시계열 예측을 위한 DNA코딩 기반의 신경망 진화

Evolutionary Neural Network based on DNA Coding Method for Time Series Prediction

*이기열, 이동욱, 심귀보
중앙대학교 전자전기공학부
Tel : 02) 820-5319, Fax : 02) 817-0553, E-mail:kbsim@cau.ac.kr

Ki-Youl Lee, Dong-Wook Lee and Kwee-Bo Sim
School of Electrical and Electronics Engineering, Chung-Ang University
Tel : 02) 820-5319, Fax : 02) 817-0553, E-mail:kbsim@cau.ac.kr

In this paper, we propose a method of constructing neural networks using bio-inspired emergent and evolutionary concepts. This method is algorithm that is based on the characteristics of the biological DNA and growth of plants. Here is, we propose a constructing method to make a DNA coding method for production rule of L-system. L-system is based on so-called the parallel rewriting mechanism. The DNA coding method has no limitation in expressing the production rule of L-system. Evolutionary algorithms motivated by Darwinian natural selection are population based searching methods and the high performance of which is highly dependent on the representation of solution space. In order to verify the effectiveness of our scheme, we apply it to one step ahead prediction of Mackey-Glass time series, Sun spot data and KOSPI data

1. 서 론

자연계에는 많은 생물들과 이 생물들이 서로 상호 작용을 하며 살아가는 생태계가 있다. 이런 생물들의 삶을 모방한 많은 인공생명의 모델들이 있다. 생물은 자신이 가진 유전자를 바탕으로 발생하고, 이렇게 발생된 개체는 생존해 나가며, 학습을 한다. 또한 자신의 유전 정보를 후손에게 전달함으로서 종족을 보존한다. 이러한 생물의 자기 조직화 현상을 모방한 인공생명의 모델은 진화모델, 발생모델, 발생/발달 모델 3가지가 있다. 진화모델로는 진화 알고리즘, 학습모델로는 신경망, 강화학습, 인공면역계 등이 있으며, 발생/발달모델로는 셀룰러 오토마타(CA) 린든마이어 시스템(L-system)[1~3] 등이 있다.

본 논문에서는 원하는 목적의 신경망을 얻기 위하여 DNA 코딩[4]을 이용하여, 규칙을 생성시킨 후 그 규칙을 통해 신경망을 구성한다. 또한 진화알고리즘을 이용하여 점점 더 우수한 개체를 선택함으로서 최종적으로 원하는 목적의 신경망을 구현한다.

DNA 코딩은 생물학적인 DNA 구조를 모방한 것으로서 DNA가 생물의 유전정보를 통해 자신을 발생시키고, 또한 다음 세대에 유전 정보를 전달하는 과정을 모방한 방법이다. 유전자에서는 같은 코돈(유전자의 최소 단위, 3개의 염기 배열로 구성)일지라도 코돈의 위치나 좌우의 상관 관계에 의하여 다른 형질로 발현되는 특징을 갖는다. 이러한 특성을 모방한 DNA코딩은 동적인 구조를 통한 중복 해석과 여분이 있다는 장점을 이용하여 하나의 해석 단위를 신경망의 노드와 그의 부수적인 요소(Weight, 입출력 범위)를 결정하게 한다. 그리고 중복해석을 통해 나온 L-system의 생성규칙을 이용하여 구성)을 이용하여 신경망의 구조를 결정한다.

이렇게 자동 생성된 신경망을 Mackey-Glass 시계열 예측 문제와 Sunspot 데이터 그리고 KOSPI(종합주가지수)의 예측에 적용시켜, 우수한 개체를 선택한 후 이들 개체의 유전자를 GA 연산자(돌연변이와 교배)를 통하여 다양화된 유전자를 다음 세대로 전달하고, 더 좋은 개체를 얻어 시계열 예측문제를 풀 수 있는 신경망을 자동적으로 생성하도록 하는 것이다.

