벡터자기회귀 모형을 이용한 건강 지표 예측

김현서 ○ 최원석, 장병탁
서울대학교
{hskhexasu, dnjstjr1030, btzhang}@snu.ac.kr

Predicting health indicator using Vector autoregression(VAR)

Hyun Seo Kim○ Won Seok Choi, Byoung-Tak Zhang
Seoul National University

요 약

건강 지표 데이터는 다차원 시계열 데이터로 데이터간의 인과관계가 복잡하고 시간 간격도 일정치 않은 경우가 많다. 이런 경우를 분석하기 위하여 일변량 분석에 주로 사용되던 자기회귀 이동평균 모형(Autoregressive moving average: ARIMA)을 발전시킨 벡터자기회귀 모형(vector autoregression, VAR)이 개발되었다. 이 연구에서는 다차원 시계열 데이터를 분석하는데 주로 사용되는 벡터자기회귀모형을 이용하여 세 종류의 건강 지표 데이터를 분석, 예측하였다.

1. 서 론

건강 지표 데이터는 다차원 시계열데이터로서 시계열 데이터 분석에서 중요한 위치를 차지하고 있다. 그동안은 심박동 신호 데이터나, 초음파 데이터 등 전파를 이용하여 짧은 시간 동안 측정한 시계열 데이터가 건강 지표 데이터 분석에 주로 등장하였다. 하지만 최근, 그보다 더 넓은 범위의, 오랫동안 한 사람을 추적해가며 건강 상태를 조사한 데이터들의 분석 필요성이 대두되고 있다.

이 연구는 인바디(Inbody), 스트레스, 양자 측정 기기로 측정한 다차원 데이터를 특정 길이 이상의 샘플들에 한해 벡터자기회귀 모형에 적용함으로써 샘플들의 그 다음 측정값을 예측한다.

이 연구에서 데이터를 얻은 기기들에 대해 먼저 살펴보자면, 첫째로 인바디 측정 기기는 생체전기저항 분석법(Bioelectric impedance analysis, BIA)을 이용하는 체성분분석기이다. 특정 주파수의 약한 전류를 몸에 흘려보내어서 전류가 잘 흐르는 근육, 혈액과 잘 흐르지 않는 체지방, 피부를 분리 분석하여 체지방의 양, 두께, 근육의 양과 두께 등을 측정할 수 있다.

두번째로, 스트레스 측정 기기는 심장이 뛸 때마다 말초혈관의 압력과 구경의 변화로 전파되는 동맥계 파동을 광센서로 측정하는 광전식 맥파계이다. 이를 이용해 심혈관기능과 자율신경 균형상태를 측정할 수 있고, 스트레스 저항력을 추론해낼 수 있다.

마지막으로 이용한 양자 분석기는 Quantum Magnetic Resonance Analyzer로도 불리며, 특정 주파수의 전파를 이용해 세포의 자기장을 측정하여 몸 상태를 알아내는 기기이다.

이렇게 세 가지의 분석 기기에서 나온 데이터를 표 1과 같이 정리하여 벡터자기회귀 모형의 입력 값으로 하였다.

표 1. 측정 데이터의 성질

측정 기기	인바디	스트레스	양자
측정 빈도	주 2~3회	주 2~3회	주 1회
데이터길이	101회 이상	101회 이상	30회 이상
샘플 수	485	862	510
차원 수	28	21	30

2. 벡터자기회귀 모형

벡터자기회귀 모형은 자기회귀 이동평균 모형(Autoregressive moving average: ARIMA)가 가지는 일변량 분석이라는 한계를 보완한 모형이다(1).

시간 t에서의 N차원 다변량 정상시계열로 구성된 $X_t = (X_{1,t}, X_{2,t}, \cdots, X_{N,t})$가 p 시차인 자기회귀과정으로 구성된 벡터자기회귀 모형 VAR(p)이라 할 때,

$$X_t = C + A_1X_{t-1} + A_2X_{t-2} + \cdots + A_pX_{t-p} + e_t$$

으로 표현할 수 있다. C는 (Nx1) 상수 벡터, A는 현시점의 변수와 시차변수들 간 시차회귀 계수 (NxN) 행렬, e는 노이즈이고, 시차 p는 시간 t로부터 p번째 뒤의(이전) 시점을 말한다(1).

