

RCGKA기반 퍼지 시스템 최적화 및 시계열 예측 응용

방영근, 심재선, 박종국, 이철희
강원대학교

Fuzzy System Optimization Based on RCGKA and its Application to Time Series Prediction

Young-Keun Bang, Jae-Sun Shim, Jong-Kuk Park, Chul-Heui Lee
Kangwon National University

Abstract -본 논문은 비정상 시계열 예측을 위한 다중모델 퍼지 시스템과, 제안된 시스템의 최적화를 위한 유전 알고리즘의 응용을 다룬다. 일반적으로, 퍼지 예측시스템의 성능은 비선형 데이터가 가지고 있는 다양한 패턴이나 법칙성, 경향 등을 잘 분석하고 시스템에 반영함으로써 개선될 수 있다. 따라서, 본 논문은 원형 시계열의 특성을 보다 잘 반영할 수 있는 그들의 차분데이터를 시스템에 적용하며, 생성 가능한 차분 데이터들 중 원형 시계열의 특징에 가까운 일부를 추출하여 다중모델 퍼지 예측 시스템을 구현함으로써 다양한 원형시계열의 패턴이나 법칙성 등이 고려될 수 있도록 하였다. 다중 모델 퍼지 시스템의 각각의 예측기에는 구조가 간단한 k-means 클러스터링 기법을 적용하여 구현의 용이성을 꾀하였으며, 성능평가를 통해 선택된 최종 예측기는 RCGKA(real-coded genetic k-means clustering algorithms)를 통해 더욱 최적화된 규칙기반을 가지게 함으로써 예측성능이 개선될 수 있도록 하였다. 본 논문에 사용된 최적화 기법인 RCGKA에는 또한 성능이 우수한 다양한 유전연산자를 도입하여 더욱 예측기 성능이 강화될 수 있도록 하였으며, 시뮬레이션을 통해 제안된 예측시스템의 효용성을 증명하였다.

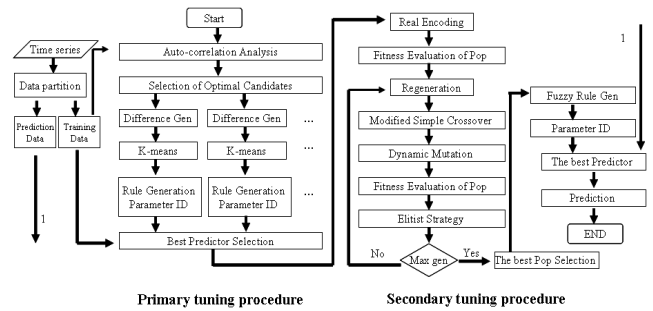
1. 서 론

유전 알고리즘은 적자생존의 법칙을 모사한 최적화 기법으로 다양한 분야에서 널리 이용되고 있다. 또한, 퍼지이론과 결합하여 다양한 분야에서 높은 성취를 이끌어 내고 있다. 하지만 일반적으로 사용되는 이진 코딩 기법의 경우에는 탐색 영역의 확장이나, 정밀도의 향상을 위해 그들의 탐색체를 증가할 필요가 있으며, 또한, 부호화와 복호화 과정으로 인해 구조적 복잡성을 초래할 수 있다. 따라서, 본 논문에서는 실수 코딩 기법과 다양한 유전연산자를 이용하여 탐색공간이나 정밀도를 위한 탐색체의 증가를 완화 하면서도 효율적으로 제안된 퍼지 시스템의 최적화를 유도할 수 있도록 하였다. 먼저, 1차 동조과정에서는 비선형데이터들을 그들의 통계적 특성이 보다 안정된 차분 데이터들로 가공하고, 이들 중 다중모델 퍼지 예측기 구현에 사용될 최적 차분 후보군을 선별한다. 또한, 하나의 차분이 아닌 다수의 차분을 이용한 다중모델예측 시스템을 구현함으로써, 원형 데이터에 내재된 다양한 특성들이 시스템을 통해 반영될 수 있도록 하였다. 다중 모델 퍼지 예측기의 구현을 위해선 k-means클러스터링 기법을 이용하여 보다 단순한 형태에서 시스템이 구현되도록 함으로써 시스템 구현의 용이성을 꾀하였고, 또한 2차 동조과정에서 유전알고리즘의 탐색공간에 대한 부담을 줄일 수 있도록 하였다. 구현된 다중모델 퍼지 예측기들은 성능검증을 통해 실제 예측을 수행할 하나의 예측기가 선택되며, 선택된 예측기는 RCGKA를 이용한 2차 동조과정을 통해 보다 최적화된 규칙기반을 구현함으로써 예측성능을 개선 하였다. 또한, 사용된 RCGKA에는 성능이 우수한 다양한 유전 연산자들을 결합하여 효용성을 증대하였으며, 시뮬레이션을 통해 제안된 시스템의 유효성을 검증 하였다.