2. DNA 코딩

2.1 생물학적 DNA

모든 생물체는 각자 고유의 DNA를 가지고 있다. DNA는 개체의 특성을 발현시키는 유전코드로서, A(아데닌) T(티민, RNA에서는 U:우라실) G(구아닌) C(시토신)의 4개의 염기배열로 이루어져 있다. 또한 염기 3개의 배열이 한 의미단위를 이루어 해석된다. 이 의미단위를 생물학적인 용어로 코돈(codon)이라 한다. 코돈의 가짓수는 4×4×4 = 64개이며 이것이 코드화 하는 아미노산은 20가지이다.[5].

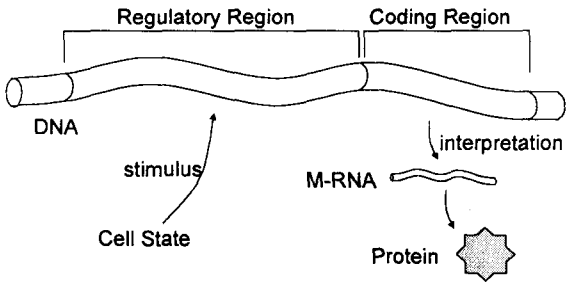

그림 1. 유전자 구조의 일반적인 구조.

생물학적 유전자의 기본적인 구조는 그림 1과 같이 단백질을 직접 코드화 하는 부위(coding region)와 그 코드 부위의 발현을 조절하는 조절부위(regulatory region)로 구성되어 있다. 코드화 부위는 조절부위의 명령에 의해 세포내의 조건이 조절부위를 자극할 때 단백질로 번역된다. 이것은 발생모델의 규칙의 표현방식과 유사하다. 즉, 조절부위는 규칙의 전건부 또는 세포의 현재 주변 상태, 코드화 부위는 후건부 또는 세포의 다음 상태에 대응된다.

2.2 DNA 코딩방법의 특징

DNA의 동작을 모방한 코딩방법에 대한 연구는 Yomohiro[4] 등에 의하여 기본적인 방법이 연구되었다. 염색체는 기본적으로 4가지 염기의 배열로 이루어져 있고 아미노산을 번역하는 것과 마찬가지로 코돈 단위로 번역한다. 또한 다음 그림 2와 같이 유전자의 번역 시작점이 일정하지 않기 때문에 중복 번역을 허용하며 때에 따라서는 번역되지 않는 부분도 존재한다. 하나의 아미노산을 생성하는 코돈이 여러 개이기 때문에 염색체의 중복을 효율적으로 이용할 수 있다. 교차점도 임의로 주어지기 때문에 염색체의 길이도 가변적이다. Yomohiro의 방법은 번역 테이블을 만듦으로서 규칙의 표현에 적합하게 구성되어 있다. 이상의 특징을 정의하면 아래와 같은 4가지로 정리할 수 있다.

◆ DNA 코딩방법의 특징
(1) 지식의 표현이 쉽다.
(2) 코딩에 여분과 중복이 있다.
(3) 염색체의 길이가 가변적이다.
(4) 교차점에 제약이 없다.

Wu[6] 등은 GA에서의 가변위치 표현법(floating representation)에 대한 스키마 분석을 통하여 유효성을 증명하였다. 가변위치 표현법이란 DNA 코딩에서와 같이 시작 기호(start tag)를 가지고 있어서 시작 기호가 발견되는 곳에서부터 유전자의 번역을 시작하는 방법이다. Wu는 비록 비트 스트링을 가지고 실험을 하였으나 이 방법도 DNA 표현 방법의 (2), (3), (4)의 특징을 모두 가지고 있다.

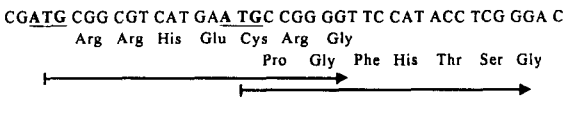

그림 2. DNA 염색체의 번역.

그림 2 에서 보듯이 위와 같은 DNA코드가 있다면, 그 안에 있는 시작코돈(ATG)부터 해석을 시작하여 각각의 아미노산을 해석하여 그에 따른 규칙을 생성한다. 이와 같은 해석 방법을 통하여 짧은 DNA코드에서 도 많은 정보를 얻을 수 있다.

