유전자 알고리즘을 이용한 변수 선택법*

노 상 규**

《目 次》

Ⅰ. 서론
Ⅱ. 유전자 알고리즘과 변수선택 문제
1. 유전자 알고리즘
2. 유전자 알고리즘을 이용한 변수선택법
Ⅲ. 변수선택법의 비교·평가
1. 시뮬레이션 데이터에 의한 평가
2. 실제데이터에 의한 평가
Ⅳ. 결론

요 약

본 논문에서는 선형회귀분석에서의 변수선택 문제에 가장 빈번히 적용되는 단계적 회귀분석 법(stepwise regression)의 대안으로 유전자 알고리즘(genetic algorithm)을 이용한 변수선택법을 소개하고 이를 단계적 회귀분석법과 최근에 제안된 베이지언(Bayesian) 변수선택법과 비교평가하였다. 시뮬레이션을 이용하여 여러 방법론을 평가한 결과 유전자 알고리즘은 단계적 회귀분석법에 비해 약간 우수하나 베이지언 변수선택법보다는 못 한 것으로 나타났다. 그러나 실제 데이터에 적용한 결과 유전자 알고리즘이 베이지언 변수선택법에 못 지 않은 것으로 나타났다.

Ⅰ. 서 론

선형회귀분석은 여러 학문 또는 실무 분야에서 가장 널리 쓰이는 데이터 분석 방법이다. 이러한 회귀분석을 하는 데 있어서 가장 어려운 문제중의 하나가 종속변수를 설명할 독립변수를 선택하는 것이라 할 수 있다. 대다수의 연구자 및 실무자들은 독립변수의 선택 문제를

* 본 연구는 서울대학교 경영대학 경영연구소의 연구비 지원에 의하여 수행되었음.

** 서울대학교 경영대학 기금조교수.

단계적 회귀분석법(stepwise regression)을 이용하여 해결하여 왔다. 그러나 단계적 회귀분석법은 실제 변수 채택 및 기각 유의수준을 통제하기 어렵다는 점과 독립변수간의 상관계수가 높을 경우 제대로 변수선택을 하기 어렵다는 점 등 여러 문제점을 지니고 있다[Berk, 1978; Draper and Smith, 1981; Hocking, 1976; Miller, 1984]. 단계적 회귀분석법의 대안으로 제시된 부분 회귀분석법(all subsets regression)은 가능한 모든 회귀모형을 사전에 정한 척도에 의해 가장 좋은 모형을 선택하는 방법으로 단계적 회귀분석법에 비해 신뢰성이 높은 것으로 알려져 있다. 그러나 부분 회귀분석법은 독립 변수의 수가 늘어남에 따라 계산량이 기하급수적으로 늘어난다는 점에서 그 사용이 제한적이다. 예를 들어 독립변수의 수가 100개인 경우 $1.27\times10^{30}(2^{100})$개의 서로 다른 회귀모형이 존재하기 때문에 부분 회귀분석법의 적용은 비현실적이다.

최근에는 단계적 회귀 분석법과 부분회귀분석법의 단점을 해결하는 변수선택법으로 베이지언 변수선택법(Bayesian variable selection [George and McCulloch, 1993; Smith and Kohn, 1996; 김병도, 1998])과 유전자알고리즘(genetic algorithms)을 이용한 변수선택법[Sudjianto et al., 1996; Wasserman and Sudjianto, 1994]이 제시되었다. 베이지언 변수선택법은 변수선택의 문제를 베이지언 혼합모형에서 부분집합을 찾는 문제로 접근하는 방법으로 진정한 회귀모형을 찾는 데 단계적 회귀분석법에 비해 우수한 방법이기는 하나 적용의 어려움이 단점이라 할 수 있다[김병도, 1998]. 유전자 알고리즘을 이용한 변수선택법은 유전자 알고리즘을 부분 회귀분석법에 적용한 것으로 이 또한 단계적 회귀분석법에 비해 우수한 것으로 알려져 있다. 본 논문에서는 유전자 알고리즘을 이용한 변수선택법을 소개하고 이를 단계적 회귀분석법과 베이지언 변수선택법과 비교 평가하고자 한다. 본 논문은 다음과 같이 구성되어 있다. 제2장에서는 유전자 알고리즘에 대해 간단히 설명하고 변수선택문제에 어떻게 적용되는가를 설명한다. 제3장에서는 유전자 알고리즘을 이용한 변수선택법을 인위적 데이터와 실제 데이터를 이용하여 평가한다. 제4장에서는 본 연구를 요약하고 향후 연구 과제를 제시한다.

