The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.627
33(5), 627–637

Predicting the number of disease occurrence using recurrent neural network

Seunghyeon Lee[a] · In-Kwon Yeo[b,1]

[a]SK C&C; [b]Department of Statistics, Sookmyung Women's University

(Received June 23, 2020; Revised July 7, 2020; Accepted July 7, 2020)

Abstract

In this paper, the 1.24 million elderly patient medical data (HIRA-APS-2014-0053) provided by the Health Insurance Review and Assessment Service and weather data are analyzed with generalized estimating equation (GEE) model and long short term memory (LSTM) based recurrent neural network (RNN) model to predict the number of disease occurrence. To this end, we estimate the patient's residence as the area of the served medical institution, and the local weather data and medical data were merged. The status of disease occurrence is divided into three categories(occurrence of disease of interest, occurrence of other disease, no occurrence) during a week. The probabilities of categories are estimated by the GEE model and the RNN model. The number of cases of categories are predicted by adding the probabilities of categories. The comparison result shows that predictions of RNN model are more accurate than that of GEE model

Keywords: elderly patient medical data, weather data, GEE, RNN

1. 서론

기상 현상은 기온, 기압, 습도, 미세먼지 등을 포함하여 대기 중에서 일어나는 각종 물리적 현상을 말하며 인간 생활 전반에 영향을 주고 있다. 현재 보건분야의 핫이슈는 코로나이지만 이 사태가 벌어지기 전까지 미세먼지로 인한 건강문제는 국민들이 가장 관심을 가지는 주요 문제 중 하나였고 정부와 지자체에서도 이를 해결하기 위한 정책을 지속적으로 세워 왔다. 국내·외적으로 기상 인자와 건강에 관련된 많은 연구가 진행되고 있으며 특히 순환기 계통 질환과 호흡기 계통 질환과 기상과의 관계분석이 많이 이루어지고 있다 (안혜연 등, 2016; 주영수, 2008; Basu와 Samet, 2002). 어떤 환경 하에서 특정 질병의 환자의 수가 얼마나 되는지 예측할 수 있다면 이는 보건산업 및 행정을 관리하는데 중요한 정보가 될 수 있다.

우리나라의 경우 평균수명의 증가와 저출산율로 인해 세계 어느 나라보다 빠르게 고령화가 되어가고 있고 세계적인 의료 체계 덕분에 고령자의 병원 접근성이 매우 높은 편이다. 고령층은 65세 이상 인구로 2014년 고령인구는 전체 인구 중 12.7%(638만6천명)를 차지했으며, 2019년 기준 고령인구는 14.9%(768만5천명)를 차지한다. 고령인구의 지속적인 증가로 인하여 사회경제적인 변화가 필요할 것으로 보여 연구를 진행하게 되었다. 이 논문에서는 우리나라 환자의 상당 부분을 차지하고 있는 고령환자

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea. E-mail: inkwon@sookmyung.ac.kr

Table 2.1. Summary of samples

연령	성별		합계(%)
	남성	여성	
60	34,383	39,236	73,619 (5.7)
65	184,439	205,798	390,237 (30.2)
70	154,864	198,408	353,272 (27.3)
75	98,334	151,202	249,536 (19.3)
80	67,454	159,420	226,874 (17.5)
합계(%)	539,474 (41.7)	754,064 (58.3)	1,293,538 (100)

들의 건강보험심사평가원(이하 심평원) 의료보험 청구자료를 분석하여 성과 연령과 같은 인구생태학적 정보와 거주 지역의 기상 자료가 특정 주상병에 걸릴 확률에 어떻게 영향을 주는지를 모형화하고 해당 지역에 발생한 환자의 수를 예측해 본다.

이 논문에서 사용된 의료자료는 심평원에서 공개한 2014년 의료보험청구(HIRA-APS-2014-0053) 자료이며 기상자료는 기상청 홈페이지에서 다운로드 받은 것으로 분석에는 Ballester 등 (2003)에서 사용되었던 기온, 습도, 기압, 바람, 강수 등의 기상인자를 고려하였다. 요즘 종속성을 가지는 자료를 분석할 때 많이 사용되고 있는 신경망 모형 중 순환신경망(recurrent neural network; RNN) 방법을 이용하여 지역별 기상정보와 인구상태학적 정보가 호흡기질환, 순환기질환, 피부질환 발병 확률에 미치는 영향을 분석하고 이를 토대로 지역별 환자수를 예측해 본다. 자료를 훈련자료(training data)와 시험자료(test data)로 나누고 훈련자료를 이용하여 질병에 걸릴 확률을 RNN으로 모형 적합을 시키고 시험자료를 이용하여 실제 환자수와 예측된 환자 수를 비교하였다. RNN에 의한 성능이 어느 정도 되는지 알아보기 위해 동일 자료를 통계적 모형인 일반화추정방정식(generalized estimating equation; GEE)으로 모형 적합과 환자 수를 예측하고 결과를 비교하였다.

2. 자료 소개 및 전처리

2.1. 건강보험심사평가원 진료데이터

본 연구에서 분석한 데이터(HIRA-APS-2014-0053)는 건강보험 청구자료로 진료 개시일 기준 1년 동안 진료를 받은 환자를 대상으로 표본 추출한 이차 자료이다. 환자의 성별과 연령구간(60세부터 5세 단위)에 따라 층화계통추출된 자료이며 추출 비율은 10%이며 연령과 성별 정보가 없거나 60세 미만인 자료를 제외하고 분석에 사용되는 데이터의 환자 수는 1,293,538명이다. Table 2.1은 해당 환자들의 성별과 연령대로 나누어 정리한 것이다.

