工學碩士學位論文

시계열 예측에 사용된 유전자 프로그래밍에서 선택 과정의 영향에 관한 연구

A Study on Effects of Selection Schemes in Genetic Programming for Time Series Prediction

2000年 2月

서울大學校 大學院
컴퓨터工學科
金 正 集

시계열 예측에 사용된 유전자 프로그래밍에서 선택 과정의 영향에 관한 연구

A Study on Effects of Selection Schemes in Genetic Programming for Time Series Prediction

指導敎授 金 榮 澤

이 論文을 工學碩士 學位論文으로 提出함

1999年 10月

서울大學校 大學院

컴퓨터工學科

金 正 集

金正集의 工學碩士 學位論文을 認准함

1999年 12月

委 員 長 ________________ 印

副委員長 ________________ 印

委　　員 ________________ 印

초 록

선택은 진화 연산의 동작을 조정하는 중요한 연산자이다. 고정 길이의 염색체를 사용하는 진화 알고리즘에 대해서 다양한 선택 연산자들이 제시되고 실험되었다. 그러나 가변 길이의 표현을 사용하는 진화 알고리즘에 대해서는 선택 과정의 영향이 상대적으로 덜 연구되어 있다. 본 논문에서는 시계열 자료에 응용한 유전자 프로그래밍에 대해 다양한 선택 방법을 비교한다. 유전자 프로그래밍은 가변 길이의 트리 표현을 염색체로 사용하는 진화 연산이다. 시계열 예측은 실제로 많은 응용이 존재하기 때문에 벤치마크로써 선택했다. 우리는 breeder genetic algorithm (BGA)에서 제시된 선택 반응과 선택 미분을 통해 유전자 프로그래밍의 진화적 동작을 분석하였다. 비교한 선택 방법은 비례 선택, 순위 선택과 토너먼트 선택이다. 레이저 시계열 데이터를 사용한 실험 결과는 강한 선택이 약한 선택보다 더 낫다는 것을 보여준다. 이것은 유전자 프로그래밍에서 유전적 상속성의 부재에 인한 것이다.

주요어 : 유전자 프로그래밍, 시계열 예측, 선택 방법, 선택 미분, 선택 반응, 유전적 상속성

1. 서론

1.1 연구 배경

선택(selection)은 진화 알고리즘에서 중요한 역할을 차지한다. 다른 유전 연산자들이 특정 지향성을 갖지 않고 새로운 탐색점을 제시하는데 반해, 선택은 탐색의 방향성을 결정한다. 선택 압력(selective pressure)는 탐색의 방향성에 영향을 끼치는 선택의 강도를 가리키는데 사용된다 [Bäck-96]. 강한 선택은 정보의 지역적 탐색을 강조하고 빠른 수렴속도를 이끈다. 반대로 약한 선택은 유전적 다양성(genetic diversity)과 전역적 탐색 경향을 강조하고, 수렴성의 안정도를 유도한다. 다양한 선택 방법의 다른 특징들은 진화 알고리즘의 효율적인 응용과 새로운 알고리즘의 설계와 분석의 필수 조건이다.

역사적으로 여러 진화 연산 학계들이 서로 다른 선택 방법을 개발해 왔다. 진화 전략(evolution strategy) 학계들은 다양한 순위 선택(ranking selection) 방법을 계획했다. 토너먼트 선택(tournament selection) 방법은 진화 프로그래밍(evolutionary programming)에서 개발되었다. 유전자 알고리즘(genetic algorithm)은 비례 선택(proportional selection)만을 사용한다. Goldberg와 Deb은 몇몇 선택 방법의 수렴 시간(convergence time)과 증가율(growth ratio)에 대해 분석했다 [Goldberg-Deb-91]. Back과 Hoffmeister는 점유 시간(takeover time)과 유전적 다양성(genetic diversity)의 관계를 연구했다 [Bäck-Hoffmeister-91].

Mühlenbein과 Schlierkamp-Voosen은 집단 유전학적인 어휘인 선택 미분(selection differential)과 선택 응답(response to selection)을 진화 연산 학계에 소개했다 [Mühlenbein-SV-93]. 그들은 이런 평균 적합도 개념에 기반하여 진화 알고리즘의 행동을 예측하게 하는 예측 모델을 제시했다. Blickle과 Thiele은 선택 후의 개체군의 적합도 분포의 평균 분산으로 선택

분산(selection variance)를 정의하고 적합도 값을 예측하기 위해 토너먼트 선택의 수학적인 분석을 완성했다 [Blickle-Thiele-95]. 그러나 선택 과정의 영향은 유전자 프로그래밍(genetic programming)과 같은 가변길이 표현을 가지는 진화 알고리즘의 분야에선 상대적으로 밝혀지지 않았다.

1.2 연구 내용

본 논문의 목적은 시계열 예측에 적용된 유전자 프로그래밍에 대한 다양한 선택방법의 성능 특성을 밝히는 것이다. 주요 선택 방법을 그 선택 압력을 변화시켜 가며 동작시켜서 완성된 해의 질을 비교했다. 여기서 비교한 선택 방법은 비례 선택, 순위 선택과 토너먼트 선택이다.

구성은 다음과 같다. 2장에서는 사용한 시계열 문제를 정의하였고, 3장에서는 이 문제에 대한 유전자 프로그래밍적인 접근 방법을 서술했다. 실험결과는 4장에 정리했고, 5장에서는 선택 방법의 진화적 동작을 선택 미분과 선택 응답을 측정함으로써 분석했다. 마지막으로 6장에서는 결론을 이야기했다.

2. 시계열 예측

2.1 시계열 자료

시계열 자료는 시간순서로 순차적으로 관측한 값들의 나열이다. 매달 공장에서 적재된 재화의 양, 매주 자동차 사고의 수, 매시간 관측한 화학 공정에서의 수율등 다양한 데이터 집합이 시계열로 존재한다. 이러한 시계열 데이터의 예는 경제학, 경영학, 공학, 자연과학(지질물리학, 기상학)등에 널리 존재한다. 시계열 자료의 내재적인 특징은 인접한 관측값이 서로 의존적이라는 것이다. 관측값 사이의 이 유사성은 실제적으로 시계열을 매우 흥미롭게 한다. 시계열 분석은 이러한 의존성의 분석을 위한 방법이라고 할 수 있다.