Ⅱ. 유전자 알고리즘과 변수선택문제

1. 유전자 알고리즘

유전자 알고리즘은 자연에서의 생명체의 진화 개념에 기초한 효율적이며 robust한 탐색 방법(search method)으로 최적화 및 인공 학습의 여러 분야에 성공적으로 적용되어 왔다

[Bennett, Ferris, and Ioannidis, 1991; Hou et al., 1994; March and Rho, 1995, Tam, 1992; Uckun, 1993]. 유전자 알고리즘의 기본적인 개념은 다음과 같다 [Goldberg, 1989; Davis, 1991, De Jong, 1990, Holland, 1975].

1) 생명체의 유전자에 해당하는 해의 표현(a representation of solutions)
2) 서로 다른 유전적 형질을 지닌 생명체들의 군에 해당하는 해집단(population)
3) 다음 세대에 자식을 생산할 부모를 선택하는 기준이 되는 적응도 개념(fitness)
4) 자식의 유전적 형질을 부모로부터 물려받는 유전적 연산(genetic operators)
5) 적자만이 살아남는 적자생존의 법칙(survival of the fittest)

위의 개념을 설명하기 위해 간단한 최적화 문제를 고려해 보자. 함수 $f(x) = -x^2 + 22x + 279$를 정수 간격 [0, 31]에서 최대화 하고자 한다고 가정하자. 해 x는 이진수로 5 bit을 이용하여 나타낼 수 있다. 예를 들어, 01001은 9를 나타낸다.

유전자 알고리즘은 무작위로 최초 해집단을 생성함으로써 시작된다. 집단의 크기는 해공간을 충분히 샘플할 수 있을 정도로 커야한다. 〈표 1〉은 크기 4인 예제의 최초 해집단을 보여주고 있다. 각 세대에서 해집단의 해는 적응도에 의해 평가 되어진다. 예제에서는 함수 값, f(x), 로 해를 평가 할 수 있다. 하지만 대부분의 최적화 문제에서는 함수 값을 그대로 사용할 수 없기 때문에 적응도를 조정하여야 한다. 예를 들어 함수 값이 음이거나 최소화(minimization) 문제인 경우는 함수 값을 그대로 사용하지 않고 조정을 하게 된다. 적응도를 어떻게 정의 하는가는 유전자 알고리즘의 효율성에 큰 영향을 미친다.

〈표 1〉 최초 해집단 (Initial Population)

해	x	$f(x)$(적응도)
11101	29	76
00101	5	364
01110	14	391
10100	20	319

해집단의 해를 적응도에 의해 평가한 후, 해집단 중의 몇 해는 자식(offspring)을 생산할 수 있는 부모(parent)로 선택되어진다. 이 때, 부모로 선택되어질 확률이 적응도에 비례하는 확률적 방법으로 부모를 선택한다.

선택된 부모들은 짝이 지워지고 유전적 연산에 의해 자식이 생성된다. 자식을 생성하기 위

해 사용되는 대표적인 유전적 연산으로 교배(crossover)와 돌연변이(mutation)가 있다. 교배는 유전자 알고리즘에서 가장 중요한 연산이다. 교배는 두 부모에 적용하며 두 부모의 일부분을 결합하여 하나 또는 두 개의 자식을 생성한다. 교배를 하는 가장 간단한 방법(1점교배)은 〈그림 1〉 a에서와 같이 교배점 (crossover point)을 무작위로 선택하고 한 부모의 왼쪽 부분과 다른 부모의 오른쪽 부분을 연결하는 것이다. 두번째 자식은 반대 부분을 결합함으로써 생성될 수 있다. 돌연변이는 〈그림 2〉 b에서와 같이 한 해의 유전형질을 무작위로 변형시키는 것이다. 돌연변이는 해공간(solution space)의 특정부분을 탐색하지 않을 확률이 0이 되지 않도록 보장하는 역할을 한다.

a. 교배(Crossover)

부모 1	00101
부모 2	*01110*
	↑ 교배점
자식 1	00*110*
자식 2	*01*101

b. 돌연변이(Mutation)

부모	00101
	↑ 돌연변이
자식	00111

〈그림 1〉 유전자 연산 (Genetic Operators)

새로운 해들이 유전적 연산에 의해 생성되면 이전 세대의 해들을 새로운 해(자식)들로 대체함으로써 새로운 세대(generation)를 형성한다. 이 때 대체되는 해는 적응도가 떨어지는 해가 된다. 〈표 2〉는 두 자식을 생성하여 적응도가 떨어지는 두 해를 대체함으로써 형성된 제 2세대이다. 마지막으로 유전자 알고리즘은 주어진 정지조건을 만족하면 끝나게 된다. 정지조건은 대부분의 경우 최대 세대 수이다.

