2000 Proceedings of the Autumn Conference, Korean Statistical Society

유전자 알고리듬을 이용한 다중이상치 탐색

고 영현[1], 이 혜선[2], 전 치혁[3]

Abstract

Genetic algorithm(GA) is applied for detecting multiple outliers. GA is a heuristic optimization tool solving for near optimal solution. We compare the performance of GA and the other diagnostic measures commonly used for detecting outliers in regression model. The results show that GA seems to have better performance than the others for the detection of multiple outliers.

Keywords

Genetic Algorithm; Multiple outliers; Outlier diagnostic; Multiple Linear Regression (MLR)

1. 서론

이상치는 다수의 데이타와 구별되는 형태의 관측치로서 이상치가 존재할 경우 왜곡된 추정치를 얻고 그에 따라 잘못된 통계적 추론을 얻을 수 있다. 단일 이상치(single outlier)만이 존재할 경우 이상치를 탐색하는 문제는 비교적 간단하고 많은 진단 방법이 알려져 있다. 그러나 하나 이상의 이상치 즉, 다중 이상치(multiple outliers)가 존재하는 경우에 마스킹(masking)과 스웜핑(swamping)의 문제로 인하여 이상치의 탐색은 보다 어려워진다. 마스킹은 다른 이상치의 존재로 또다른 이상치의 발견이 어려워지는 현상을 말하고, 스웜핑은 어떤 이상치의 존재가 정상적인 관측개체를 이상치로 잘못 판단하게 만드는 현상을 말한다[1, 3].

이러한 다중 이상치 탐색을 위해 많은 연구가 이루어졌지만 이 중 몇가지 방법은 조합문제(combinatorial problem)을 풀어야하는 어려운 점이 있다. 본 연구에서는 다중회귀분석(multiple linear regression; MLR)에서 다중 이상치의 탐색을 위해 복잡한 최적화 문제를 생태학적 유전 원리를 이용하여 해결하는 유전자 알고리듬(genetic algorithm)을 통해 해결하도록 한다.

2. 이론 및 성능 지표

2.1. 단일 이상치의 판별

본 연구에서는 유전자 알고리듬에 의한 진단과 다음 4가지 형태의 진단(diagnostic) 방법

1) 포항공과대학교 산업공학과 박사과정, 경북 포항시 남구 효자동 산 31번지.
2) 포항공과대학교 산업공학과 연구원, 경북 포항시 남구 효자동 산 31번지.
3) 포항공과대학교 산업공학과 교수, 경북 포항시 남구 · 효자동 산 31번지.

을 비교한다[1 ,2, 5]. 여기서 n은 표본의 수, k는 독립변수의 수, $e_i = y_i - \hat{y}_i$ 는 MLR 모델에서의 예측 오차, 즉 잔차(residual)를 의미한다.

a) Prediction matrix에 의한 진단

$p_i = x_i(X'X)^{-1}x_i'$, 기각역 : $2k/n$

b) 잔차에 의한 진단

$t_i^* = t_i\sqrt{(n-k-1)/(n-k-t_i^2)}$

where $t_i = e_i/\hat{\sigma}\sqrt{1-p_i}$ 기각역 : $t(n-k-1)$

c) Volume of confidence ellipsoids에 의한 진단

$p_i^* = p_i + e_i^2/e'e$ 기각역 : $2(k+1)/n$

$CVR_i = [(n-k-t_i^2)/(n-k-1)]^2/(1-p_i)$ 기각역 : $|CVR_i - 1| > 3k/n$

d) Influence function에 의한 진단

$C_i = p_i t_i^2/k(1-p_i)$ 기각역 : $F(k, n-k)$

$WK_i = |t_i^2|\sqrt{[p_i(1-p_i)]}$ 기각역 : $2\sqrt{k/n}$

$C_i^* = WK_i\sqrt{[(n-k)/k]}$ 기각역 : $2\sqrt{(n-k/n)}$

2.2. 진단척도의 성능 지표

진단 척도의 성능 지표는 RMSE(root mean squared error)를 사용하였다.

$$\text{RMSE} = \sqrt{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2/n}$$

모델링 데이터에 적용한 경우엔 RMSEC(root mean squared error of calibration)로 검증 데이터에 적용한 경우엔 RMSEV(root mean squared error of validation)으로 정의하도록 한다.