표 3 생물학적 DNA와 DNA 코딩의 비교

	생물학적 DNA	DNA 코딩
코돈	아미노산을 암호화하는 최소단위	규칙의 최소 의미단위
시작 코돈	번역의 시작점	번역의 시작점
종료 코돈	번역의 종료점	번역의 종료점 반드시 필요하진 않음
번역 결과물	단백질, 효소	규칙

3. L-System

L-system[1~3]은 병렬적인 문자열의 재조합 속성을 갖는 일종의 문법으로서 1968년 Aristid Lendenmayer에 의해 제안되었다. 이것은 다세포 생물의 성장 과정의 모델링이 가능하기 때문에, 이후 컴퓨터 그래픽 등에서 식물을 모델링하는데 많이 이용되고있다.

3.1 Simple L-system

L-system은 초기문자열(axiom)로부터 생성규칙(production rule)의 반복적인 적용에 의하여 생성된 최종 문자열은 심벌(symbol)의 문맥에 따라 여러 가지 방식으로 해석되며, 일반적으로 선을 그리는 방식을 택하여 나무 모양을 만들어 낸다. 간단한 L-system의 구성요소들은 다음과 같이 정의한다.

· **문자(Alphabet)** Σ : 심벌들의 유한집합으로, 주로 a, b, c 같은 문자들이 쓰지만 다른 문자라도 상관이 없다.

ex) $\Sigma=\{a,b,c\}$

· **초기문자열(Axiom)** α : 집합 V에서 정의된 심벌들의 연속된 문자열의 집합을 Σ^*라고 하면 초기문자열은 집합 Σ^*의 ϕ가 아닌 한 원소이며, 초기문자열부터 생성규칙에 의해 생장한다.

ex) Axiom = b

· **생성규칙(Producton Rule)** Π : 하나의 심벌 a ($a\in\Sigma$)을 하나의 문자열 w($w\in\Sigma^*$)로 대응시키는 것. 만약 특정 심벌에 대하여 어떤 생성규칙도 주어지지 않으면 자기 자신으로 대응시키는 것을 기본으로 한다.

ex) a → ab
b → a

4. 신경망의 DNA 코딩방법

4.1 신경망 구성을 위한 DNA 코딩방법

본 절에서는 신경망을 진화시키기 위한 DNA코딩방법을 설명한다. 신경망을 구성하기 위해 DNA 코드에서 규칙(L-system의 규칙)을 생성한 후, 그 규칙을 바탕으로 신경망을 구성한다.

우선 임의의 DNA 코드를 발생시킨다. 이 코드를 위 표에 의하여 해석을 하여 여러 개의 규칙을 만든다. 시작코돈(ATG)이 나오면 해석을 시작한다. 처음에 나오는 첫 코돈은 신경망의 문자로 해석을 한다. 두 번째 코돈은 신경망의

연결 범위(connecting raing)[1]를 결정한다. 예를 들어 숫자가 3이라면 최종적으로 만들어진 문자열에서, 그 노드로 3번째 노드까지 모두 연결이 된다. 단, 쉼표(,)가 나오면 그 옆의 노드는 연결하지 않는다. 세 번째 코돈은 신경망의 노드에 포함되는 Bias로 해석을 한다. 값의 계산은 DNA A=0, G=1, T=2, C=3으로 하여 식 (1)과 같이 계산을 한다. 이 후에 나타나는 코돈은 두 번째 코돈에서 정해진 숫자만큼 Weight 값으로 해석을 한다. 계산은 Bias와 같은 방법으로 한다. 여기까지 해석을 하면 하나의 신경망노드를 구성한 것이다. 같은 방법으로 정지 코돈이 나올 때까지 해석을 계속한다.