2.2 시계열 예측

시계열 예측은 주어진 임의 변수의 과거값의 순열로부터 해당 변수의 미래값을 예측하려는 것이다. 과거값 $X(t)$은 다음과 같은 벡터 형태로 표현한다.

$$X(t) = (x(t), x(t-1), \ldots, x(t-\phi)). \tag{1}$$

미래값 $x(t+\tau)$는 과거값들의 함수 $f(X(t))$로 예측한다. τ의 값은 과거의 값들로 얼마나 먼 미래의 값까지 예측할 지를 정하는 인자인데, $\tau = 1$일 경우를 단기 예측이라 하고 1이상일 경우 장기 예측이라고 한다. ϕ는 미래값의 예측을 위해 얼마나 먼 과거의 값까지 사용할 지를 나타내는 인자이다. 본 논문에서는 $\phi = 2$와 $\tau = 1$의 단기 예측 문제만를 고려한다. 이 경우 시간 $t+1$에서 입력값

$$X(t) = (x(t), x(t-1), x(t-2)) \tag{2}$$

으로부터 예측값 $x(t+1)$을 계산한다.

모델 f_k의 예측 정확도는 다음과 같은 정규화된 평균 제곱 에러 (normalzed mean squared error, NMSE)를 예측함으로써 계산가능하다.

$$E_k = NMSE(k) = \frac{\sum_{t=1}^{N} (x(t+1) - f_k(X(t)))^2}{\sum_{t=1}^{N} (x(t+1) - \bar{x})^2}$$
$$\approx \frac{1}{N \times Var} \times \sum_{t=1}^{N} |x(t+1) - f_k(X(t))|^2 \quad (3)$$

$$Var = \frac{1}{N} \times \sum_{t=1}^{N} \left| x(t+1) - \frac{1}{N} \times \sum_{t=1}^{N} x(t+1) \right|^2 \quad (4)$$

시계열 예측 문제는 NMSE**값을 최소화하는 예측자**(predictor) $f_k(X(t))$ **를 찾는 문제로 귀결된다**.

2.3 실험 데이타

본 논문의 실험 데이터들은 카오스 상태에 있는 원적외선 레이저에서 기록한 것이다.

다음은 Huebner박사의 기술이다 [Huebner-89-1, Huebner-89-2, Huebner-89-3]. [Zhang-et-al-97].

측정은 81.5미크론 14NH3 cw (FIR) 레이저에서 이루어 졌다. 진동성 aQ(8,7) NH3 전이를 통한 N2O 레이저의 P(13)라인으로 광학적으로 펌프된 것이다. 기본적인 laser 설정은 참고문헌 1에서 찾을 수 있다. 강도 데이터는 LeCroy 오실로스코프로 기록한 것이다. 더 이상의 처리는 하지 않았다. 실험시의 s/n비는 300정도이었다. 이것은 a/d변환 사이에 반 bit정도의 불확실성을 의미한다.

이 데이터는 레이저의 카오스적인 강도 맥동(intensity pulsation)의 주기에서 추출한 것이다. 카오스적인 맥동은 다소 이단계 시스템의 이론적인 로렌츠 모델을 따른다.

실험 데이터로 레이저 데이터를 선택한 이유는 크게 두 가지이다. 첫째는 이 데이터가 겉으로는 내재하는 동작구조가 잘 이해되는 것처럼 보이지만 실제로는 복잡한 행동을 나타내는 좋은 예이기 때문이다. 실제로 상대적으로 단순한 진동이 있어서 짧은 시간 간격에서는 쉽게 예측 가능하지만, 진동의 빠른 붕괴 때문에 전체적인 추이는 예측하기 어려운 것이다.

두 번째는 데이터의 크기가 작기 때문이다. 경제학과 같은 많은 영역에서 데이터 집합은 수백개 정도의 점들로 구성되어 있어서, 이런 짧은 데이터 집합을 분석하는데 많은 전문적인 기술이 개발되었다. 데이터 집합의 크

기는 1000개로 사람들이 흔히 분석하는 짧은 시계열보다는 길지만, 다른 기술을 요구하는 길이보다는 짧다. 작업을 더 관리하기 쉽게 하기 위해 짧은 데이터 집합을 분석했는데, 저차원의 역학구조를 가졌다고 알려진 데이터의 부분집합을 테스트 케이스로 선택했다.

[Zhang-et-al-97] 에 따라 예측자를 진화시키기 위해 짝수번째의 데이터 점들만을 사용하였다. 그림 1과 2는 각각 실험에 사용된 학습 데이터와 검사 데이터를 보여준다.