〈표 2〉 제2세대 해집단(Second-generation Population)

해	x	$f(x)$(적응도)
00101	5	364
01110	14	391
00110	6	375
01101	13	396

단순하기는 하지만 위의 예는 유전자 알고리즘이 왜 효과적인가를 잘 나타내고 있다. 교배가 해를 결합하면서, 높은 적응도를 지닌 스키마(schema) 또는 부분해(partial solution)가 (예를 들어 0****, *1**1, 01***) 여러 해에 나타나기 시작한다. 평균이상의 적응도를 지닌 부모들은 좋은 스키마를 가지고 있을 것이고 그렇지 못한 부모들은 좋은 스키마를 가지고 있지 않을 것이다. 확률적 선택과정을 통해 높은 적응도를 지닌 부모들이 그렇지 못한 부모들 보다 많은 자식을 생성하게 될 것이고 세대가 지나면서 좋은 스키마의 수자는 늘고 나쁜 스키마의 수자는 줄어 들 것이다. 따라서 해집단의 평균 적응도는 향상되고 최적해를 찾게 될 가능성이 높아지는 것이다.

2. 유전자 알고리즘을 이용한 변수선택법

본 논문에서 다루고자 하는 변수선택 문제를 다음과 같이 정형화하여 표현할 수 있다. 종속변수 Y 와 이를 설명하기 위한 일군의 독립변수 $X_1, \ldots, X_P$가 주어졌다고 할 때 모든 가능한 독립변수를 포함한 선형회귀모형은 다음과 같이 표현할 수 있다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (1)$$

위의 회귀모형 식(1)에서 종속변수 Y를 진실로 설명하는 일군의 독립변수를 선택하는 문제가 변수선택의 문제이다.

유전자 알고리즘에서는 회귀모형 식(1)의 문제의 해를 (p+1)개의 이진수로 나타낼 수 있다. 변수가 선택되는 경우는 1로 표현하고 그렇지 않은 경우는 0으로 표현할 수 있다. 예를 들어 해 {10100...001}은 다음의 회귀모형에 해당된다.

$$Y = \beta_0 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (2)$$

각 모형의 적응도로는 회귀모형을 평가하는 데 자주 사용되는 척도인 R^2, Adjusted R^2, *AIC* (Akaike Information Criteria [Akaike, 1974]), corrected *AIC*, 그리고 *BIC* (Bayesian Information Criteria [Schwarz, 1978])등을 사용할 수 있는 데 본 연구에서는 Adjusted *R2*, *AIC*, 및 *BIC*를 사용하였다. 이들 척도는 다음과 같이 정의된다.

$$\text{Adjusted } R_2 = 1 - ((n-1)SSE)/((n-p)SST)$$

$$AIC = n \log(SSE/n) + 2p$$

$$BIC = \log(SSE/n) + p \log n/n$$

위의 식에서 n은 표본의 수이고, p는 추정하는 모수의 수이고, SSE와 SST는 각각 오차제곱합(error sum of squares)과 총제곱합(total sum of squares)을 나타낸다.

본 연구의 유전자 알고리즘에서는 최초 해집단은 무작위로 생성된다. 각 세대에서 부모해 두개가 확률적으로 선택되고 이 부모들로부터 하나의 자식이 평등교배(uniform crossover [Syswerda, 1989])에 의해 생성된다. 평등교배란 전 절에서 설명한 1점 교배와는 달리 두 z부모로부터 각 유전형질을 동일한 확률로 물려받는 방법으로 1점 교배 보다 우수하다 [Syswerda, 1989]. 그리고 생성된 자식이 전 세대에서 적응도가 가장 나쁜 해를 대체함으로써 새로운 세대를 형성한다. 이렇게 한 세대에서 하나 또는 둘의 자식을 생성하여 대체하는 방식을 steady state 접근법이라고 하며 이는 해집단이 자식들에 의해 완전히 대체되는 단순 유전자 알고리즘(simple genetic algorithm)에 비해 우수하다[Davis, 1991; Whitley, 1988]. 본 연구의 유전자 알고리즘은 한 해가 해집단의 대부분을 차지 할 때 중단하도록 하였다. 이 경우에는 세대가 지속되더라도 나은 해가 나오기 어렵기 때문이다.