3. 유전자 알고리듬에 의한 다중 이상치 탐색

최적화 과정에서, 보통 해집합이 작은 경우에는 고전적인 완전탐색 방법으로 충분하지만, 해집합이 아주 많은 경우에는 특수한 인공지능 기법이 사용되어야 한다. 유전자 알고리즘(GA)은 그러한 기법들 중 하나로써, 유전적 계승과 다윈의 생존 경쟁이라는 자연의 현상을 모델화한 확률적인 검색(stochastic search) 방법으로 조합문제와 같이 해집합이 무수한 경우에도 적용이 가능하다[4].

3.1. 스키마 정리와 이상치의 탐색

유전자 알고리듬의 이론은 스키마 정리(schema theorem)에 기초한다. 이는 유전자 알고리듬에서 선택(selection), 교배(crossover), 돌연변이(mutation)의 일련의 과정에서 개체들에 존재하는 유산한 인자들이 어떻게 다른 개체로 전파 또는 소멸되면서 좋은 해를 찾아가는가를 보여준다. 스키마는 개체들간의 유사성을 나타내는 일종의 전형이다. 스키마는 개체의 특정위치들에 존재하는 인자들의 값에 의해 나타난다. 스키마에서 특정인자의 값을 정의할 때 어느 인자

값이나 가질수 있다는 표시로 '*'를 사용한다. 예로 이진표현에서 스키마 (**011)에는 (00011), (01011), (10011), (11011)이 포함된다. 따라서 이진표현에서 r개의 *를 갖는 스키마에 부합되는 개체의 수는 2^r개가 된다. 유전자 알고리듬의 선택, 교배, 돌연변이의 특성에 따라 스키마 중 해의 결과에 좋은 영향을 미치는 스키마는 평균적으로 그 수가 증가하고 나쁜 영향을 주는 스키마는 평균적으로 그 수가 감소하므로 최적해를 찾아낸다는 이론이다.

따라서 이러한 이진 표현에서 0은 선택된 관측개체 1은 이상치라 표현한다면 유전자 알고리듬은 어떤 목적함수에 악영향을 주는 이상치를 소멸해 가는 방향으로 진행하게 된다. 이에 다중 이상치가 존재할 경우 몇 개가 동시에 존재하는지를 모르거나 존재한다는 것 자체를 모른다 하더라도 유전자 알고리듬의 스키마 정리는 이를 해결할 수 있게 된다.

3.2. 이상치 탐색을 위한 적합도 함수(fitness function) 및 파라미터

개체집단에서 각 관측개체 조합을 미리 결정된 다변수 회귀 분석 모델을 적용하여 적합도 값을 구하고, 그 값이 최적에 가까운 염색체가 확률적으로 더 많이 선택되도록 한다. 이 과정에서 초기 개체집단으로부터 일부 열성 관측개체들의 조합은 제거되고, 일부 우성 관측개체들의 조합은 여러 번 선택된다. 이상치 선택의 방법에 적용할 때 적합도함수 함수는 아래와 같이 정의된다.

$$F = PRESS + Penalty = \sum_{i=0}^{n}(y_i - \hat{y}_{i,-i})^2 + \infty \text{ if outlier}>m, \text{ 0 otherwise}$$

$\hat{y}_{i,-i}$ 는 i번째 관측치를 뺀 후 만들어진 모델로 예측된 값이다. 여기서 Penalty는 특정개수(m)의 이상치만을 뽑아내기 위한 것으로 유전자 알고리듬의 적합도 함수값이 나빠져 더 많은 이상치를 뽑아내는 것을 방지하도록 한것이다. 또한 유전자 알고리듬을 적용하기 위한 파라미터는 다음과 같다.

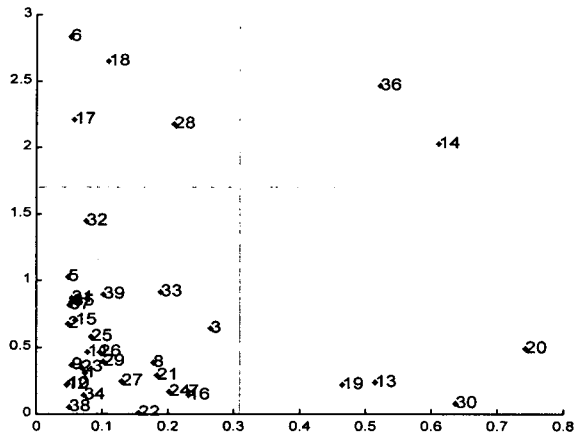
- 순위에 의한 확률바퀴선택(roulette wheel selection)
- 모집단크기(population size) : 100
- 교배율(crossover rate) : 0.3
- 돌연변이율(mutation rate) : 0.01