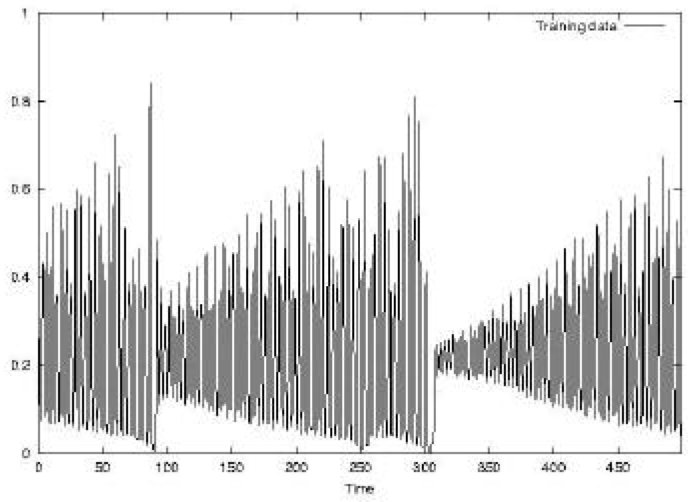

그림 1 학습 데이터

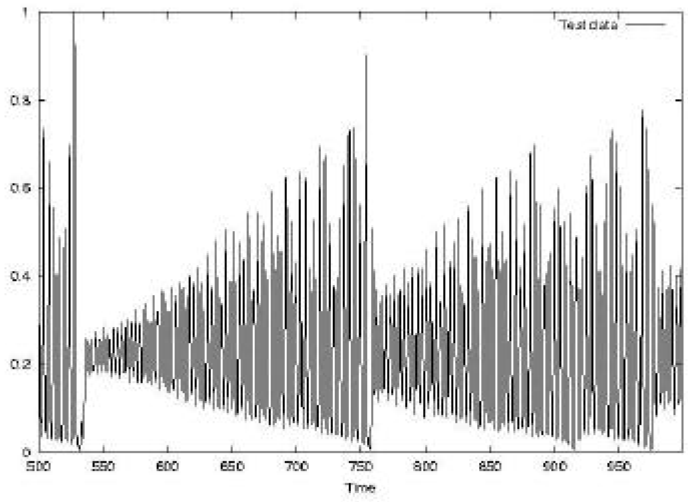

그림 2 검사 데이터

3. 시계열 예측에 사용한 유전자 프로그래밍

3.1 유전자 프로그래밍

본 논문에서는 시계열 예측 문제를 푸는데 유전자 프로그래밍을 사용하였다. 유전자 프로그래밍은 유전 알고리즘의 한 분야로써 Koza에 의해 완성되었다 [Koza-92]. 기존의 유전 알고리즘은 고정 길이의 염색체를 대상으로 유전 연산자를 적용하는데 비해, 유전자 프로그래밍은 가변 길이의 염색체를 사용할 수 있다.

일반적으로 트리 형태의 염색체를 많이 사용하는데 일반적인 프로그램을 트리 형태로 변환가능하므로 유전자 프로그래밍의 표현 능력은 일반적인 프로그램과 동일하다고 할 수 있다. 유전자 프로그래밍의 염색체가 유전자 프로그램이라 불리기도 하는 이유는 바로 이것 때문이다.

유전자 프로그래밍의 성능을 높이기 위해서는 문제에 적합한 트리의 노드의 선택이 중요하다.

염색체의 모양이 트리모양이 되므로 유전 연산자도 달라지게 된다. 선택의 경우 염색체의 모양에 의존하는 부분이 없고 단지 그 적합도에 의존하므로 변화가 없게 된다. 흔히 쓰이는 돌연변이와 교차의 경우 여러가지 변이가 존재한다. 돌연변이는 임의의 서브트리를 임의의 트리로 바꾸는 과정을 많이 사용한다. 교차는 두 부모의 임의의 서브트리를 교환하는 과정을 사용한다.

유전자 프로그래밍의 기본적인 과정은 표 1에 요약되어 있다. 최초의 개체군은 임의로 생성된다. 유전자 연산자들은 개체들을 수정하고 비교적 좋은 개체들만이 다음 세대로 선택된다.

1. 함수의 집합 F과 단말의 집합 T을 사용하여 임의의 트리들로 임의의 개체군 $A(0)$을 생성한다. $g \leftarrow 1$. 2. 적합도 사례의 집합 D에 동작시켜서 $A(g)$에 있는 모든 프로그램의 적합도를 평가한다. 3. 개체군 $A(g)$에 선택, 교차, 돌연변이 연산자를 적용하여 새 개체군 $A(g+1)$을 만든다. 4. 적합한 해가 만들어지면 동작을 마친다. $g \leftarrow g+1$. 2번 과정으로 간다.

표 1 유전자 프로그래밍의 동작

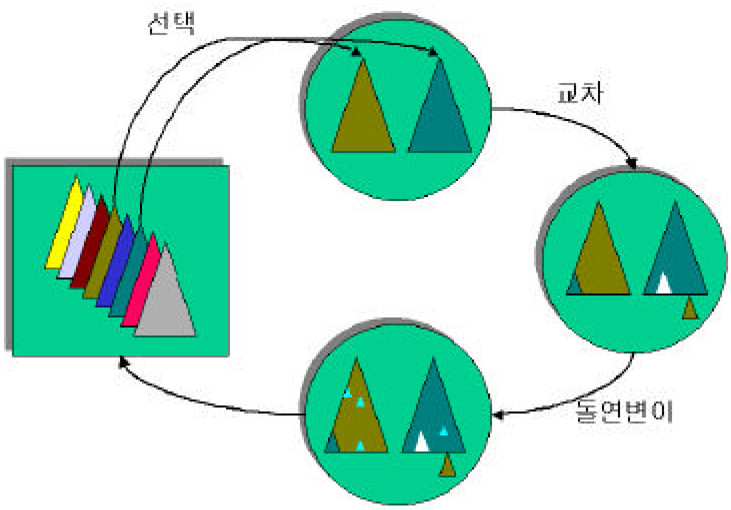

그림 3 유전자 프로그래밍의 일반적인 동작 과정

3.2 시계열 예측에 사용한 유전자 프로그래밍

여기서 유전자 프로그램들은 시계열 데이터에 대한 모델로 사용되었다. 유전자 프로그램 A_i의 집합은 개체군 A를 구성한다.

$$A = \{A_1, A_2, \ldots, A_M\} \tag{5}$$

유전자 프로그램은 기본 함수와 단말 심볼로 구성된다. 기본함수와 단말 심볼은 각각 함수 집합 $F = \{+, -, \times, \%\}$과 단말 집합 $T = \{x(t), x(t-1), x(t-2), R\}$의 원소들을 사용하였다. %는 0으로 나눌 때 1을 반환하는 보호 나눗셈을 가리키고, R은 $[0, 1]$의 범위에서 임의로 뽑는 난수의 단말을 말한다. 그림 4은 유전자 프로그램의 한 예를 보여준다. 이 프로그램이 나타내는 함수는 다음과 같다.

$$\begin{aligned} f(t) = {} & (x(t-2) + 0.48) \times x(t-2) \times 0.48 \\ & \% \; x(t-1) \; \% \; (x(t-2) + x(t)) \times x(t) \end{aligned} \tag{6}$$