벡터자기회귀 모형을 구현하는 것은 검증-학습-예측의 3단계로 나뉠 수 있다. 먼저 검증단계에서는 원본 시계열 데이터의 특성을 개형을 통해 대략적으로 파악하고 시계열 데이터 차원(변수)간의 그레인저 인과관계(Granger Causality)를 파악, Johanson's Cointegration Test, Augmented Dickey-Fuller Test(ADF test)를 진행해야한다.

그레인저 인과관계는 두 개의 시계열 데이터가 있을 때, 한 시계열이 다른 하나 시계열을 예측할 때 도움이 되는지를 판단하는 테스트이다(2). 그레인저 인과관계 테스트를 하기 위해서는 두 개의 시계열과 시차값이 필

요하다. 그리고 그 두 변수가 서로에게 영향을 주는지 양방향으로 테스트한다. 또한 여느 시계열 분석과 유사하게 정상성(stationary)를 만족하는 것을 전제조건으로 하고 있다. 시계열 데이터가 정상성을 만족하는지 알기 위해서는 Augmented Dickey-Fuller Test(ADF test)를 해보면 된다(3). 만약 시계열 데이터가 정상성을 만족하지 않는다면 두 시계열 간의 차이(difference)를 구해 정상성을 만족하도록 할 수 있다.

그레인저 인과관계 테스트를 끝내면 Johanson's Cointegration Test를 거친다. 테스트의 결과값인 공적분 벡터(cointegrating vector)는 입력 값으로 들어왔던 시계열들이 같이 움직이는 경향성을 나타내게 된다(4). 이런 과정을 통해 벡터자기회귀모형에 넣을 입력 시계열 데이터들을 정제하고 나면 시차 p를 무엇으로 할지 결정해야 한다. p를 결정하는 데는 4가지 기준이 있다. Akaike information criterion(AIC), Bayesian information criterion(BIC), Hannan-Quinn Criterion(HQ), Final Prediction error criterion(FPE)이다. 각각 장단점이 있기 때문에 상황에 맞는 기준을 고르면 된다. 보통 AIC를 많이 쓰고, AIC를 최소화시키는, 즉 우도(likelihood)를 가장 크게 하면서 변수 개수는 가장 적은, p값을 고르게 된다.

고른 p값을 이용해 VAR 모델을 학습시키고 그 후 예측을 하기 전 Durbin Watson test를 거쳐 학습시키고 남은 데이터들 간에도 연관성이 있나 살펴보게 된다. 만약 남은 데이터 간 연관성이 있다면 모델에 학습시킬 수 있는 것이 남아있는 것이므로 다시 한번 학습 파라미터를 조정해야 하고, 아니라면 그대로 진행하여 예측을 하고 예측 오차를 비교하면 된다(5). 예측 오차를 보는 방법에는 주로 mean average percentage error(mape), mean error(me), mean average error(mae), mean percentage error(mpe), root mean square error(rmse), correlation(corr), minmax를 보는데, 이번 실험에서는 mape를 이용하였다.

3. 분석 결과

먼저, 모델의 적합성을 평가하기에 가장 유용한 인바디 데이터를 가지고 실험을 해보았다. 인바디 데이터는 스트레스 데이터나 양자 데이터보다 접하는 빈도가 높아 결과예측이나 요소 간 영향을 측정하였을 때 결과를 이해하기가 쉽다. 또한 다른 데이터에 비해 다양한 성격의 데이터를 다루기 때문에 경향성이 더 예측하기 어렵다는 것이 흥미로웠다.

그리하여 그림 1에서 101회 이상 측정한 고객의 데이터를 시간 간격이 일정하도록 적절히 interpolation한 후 개형을 살펴보았다. 상당히 유사한 경향을 보이는 데이터들도 있고, wc_mc 같이 다른 데이터들과 확연히 다른 개형을 보이는 데이터도 있었다. 그림 1의 항목 설명은 표 2에 하였다.