4. 데이터 분석

4.1 분석 데이터

이 절에서는 본 논문에서 제안된 유전자 알고리듬을 이용한 다중 이상치 탐색을 Sensitivity analysis in linear regression(Hadi, 1988)에 수록된 자료에 적용하였다. 이 데이터는 독립 변수 5개와 종속변수 1개, 관측개체는 총 49개로서 이상치의 탐색 및 각 방법의 성능 비교를 위해 39개의 모델링 데이터와 탐색 결과의 검증의 위해 10개의 검증 데이터로 나누어 분석을 실시하였다. 모든 데이터를 사용한 경우 RMSEC는 340, RMSEV는 293 이다.

[그림 1]은 이상치 진단의 가장 기본이 되는 지렛대값(leverage value)와 잔차(residual)을 그린 것으로 보통 이상치의 종류를 구분하는 데 사용된다. 여기서는 36, 14번 관측개체가 포함된 부분이 큰 지렛대값과 큰 잔차값을 가지는 것으로 보아 가장 모델에 많은 영향을 주는 이상치라 짐작할 수 있고, 13, 19, 20, 30은 큰 지렛대값과 작은 잔차값을 가지므로 모델에 좋은 영향을 주는 관측 개체임을 짐작할 수 있다.



[그림 1] Leverage(p^*) vs Residual(t^*)

또한 6, 17, 18, 28번 관측개체는 큰 잔차만을 가지므로 이상치일 가능성이 있다. [그림 1]에 의한 판단은 대략적인 윤곽은 보여주지만 구체적 진단은 [표 1]과 같은 진단치들을 이용해서 이루어진다. [표 1] 은 39 개의 데이터에 하나의 이상치 탐색을 위한 7가지 진단 방법을 적용한 것으로 굵은 글씨로 된 것이 기각역에 의해 이상치로 판단된 것이다. 각 진단 방법에 따라 다소 차이가 있는 결과를 주는 것을 볼 수 있다.

4.2. 단일 이상치 진단에 의한 순차적 방법

2장에서 설명한 7가지 방법 중 X 공간상에서의 이상치를 나타내는 지표인 p, p^*와 이상치를 발견하지 못한 C를 제외한 4가지 방법을 적용해 보기로 한다. 단일 이상치에 의한 방법으로 여러개의 이상치를 검색하기 위해서는 각 진단 방법을 통해 진단치의 값이 가장 큰 이상치 하나를 제거하고, 그 다음 새로운 진단치를 구해 이상치를 찾는 순차적 과정을 반복한다.

[표 2]는 CVR에 의한 방법으로 이상치를 하나씩 제거해 나간 결과로 모델의 성능을 나타낸 것이다. CVR은 이상치가 회귀 분석 계수의 분산에 영향을 미치는 정도로 생각할 수 있는데, 이 데이터에서 RMSEV의 개선이 별로 이루어지지 않은 것으로 보아 실제적인 모델 성능의 향상을 위한 지표로는 적합하지 않은 것을 알 수 있다.

[표 3]은 표준화된 잔차(residual)의 크기에 대한 진단치인 t^*로 이상치를 판단한 모델의 성능을 나타낸 것이고 6, 36번의 두 개의 이상치를 제거했을 때 RMSEV가 266으로 CVR에 의한 탐색 결과 보다 다소 좋은 결과를 보여주고 있다. 이 방법은 특히 잔차가 가장 큰 것부터 제거하는 방법으로 모델 내에서의 오차인 RMSEC에서는 확연한 증가를 가져오지만 모델의 안정성에 있어서는 많이 부족한 결과를 보여준다.

[표 4]는 WK, C^* 각각에 분석한 결과이지만 표에서 처럼 같은 결과를 보여 주고 있다. 두 방법 모두 영향 함수(influence function)에 근거한 진단치로서 WK는 회귀분석에서 어떤 관측 개체가 예측치 $\hat{y}$의 표준오차에 영향을 주는 정도로 설명이 되고 C^*는 2장의 식에서도 알 수 있듯이 WK와 거의 유사하고 분석결과도 동일하게 나타났다.