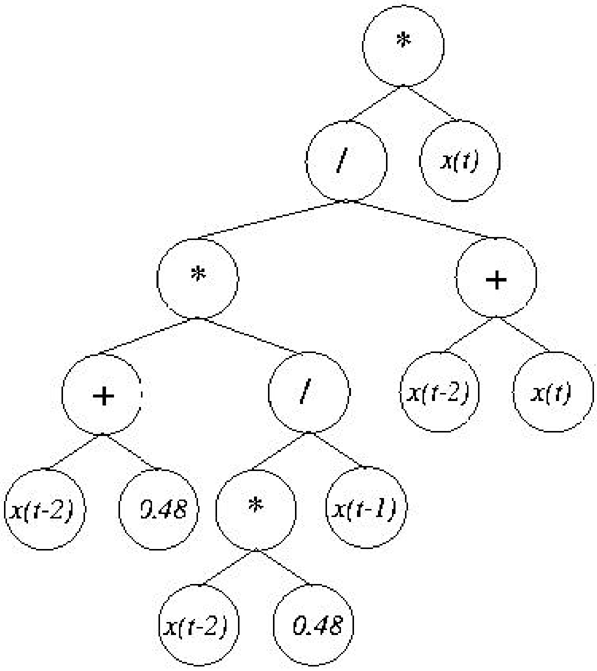

그림 4 시계열 예측에 사용된 유전자 프로그램

적합도 평가를 위해서, 유전자 프로그램 A_k는 학습 데이타

$$D = \{(X(t), x(t+1))\}_{t=1}^{N} \tag{7}$$

로부터 $f_k(X(t))$를 계산하기 위해 사용된다. 이 f_k값을 식 3에 대입하여 NMSE 에러 E_k를 계산한다. 개체 A_k의 적합도는 다음과 같이 평가된다.

$$F_k = \frac{1}{1 + E_k + (K \cdot S_k)}, k = 1, 2, \ldots, M \quad (8)$$

여기서 S_k는 노드 수로 계산되는 트리의 크기이고 K는 트리의 복잡도에 불이익을 주기 위한 작은 크기의 상수이다.

염색체의 적합도 값이 평가되면, 비교적 적합한 개체들이 다음 세대로 선택된다. 이 선택 방법은 다음 장에서 설명된다.

선택 과정 다음에는 유전자 프로그램들이 돌연변이나 교차같은 유전 연산자들에 의해 변경된다. 돌연변이는 트리의 노드들의 값을 돌연변이 확률에 따라 바꾼다. 교차는 두 부모 개체의 서브트리를 교환한다. 부모의 선택은 각 선택 방법에 따라 다르게 된다. 이 과정은 미리 지정한 최대 세대까지 세대가 진행하거나 다른 종료 조건이 만족할 때만 종료한다.

모든 선택 방법에 대해 엘리트 전략을 채용하였다. 이 방법은 최적의 개체는 항상 다음 세대로 유지하는 것이다. 또한 가중 데이터 상속(incremental data inheritance, IDI)를 사용하였다 [Zhang-Joung-99-GP]. 이 방법은 모든 선택 방법의 적합도 평가 과정을 빠르게 하지만 선택방법의 상대적인 질에는 영향을 끼치지 않는다.

매개 변수	값
개체군의 크기 M	200
최대 세대수 G_{max}	500
실행 횟수 R	20
교차 확률 P_c	0.9
돌연변이 확률 P_m	0.1
복잡도 페널티 K	0.00001
트리의 최대 깊이 L	20
단말	$x(t), x(t-1), x(t-2), R$
함수	$+, -, \times, \%$

표 2 **실험에 사용한 매개 변수값**

4. 실험및 결과

4.1 선택 방법

본 논문에서는 유전자 프로그래밍에서 3가지 주요한 선택 방법인 비례, 토너먼트, 순위선택의 영향을 비교하였다. 비례선택의 경우, 각 개체들은 그들의 상대적인 적합도 값에 따라 선택된다. g번째 세대에서 i번째 개체의 선택될 확률(selection probalilty)는 다음과 같이 정의 된다.

$$p_s(A_i^g) = \frac{f(A_i^g)}{\sum_{j=1}^{\lambda} f(A_j^g)} \quad (9)$$

이 방법은 확률적인 선택 방법으로 0이 아닌 적합도를 가지는 모든 개체들은 복제되기 위해 선택될 가능성이 있다. 이 선택 방법은 단순 유전자 알고리즘에 채택되었다. 그리고 자연에서 일어나는 선택 과정과 가장 유사하다고 믿고 있다.

토너먼트 선택 [Blickle-Thiele-95] 은 한 그룹의 부모를 임의로 선택하고 이 중에서 가장 좋은 개체만을 복제하는 방법이다. 부모의 수가 q일 때, 이 방법을 q-토너먼트 선택이라고 부른다. 이 과정을 λ번 반복하면 다음 세대의 개체군을 형성하게 된다. q-토너먼트 선택의 선택 확률은 다음과 같이 주어진다.

$$p_s(A_i^g) = \frac{1}{\lambda^q}((\lambda - i + 1)^q - (\lambda - i)^q) \quad (10)$$

순위 선택은 선택 확률을 개체의 적합도값은 무시하고, 그 순위에 기초해서 정하는 선택 방법이다. (μ,λ) 균일 선택 방법 [Schwefel-95] 에서는

가장 좋은 μ개의 개체들이 $\frac{1}{\mu}$의 선택 확률을 갖고, 나머지는 무시된다.

$$p_s(A_i^g) = \begin{cases} \frac{1}{\mu}, & 1 \le i \le \mu \\ 0, & \mu < i \le \lambda \end{cases} \tag{11}$$

4.2 실험 결과

선택 방법의 성능은 해의 질로 비교하였다. 표 3**은 각 선택 방법마다** 20**번의 실행을 해서 얻은 최적의** NMSE**값을 나타낸다.**

	비례 선택	**토너먼트 선택**	**순위 선택**	MIP
학습	0.133	0.063	0.070	0.060
검사	0.204	0.116	0.131	0.097

표 3 NMSE**의 비교**(MIP**는** [Angeline-98] **의 결과이다.**)

그림 4**와** 5**는 다음과 같은 식을 표현하는 유전자 프로그램의 예측값과 예측 에러를 보여준다.**

$$\begin{aligned} f(t) = {} & (((v0 \% (((((((v1 + v0 \% ((v2 \% v0) \\ & \times (v1 \% (v2 - v1)))))) \% ((((v1 + (v1 \\ & + (v0 \% ((v2 \% v0) \times 0.67)))) + v1) \\ & \times 0.81)) \% (v1 \% (v1 \times 0.79))) + v2) \\ & + 0.79) \times v2) + v0)) \% (v1 \% 0.51)) \times v2) \end{aligned} \tag{12}$$