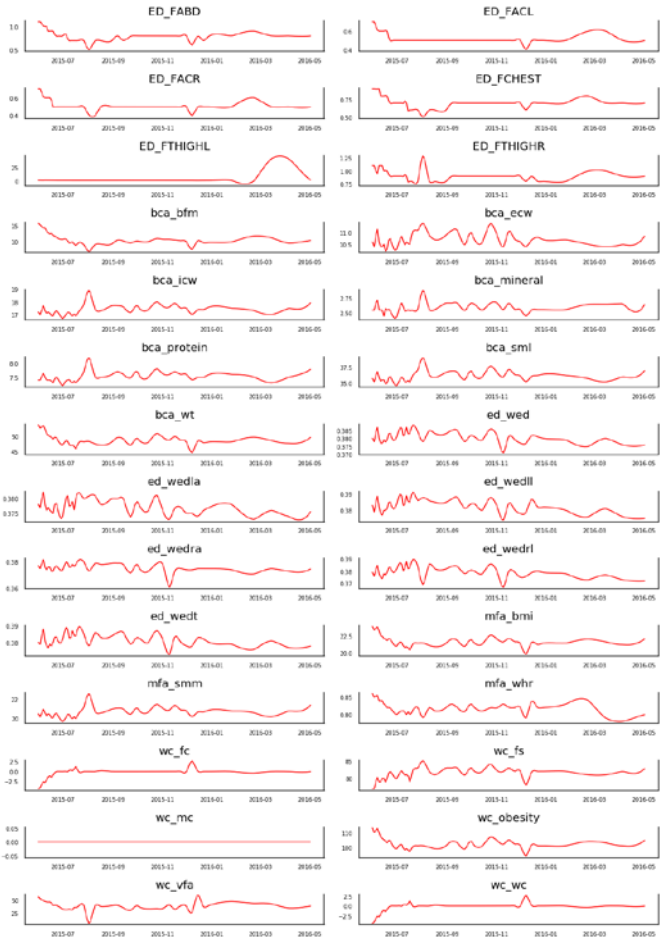

그림 1. 28차원 인바디 원본 데이터의 개형

표 2. 28차원 인바디 항목 설명

항목	설명	항목	설명
bca_icw	내수분	Wc_vfa	내장지방
Bca_ecw	외수분	Wc_wc	체중조절
Bca_protein	단백질	Wc_fc	지방조절
Bca_mineral	미네랄	Wc_mc	근육조절
Bca_sml	근육량	Wc_fs	종합점수
Bca_wt	체중	Wc_obesity	지방도
Bca_bfm	체지방	Mfa_bmi	BMI
Ed_wed	부종	Mfa_whr	복부지방률
Ed_wedra	부종오른팔	Mfa_smm	골격근량
Ed_wedla	부종왼팔	Ed_fchest	가슴지방둘레
Ed_wedt	부종몸통	Ed_fabd	몸통지방둘레
Ed_wedrl	부종오른다리	Ed_facr	오른팔지방둘레
Ed_wedll	부종왼다리	Ed_facl	왼팔지방둘레
Ed_fthighr	오른허벅지지방둘레	Ed_fthighl	왼허벅지지방둘레

이런 시계열 데이터들 사이 관계를 검증하고, 서로 연관성 있는 12개만을 뽑아 학습을 시키고 예측을 해보았다. 테스트로 쓰는 데이터는 원본 데이터의 10% 길이가 되도록 설정하였다.

