

DNA 코드 유전자 알고리즘을 이용한 퍼지 모델링

유진영 · 이연우 · 주영훈 · 박진배
 *군산대학교 전자정보공학부, **연세대학교 전기전자공학과

Fuzzy Modeling Using DNA-Coded Genetic Algorithm

Jin-Young Yu*, Yeun-Woo Lee*, Young-Hoon Joo*, Jin-Bae Park**
 *School of Electronic & Information Eng. Kunsan National University,
 Department of Electrical & Electronic Eng. Yonsei University.

Abstract - 본 논문에서는 발생모델인 DNA 코딩 기법과 진화 모델인 유전자 알고리즘을 이용한 비선형 시스템의 퍼지 모델링에 대한 새로운 방법을 제안한다. DNA 코딩 기법은 실제 생체 분자(bio molecule)를 계산의 도구로 사용하는 새로운 계산 방법으로, 진화 연산과 결합하여 인공지능의 새로운 분야로 부각되고 있다. 그러나, 실제 생체 분자를 계산의 도구로 사용하기 때문에 기존의 컴퓨터에 적용하기 어렵고, 단순히 합성과 분리라는 간단한 방법으로 해를 구하기 때문에 보다 효과적인 알고리즘을 개발하여야 할 필요성이 있다. 따라서 본 논문에서는 DNA 코드 유전자 알고리즘을 제안하며, 제안된 방법은 비선형 시스템의 퍼지 모델링에 적용하였으며, 기존의 유전자 알고리즘과 비교를 통하여 그 우수성을 입증하였다.

코드와 하는데 적용되었다. DNA 코딩은 DNA 염기 배열이 단백질로 번역되는 것과 유사한 과정을 통하여 규칙으로 번역된다. 4가지 염기, 즉 A(Adenine), C(Cytosine), G(Guanine), T(Thymine)으로 정보를 표현하고, 또한 막대한 병렬성을 이용하여 주어진 탐색 공간을 효율적으로 탐색할 수 있다. 그러나 아직 실제 생체 분자의 정보 처리 원리에 바탕을 둔 새로운 컴퓨터 기술이 개발되지 않고 있기 때문에 DNA 코딩 기법은 기존의 컴퓨터에 적용하기 어렵고, 단순히 합성과 분리라는 간단한 방법으로 해를 구하기 때문에 효과적인 알고리즘을 개발해야 할 필요성이 있다. 따라서 본 논문에서는 진화 알고리즘인 유전자 알고리즘의 코드화 부분을 발생모델인 DNA구조의 코드 합성 기법으로 함으로써 보다 유연한 구조의 코딩 방법을 제안하며, 비선형 시스템인 가스로의 퍼지 모델링 결과를 가지고 DNA 코드 유전자 알고리즘과 기존 유전자 알고리즘의 결과를 비교함으로써 DNA 코드 유전자 알고리즘의 우수성을 입증하려 한다.