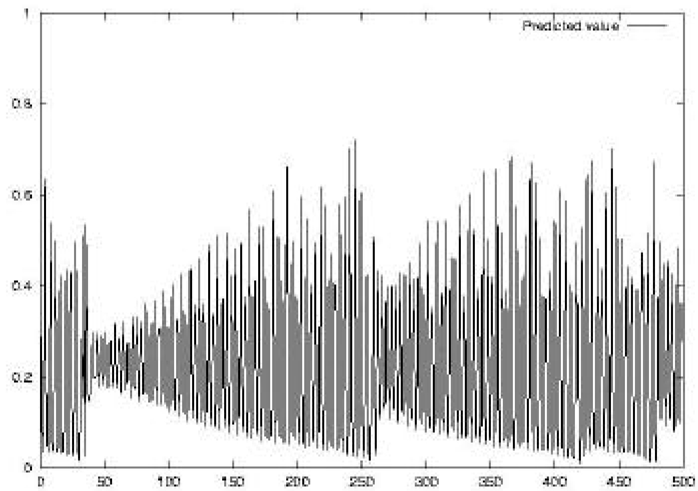

그림 5 예측 값

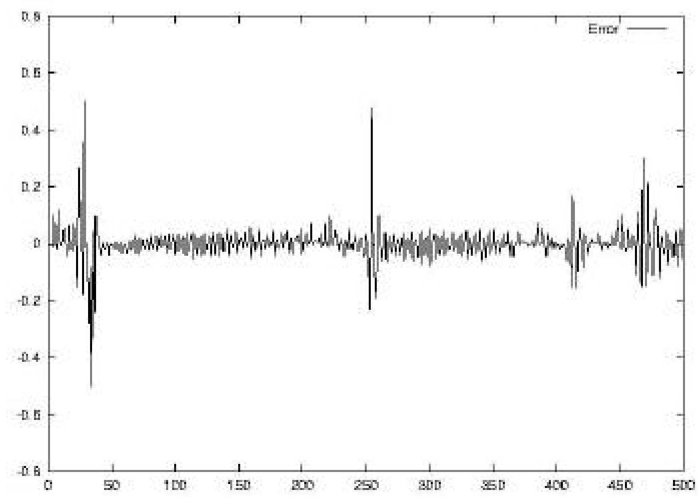

그림 6 예측 에러

그림 7-9는 세가지 선택 방법의 NMSE 에러 값을 보여준다. 각각의 선택 방법에 대해 선택 압력을 조절하는 알고리즘에 한정된 매개변수들을 변화시켰다. 이 인자들은 각기 적합도 스케일 (fitness scaling), 토너먼트 크기 (tournament size)와 절삭 한계치 (truncation threshold)이다. 결과적으로 강한 선택이 모든 선택 방법중에서 가장 좋은 성능을 얻었다. 예를 들어, 토너먼트 크기 100은 토너먼트 크기 5와 2의 것보다 좋은 성능을 나타냈다 (그림 9).