$$Bias = \frac{(DNA \times 4^2 + DNA \times 4^1 + DNA \times 4^0) - 32}{10} \quad (1)$$

$$(-3.2 \le W \le 3.2 \quad 0.1\text{간격})$$

이렇게 DNA 코드를 해석한 모드의 배열에서 첫 번째 코돈의 문자만을 뽑아 문자열을 만든다. 맨 처음에 나오는 문자를 L-system 규칙의 전건부로, 그 이후에 나오는 문자열은 후건부로 해석을 하여 L-system의 규칙을 생성한다. 이렇게 해석된 규칙은 문맥자유 L-system의 규칙과 같은 특성을 지닌다. 만약 같은 전건부를 갖는 규칙이 여러 개 나올 경우는 먼저 나온 규칙을 적용한다. 그 규칙을 가지고, 정해진 초기 문자열에 따라 규칙을 적용해 그 다음 단계의 문자열을 만들고, 맞는 규칙이 없을 경우에는 그 문자를 그대로 유지한다. 정해진 수 만큼 반복하여 문자열을 치환한 후 생긴 최종 문자열을 가지고, 신경망을 구성한다.

이렇게 생성된 신경망은 반드시 한 개이상의 입력노드와 한 개이상의 출력노드를 갖는 무정형의 신경망이 된다. 문제에 따라 적당한 수의 입력노드와 출력노드를 갖는 신경망을 선택하여, 주어진 입력을 넣고, 그 출력을 검사하여 원하는 결과와 비교한 후, 유전자 알고리즘에 의해 신경망을 진화시킨다. 신경망의 적합도는 식 (2), (3)에 의해 구한다.

$$Fitness = \frac{1}{1 + e^{-\lambda \cdot E}} \quad (2)$$

$$E = \sum_{i=1}^{n} (R_i - O_i)^2 \quad (3)$$

여기서 λ는 비례상수이며, R은 원하는 출력, O는 신경망의 출력이다.

이 집단을 진화시키기 위하여 교배와 돌연변이, 진화전략(Evolution Strategy)의 $(\mu + \lambda)$선택 방법을 사용하였다. 생성된 신경망의 DNA를 돌연변이와 교배를 통하여 3배수의 자손을 생성한 뒤, 원래 부모세대의 신경망과 같이 평가하여 순위선택으로 우수한 개체를 뽑아 다음세대의 부모 개체로 삼는다. 이 과정을 반복하여 점점 좋은 개체를 얻는다.

이 방법의 유효성을 검증하기 위해 Mackey-Glass 시계열 예측문제와 Sun-Spot 예측문제 그리고 KOSPI 예측문제를 해결할 수 있는 신경망을 구성하였다.

1) Connecting Range (x,y) : 문자열(노드의 배열)에서, 현재 노드에서 연결할 수 있는 범위를 x 번째 노드부터 y 번째 노드까지 정함. 단 쉼표(,)가 나올 경우 쉼표 뒤에 있는 노드는 연결하지 않으나 연결 범위에는 들어감.

표 2. DNA 코드표

Amino Acid	# of Amino Acid	Node's Name	Connecting Range
Leu	6	A	1,1
Arg	6	B	2,2
Ser	6	C	3,3
Thr	4	D	1,2
Ala	4	A	1,3
Gly	4	B	1,4
Val	4	C	2,3
Pro	4	D	2,4
Stop	3		
Ile	3	A	3,4
Tyr	2	B	4,4
Gln	2	C	1,1
Phe	2	D	2,2
Asp	2	,	3,3
Cys	2	,	1,2
Asn	2	,	1,3
Glu	2	,	1,4
His	2	,	2,3
Lys	2	,	2,4
Trp	1	C	3,4
Met	1	D	4,4

4. 시계열예측을 위한 DNA 코딩 방법

4.1 시계열 예측

시계열 예측이란 과거에 얻어진 데이터를 가지고, 미래의 값을 예측하는 것이다. 시계열 예측방법으로는 신경망을 비롯해 Genetic Programming 등 다양한 방법이 있다.