2. 제안된 예측 시스템의 구현

2.1 1차 동조 과정

그림 1은 제안된 다중 모델 퍼지 예측 시스템의 전체적인 구조를 보여준다. 제안된 시스템의 다중모델 퍼지 예측기의 수는 상관분석을 통해 선별된 최적 차분 후보군의 개수만큼 구현되며, 수식(1)은 본 논문에 사용된 자기상관 함수이다. 최적 차분 후보군의 개수는 원형 시계열 데이터의 특성을 적절히 반영할 수 있어야 함으로, 본 논문에서는 6개 이상의 차분 후보군이 선택될 수 있도록 하였다. 이를 위하여 먼저 상관 계수 값이 높은 5개의 차분 간격을 선별하고, 나머지 계수들의 차를 이용하여 가장 큰 폭의 변화를 보이는 차 연산 값보다 높은 계수 값들에 상응하는 차분 간격 값을 다시 최적차분 후보군으로 선정하였다.



〈그림 1〉 RCGKA 기반 다중 모델 퍼지 예측 시스템의 순서도

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

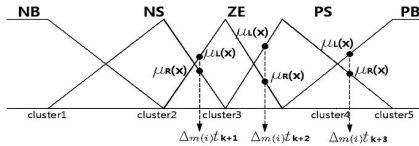
수식(1) 자기상관함수에서, N 은 훈련데이터의 길이이고, j 는 차분 간격 값을 의미한다. 또한, $y(i)$ 는 i 번째 훈련데이터이며, \bar{y} 는 훈련데이터의 평균이다. 1차 동조과정에서 다중 모델 퍼지 예측 시스템을 구성하는 예측기의 수가 정의 되면, 그들을 위한 차분 데이터들은 다음에 의해 생성된다.

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

여기서, $m(i)$ 는 선정된 차분 간격 값들을 의미한다. 이렇게 생성된 차분 데이터들은 각각의 퍼지 예측기들의 입력으로 사용되며 k-means 클러스터링 기법에 의해 퍼지 분할된다. 본 논문은 각각의 예측기의 입력공간을 삼각형 소속함수를 이용하여 5개로 분할하였으며 각 퍼지집합을 만족하는 입력에 대한 소속 정도는 (3)에 의해 얻어진다.

$$\begin{aligned} \text{if } x \in \text{duster}(1) \\ \mu_L(x) &= 1 \\ \mu_R(x) &= \frac{\text{duster}(z+1) - x}{z\text{cluster}(z+1) - \text{duster}(z)} \\ \text{elseif } x \in \text{duster}(z) \\ \mu_R(x) &= \frac{\text{duster}(z+1) - x}{z\text{cluster}(z+1) - \text{duster}(z)} \\ \mu_L(x) &= \frac{x - \text{duster}(z+1)}{z\text{cluster}(z) - \text{duster}(z-1)} \\ \text{elseif } x \in \text{duster}(5) \\ \mu_R(x) &= 1 \\ \mu_L(x) &= \frac{x - \text{duster}(z+1)}{z\text{cluster}(z) - \text{duster}(z-1)} \end{aligned} \quad (3)$$

아래의 그림 2는 본 논문에 사용된 퍼지분할의 한 예이다.