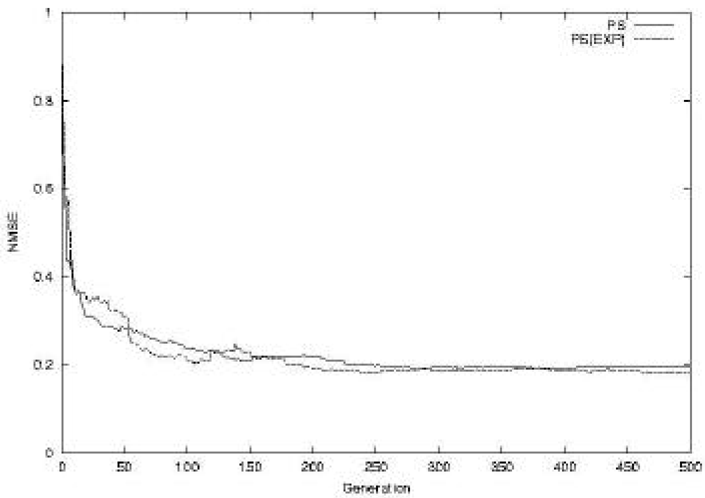

그림 7 비례 선택에 대한 NMSE값의 진화

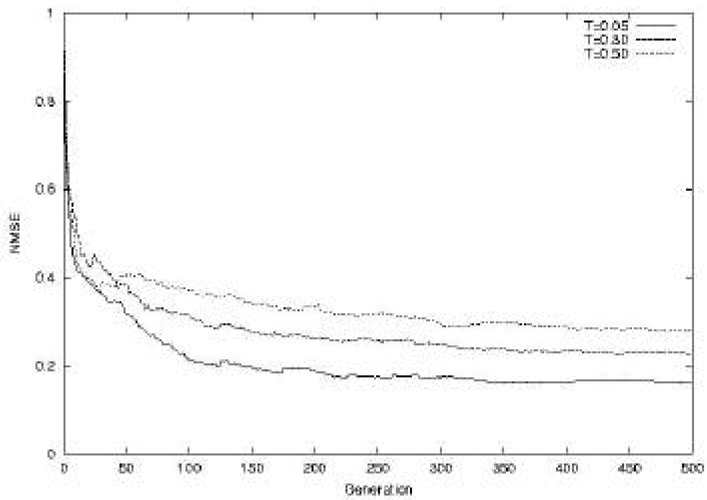

그림 8 **순위 선택에 대한** NMSE**값의 진화**

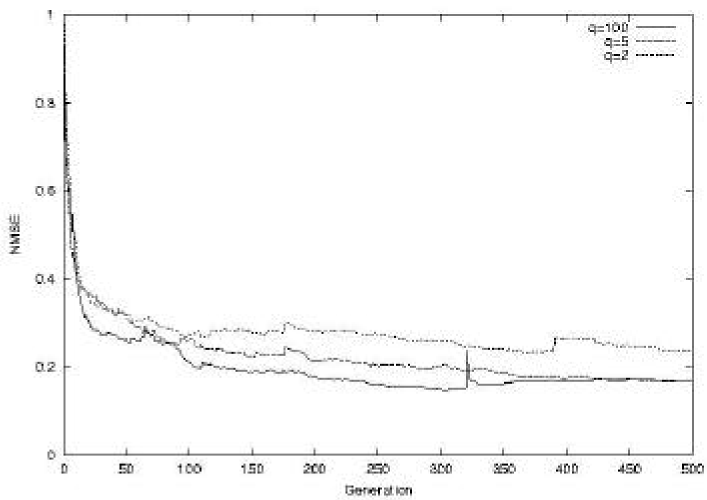

그림 9 **토너먼트 선택에 대한** NMSE**값의 진화**

5. 분석

진화 알고리즘에서의 선택은 최적의 개체의 적합도 [Goldberg-Deb-91, Bäck-96], 개체군의 평균 적합도 [Mühlenbein-SV-93], 개체군의 적합도 값의 분포 [Blickle-Thiele-95] 라는 3가지 다른 수준에서 분석할 수 있다. 우리는 다음 과정에서 실험결과를 이론적인 값과 비교하여 분석할 것이다. 이 분석은 최적의 개체에 기반을 둔다. 또한 선택 미분과 선택 반응을 통해 선택 과정의 행동을 검사할 수 있다. 이 들은 평균 적합도 값에 기반을 선택 압력을 측정하는 것이다.

5.1 선택 반응

우리는 레이저 시계열 자료에 적용된 유전자 프로그래밍의 개체군 적합도의 동작을 분석하였다. 선택 미분과 선택 반응의 개념은 Mühlenbein과 Schlierkamp-Voosen에 의해 제시된 것이다 [Mühlenbein-SV-93]. 선택 반응 $R(g)$는 $g+1$세대의 개체군의 평균 적합도와 g세대의 것의 차로 정의된다. $R(g)$는 개체군의 평균 향상을 측정한다.

$$R(g) = M(g+1) - M(g) \qquad (13)$$

여기서 $M(g)$는 g세대의 개체군의 평균 적합도를 나타낸다.

선택 미분은 선택된 부모들의 평균 적합도 $M_s(g)$와 개체군의 평균 적합도 차이로 정의된다.

$$S(g) = M_s(g) - M(g) \qquad (14)$$

이 측정값들은 진화 알고리즘의 동작을 예측하는데 유용하다. 다시 말해

서, 알고리즘의 설계자는 다음 식에 기반해서 $S(g)$의 값에서 $R(g)$값을 예측하려고 시도할 수 있다.

$$R(g) = b_g S(g) \tag{15}$$

b_g는 정량 유전학(quantitative genetics)에서 realized heritability라고 불린다.

b_g를 일정 세대 동안 상수라고 가정하면 s세대 동안의 누적 반응 R_s는 다음과 같이 예측가능 하다.

$$R_s = \sum_{g=1}^{s} R(g) = b \sum_{g=1}^{s} S(g) \tag{16}$$

선택 반응은 상속성(heritability)과 선택 미분의 곱이 된다.

그림 10-12는 각각의 선택 방법을 통해 20번의 실행동안의 $S(g)$값의 평균값을 표시한 것이다. 이 결과와 그림 6-8에 나타난 NMSE 결과를 비교하면 일반적으로 유전자 프로그래밍에서 강한 선택이 약한 선택보다 낫다는 것을 보여준다. 이유는 낮은 $R(g)$값을 관측함으로써 알 수 있다 (그림 13-15). 선택 반응이 상속성과 선택 미분의 곱이므로, 이 결과는 적어도 여기서 구현한 유전자 알고리즘은 낮은 상속성을 갖는다는 것을 의미한다.