1. 서 론

비선형 시스템을 수학적으로 모델을 구하려 하는 경우 일반적으로 미분 방정식이 광범위하게 사용되어왔다. 하지만, 많은 실질적인 문제는 많은 비선형성과 시변수를 포함하고 있기 때문에 정확한 미분방정식을 구하는 것이 쉽지 않을 뿐만 아니라 비교적 정확한 모델을 구하더라도 모델 근사화에 의한 효용성이 감소한다. 이러한 시스템을 다루기 위해 if then 규칙 형식으로 표현되는 퍼지 모델링 기법이 활발히 연구되어지고 있다. Zadeh[1]에 의해 최초로 도입된 퍼지 논리 개념은 이런 문제를 극복할 수 있는 해를 제시하였고 각 계에서 새로운 가능성을 보여주었다. Tong[2]은 퍼지 논리를 바탕으로 시스템 묘사하는 퍼지 모델링을 제안하여 복잡한 비선형 시스템의 모델링에 새로운 접근 방법을 제시하였다. Sugeno[3]는 비선형 시스템의 입출력 데이터만으로 시스템의 정성적 모델을 구하는 방법과 결론부를 선형식으로 대체하는 퍼지 모델의 동정 방법을 제안하였다. 상기 연구에서 알 수 있듯이 퍼지 모델링의 주 관심사는 전문가 경험은 이용하여 마지막 시스템을 퍼지 규칙으로 정확히 묘사하는데 있다. 최근에는 진역 최적화 탐색 기법인 유전자 알고리즘을 이용한 새로운 퍼지 모델링 방법이 제시되어 많은 연구가 진행되어왔다. 유전자 알고리즘은 생물학적 유전학과 자연 선택설을 바탕으로 한 탐색 알고리즘이다. 그러나 유전 알고리즘은 비선형성이 강한 시스템에 대하여 최적의 해를 찾지 못하고 국소적인 값에 빠질 위험이 있으며 최적의 해를 찾기 위해 많은 시간을 필요로 하게 된다. 이에 Joo [4, 5]는 유전 알고리즘과 클러스터링 기법을 이용하여 퍼지 모델의 파라미터와 구조를 동정하는 자동동정 기법과 유전 알고리즘의 변형된 형태의 mGA (messy Genetic Algorithm)를 이용한 퍼지 모델링 기법을 제안하였다. 유전 알고리즘은 최적 해를 찾기 전에 국소 값에 도달하는 것과 최적 해를 찾기 위해서는 많은 시간이 필요하다는 단점이 있다. 이러한 유전 알고리즘의 문제점을 해결하기 위한 방법으로 생물학적 DNA 구조를 모방한 DNA 코딩 기법의 연구가 활발해지기 시작했다. DNA 코딩 기법은 실제 생체 분자인 DNA를 계산의 도구 및 정보 저장 도구로 사용한다. 이 생물학적 DNA구조를 모방한 것을 DNA 코딩이라 명명되었다. DNA 코딩 기법은 염색체의 가변적인 형태와 비트 단위가 아닌 생물의 DNA 구조에 의한 해석으로 유전 알고리즘보다 빠르게 최적 해를 찾을 수 있으며 다변적인 염색체의 구조와 지식의 융통성이 있는 표현으로 최적 해를 구하는데 유전 알고리즘보다 유리하다고 알려져 있다. 이 방법은 Yoshikawa[6]등에 의해서 퍼지 규칙을

2. DNA 코딩기법 및 알고리즘

2.1 생물학적 DNA

모든 생명체는 각각 고유의 DNA를 가지고 있다. DNA는 개체의 특성을 발현시키는 유전코드로서, A, U(T), G, C, 4 종류의 염기 배열로 이루어져 있다. 이들 중 세 개의 배열이 한 의미 단위를 이루어 해석되며, 이 의미 단위를 생물학적인 용어로 코돈(codon)이라 한다. 총 64종류의 코돈이 있으며, 이 코돈에 의해 20종류의 아미노산이 만들어진다. 그 이유는 서로 다른 코돈이 같은 아미노산을 만들기 때문이다. DNA는 RAN로 전사되어 리보솜에서 단백질로 번역된다. 즉 아미노산을 암호화하는 DNA의 배열에 따라 아미노산의 합성순서를 결정하여 여러 종류의 단백질을 만들어 낸다. RNA의 단백질로의 번역은 AUG에서 시작되어 UGA에서 번역이 끝나며, DNA 코드중 단백질로 번역되는 부분은 시작 코돈과 종료 코돈 사이에 존재하는 염기들이다. DNA에서는 U대신 T를 사용한다. 표 1은 RNA(DNA) 코돈과 생성하는 아미노산을 나타낸다. 생물학적 유전자 코드의 일반적인 구조는 그림 1과 같다. 즉 단백질을 직접 코드화 영역(coding region)과 그 코드부위의 발현을 조절하는 조절부위(regulatory region)로 구성된다. 코드부위는 조절부위의 명령에 의해 세포내의 조건이 조절된다. 즉, 조절부위는 규칙의 조건부 또는 세포의 현재 주변상태, 코드부위는 결론부 또는 세포의 다음상태에 대응된다.