Ⅲ. 변수선택법의 비교 · 평가

이 장에서는 지금까지 설명한 유전자 알고리즘을 이용한 변수선택법이 다른 방법에 비해 어느 정도 우수한가를 평가하고자 한다. 우선 시뮬레이션을 통한 인위적 데이터에 의해 평가하고 다음 절에서는 실제 데이터에 의해서도 평가하고자 한다. 다음의 모든 평가에서 유전자 알고리즘의 해집단 크기는 1000을 사용하였고 알고리즘은 한 해가 해집단의 99%이상을 차지할 때 중지하였다.

1. 시뮬레이션 데이터에 의한 평가

이 절에서는 김병도[1998]에서 사용한 3개의 인위적 데이터를 이용하여 유전자 알고리즘의 우수성을 평가하고자 한다. 첫번째 데이터는 상호 독립적인 50개의 독립변수, $X_1, \ldots, X_{50}$을 표준정규분표 $N(0,1)$에서 100개씩 추출한다. 그리고 ε을 독립적으로 표준정규분포 $N(0,1)$에서 100개 추출하고 $Y = \varepsilon$ 이라고 정의한다. 즉 종속변수 Y를 진정으로 설명하

는 회귀식은 50개의 독립변수 중 어느 것도 포함하지 않아야 하는 경우이다.

〈표 3〉에서 나타난 것과 같이 단계적 회귀분석법은 5개의 변수를 선택하였으며 회귀식의 R^2도 0.18로 α = 0.05에서 유의하였다. 단계적 회귀분석법은 실제로는 아무 설명력이 없는 다수의 독립변수를 선택한 것이다. 베이지언 변수선택법은 정확하게 아무 변수도 선택하지 않았다. 유전자 알고리즘의 경우는 척도에 따라 결과가 상당히 차이가 났다. Adjusted R^2와 *AIC*를 사용한 경우 각각 21개와 14개의 변수를 선택하여 단계적 회귀분석법보다도 못 한 결과가 나왔다. 그러나 *BIC*를 척도로 사용한 경우에는 2개의 변수만 선택하여 단계적 회귀분석법보다 우수하였다. 하지만 아무 변수도 선택하지 않은 베이지언 선택법보다는 못 하였다. 여기서 주목할 점은 베이지언 선택법이 선택한 진정한 모형의 *AIC*나 *BIC*는 유전자 알고리즘이 선택한 모형에 비해 떨어진다는 점이다. 진정한 모형의 *AIC*와 *BIC*는 0.0732인데 비해 *AIC*를 척도로 이용하여 유전자 알고리즘이 선택한 모형의 *AIC*는 -0.1353, *BIC*를 척도로 선택한 모형의 *BIC*는 0.0514로 진정한 모형의 척도보다 우수하였다. 따라서 유전자 알고리즘이 진정한 모형을 선택하지 못한 것은 유전자 알고리즘 자체의 문제라기 보다는 회귀모형을 평가하는 척도가 불완전 한데서 기인한다고 볼 수 있다.

〈표 3〉 시뮬레이션 1($Y = \varepsilon$)

Method	Selected Variables	R^2	Adj. R^2	*AIC*	*BIC*
All Variables	X_0, . X_{50}	**0.4941**	-0.0220	0.4102	1.7389
Stepwise	X_0, X_2, X_{27}, X_{36}, X_{45}	0.1805	0.1460	-0.0273	0.1029
Bayesian	No Variables Selected	-	-	0.0732	0.0732
Genetic (Adj. R^2)	X_2, X_6, X_{11}, X_{12}, X_{13}, X_{14}, X_{18}, X_{19}, X_{21}, X_{23}, X_{25}, X_{26}, X_{27}, X_{28}, X_{33}, X_{35}, X_{36}, X_{40}, X_{44}, X_{45}, X_{46}	0.4511	**0.3121**	-0.1081	0.4390
Genetic(*AIC*)	X_2, X_{11}, X_{18}, X_{21}, X_{23}, X_{25}, X_{27}, X_{28}, X_{33}, X_{35}, X_{36}, X_{44}, X_{45}, X_{46}	0.3856	0.2927	**-0.1353**	0.2294
Genetic(*BIC*)	X_{27}, X_{45}	0.1063	0.0972	-0.0007	**0.0514**

두번째 데이터는 첫번째 데이터와 같은 방식으로 상호 독립적인 50개의 독립변수 X_1, ...,X_{50}을 표준정규분표 $N(0,1)$에서 100개씩 추출한다. 그리고 ε을 독립적으로 표준정규분포 $N(0,1)$에서 100개 추출하고 $Y = X_1 + 2X_2 + \varepsilon$ 이라고 정의한다.

