된다. k번째 클러스터 중심 X_k^*가 구해졌을 때, k번째 클러스터 중심의 영향을 제거한 포텐셜 값은 식 (10)으로 표현할 수 있으며 $(k+1)$번째 클러스터 중심을 구하게 된다.

$$P_i'=P_i-P_k^*\cdot\exp\left(-\beta\cdot\|X_i-X_k^*\|^2\right) \quad (10)$$

여섯째, $d_{\min}$을 $(X_1^*, X_2^*,\cdots, X_{k-1}^*)$과 X_k^*의 거리 중에서 가장 짧은 거리로 정의하였을 때, $P_k^*\geq\bar{\varepsilon}\cdot P_1^*$이면 클러스터 중심으로 채택하고 위의 과정을 반복한다. $P_k^*\geq\underline{\varepsilon}\cdot P_1^*$이고 $\frac{d_{\min}}{r_a}+\frac{P_k^*}{P_1^*}\geq 1$이면 클러스터 중심으로 채택하고, $P_k^*>\underline{\varepsilon}\cdot P_1^*$이고 $\frac{d_{\min}}{r_a}+\frac{P_k^*}{P_1^*}<1$ 이면 그때의 X_k^*를 0으로 하고 다음으로 높은 포텐셜 값을 선정한 후, $d_{\min}$을 구하여 $\frac{d_{\min}}{r_a}+\frac{P_k^*}{P_1^*}\geq 1$이면 새로운 클러스터 중심으로 채택한다. $P_k^*<\underline{\varepsilon}\cdot P_1^*$이면 위의 반복을 중단한다. 이때 $\bar{\varepsilon}$와 $\underline{\varepsilon}$은 클러스터 중심 선정의 상하 기준이 되는 상수이다.

$$d_{\min}(k)=\min_i\{\sqrt{(\|X_i^*-X_k^*\|^2)}\}\ i=1,2,\cdots,k-1 \quad (11)$$

위와 같은 방법으로 구한 $(n+1)$ 차원인 M개의 클러스터 중심 $\{X_1^*, X_2^*,\cdots, X_M^*\}$을 입출력 공간으로 나누었을 때 X_i^*의 입력공간은 n차원의 x_i^*로, 출력공간은 1차원의 z_i^*이 된다.

여기에서 x_i^*는 퍼지규칙이 되며 퍼지 규칙에 의한 적합도 μ_i 및 최종 출력값 y^*를 구하기 위해서 z_i^*를 $G_i\cdot x+h_i$로 나누어 식 (5) 형태를 만든 후 선형최소자승법(linear least-squares estimation)알고리즘을 이용하여 변수 G_i, h_i값을 구한다. 적합도 μ_i는 식 (12)를 통해서 그리고 최종 출력값 y^*은 식 (13)을 통해 구해지며 이는 식 (6), (7)과 동일한 결과를 얻게 된다.

$$\mu_i=\exp\left(-\alpha\cdot\|x-x_i^*\|^2\right) \quad (12)$$

$$y^*=\frac{\sum_{i=1}^{M}\mu_i\cdot(G_i\cdot x+h_i)}{\sum_{i=1}^{M}\mu_i} \quad (13)$$

여기에서 x는 입력 벡터를 나타내며 α는 식 (8)에서와 동일한 값이다.

4. 시뮬레이션

4.1 태양 흑점 시계열 데이터

태양 흑점 시계열 데이터는 1749년 1월부터 2003년 4월까지 수집된 데이터로서 여러 지역에서 관측을 통해 식 (14)로 표현되었다.

$$R=k(10g+s) \quad (14)$$

여기에서 g는 흑점을 관측한 지역의 수를, s는 각 지역에서 관측된 흑점의 전체 수를 각각 나타내며, k는 1보다 작은 상수이다.

표1. 태양 흑점 시계열 데이터

Index	Year/Month	Sunspot Number (R)
1	1749/1	58.0
2	1749/2	62.6
3	1749/3	70.0
...	...	...
3050	2003/2	21.6
3051	2003/3	24.7
3052	2003/4	29.1

수집된 데이터를 표 1에 나타내었으며, 그림 2에는 태양흑점 시계열 데이터를 나타내었다.

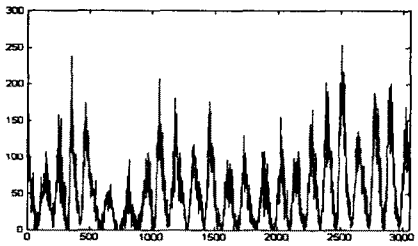

그림 2: 태양 흑점 시계열 데이터

Subtractive 클러스터링 알고리즘에 사용된 파라미터로는 r_a=0.3, r_b=0.75, $\bar{\epsilon}$=0.3, $\underline{\epsilon}$=0.1을 각각 사용하였다.