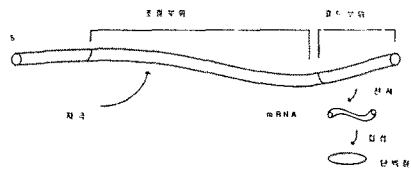


그림 1. 유전자 코드의 일반적인 구조

표 1. 코돈에서 아미노산의 번역

	U (T)	C	A	G				
U (T)	UUU	Phe	UCU	UAU	Tyr	UGU	Cys	U
	UUC	(1)	UCC	UAC	(4)	UGC	(19)	C
	UUA	Leu (9)	UCA	UAA	(4)	UGA	결외 (20)	A
	UUG		UCG	UAG	(5)	UGG	Trp (20)	G
C	CUU	Ile (2)	CCU	CAU	His	CGU	Arg (11)	U
	CUC		CCC	CAC	(5)	CGC		C
	CUA		CCA	CAA	Gln	CGA		A
	CUG		CCG	CAG	(14)	CGG		G
A	AUU	Met (3)	ACU	AAU	Asn	AGU	Ser	U
	AUC		ACC	AAC	(15)	AGC	(10)	C
	AUA	ACA	AAG	Lys	AGA	Arg (11)	A	
	AUG	ACG	AAG	(16)	AGG	(11)	G	
G	GUU	Val (6)	GCU	GAU	Asp	GGU	Gly (13)	U
	GUC		GCC	GAC	(17)	GGC		C
	GUA		GCA	GAA	Glu	GGA		A
	GUG		GCG	GAG	(18)	GGG		G

2.2 DNA 코딩 방법 및 유전자 알고리즘

본 논문에서 제안한 DNA코드 기반 유전자 알고리즘의 코돈의 의미는 표1의 괄호안의 값과 같다. 괄호안의 값은 입력 및 출력 학습 데이터 상의 임의로 분할된 값을 지정하는 역할을 한다. 코돈은 아미노산으로 번역되고, 아미노산이 가지는 유전 정보에 따라 퍼지규칙이 생성된다. 또한 각각의 개체들은 유전 연산자를 이용하여 진화하게 되며 본 논문에서는 교차, 돌연변이, 삭제(deletion) 및 도치(inversion) 연산자를 사용하였고, 이러한 과정을 거쳐 적합도가 높은 개체는 다음 세대로 재생된다. 다음은 연산자에 대한 설명이다.

재생 연산자 : 재생은 적합도 함수 값에 의해서 개별적인 스트링의 복제가 이루어지고 복제된 스트링은 다음의 유전 연산을 위해 교배 풀에 놓이게 된다.

교차 연산자 : 교배 풀에서 두 스트링은 임의적으로 자신의 짝을 결정한다. 유전 알고리즘의 교차 연산자와 다른 점은 교차 지점의 위치가 가변적이라는 것이다. 그러므로 스트링 A와 스트링 B는 서로 중복될 수도 있다.

돌연변이 연산자 : DNA 코딩 기법도 유전 알고리즘과 마찬가지로 돌연변이 연산자를 사용한다. 돌연변이는 재생 및 교배와 함께 사용될 때 중요한 개체의 조기 손실을 막아준다. 본 논문에서 쓰인 스트링은 1~4까지의 4가지이기 때문에 임의의 스트링 위치에서 랜덤하게 바뀌게 된다. 이 주 연산자 이외에 삭제 연산자(deletion operator)와 도치 연산자(inversion operator)를 부 연산자로 사용하였다.

3. 퍼지 모델

일반적인 퍼지 모델의 특징으로는, 언어 규칙으로 표현하기가 쉽고, 제어 대상 플랜트의 정확한 수학적 모델이 필요 없으며, 강인하고 구현이 용이한 장점을 가지고 있다. 그러나, 시행착오적이며, 적용적이지 못하고, 규칙 베이스 설정과 최적의 소속함수 파라미터 설정의 어려움을 갖는 단점이 있다. 그림 2은 일반적인 퍼지 모델의 구조를 보여준다.

본 논문에서는 시스템의 입출력 데이터만을 이용하여 식 (1)과 같은 TS(Takagi-Sugeno) 퍼지 모델을 동정하는 방법을 제안한다.

$$IF \ x_1(t) \text{ is } A_1^i \text{ and } x_2(t) \text{ is } A_2^i, \text{ THEN } y \text{ is } w_i; \quad (i=1,2,\dots,r) \quad (1)$$

식 (1)에서 A_1^i, A_2^i 는 전방부 소속함수, w_i 는 후방부 소속함수로서 실수값을 갖는다. 그리고, $x_j(t)$ 는 입력변수, y 는 출력변수, r 는 규칙들의 총 개수를 나타낸다.

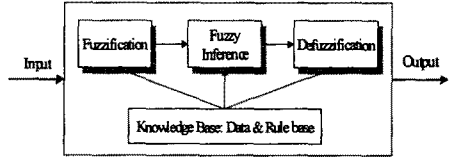


그림 2. 퍼지 모델의 일반적인 구조

본 논문에서는 멤버십 함수로 가우시안 함수를 사용하였다. 그림 3은 본 논문에 제안한 알고리즘을 이용하여 코돈에서 아미노산으로의 번역 및 암호화 과정을 보여준다.