두번째 시뮬레이션 결과도 첫번째 결과와 유사하게 나타났다. 〈표 4〉에서 나타난 것과 같이 단계적 회귀분석법은 X_1과 X_2를 포함한 9개의 변수를 선택하였으며 회귀식의 R^2도 0.85로 유의하였다. 단계적 회귀분석법은 실제로는 아무 설명력이 없는 다수의 독립변수를 선택한 것이다. 베이지언 변수선택법은 정확하게 X_1과 X_2만 선택하였다. 유전자 알고리즘의 경우는 척도에 따라 결과가 상당히 차이가 났다. Adjusted R^2와 AIC를 사용한 경우 각각 19개와 17개의 변수를 선택하여 단계적 회귀분석법보다 좋지 않은 결과가 나왔다. 그러나 BIC를 척도로 사용한 경우에는 8개의 변수를 선택하여 단계적 회귀분석법보다 약간 우수하였지만 베이지언 선택법보다는 우수하지 못 하였다. 첫번째 시뮬레이션 결과와 마찬가지로 베이지언 선택법이 선택한 진정한 모형은 AIC나 BIC가 가장 낮은 모형은 아니었다.

〈표 4〉 시뮬레이션 2($Y=X_1+2X_2+\varepsilon$)

Method	Selected Variables	R^2	Adj. R^2	AIC	BIC
All Variables	X_0, . X_{50}	**0.4941**	-0.0220	0.4102	1.7389
Stepwise	X_0, X_1, X_2, X_{11}, X_{14}, X_{28}, X_{38}, X_{42}, X_{50}	0.1805	0.1460	-0.0273	0.1029
Bayesian	X_1, X_2	-	-	0.0732	0.0732
Genetic (Adj. R^2)	X_0, X_1, X_2, X_4, X_6, X_8, X_9, X_{13}, X_{17}, X_{20}, X_{24}, X_{30}, X_{32}, X_{38}, X_{39}, X_{41}, X_{42}, X_{44}, X_{47}	0.4511	**0.3121**	-0.1081	0.4390
Genetic(AIC)	X_0, X_1, X_2, X_4, X_6, X_8, X_9, X_{13}, X_{20}, X_{24}, X_{30}, X_{32}, X_{39}, X_{41}, X_{42}, X_{44}, X_{47}	0.3856	0.2927	**-0.1353**	0.2294
Genetic(BIC)	X_1, X_2, X_{11}, X_{14}, X_{28}, X_{38}, X_{42}, X_{50}	0.1063	0.0972	-0.0007	**0.0514**

과거의 연구에 의하면 독립변수간의 상관계수가 높은 경우 단계적 회귀분석법으로 변수를 선택하는 것에 문제가 있는 것으로 지적되었다[Berk, 1978]. 따라서 세번째 데이터는 독립변수간의 상관계수가 높은 경우를 선정하였다. 우선 상호 독립적인 51개의 독립변수 Z_1, ..., Z_{51}을 표준정규분포 $N(0,1)$에서 100개를 추출한다. 그리고 $X_i = Z_i + 2Z_{51}$ (i = 1, ..., 50)이라고 정의하면 X_i간의 상관계수가 약 0.8정도의 50개의 독립변수를 얻을 수 있다. 그 다음 을 독립적으로 표준정규분포 $N(0,1)$에서 100개 추출하고 $Y = X_1 + 2X_2 + \varepsilon$ 이라고 정의한다.

〈표 5〉에서 나타난 것과 같이 단계적 회귀분석법은 X_1과 X_2를 포함한 7개의 변수를 선택하였으며 회귀식의 R^2도 0.98로 유의하였다. 앞에서와 마찬가지로 단계적 회귀분석법은 실제로

는 아무 설명력이 없는 다수의 독립변수를 선택한 것이다. 베이지언 변수선택법은 X_1과 X_2만 선택하여 진정한 모형을 찾았다. 유전자 알고리즘의 경우는 척도에 관계 없이 좋지 않은 결과를 나타냈다. Adjusted R^2, *AIC*, *BIC*를 사용한 경우 모두 각각 18, 15, 8개의 변수를 선택하여 단계적 회귀분석법보다 좋지 않은 결과가 나왔다. 앞의 시뮬레이션 결과와 마찬가지로 베이지언 선택법이 선택한 진정한 모형은 *AIC*나 *BIC*가 가장 낮은 모형은 아니었다

〈표 5〉 시뮬레이션 3($Y=X_1+2X_2+\varepsilon$)