주어진 과거 데이터를 x, 예측기를 통해 얻어진 결과를 $\tilde{x}$, 과거 데이터의 집합을 벡터 x라고 하면,

$$\mathrm{x}(t) = (x(t), x(t-1), \ldots, x(t-\phi)) \quad (4)$$

로 표현 할 수 있다. ϕ개 만큼의 과거 데이터를 이용하여, x(t+1)를 예측한 값, $\tilde{x}(\mathrm{x}(t))$를 얻을 수 있다. 즉, 미래의 값, $x(t+\tau)$는 과거 값을 이용해 예측한 값 $\tilde{x}(\mathrm{x}(t))$로 얻을 수 있다. 여기서 $\tau = 1$이면 shot-term prediction 이라하고, τ가 2이상이면 long-term prediction이라고 한다.

4.2 Mackey-Glass 시계열 예측문제

Mackey-Glass 함수는 카오스시스템의 대표적인 예로 식 (5)와 같이 표현된다.

$$\frac{dx(t)}{t} = \frac{bx(t-\tau)}{1 + x(t-\tau)^c} - ax(t) \quad (5)$$

Mackey-Glass 시계열 예측 문제의 경우는 DNA 길이 500으로 무작위로 발생된 초기개체군을 시작으로 진화를 시작하였다. 교배율은 0.9, 돌연변이율은 0.3으로 설정하였다. 교배방법은 일점(one point)교배 방식을 사용하였고, 2개 부모개체의 DNA에서 같은 위치에서 교차가 이루어지게 하여 DNA의 길이는 유지하도록 하였다. 그리고 개체의 선택방법은 Ranking Selection과 ES의 $(\lambda + \mu)$ 선택법을 혼합

하여 선택하였다.

신경망의 입력은 과거 데이터 값으로 $x(t)$의 값을 구하는데 $x(t-1)\sim x(t-19)$의 값 중에서 구성된 신경망의 입력노드 수만큼 사용하였으며, 신경망 노드에서의 출력함수는 식 (6)을 사용하였다.

$$f(i)=(\frac{2}{1+e^{\eta\cdot i}}-1)\times 2 \qquad (6)$$

입력노드는 5개에서 19까지 가질 수 있으며, 출력노드는 하나만 갖는 신경망을 평가 대상으로 하여 250개의 데이터를 넣고, 출력 값을 계산하여 오차가 적은 신경망을 선택하여 다음 세대의 부모 개체로 삼는다.

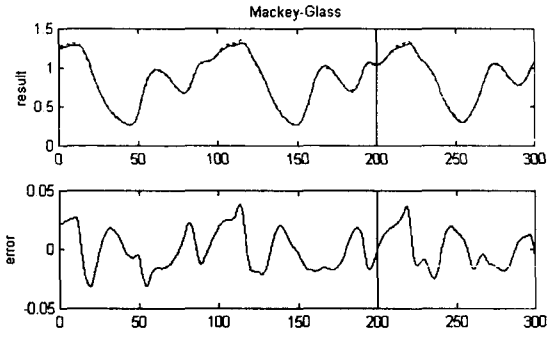

그림 3 Mackey-Glass 시계열 예측 결과
(… ideal - predicted)

4.3 Sun Spot 시계열 예측문제

Sun Spot 데이터는 1년 동안 태양의 흑점 수를 기록한 것으로, 일정한 규칙이나 흐름이 없이 관측된 과거 데이터만이 존재를 한다. 그러므로 예측방법도 과거의 데이터를 이용할 수밖에 없다. 본 논문에서는 Sun Spot 예측문제를 제안한 신경망에 적용시켜보았다. 200개의 학습데이터와 50개의 테스트 데이터를 사용하여 예측한 결과이다. 신경망의 발생을 위한 요소들은 Mackey-Glass문제와 동일하게 구성을 하였다.