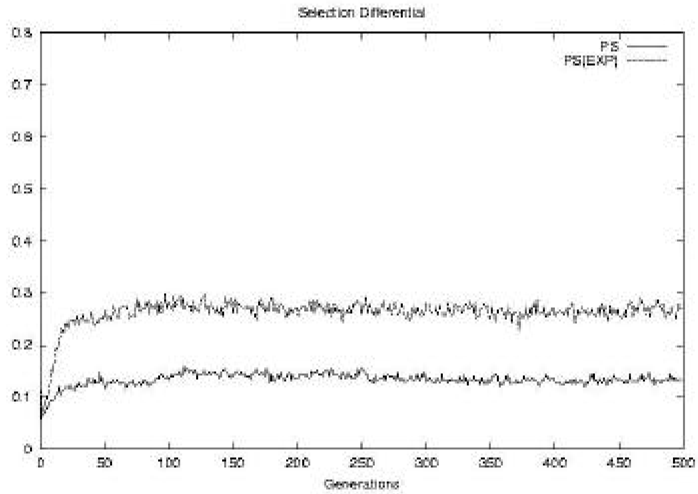

그림 10 비례 선택에 대한 선택 미분

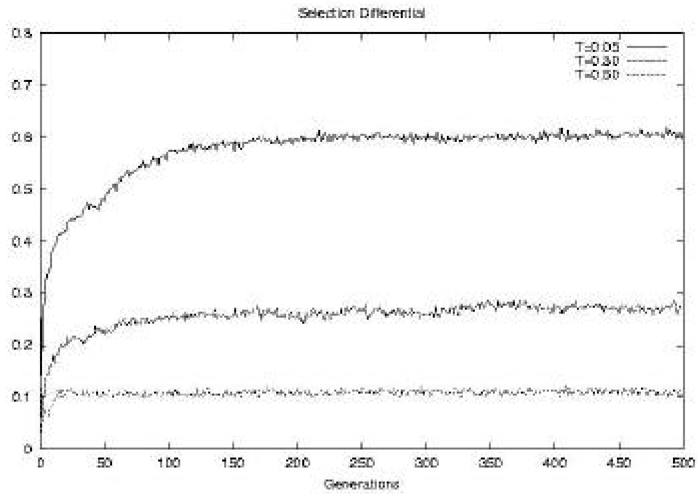

그림 11 순위 선택에 대한 선택 미분

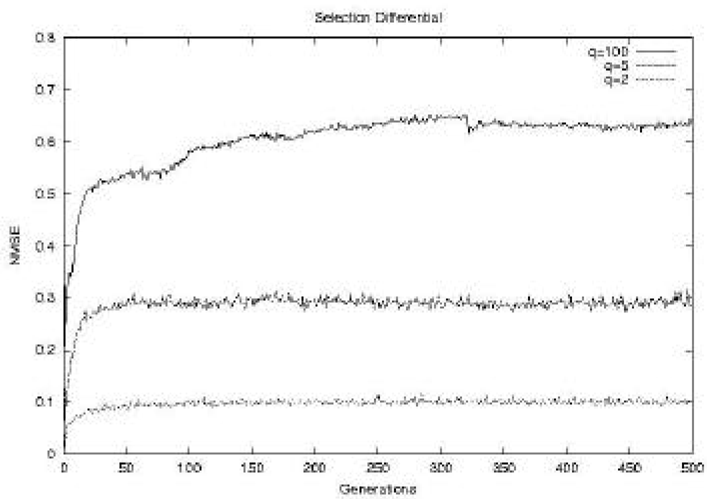

그림 12 **토너먼트 선택에 대한 선택 미분**

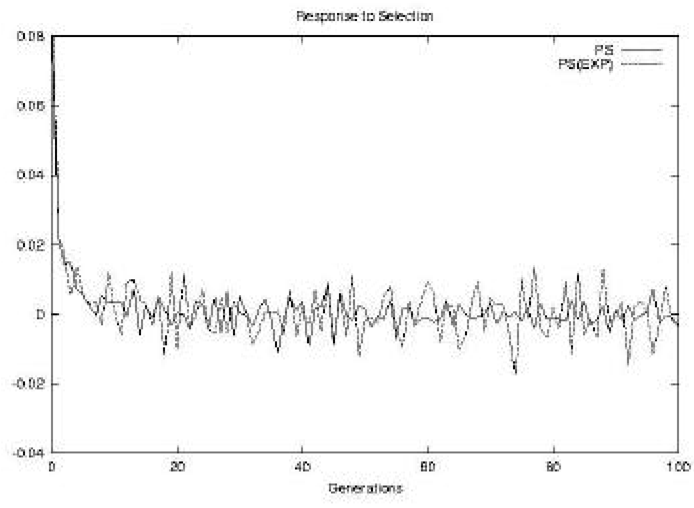

그림 13 **비례 선택에 대한 선택 응답**

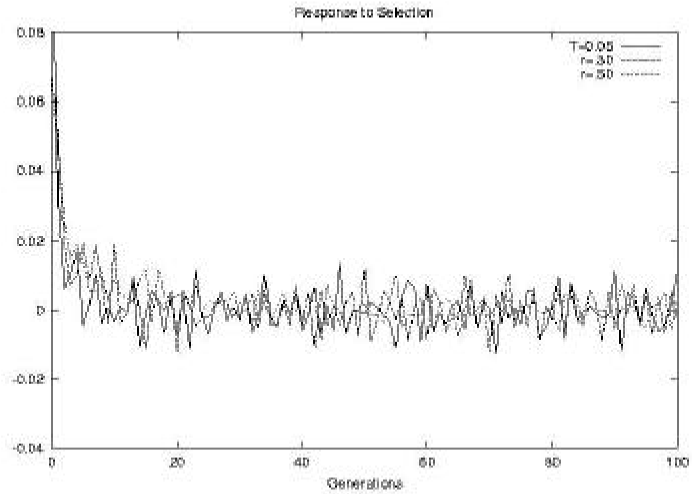

그림 14 **순위 선택에 대한 선택 응답**

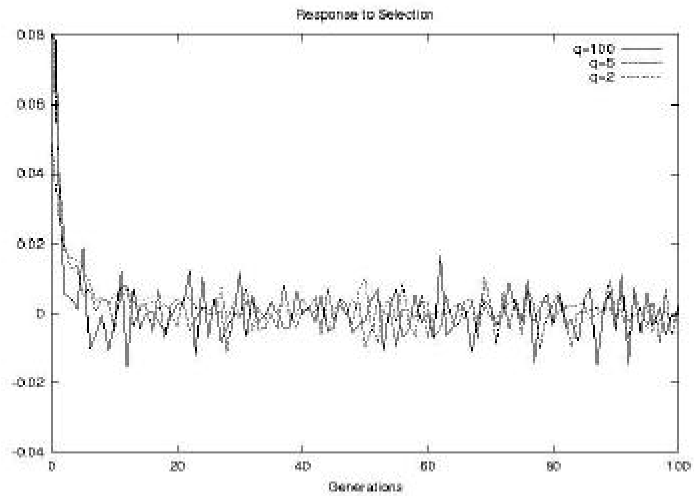

그림 15 **토너먼트 선택에 대한 선택 응답**

6. 결론

본 논문에서는 시계열 예측 문제의 문맥에서 유전자 프로그래밍에 대한 세가지 주요한 선택 과정을 비교해 보았다. 다른 유전 연산자는 동일하게 한 후에 선택 압력의 해의 질에 대한 영향력도 비교하였다.

선택 방법의 특성을 이해하기 위해, 선택 미분과 선택 반응을 측정함으로써 다른 선택 방법의 유전자 프로그래밍의 동작을 분석하였다. 각각의 선택 방법마다 강한 선택 압력을 가진 유전자 프로그래밍이 제일 좋은 결과를 낸다는 것이 판명되었다. 그 이유는 현재의 유전자 프로그래밍이 낮은 선택 반응값에 의해 표시되는 낮은 상속성을 가지기 때문이다. 큰 선택 미분과 작은 상속성의 결합은 효과적인 유전자 프로그래밍 탐색을 위한 중요한 필요조건이다.

참고문헌

[Angeline-98] Angeline, P. J., Evolving predictors for chaotic time series, *Applications and Science of Computational Intelligence: Proceedings of SPIE*, Volume 3390, S. Rogets, D. Fogel, J. Bezdek, and B. Bosacci (Eds.), SPIE, Bellingham, WA, pp. 170-180, 1998.

[Bäck-96] Bäck, T., *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, New York, 1996.

[Bäck-Hoffmeister-91] Bäck, T. and Hoffmeister, F., Extended selection mechanisms in genetic algorithms, In Belew, R.K. and Booker, L.B. (eds.), *Proc. 4th Int. Conf. on Genetic Algorithms*, Morgan Kaufmann, pp. 92-99, 1991.

[Blickle-Thiele-95] Blickle, T. and Thiele, L., A mathematical analysis of tournament selection, In *Proc. Sixth Int. Conf. on Genetic Algorithms*, Morgan Kaufmann, pp. 9-16, 1995.