Method	Selected Variables	R^2	Adj. R^2	*AIC*	*BIC*
All Variables	X_0, , X_{50}	**0.9910**	0.9817	-0.0167	1.3119
Stepwise	X_0, X_1, X_2, X_7, X_{20}, X_{23}, X_{42}	0.9851	0.9842	-0.3962	-0.2139
Bayesian	X_1, X_2	0.9799	0.9797	-0.1968	-0.1447
Genetic (Adj. R^2)	X_0, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_{13}, X_{20}, X_{21}, X_{23}, X_{25}, X_{34}, X_{37}, X_{40}, X_{42}, X_{48}	0.9888	**0.9865**	-0.4640	0.0050
Genetic(*AIC*)	X_0, X_1, X_2, X_3, X_5, X_6, X_7, X_{13}, X_{20}, X_{21}, X_{23}, X_{25}, X_{37}, X_{42}, X_{48}	0.9882	0.9863	**-0.4713**	-0.0805
Genetic(*BIC*)	X_0, X_1, X_2, X_3, X_5, X_7, X_{23}, X_{42}	0.9860	0.9849	-0.4370	**-0.2286**

시뮬레이션 결과 *BIC*를 척도로 이용한 유전자 알고리즘은 단계적 회귀분석법에 비해 비슷하거나 우수하지만 베이지언 변수선택법에 비해서는 우수하지 못한 것으로 나타났다. 그리고 유전자 알고리즘의 효과성은 척도에 의해 좌우되는 것으로 나타났다.

2. 실제데이터에 의한 평가

이 절에서는 유전자 알고리즘을 이용한 변수선택법을 보다 현실적인 상황에서 평가하기 위해 Kim and Park[1997]과 김병도[1998]에서 사용한 IRI (Information Resources, Inc.) 데이터에 적용하였다. 이 데이터는 미국의 대표적 시장조사업체인 IRI가 1984년 1월부터 약2년 동안 1,398명의 소비자가 슈퍼마켓에서 얼마를 지출하였고 그들이 어떤 인구통계학적 특성을 지니고 있는가를 기록한 것이다. 인구통계적 변수(즉 독립변수)의 수는 가구구성원수, 소득, 인종 등 18개이다.

평가를 보다 체계적으로 하기위해 1,398명의 소비자를 무작위로 1,000명과 398명의 두 집단으로 나누고 1,000명의 소비자 샘플에 변수선택법을 적용하여 모수를 추정하고 나머지

398명의 소비자에게 추정된 모형을 적용하여 모형의 예측력을 테스트하였다.

우선 단계적 회귀분석법을 1,000명의 소비자 데이터에 적용한 결과 11개의 변수를 선택하였다(〈표 6〉 참조). 회귀모형의 R^2는 0.29이고 추정된 11개의 모수 모두가 α = 0.05에서 유의하였다. 베이지언 선택법을 동일한 데이터에 적용한 결과 절편을 포함하여 8개의 변수를 선택하였다. 유전자 알고리즘을 적용한 결과 adjusted R^2를 사용하였을 때는 13개의 변수를 *AIC*를 사용하였을 때는 12개의 변수를 선택하여 단계적 회귀분석법에 비해 많은 변수를 선택하였다. 그러나 *BIC*를 사용한 경우는 베이지언 선택법과 동일한 8개의 변수를 선택하였다. 그렇다면 어떤 모형이 진정한 회귀모형인가? 시뮬레이션의 경우 사전에 어떤 변수가 설명력이 있는가를 알고 있기 때문에 비교가 쉽지만, 실제 데이터의 경우에는 비교가 단순하지 않다.

〈표 6〉 IRI 데이터 추정 모형

Method	Selected Variables	R^2	Adj. R^2	*AIC*	*BIC*
All Variables	X_0, , X_{18}	**0.2933**	0.2805	5.9735	6.0668
Stepwise	X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{18}	0.2888	0.2816	5.9640	6.0180
Bayesian	X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{15}	0.2771	0.2720	5.9744	6.0136
Genetic (Adj. R^2)	X_0, X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{16}, X_{18}	0.2926	**0.2840**	5.9627	6.0265
Genetic(*AIC*)	X_0, X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_{11}, X_{13}, X_{15}, X_{18}	0.2915	0.2836	**5.9623**	6.0212
Genetic(*BIC*)	X_0, X_1, X_2, X_4, X_5, X_7, X_8, X_{15}	0.2771	0.2720	5.9744	**6.0136**

추정샘플내에서의 모형을 평가하는 데 따르는 어려움을 극복하기 위해 추정하는데 이용하지 않은 398명의 샘플에 위의 여섯가지 모형을 적용하여 예측력을 평가하였다. 모형의 예측력을 측정하기 위해 두가지 척도를 사용하였다. 하나는 평균자승오차(Mean Squared Error)이고 다른 하나는 평균절대오차(Mean Absolute Error)이다. 평균자승오차와 평균절대오차는 다음과 같이 정의한다.

$$MSE = \sum_{i=1}^{n} (Y_i - Y_i)^2 / n$$

$$MAE = \sum_{i=1}^{n} | Y_i = Y_i | / n$$

위의 식에서 Y_i는 i번째 표본의 실제 관측치이고 Y_i는 각 모형에 의한 이의 예측치를 나타내며 n은 표본의 수를 나타낸다.