[표 1] 7가지 진단 방법의 적용

	p	t*	p*	CVR	C	WK	C*
1	0.0736	0.3051	0.0763	1.2759	0.0013	0.0797	0.1868
2	0.0487	0.6749	0.0620	1.1615	0.0040	0.1453	0.3406
3	0.2647	0.6405	0.2740	**1.5154**	0.0251	0.2825	0.6626
4	0.0538	0.8426	0.0743	1.1143	0.0068	0.1901	0.4459
5	0.0481	1.0261	0.0784	1.0405	0.0089	0.2196	0.5150
6	0.0508	**2.8369***	0.2416	**0.3298**	0.0592	0.6230	1.4611
7	0.2318	0.1650	0.2325	**1.5577**	0.0014	0.0696	0.1633
8	0.1776	0.3857	0.1814	1.4223	0.0055	0.1474	0.3456
9	0.0529	0.3696	0.0569	1.2379	0.0013	0.0827	0.1940
10	0.0476	0.2275	0.0491	1.2507	0.0004	0.0484	0.1136
11	0.0770	0.4703	0.0833	1.2503	0.0031	0.1254	0.2940
12	0.0466	0.2215	0.0481	1.2500	0.0004	0.0467	0.1095
13	**0.5149**	0.2367	**0.5157**	**2.4535**	0.0102	0.1183	0.2774
14	**0.6107**	**2.0224**	**0.6548**	**1.5013**	0.9777	**0.9861**	**2.3126**
15	0.0587	0.7007	0.0729	1.1662	0.0052	0.1647	0.3862
16	0.2317	0.1374	0.2321	1.5599	0.0010	0.0580	0.1360
17	0.0578	**2.2120**	0.1828	0.5436	0.0448	0.5163	1.2107
18	0.1091	**2.6477**	0.2692	**0.4113**	0.1211	**0.8256**	**1.9361**
19	**0.4664**	0.2221	**0.4673**	**2.2335**	0.0074	0.1108	0.2599
20	**0.7429***	0.4941	**0.7449***	**4.4706***	0.1203	0.2159	0.5064
21	0.1830	0.2871	0.1851	1.4497	0.0032	0.1110	0.2604
22	0.1561	0.0098	0.1561	1.4252	0.0000	0.0035	0.0083
23	0.0688	0.3549	0.0725	1.2616	0.0016	0.0899	0.2107
24	0.2018	0.1677	0.2025	**1.4989**	0.0012	0.0673	0.1578
25	0.0835	0.5843	0.0932	1.2314	0.0053	0.1616	0.3790
26	0.0961	0.4662	0.1022	1.2777	0.0039	0.1374	0.3223
27	0.1287	0.2437	0.1303	1.3651	0.0015	0.0816	0.1914
28	0.2079	**2.1704**	0.3095	0.6661	0.1852	**0.8807**	**2.0655**
29	0.1022	0.3986	0.1067	1.3005	0.0031	0.1208	0.2832
30	**0.6366**	0.0809	**0.6366**	**3.3054**	0.0020	0.0389	0.0913
31	0.0533	0.8726	0.0753	1.1033	0.0072	0.1961	0.4598
32	0.0760	1.4430	0.1325	0.8918	0.0277	0.3825	0.8970
33	0.1876	0.9106	0.2081	1.2699	0.0321	0.3555	0.8337
34	0.0716	0.1414	0.0722	1.2907	0.0003	0.0365	0.0855
35	0.0562	0.8413	0.0766	1.1176	0.0071	0.1937	0.4542
36	**0.5221**	**2.4639**	**0.5983**	0.8876	0.9583	**1.2308***	**2.8864***
37	0.0506	0.8185	0.0701	1.1188	0.0060	0.1795	0.4209
38	0.0499	0.0555	0.0500	1.2652	0.0000	0.0121	0.0283
39	0.1025	0.8973	0.1246	1.1545	0.0154	0.2722	0.6384

[표 2] CVR에 의한 이상치의 순차적 탐색

Step	Outlier by CVR	RMSEC	RMSEV
1	20	344	284
2	20, 30	348	282
3	20, 30, 13	336	315
4	20, 30, 13, 33	338	339

[표 3] t^* 에 의한 이상치의 순차적 탐색

Step	Outlier by t*	RMSEC	RMSEV
1	6	308	293
2	6, 36	**270**	**266**
3	6, 36, 18	242	276
4	6, 36, 18, 32	222	290
5	6, 36, 18, 32, 17	203	291

[표 4]에서 보면 36, 14, 6번의 3개의 관측개체가 제거되었을 때 RMSEV가 261로서 가장 좋았고, 여러 가지 진단 방법 중 가장 좋은 결과를 나타냈다.