〈그림 2〉 제안된 시스템의 퍼지 분할 예

이러한 퍼지 분할은 예측기들 각각에 적용되며, 각각의 예측기의 규칙 기반을 생성하기 위해 이용된다. 또한 본 논문에서는 퍼지 예측기의 모델로 TSK퍼지 모델을 사용하였으며, 그림 2와 같이 3개의 연속된 차분 데이터를 1개의 입력집합으로 사용하였다. 따라서, 제안된 논문의 차분 간격 $m(i)$ 에 상응하는 예측기의 j 번째 퍼지 규칙은 다음과 같이 정의된다.

$$R_j : \text{If } \Delta_{m(i)}^j t_{k+1} \text{ is } A_1 \text{ and } \Delta_{m(i)}^j t_{k+2} \text{ is } A_2 \text{ and } \Delta_{m(i)}^j t_{k+3} \text{ is } A_3 \quad (4)$$

$$\text{Then } \nabla_j = p_o + p_1 \Delta_{m(i)}^j t_{k+1} + p_2 \Delta_{m(i)}^j t_{k+2} + p_3 \Delta_{m(i)}^j t_{k+3}$$

그리고 각각의 규칙의 파라미터들은 구조가 간단하고 효율적인 LSM(least square method)를 이용하여 추정하였다. 마지막으로 추정된 파라미터를 이용하여 예측을 수행할 경우 입력 데이터 쌍이 만족하는 퍼지 규칙의 수가 q 개라면 차분 예측 값은 (5)와 같이 구할 수 있으며, 최종 예측 값은 (6)과 같이 구할 수 있다.

$$\hat{\nabla}(t) = \frac{\sum_{i=1}^q \mu_i \hat{\nabla}_i}{\sum_{i=1}^q \mu_i} \quad (5) \quad \hat{y}(t+p) = y(t) + \hat{\nabla}(t) \quad (6)$$

이렇게 구현된 다중 예측모델 중에서 예측기 선택과정에서는 식 (7)을 최소화는 최적 예측기를 선택하게 되고 선택된 최종 예측기는 2차 동조 과정에서 더욱 세밀히 최적화 된다.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y(i) - \hat{y}(i))^2 \quad (7)$$

2.2 2차 동조 과정

그림 1에서 보여주듯이 2차 동조 과정은 1차 동조과정의 퍼지 분할을 더욱 최적화 시키는 과정이다. 이를 위해 본 논문은 1차 동조과정에서 k-means 클러스터링에 의해 분할된 퍼지 집합 중심 값들을 유전 알고리즘의 실수 코딩 기법, 최적개체의 추적을 돕는 재생산 연산자, 교배, 돌연변이연산자를 이용하여 더욱 적합화 될 수 있도록 하였다. g 세대에서의 집단 $p(g)$ 는 (8)와 같이 염색체로 특정 지어지는 L 개의 개체들의 집합으로 이루어지며 각각의 염색체들은 (9)과 같이 실수 형으로 표현된다.

$$P(g) = \{s_{1(g)}, s_2(g), \dots, s_L(g)\} \quad (8)$$

$$s_i(g) = (v_{i1}(g) \ v_{i2}(g) \ \dots \ v_{iz}(g)) \quad (9)$$

수식 (9)의 염색체의 성분은 정확하게 5개의 퍼지집합의 중심과 일치하게 된다. 이는 1차 동조 과정 k-means에 의해 분할된 퍼지 집합의 중심값을 2차 동조과정의 유전 알고리즘에 바로 적용함으로써, 탐색공간의 부담을 줄이기 위한 것이다. 또한, 각각의 개체들의 적합도 평가를 위한 목적함수로는 (10)과 같이 클러스터내 분산을 최소화 하는 방향으로 정의 되었다.

$$OF = \sum_{c=1}^z \left(\frac{1}{h} \sum_{u=1}^h (x_u - v_c)^2 \right) \quad (10)$$