〈표 7〉 추정의 샘플에서의 예측결과

방법론	*MSE*	*MAE*
All Variables	294.60	13.28
Stepwise	299.90	13.38
Bayesian	285.60	13.15
Genetic(Adj. R^2)	294.20	13.26
Genetic(*AIC*)	296.80	13.34
Genetic(*BIC*)	285.60	13.15

예측 결과 정도의 차이는 크지는 않지만 베이지언 변수선택법과 *BIC*를 이용한 선택법이 가장 우수한 방법이고 단계적 회귀분석법이 가장 좋지 않은 방법으로 평가되었다(표 7 참조). 평가를 보다 객관적으로 하기 위해 앞에서 언급한 방법으로 추정과 예측을 추가적으로 6번 반복하였다. 그 결과는 〈부록 1〉에 요약하였다. 베이지언 선택법과 *BIC*를 이용한 선택법은 모두 같은 모형을 선택하였으며 변수의 수는 평균 8.43개로 단계적 회귀분석법의 평균 9.43개에 비해 하나 정도 적은 수의 변수를 선택하였다. 그러나 평균자승오차나 평균절대오차에 있어서는 거의 차이가 나지 않았다. 결론적으로 유전자 알고리즘을 이용한 변수선택법은 실제 데이터에 의한 평가에서는 베이지언 변수선택법에 비해 떨어지지 않는 것으로 나타났다.

Ⅳ. 결 론

이 논문에서는 회귀분석에서 변수선택의 방법으로 가장 흔히 사용되는 단계적 회귀분석법의 대안으로 유전자 알고리즘을 이용한 변수선택법을 소개하였고 이를 단계적 회귀분석법과 최근에 제안되어 효과적인 것으로 알려진 베이지언 변수선택법과 비교 평가하였다. 유전자 알고리즘을 이용한 방법론을 인위적 데이터에 적용한 결과 단계적 회귀분석법보다는 약간 우수하였고 베이지언 변수선택법보다는 나쁜 것으로 나타났다. 유전자 알고리즘이 베이지언 방

법론보다 못 한 것은 유전자 알고리즘에서 이용한 척도가 불완전하기 때문으로 볼 수 있다. 하지만 실제 데이터에 적용한 결과 베이지언 선택법과 동일한 결과를 나타냈다. 결론적으로 유전자 알고리즘을 이용한 변수선택법은 베이지언 변수선택법에는 미치지 못 하지만 충분한 가능성을 지닌 방법론이라 할 수 있다.

향후에는 유전자 알고리즘에서 이용할 척도를 보완하는 연구가 진행되어야 할 것이다. 앞에서 언급한 바와 같이 유전자 알고리즘의 효과성은 사용하는 척도에 의해 좌우된다고 볼 수 있다. 척도의 보완은 유전자 알고리즘의 효과성을 높이는 데 있어 가장 우선 되어야 할 연구 과제라 할 수 있다. 두번째로는 유전자 알고리즘을 비모수 회귀분석[Hastie and Tibshirani, 1990; Fahrmeir and Tutz, 1994]에 적용하는 것이다. 본 연구에서는 변수선택의 문제에 유전자 알고리즘을 적용하였는 데 이를 모형선택의 문제에 적용하는 것이다. 즉 변수뿐 아니라 그 변환방법까지 유전자 알고리즘을 이용하여 선택하는 것이다.

참 고 문 헌

Bennett, K., Ferris, M. C., and Ioannidis, Y. E. (1991), "A Genetic Algorithm for Database Query Optimization," *Proceedings of the 4th International Conference on Genetic Algorithms*, San Diego, CA, 400-407.

Berk, K. (1978), "Comparing Subset Regression Procedures," *Technometrics*, 20, 1, 1-6.

Davis, L. (1991), ed., *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.

De Jong, K. A. (1990), "Genetic-Algorithm-Based Learning," in Kodratoff, Y. and Michalski, R. S. (eds.), *Machine Learning*, Morgan Kaufmann Publishers, 611-638.

Draper, N. and Smith, H. (1981), *Applied Regression Analysis*, 2nd edition, John Wiley, New York.