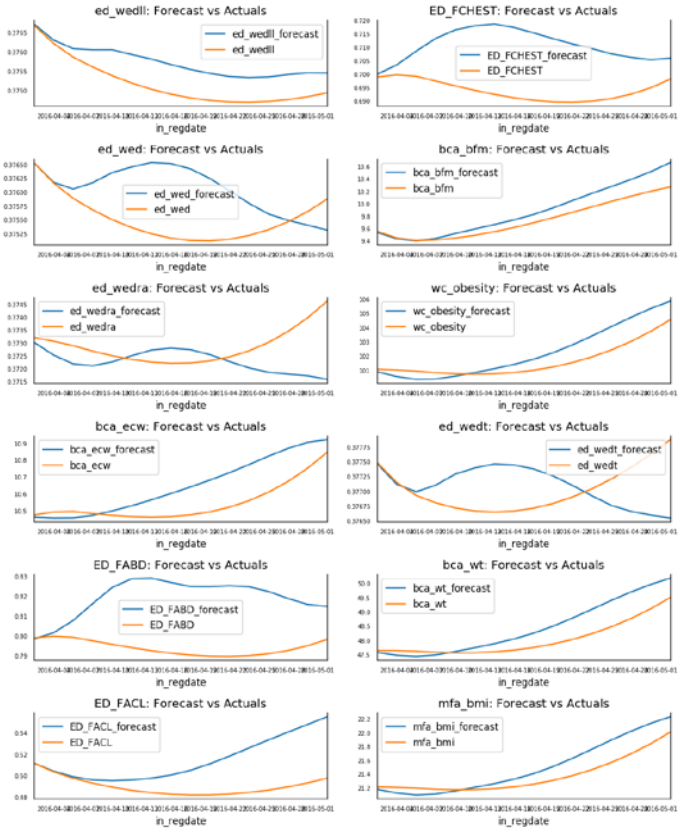

그림 2. 인바디를 101회이상 측정한 샘플의 인바디 데이터 예측 그래프

그림 2에서 나타난 파란 선은 예측값이고 주황색 선은 원래 값이다. 절반 이상이 매우 유사하게 예측된 것을 알 수 있고, 나머지도 예측 오차를 표3을 통해 살펴보았을 때 전부 0.05% 이하의 오차를 갖는 것을 알 수 있다.

표 3. 예측 값들의 MAPE 오차

항목	Mape (%)	항목	Mape (%)
Ed_wedll	0.0015	Wc_obesity	0.0083
Ed_fchest	0.0239	Bca_ecw	0.0109
Ed_wed	0.0017	Ed_wedt	0.0014
Bca_bfm	0.0151	Ed_fabd	0.0315
Ed_wedra	0.0022	Bca_wt	0.0097
Ed_facl	0.0498	Mfa_bmi	0.0076

4. 결론 및 향후 연구

이번 실험에서 다루었던 인바디 데이터를 485개 샘플에 대해 모두 학습 및 예측을 해보았을 때, 많은 수의 샘플들이 데이터에 역행렬이 존재하지 않거나 non positive definite matrix 성격이 있어 VAR 모델을 적용하지 못하였다. 향후 연구에서는 그런 특성을 갖는 샘플들에 대해 임의의 매우 작은 수가 곱해진 단위 행렬을 더해 역행렬을 가지도록 하거나 non positive definite matrix가 positive definite해지도록 변환할 수 있을 것이다. 또, 그렇게 변환하였을 때 원본 데이터의 결과와 큰 차이가 나지 않도록 하는 것도 중요할 것이다.

그리고 이번 연구에서는 다루기 쉬웠던 인바디 데이터만을 가지고 실험해본 만큼, 다른 두 데이터, 스트레스와 양자 데이터를 가지고 실험해보고, 마지막에는 세 종류의 데이터를 모두 합쳐서 가장 영향을 많이 주는 요소를 고르고 그것들의 예측 값을 계산해볼 계획이다.

5. 감사의 글

이 논문은 2019년도 ㈜쥬비스의 지원을 받아 수행된 연구이며, 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind)와 한국산업기술진흥원(P0006720-GENKO)의 지원을 일부 받았음. 그리고 서울대학교 컴퓨터연구소로부터 연구장비 및 공간을 지원 받았음.

참 고 문 헌

(1) 문권순, “벡터자기회귀(VAR)모형의 이해”, 통계분석연구, 제2권, 제1호, 23-56, 1997.

(2) Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". Econometrica. 37 (3): 424-438. 1969.

(3) 조성일, 최종수, “몬테 카를로 실험에 의한 Augmented Dickey-Fuller 단위근 검정법의 검정력에 관한 연구”, 통계연구, 제10권, 제1호, 165-188, 2005.

(4) Johansen, Søren. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models". Econometrica. 59 (6): 1551-1580. 1991.

(5) 남준우, 이한식, 허인, 계량경제학(4판), 294-296, 397-400. 2016.