전체 3052개의 데이터 중에서 테스트에 사용될 150개의 데이터를 제외한 나머지 데이터 중 80%를 PSFS를 모델링 및 최적 m을 결정하는데 사용하였으며, 나머지 20%의 데이터는 최적 m을 결정할 때만 사용하였다.

그림 3은 최적 m을 결정하기 위한 과정으로서 τ가 12로 고정되었을 경우, m이 2에서 8까지 변할 때의 클러스터 중심의 수와 MAE 그리고 MSE를 각각 보여주고 있다.

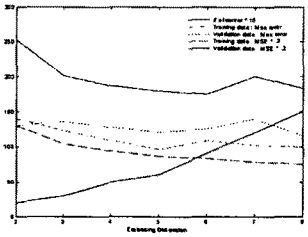

그림 3: 최적 m의 결정

그림 4에서 알 수 있듯이 τ가 12일 때, 최적 m은 3,4,5임을 알 수 있다.

4.2 개선된 PSFS을 이용한 예측

개선된 PSFS는 세 개의 퍼지시스템(N=3)을 가지며, τ가 12로 고정된 값에서 m이 3,4,5를 갖는 시스템이다. (τ,m)을 갖는 각 퍼지시스템은 클러스터링을 통해 모델링되었으며 각 퍼지시스템의 출력들의 평균을 최종 예측값으로 사용하였다.

제안한 시스템의 성능을 검증하기 위해 3,052개의 태양흑점 시계열 데이터가 사용되었다. 2,902개의 데이터 중에서 2,321개의 데이터는 퍼지시스템을 모델링하기 위해 사용되었으며, 다음 581개 데이터는 최적 m을 결정하기 위해 사용되었다. 그리고 나머지 150개의 데이터는 제안한 예측시스템을 검증하기 위해 사용된 테스트 데이터이다. 테스트 데이터인 150개 데이터를 제안한 예측시스템에 적용하여 예측한 결과를 그림 4와 같이 나타났다.

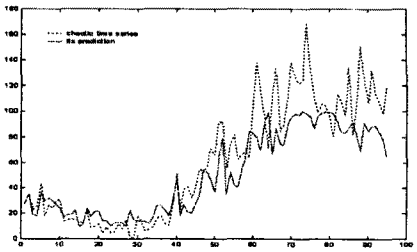

그림 4. 테스트 데이터의 예측

다음으로, 제안한 예측시스템에 2003년 5월부터 2011년 10월까지의 미래 데이터를 적용하여 예측하였고 그 결과를 그림 5에 나타내었다.

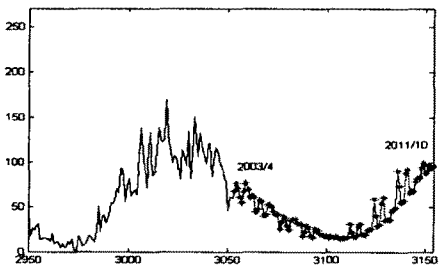

그림 5: 미래데이터의 예측

5. 결 론

본 논문에서는 태양 흑점 시계열 데이터의 예측을 위한 개선된 병렬구조 퍼지시스템(PSFS)에 대해 다루었다. PSFS는 여러 개의 퍼지시스템들을 병렬 연결형태로 구성되어 있으며, 각 퍼지시스템은 입출력 데이터 형태로 재구성된 시계열 데이터를 클러스터링함으로써 모델링되었다. 각 퍼지시스템의 출력값들은 최대, 최소값을 제외한 나머지 값들의 평균을 최종 출력인 예측값으로 사용하였으며 이 값은 다음 예측을 위해 입력으로 사용된다.

제안한 개선된 시스템은 시간지연τ에 따른 최적 임베딩 차원 m을 입력으로 사용했던 기존 PSFS의 입력 조건을 보다 유연하게 설정하여 시간지연 τ의 중복을 허용하도록 구성하였다. 시뮬레이션에서는 제안한 예측시스템에 태양의 흑점 시계열데이터를 적용하여 제안한 시스템이 가지는 예측성능을 확인하였으며, 그 결과 비교적 정확한 예측성능을 보여줌을 알 수 있었다.

[참 고 문 헌]

[1] S. Chiu, "Fuzzy Model Identification based on Cluster Estimation," *Journal of Intelligent and Fuzzy systems*, Vol. 2, No. 3, sept. 1994.

[2] M.S. Kim and S.G. Kong, "Parallel Structure Fuzzy Systems for Time Series Prediction," *Int. Jurnal of Fuzzy Systems*, Vol. 3, No1, March 2001.

[3] M.S. Kim, H.S, Lee, C.H. You, and C.S. Chung, "Chaotic Time Series Prediction using PSFS2," *2002 41st Annual Conference on SICE*, August 2002.