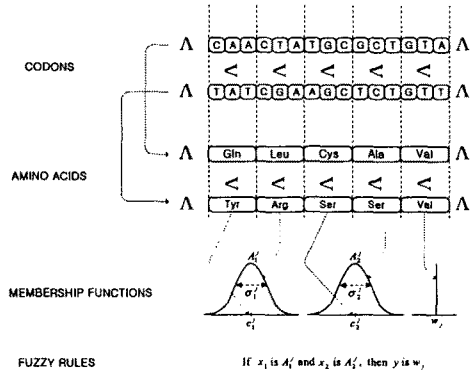


그림 3. 염색체로부터 퍼지규칙의 생성

비퍼지화 방법으로는 무게 중심법을 사용하였으며, 식 (2)과 같이 표현된다.

$$y^* = \frac{\sum_{i=1}^r \mu_i(x(t))w_i}{\sum_{i=1}^r \mu_i(x(t))} \quad (2)$$

$$\mu_i(x(t)) = \prod_{j=1}^n A_j^i(x_j(t)) \quad (3)$$

여기서, $A_j^i(x_j(t))$ 는 A_j^i 에 대한 $x_j(t)$ 의 소속 정도이다.

일반적으로 퍼지모델링의 최적화는 평가 함수값을 최소화시키는 퍼지규칙의 수와 소속함수의 파라미터를 찾는 것이다. 한편 GA에서는 적합도 함수를 최대화시키도록 진화가 이루어지므로 최소화 문제를 최대화 문제로 변경하여야 한다. 식 (4)는 본 논문에서 사용된 적합도 함수를 나타낸다.

$$f(PI_{mse}, PI_{rule_*}) = \lambda \frac{1}{1 + PI_{mse}} + (1 - \lambda) \frac{1}{PI_{rule_*}} \quad (4)$$

식 (4)의 적합도 함수의 사용은 모델의 정확성 (PI_{mse})과 규칙의 수 (PI_{rule_*})를 감소시키는데 있다. 또한 $\lambda(0 \leq \lambda \leq 1)$ 는 하중 상수로써, 그 값이 크면 퍼지 추론 시스템은 좀더 정확해지고, 반대로 작으면 적은수의 규칙수가 결정된다.

4. 시뮬레이션 및 결과

본 논문에서 제안한 DNA 코드 유전자 알고리즘의 우수성을 입증하기 위하여 Box와 Jenkins [7]가 제시한 가스로의 시계열 입력력 데이터를 사용하여 가스로의 연소된 이산화탄소의 농도를 퍼지 모델링한다. 이 데이터들은 입력으로 가스 유입률($u(t)$)와 출력으로 CO_2 밀도 ($y(t)$)로 나타내는 296개의 데이터 쌍으로 구성되어 있다. 본 논문에서는 시뮬레이션을 위해 현재의 출력 $y(t)$ 에 영향을 미치는 입력들로서는 $u(t-4)$ 와 $y(t-1)$ 의 두 개를 택하였다. 시뮬레이션을 위한 초기 파라미터 값들은 표 2와 같이 하였다.

표 2. 퍼지 모델링을 위한 초기 파라미터

Parameter(DNA,GA)	Value
generation	200
Population	300
Crossover	0.9
Mutation	0.01
Deletion rate(DNA)	0.01
Inversion rate(DNA)	

