Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

# 공간데이터마이닝에서의 유전자알고리즘을 이용한 예측방법연구

김효정[1], 강한구[2], 강창완[3]

## 요 약

공간자료의 예측문제에 있어 전통적 예측방법인 크리깅방법과 최근 통계적문제 적용되기 시작한 신경망분석방법 간의 비교를 사례연구를 통해 행하였다. 일반적으로 크리깅에 의한 선형예측은 공간자료에 대한 일반적 통계모형으로서 간주되어 왔다. 한편 예측문제에 있어 뉴럴네트워크에 기초한 비모수적 방법이 관심의 대상이 되고 있으며 특히 대용량 자료의 경우 데이터마이닝 기법의 한 분야로 널리 사용되고 있는 실정이다. 본 연구에서는 공간 자료의 예측에 있어 유전자 알고리즘을 신경망분석 모형을 결합하여 기존의 크리깅방법과의 예측력을 비교한다.

주요용어: 공간자료, 크리깅, 유전자알고리즘, 신경망분석

## 1. 서론

신경망분석은 광범위하게 과학적 문제를 해결하는데 컴퓨터과학 분야에서 발전되어 왔다. 최근들어 일반적인 통계적 문제에 신경망분석이 유용하게 적용되는 연구가 나타나기 시작했으며 이러한 신경망은 복잡한 구조를 가진 자료에서의 예측문제를 해결하기 위해서 사용되는 비선형모형의 하나로 분류할 수 있다. 통계적 예측문제에 적용한 예로 시계열자료에 대하여 Box-Jenkins 방법과 신경망분석의 비교(Faraway,J et el.,1995), 공간자료 예측 문제에서의 Kohonen 신경망을 적용된 연구(Sarzeaud et el., 1999 ) 등 여러 가지 사례연구가 있다.

한편 공간자료의 예측에 있어 전통적 통계분석 방법인 크리깅은 몇 가지 제한점을 가지고 있다. 첫째는 대용량의 공간자료가 주어졌을 때 각 점에서의 크리깅은 방대한 선형시스템으로 말미암아 모든 자료를 이용하는데 문제가 발생한다. 둘째로는 확률변수에 대한 가정이 너무 강하다는 것이며 셋째로는 크리깅에 필요한 공간상관 구조에 대한 모형 구축이 주관적이어서 크리깅자체에 대한 확신이 어렵다는 점을 들 수 있다. 반면, 최근 통계적 문제에 시도되기 시작한 신경망 분석은 위에서 언급한 문제점에 상관없이 적용할 수 있다는 장점이 있으나 추정의 비수렴성 문제라든지 초기값 선택 문제, 그리고 전체적 해가 아닌 국부적 해를 구하는 가능성이 많다는 단점이 있다. 이러한 이유로 해서 본 연구에서는 전체적 해를 구하는데 효율적으로 알려진 유전자알고리즘을 신경망분석모형에 접목시켜 공간자료의 예측문제를 다루고자 하며 이들

1) (614-714) 부산시 부산진구 가야동 동의대학교 전산통계학과, 석사
2) (614-714) 부산시 부산진구 가야동 동의대학교 전산통계학과 석사
3) (614-714) 부산시 부산진구 가야동, 동의대학교 전산통계학과, 조교수

방법간의 예측력 비교를 할 것이다.

## 2. 공간자료에 대한 크리깅

선형공간자료분석에서 관측된 자료를 근거로 하여 관측되지 않은 위치의 값을 예측하는 방법으로 공간의존관계를 고려하는 방법을 크리깅(kriging)이라 하며 이러한 크리깅에는 ordinary kriging, universal kriging, simple kriging 등 여러 가지가 알려져 있다. 본 연구에서는 이 중 ordinary kriging을 고려하며 이에 대한 정식화는 다음과 같다.(Cressie,N.,1991)

먼저 공간 자료가 확률과정의 실현치로 간주했을 때 $s$가 D의 위치이고 $R^d$는 $d$차원의 유클리드공간이라 하면 확률벡터 $z(s)$는

$$\{ z(s) : s \in D \subset R^d \}$$

로 표현할 수 있다. 이때 공간자료를 $(z_i, s_i)$, $i = 1, \cdots, n$ 라 표기하면 $z_i$는 $s_i$ 위치에서의 관측치를 말한다. 여기서 $z(s)$가 2차 정상성(second-order stationary)이라고 가정하면 미지의 위치 $s_0$에서 예측치 $\hat{Z}(s_0)$는

$$\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i)$$

이고 문제는 추정오차를 최소화하는 가중치 $w_i$를 구하는 것이 된다. 이러한 경우 공간상관을 고려하여 결합시스템의 해를 구함으로써 가중치를 구할 수 있는데 이를 크리깅방법이라 한다.

## 3. 다층인식자(MLP)모형과 유전자 알고리즘

신경망에는 여러 가지 다양한 모형이 있으나 자료분석을 위해 가장 널리 사용되는 모형은 다층인식자(multilayer perceptron) 신경망이다. 그리고 이 다층인식자신경망은 입력층과 은닉마디로 구성된 은닉층, 그리고 출력층으로 구성된 전방향 신경망이라 할 수 있다.

한편, 유전자 알고리즘이란 개체군이 다음세대의 집단을 형성하는 과정에서 적자생존을 확률적으로 알고리즘화 한 것이다. 즉 유전자 알고리즘은 자연도태의 유전적인 메카니즘에 기초한 탐색알고리즘이며, 그 목적은 세대가 지날수록 주어진 환경에 잘 적응한 개체가 발생한다는 가설에 입각하여 최적 개체를 찾는 것이다.

## 4. 공기오염자료에 대한 예측방법의 비교

비교분석에 사용될 자료는 한국의 환경청(www.me.go.kr)으로부터 구한 한국의 116개 관측소에서 측정된 1년간 월평균 일산화탄소 자료이다. 예측력비교를 위해서 원자료를 분석용($n_1 = 100$)과 평가용 데이터 세트($n_2 = 16$)로 분할하여 예측치에 대한 평균제곱오차(MSE)를

구하여 비교하였다. 아래 <표 1>은 전통적 공간예측 방법인 크리깅과 다층인식자모형을 이용한 신경망분석, 그리고 유전자 알고리즘을 이용한 신경망분석 결과를 보여준다.

| | Kriging | NN(Neural Networks) | GA(Genetic Algolism) |
|---|---|---|---|
| Training Set | 0.15 | 0.081713249 | 0.077322 |
| Validation Set | 0.035 | 0.0547143808 | 0.054718 |

<표 1 MSE 기준 결과 비교 >

## 5. 결론.

본 논문에서는 유전자 알고리즘을 신경망의 모형선택 즉 신경망 구조를 결정하는 모수들의 최적해를 찾기위하여 이용하였다. 그래서 유전자 알고리즘 수행결과로 결정되어진 신경망의 모형에서 함수추정을 위한 학습을 시킨 결과가 기존의 MLP 신경망 방법에 비해 비교적 좋음을 알 수 있다.

## 참고문헌

강현철, 한상태, 최종후, 김차용, 김은석, 김미경(1999) SAS EnterpriseMiner를 이용한 데이터마이닝, 서울, 자유아카데미.

Cressie, N.(1991) Statistics for spatial data. John Wiley & Sons, Inc, New York.

Faraway, J.and Chatfield, C.(1998) Time Series Forecasting with Neural Networks: A Case Study, Applied Statistics. 47, 231-250.

Sarzeaud, Stephan (1999) Fast Interpolation using Kohonen Self-Organizing Neural Networks. Technical report OS/99003.