[George-et-al-94] George E.P. Box, Gwilym M. Jenkins and Gregory C. Reinsel, *Time series analysis : forecasting and control*, Prentice Hall 1994.

[Goldberg-Deb-91] Goldberg, D. E. and Deb, D., A comparative analysis of selection schemes used in genetic algorithms, In Rawlins, G.J.E. (ed.) *Foundations of Genetic Algorithms*, pp. 69-93, 1991.

[Huebner-89-1] U. Huebner, N. B. Abraham, and C. O. Weiss: ``Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared NH3 laser.'' *Phys. Rev.* A 40, p. 6354 (1989)

[Huebner-89-2] U. Huebner, W. Klische, N. B. Abraham, and C. O. Weiss: On problems encountered with dimension calculations. *Measures of Complexity and Chaos*; Ed. by N. B. Abraham et. al., Plenum Press, New York 1989, p. 133

[Huebner-89-3] U. Huebner, W. Klische, N. B. Abraham, and C. O. Weiss: Comparison of Lorenz-like laser behavior with the Lorenz model. *Coherence and Quantum Optics VI*; Ed. by J. Eberly et. al., Plenum Press, New York 1989, p. 517

[Kim-Zhang-98] Kim, J.-J. and Zhang, B.-T., Comparison of selection schemes for machine layout design, In *Proc. Asia-Pacific Conf. on Simulated Evolution and Learning* (SEAL-98), Canberra, Australia, 1998.

[Koza-92] Koza J. R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.

[Mühlenbein-SV-93] Mühlenbein, H. and Schlierkamp-Voosen, D., Predictive models for the breeder genetic algorithm: Continuous parameter optimization, *Evolutionary Computation*, 1(1):25-49, 1993.

[Mulloy-Riolo-Savit-96] Mulloy, B.S., Riolo, R.L. and Savit, R.S., Dynamics of genetic programming and chaotic time series prediction, In *Genetic Programming 1996*, pp. 166-174, 1996.

[Schwefel-95] Schwefel, H.-P., *Evolution and Optimum Seeking*, Wiley, New York, 1995.

[Weigend-Gershenfeld-94] Weigend, A. and Gershenfeld, N., *Time Series Prediction: Forecasting the Future and Understanding the Past*, New York: Addison-Wesley,1994

[Whitley-89] Whitley, D., The Genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best, In *Proc. Third Int. Conf. on Genetic Algorithms*, Morgan Kaufmann, pp. 116-121, 1989.

[Zhang-Joung-99-GP] Zhang, B.-T. and Joung, J.-G., Genetic programming with incremental data inheritance, *Genetic and Evolutionary Computation Conference* (GECCO-99), Orlando, Florida, Morgan Kaufmann, 1999.

[Zhang-et-al-97] Zhang, B.T., Ohm, P., and Mühlenbein, H., Evolutionary induction of sparse neural trees, *Evolutionary Computation*, 5(2):213-236, 1997.

Abstract

Selection is an important operator that controls the dynamics of evolutionary computation. A number of selection operators have been so far proposed and tested in evolutionary algorithms with fixed-size chromosomes. However, the effect of selection schemes remains relatively unexplored in evolutionary algorithms with variable-size representations. In this thesis, we compare various selection methods for genetic programming, an evolutionary algorithm with variable-size tree representations, with application to time series data. Time series prediction is chosen as a benchmark domain since it has many applications in practice. We analyze the evolutionary dynamics of genetic programming by means of the selection to response and the selection differential proposed in the breeder generic algorithm (BGA). The repertoire of selection methods compared includes proportional selection, truncation selection and tournament selection. Empirical analyses using the laser time-series data suggests that hard selection is in general more preferable than soft selection. This seems due to the lack of heritability in genetic programming.

Keyword: genetic programming, time series prediction, selection methods, selection differential, response to selection, heritability

감사의 글

대학원에 들어왔을 때처럼 대학원에서 나갈 때에도 빠른 시간을 느낍니다. 짧은 생활이었지만, 학문의 어려움과 저 자신의 부족함을 느끼게 되었습니다. 그런 제가 석사 논문을 쓰고 나가게 된다니 참으로 감개무량합니다.

일단 낳아주시고 길러주신 부모님께 감사드립니다. 컴퓨터 공학자가 가야할 길을 일러주신 김영택 교수님, 부족한 제자를 지도해주시느라 고생하신 장병탁 교수님과 그 외 컴퓨터 공학과 교수님들께도 감사드립니다.

연구실 생활을 같이한 피파의 달인 이재원 교수님, 자상하신 성동이형, 밤의 제왕 유섭이형, 아벨라 태워 주신 성배형, 친목의 자리를 많이 만들어 주신 종우형, 언제나 어려운 정호형, 연구실 엔터테이너 장민이형, 거짓말을 모르는 남자 규백이형 모두에게 감사드립니다. 케이블 가이 수용, shlim 승환, 성실한 동연, 언제나 진지한 동렬, 공공칠 준용, 류~ 정우, 웹마스터 인영 및 기타 94학번 친구들 모두에게 감사드립니다. 박사장 상욱, 터프가이 유환, 돌고래 호식, 바니 형주들에겐 선배로써 별로 잘 못해줘서 미안하군요. 옆자리에서 많은 도움을 준 영어의 제왕 상윤이형과 제균이형, 인지과학의 폭넓음을 보여주신 동호형과 정우형, 학문하는 사람의 자세를 일깨워준 석준이형에게도 감사를 드립니다. 마지막 학기를 즐겁게 보내도록 도와주신 선이형과 재홍이, 소영이 누나와 광주형에게도 감사드립니다. 짧은 기간이었지만 같이 연구한 Hubert Wang에게도 感謝드립니다.

변삼량 친구인 인건, 득식, 신엽, 희용의 오랜 우정도 큰 도움이 되었습니다. 물을 마시면 그 근원을 생각한다는 말처럼, 컴퓨터란 것을 알게 해주신 손사부님께 감사드립니다. 그 밖에 그들은 나를 모르지만 감사하고 싶은 사람들: RATM, 델리 스파이스, 김현식, 심수봉, Matt Groening, 미노루 후루야, 우스타 쿄스케, 이가리시 미키오, 아야하나 민, 아무로 레이, 까뮤 비던, l'arc en ciel, rancid, bob marley, doors, X-japan, 金庸...