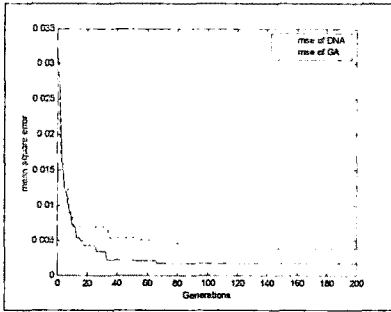


그림 4 DNA와 GA의 MSE 비교

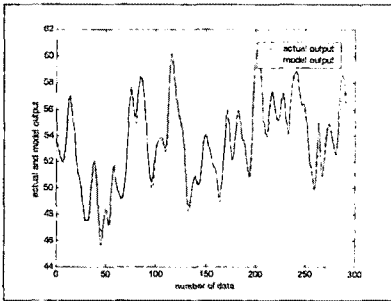


그림 5. DNA코드의 퍼지 모델링 결과

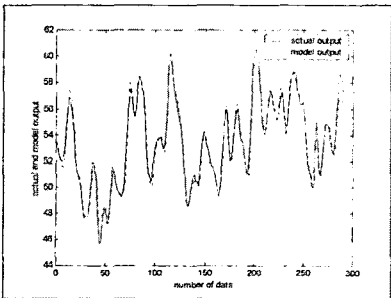


그림 6. GA의 퍼지 모델링 결과

그림 4는 DNA 코드 유전자 알고리즘과 기존 유전자 알고리즘의 MSE를 비교한 그림이며, 그림 5와 6은 Gas Furnace의 실제 출력값과 각각의 퍼지 모델링 값의 결과를 비교한 그림이다. 표 3은 제안된 방법과 같은 조건에서의 GA의 결과 비교 및 가스조에 대한 다른 모델들과의 비교를 보여준다. 표 3에서 알 수 있듯이 본 논문에서 제안한 퍼지 모델링 기법이 다른 기법보다 규척수와 정확도 면에서 우수한 퍼지 모델을 생성하였음을 알 수 있다.

표 3. 가스조에 대한 모델들과의 비교

Model Name	MSE	Rules	Note
[7]	0.71		ARMA model
[8]	0.469	19	Fuzzy model
[9]	0.320	81	Fuzzy model
[10]	0.328	25	Fuzzy model
[11]	0.190	6	Fuzzy model
GA	0.0038	13	Fuzzy model
Our result	0.0017	10	Fuzzy model

5. 결론

본 논문에서는 복잡하고 불확실한 비선형 시스템의 입력력 데이터만으로 효율적으로 퍼지 모델링하는 기법을 제안하고 이를 실험을 통하여 검증하였다. 제안된 방법은 기존의 유전자 알고리즘에 비하여 유연한 암호화 기법인 DNA 코딩 기법과 진화 알고리즘인 유전자 알고리즘의 장점을 이용하여, 비선형 시스템인 Gas Furnace 시스템의 퍼지 모델링에 적용하였고, 기존 유전자 알고리즘과 다른 퍼지 모델들과 비교를 통하여 그 우수성을 입증하였다.

[참고 문헌]

- [1] L. A. Zadeh, "Fuzzy Sets," Information Control, Vol. 8, pp.338-353, 1965.
- [2] R. M. Tong, "The Evaluation of Fuzzy Models Derived from Experimental Data", Fuzzy Sets and Systems, Vol. 4, pp. 1-12, 1980.
- [3] D. E. Goldberg, B. Korb, and K. Deb, "Messy Genetic Algorithms: Motivation, Analysis, and First Results", Complex Systems, Vol. 3, No. 5, pp. 493-530, 1989.
- [4] Y. H. Joo, H. S. Hwang, K. B. Kim, and K. B. Woo, "Fuzzy System Modeling by Fuzzy Partition and GA Hybrid Schemes", Fuzzy Sets and Systems, Vol. 86, No. 3, pp.279-288, 1997.
- [5] 주영훈, 이연우, 박진배, "메시 유전 알고리즘 기반 퍼지 모델 동정", 대한 전기 학회 논문지, Vol. 49D, No. 8, pp. 407-414, 2000년 8월.
- [6] T. Yoshikawa, T. Furuhashi, Y. Uchikawa, "A Combination of DNA Coding Method with Pseudo Bacterial GA for Acquisition of Fuzzy Control Rules", The 1st Online Workshop on Soft Computing(WCSD), 1996.8
- [7] G. E. P. Box and G. M. Jenkins, "Time Series Analysis, Forecasting and Control", Holden Day, San Francisco, 1970.
- [8] R. M. Tong, "The Evaluation of Fuzzy Models Derived from Experimental Data", Fuzzy Sets and Systems, Vol. 4, pp.1-12, 1980.
- [9] W. Pedrycz, "An Identification Algorithm in Fuzzy Relational System", Fuzzy Sets and System, Vol. 13, pp. 153-167, 1984.
- [10] C. W. Xu, "Fuzzy Model Identification and Self-learning for Dynamic Systems", IEEE Trans. Sys. Man, Cybern., Vol. 17, No. 4, pp. 683-689, 1987.
- [11] M. Sugeno and T. Yasukawa, "A Fuzzy logic based Approach to Qualitative Modeling", IEEE Trans. Fuzzy sys., Vol. 1, No. 1, pp. 7-31, Feb. 1993.