[표 4] WK, C^*에 이상치의 순차적 탐색

Step	Outlier by C*, WK	RMSEC	RMSEV
1	36	316	268
2	36, 14	294	271
3	36, 14, 6	**248**	**261**
4	36, 14, 6, 18	222	265
5	36, 14, 6, 18, 7	220	265
6	36, 14, 6, 18, 7, 17	198	273

이제까지 하나의 이상치를 가정한 진단치에 의해서 이상치를 탐색하는 방법을 적용하여 보았는데 이러한 방법은 모두 통계적 가정 및 추론에 의한 방법으로서 다음과 같은 문제를 가지고 있다. 첫째로 이상치의 탐색이 순차적으로 이루어지므로 다중이상치에 의한 상호 관계를 고려할 수 없고 둘째로, 진단의 유의 수준 혹은 신뢰도를 어떻게 결정해야 하는가에 대한 문제, 즉 몇 개 까지의 이상치를 찾아내야 하는가에 대한 문제가 있다. 또한 어떤 진단 방법을 사용해야 하는가에 대한 문제도 있다.

4.3 유전자 알고리듬에 의한 이상치의 탐색

[표 5]는 유전자 알고리듬을 통한 이상치 탐색의 결과를 보여주고 있는데 4개까지의 결과는 여러 방법 중 가장 좋은 결과를 보여 주었던 WK, C^*에 의한 방법과 동일한 결과를 보여주고 있고 그 이후에는 WK, C^*에 의한 결과 보다 좋은 결과를 보여 주고 있다. WK, C^*의 방법과 같은 순차적 방법은 한번 제거된 이상치는 다시 모델에 돌아 올 수 없다는 점 때문에 해집합이 작아지지만 유전자 알고리듬은 모든 가능한 조합을 해집합으로 최적의 결과를 향해 진화함으로서 최적해에 근접하는 경향을 보여주고 본 연구의 결과에서도 가장 좋은 결과를 보여주고 있음을 알 수 있다.

[표 5] GA에 의한 이상치 탐색

탐색갯수	Outlier by GA	RMSEC	RMSEV
1	36	316	268
2	14, 36	294	271
3	6, 14, 36	**248**	**261**
4	6, 14, 18, 36	222	265
5	6, 7, 14, 33, 36	**244**	**259**
6	4, 6, 7, 14, 18, 36	211	261

5. 결론 및 추후 연구 과제

여러 방법에 비해 유전자 알고리듬을 이용한 다중 이상치의 탐색이 본 논문의 데이타에서 적합한 이상치를 찾아 주는 것을 확인할 수 있었다. 유전자 알고리듬을 이용한 이상치탐색 성과의 검증을 위해서는 몬테카를로 시뮬레이션(Monte Carlo simulation)이나 보다 많은 데이타에 대해 적용해볼 필요가 있다. 또한 최근의 다른 연구와 비교하고, 유전자 알고리듬의 성능

에 가장 중요한 역할을 하는 적합도 함수의 설계에 기존 이상치탐색에 대한 연구결과를 응용하여 적용해 봄으로써 그 개선안을 찾을수 있을 것이다.

참고문헌

[1] B. Walczak, D.L. Massart (1998), Multiple oulier detection recisited, *Chemometr. Intelligent Lab. Syst.* 41, 1-15.

[2] Andrzej S. Kosinski (1999), Aprocedure for the detection of multivariate outliers, *Computational Statistics & Data Analysis*, 29, 141-161.

[3] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord, Detection of preciction outliers and inliers in multivariate calibration, *Analytica chimica Acta*, 388, 283-301.

[4] Sangit Chatterjee, Matthew Laudato, Lucy A. Lynch, (1996), Genetic algorithms and their statistical applications : an introduction, *Computational Statistics & Data Analysis*, 22, 633-651.

[5] Samprit Chatterjee, Ali S. Hadi (1988), *Sensitivity Analysis in linear regression*, John Wiley & Sons.