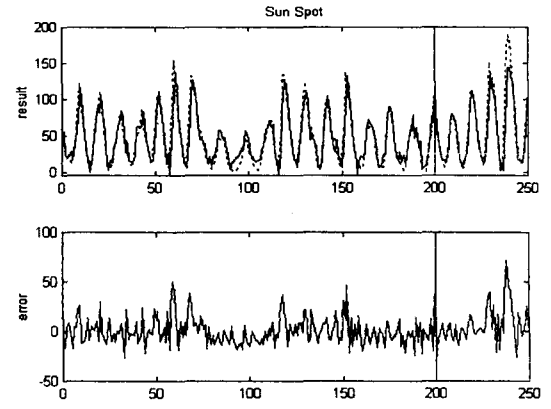

그림 4 Sun Spot 시계열 예측 결과
(… ideal - predicted)

4.4 KOSPI 예측

세 번째 적용문제는 국내 주식시세 예측문제이다. 예측을 위한 데이터는 1998년 2월부터 2000년 2월까지 얻은 350개 KOSPI(Korea Stock Price Index) 종합주가를 이용하였다. 250개를 진화를 위한 학습데이터로 나머지 100개를 테스트 데이터로 사용하였다. 그림 5는 L-system 신경망을 이용하여 KOSPI 데이터의 시계열 예측한 결과 및 오차 그래프이다.

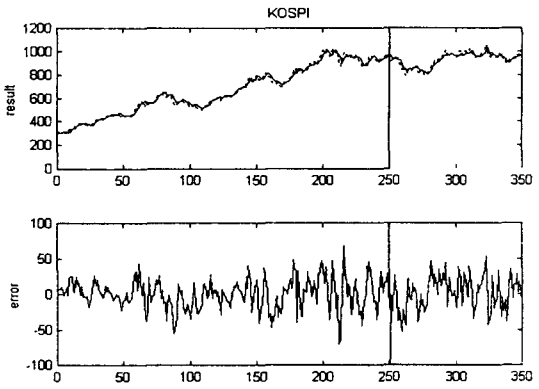

그림 5 KOSPI 데이터의 시계열 예측 결과
(…ideal - predicted)

5. 결 론

이 논문에서는 발생/발달 모델의 하나인 L-system과 DNA 코딩 방법을 이용하여 신경망을 진화시키는 방법을 제안하였다. DNA 코딩을 이용함으로서 L-system의 생성규칙을 짧은 DNA로 표현하였고, 생성된 규칙을 바탕으로 문자열을 구성한 후 이를 신경망으로 구성하였다. 이렇게 구성된 신경망을 Mackey-Glass 시계열 예측문제와 SunSpot 및 KOSPI 예측 문제에 적용하여 보았다.

이렇게 규칙을 통해 신경망을 구성함으로서, 작은 DNA의 변화로 신경망 전체에 큰 변화를 유도 할 수도 있다.

감사의 글
본 연구는 과학기술부의 뇌과학 프로젝트(Brantech21)의 지원으로 이루어진 결과임 (과제번호 : 98-J54-01-01-A-07)

참고문헌

[1] P. Prusinkiewicz, M. Hammel, J. Wolters, R. Mech, "Visual Models of Plant Development," *Hand Book of Formal Languages*, Springer-Verlag, 1996

[2] Aristid Lindenmayer, Prezemyslaw Prusinkiewicz, "Developmental Models of Multicellular Organisms : A Computer Graphics Perspective," *Artificial Life VI*, pp. 221-249, 1987

[3] Aristid Lindenmayer, "Mathematical Models for Cellular Interaction in Development, Part I, II," *Journal of Theoretical Biology*, vol.18, pp. 280-315, 1968.

[4] T. Yomohiro, T. Furuhashi, Y. Uchikawa, "A Combination of DNA Coding Method with Pseudo-Bacterial GA for Acquisition of Fuzzy Control Rules," *Proc. of 1st Online Workshop on Soft Computing*, Aug. pp. 19-30, 1996.

[5] R.A. Wallace, G.P. Sanders, R.J. Ferl, *BIOLOGY : The Science of Life 3rd eds.*, HarperCollins Publishers Inc., 1991.

[6] A.S. Wu, R.K. Lindsay, "A Comparison of the Fixed and Floating Building Block Representation in Genetic Algorithm," *Evolutionary Computation*, vol. 4, no. 2, pp. 169-193, 1996