Fahrmeir, L. and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer Series in Statistics, Springer-Verlag:

New York.

George, E. and McCulloch, R. (1993), "Variable Selection Via Gibbs Sampling," Journal of the American Statistical Association, 88, 423, 881-889.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Monographs on Statistics and Applied Probability 43, Chapman & Hall: London, UK.

Hoching, R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

Hou, E. S. H., Ansari, N., and Ren, H. (1994), "A Genetic Algorithm for Multiprocessor Scheduling," *IEEE Transactions on Parallel and Distributed Systems*, 5, 2, 113-120.

Kim, B. and Park, K. (1997), "Studying Patterns of Consumer's Grocery Shopping Trips," Journal of Retailing, 73, 4, 501-517.

March, S. T. and Rho, S. (1995), "Allocating Data and Operations to Nodes in Distributed Database Design," *IEEE Transactions on Knowledge and Data Engineering*, 7, 2, 305-317.

Miller, A. (1984), "Selection of Subsets and Regression Variables," Journal of Royal Statistical Society, A, 147, 389-429.

Smith, M. and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317-343.

Sudjianto, A., Wasserman, G. S., and Sudarbo, H. (1996), "Genetic Subsets Regression," *Computers and Industrial Engineering*, 30, 4, 839-849.

Syswerda, G. (1989), "Uniform Crossover in Genetic Algorithm," *Proceedings of the 3rd International Conference on Genetic Algorithms*, 2-9.

Tam, K. Y. (1992), "Genetic Algorithms, Function Optimization, and Facility Layout Design," *European Journal of Operational Research*, 63, 2, 322-346.

Uckun, S. Bagci, S., Kawamura, K., and Miyabe, Y. (1993), "Managing Genetic

Search in Job Shop Scheduling," *IEEE Expert*, 15-25.

Wasserman, G. S., and Sudjianto, A. (1994), "All Subsets Regression Using a Genetic Search Algorithm," *Computers and Industrial Engineering*, 27, 489-492.

Whitley, D. (1988), "GENITOR: A Different Genetic Algorithm," *Proceedings of the Rocky Mountain Conference on Artificial Intelligence*.

김병도 (1998), "선형회귀모형에서의 효과적 변수선택법," 경영논집, 32, 3, 1-19.

〈부록 1〉 IRI 데이터 예측 결과

a. Mean Squared Error

Data Set	1	2	3	4	5	6	7	*MIN*	*MAX*	*AVG*
All Variables	294.60	339.47	350.15	383.07	383.16	390.91	376.85	294.60	390.91	353.32
Stepwise	299.90	339.20	345.79	383.04	339.14	386.55	371.87	299.90	386.55	352.21
Bayesian	285.60	339.20	345.79	389.66	330.00	392.87	377.63	285.60	392.87	351.54
Genetic-Adj R^2	294.20	341.02	345.79	383.55	339.42	392.49	372.79	294.20	392.49	352.75
Genetic-*AIC*	296.80	340.57	350.64	386.14	339.23	391.31	374.28	296.80	391.31	354.14
Genetic-*BIC*	285.60	339.20	345.79	389.66	330.00	392.87	377.63	285.60	392.87	351.54

b. Mean Absolute Errors

Data Set	1	2	3	4	5	6	7	*MIN*	*MAX*	*AVG*
All Variables	13.28	13.94	14.41	14.82	14.66	15.33	14.48	13.28	15.33	14.42
Stepwise	13.38	14.05	14.30	14.89	14.71	15.30	14.45	13.38	15.30	14.44
Bayesian	13.15	14.05	14.30	14.97	14.51	15.45	14.56	13.15	15.45	14.43
Genetic-Adj R^2	13.26	14.00	14.30	14.83	14.71	15.37	14.38	13.26	15.37	14.41
Genetic-*AIC*	13.34	14.03	14.40	14.91	14.71	15.28	14.48	13.34	15.28	14.45
Genetic-*BIC*	13.15	14.05	14.30	14.97	14.51	15.45	14.56	13.15	15.45	14.43

c. Nember Of Variables

Data Set	1	2	3	4	5	6	7	*MIN*	*MAX*	*AVG*
All Variables	19	19	19	19	19	19	19	19	19	19.00
Stepwise	11	9	9	9	10	9	9	9	11	9.43
Bayesian	8	9	9	8	9	8	8	8	9	8.43
Genetic-Adj R^2	13	14	9	13	14	14	13	9	14	12.86
Genetic-*AIC*	12	11	11	11	11	11	11	11	12	11.14
Genetic-*BIC*	8	9	9	8	9	8	8	8	9	8.43