재생산 연산자는 각 개체들의 적합도와 가장 우수한 개체의 적합도를 이용하여 약한 개체는 최적 개체쪽으로 더 많은 가중치를 부여하고 강한 개체는 작은 가중치를 부여하여 최적 개체쪽으로 유도되게 하였으며, 수정단순 교배와 미세 조정이 가능한 동적 돌연변이를 적용하여 유전 알고리즘의 정밀도를 높일 수 있도록 하였다. 또한, 마지막으로, 전 세대의 우수한 개체가 다음 세대의 자손 생성에 기여하지 못하고 소멸 되는 것을 방지하기 위하여 엘리트 전략을 채용하여 유전 알고리즘의 성능을 개선하였다.

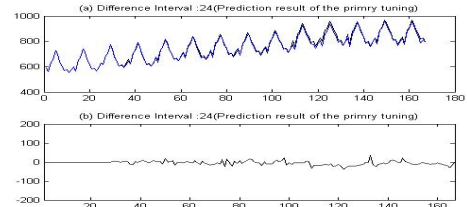
3. 시뮬레이션 및 결과 고찰

제안된 시스템의 성능을 검증하기 위하여 후주의 유유생산량 데이터를

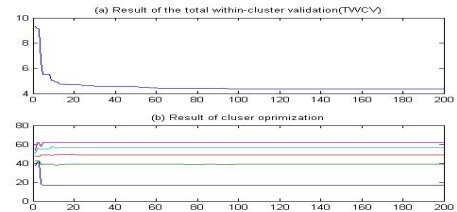
를 적용하였으며, 이 데이터는 166개의 값으로 구성되며 이중 60개를 훈련 데이터, 나머지를 검증 데이터로 사용하였다. 또한 성능 평가 지수는 다음과 같이 정의 되는 상대 오차 평균(MRE: mean relative error)을 이용하였다.

$$MRE = \frac{1}{N} \sum_{i=1}^N \left(\frac{|y(i) - \hat{y}(i)|}{|y(i)|} \times 100 \right) \% \quad (11)$$

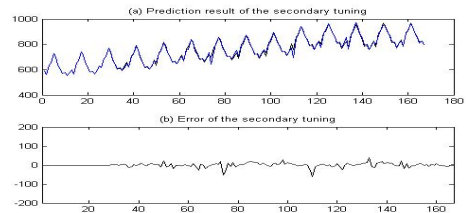
그림 3은 1차 동조과정 후의 예측 결과를 보여주며, 그림 4는 2차 동조 과정에서 퍼지집합의 중심이 최적화 되는 과정을 보여준다. 또한, 그림 5는 제안된 시스템의 최종 예측 결과를 보여 준다.



〈그림 3〉 1차 동조 후의 예측 결과 (최적 차분 간격 24)



〈그림 4〉 2차 동조의 파라미터 최적화



〈그림 5〉 제안된 퍼지 예측 시스템의 최종 결과

아래의 표 1은 1차 동조과정 후의 결과와 2차 동조과정 후의 결과를 비교한 것이다.

〈표1〉 1,2차 동조과정의 성능 비교

과정	1차 동조 과정 후	2차 동조과정 후
Data		
MRE	1.3687	1.1145

그림 3을 살펴보면 제안된 예측 시스템의 성능이 1차 동조과정에서도 우수한 특성을 보임을 알 수 있으며, 이는 적용된 차분 데이터들이 원시 계열의 특성을 잘 반영한 결과로 볼 수 있다. 또한, 그림 5와 표1을 살펴보면, 2차 동조 과정이 제안된 예측 시스템의 성능을 더욱 개선할 수 있음을 알 수 있다.

감사의 글

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

[참고 문헌]

- [1] 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
- [2] K. Venkatalakshmi. P. A. Praisay, R. Maragathavalli, S. M. Shalinie, "Multispectral Image Clustering Using Enhanced Genetic k-means Algorithm," Information Technology Journal, Vol. 6, pp. 554-560, 2007