이 데이터는 명세서 일반내역, 진료내역, 상병, 처방전 상세내역 그리고 요양기관 현황 파일이 별도로 있으나 개인정보보호를 위해 각 환자에 대해 '명세서 연결코드'라는 변수에 일련번호를 무작위 부여하고 이 번호를 통해 각 환자들의 데이터를 연결할 수 있도록 하였다. 명세서 일반내역에 '요양기관 대체키'를 포함시켜 진료를 받은 요양병원의 현황과 연결시킬 수 있도록 하였다. 자세한 변수 내역은 Table 2.2에 표시하였다. 전체 명세서는 26,443,871개이고 이 중 한 주에 두 번 이상의 내원 기록이 있는 환자들의 전체 명세서는 10,125,584개로 약 38.29%를 차지한다.

요양기관 현황의 경우 해당기관이 어느 병원인지는 알 수 없으나 소재 시도는 알 수 있다. 이를 통해 환자가 어느 지역의 병원에서 진료를 받았는지는 알 수 있으나 환자의 실제 거주지를 알 수 없다. 지역별로 기상자료와 연결하려면 적절한 가정 하에서 환자의 거주지 추정하는 작업이 필요하다. 이에 대한 자세한 내용은 아래의 전처리 과정에서 설명한다. 상병에는 치료나 검사에 대한 환자의 요구가 가장 컸

Table 2.2. List of HIRA's variables

파일	내역
명세서 일반내역	*명세서 연결코드*, **청구구분코드, 청구형태코드, 서식코드, 수진자 고유번호, 연령군, 추출확률, 샘플가중치, 수진자연령, 성별구분코드, 보험자코드, 주상병코드, 부상병코드, 요양개시일자, 요양만료일자, 진료결과구분코드, 진료과목코드, 청구DRG코드, 최초입원일자,** 입원도착경로구분코드, 공상구분코드, 요양일수, 내원 일수, 심결요양급여비용총액, 심결본인부담금, 심결보험자 부담금, 수술 여부, 특정기호 구분코드, 의료급여종별코드, 방사선 진단 여부, 방사선 치료 여부, *요양기관 대체키*
진료내역	*명세서 연결코드*, **항코드, 목코드, 분류유형코드, 분류코드, 1-2구분코드, 단가, 1회 투약량, 1일 투약량, 1일 투여량 실시횟수, 총투여일수 또는 실시횟수, 총 사용량 또는 실시횟수, 금액, 가산적용금액, 출번호, 일반명 코드**
상병	*명세서 연결코드*, **상병일련번호, 상병진료과목코드, 청구상병기호, 청구진료과목코드**
처방전 상세내역	*명세서 연결코드*, **줄번호, 처방전 교부번호, 분류유형코드, 단가, 1회 투약량, 1일 투약량,** 총투여일수 또는 실시횟수, 총사용량 또는 실시횟수, 금액, 일반명 코드
요양기관 현황	*요양기관 대체키*, **요양기관 종별코드, 설립구분, 시·도코드, CT유무, MRI 유무, PET** 유무, 최종 청구월(월말기준), 병상수준, 50병상 당 의사수, 50병상 당 치과의사 수, 50병상 당 한의사 수, 50병상 당 간호사 수

Table 2.3. The number of major disease patients

연령	성별	호흡계통	소화계통	피부·피하
60	여성	144,107	144,385	70,578
	남성	119,480	130,252	63,511
65	여성	134,596	132,519	68,259
	남성	103,569	105,844	55,649
70	여성	97,669	96,897	50,193
	남성	67,286	66,930	36,321
75	여성	86,253	80,023	43,747
	남성	43,908	41,245	24,073
80	여성	20,363	18,365	7,661
	남성	15,967	16,495	7,126
합계		833,198	832,955	427,118

던 주상병이 있으며 이는 질병의 유사성에 따라 체계화된 코드를 모아둔 한국표준질병사인분류(Korean Standard Classification of Diseases; KCD)를 참고한 주상병 코드로 표시되어 있다.

기존 문헌 연구를 통해 이 주상병 중 기상 요인과 밀접하게 관련이 있을 것이라 예상되는 질병은 호흡계통 질환, 순환계통 질환, 피부 및 피하조직 질환이다. Table 2.3은 이들 주상병에 대해 성별과 연령을 구분하여 표시한 것으로 각 칸의 빈도는 해당 주상병이 한 번 이상 발병한 환자들의 수를 의미한다. 자료분석은 각 질환별로 나누어 분석하였으며 각 질환에 대해 반응변수를 3개의 범주(질병에 걸리지 않음, 해당 질병 발생, 그 외의 질병 발생)로 나누어 분석하였다. 호흡계총 질환은 질병코드 I에 해당하고 질병코드 J와 L은 순환계통 질환과 피부 및 피하조직의 질환 코드로 자세한 상세내역은 한국표준질병사인분류(www.koicd.kr)에서 확인할 수 있다.

2.2. 기상 데이터

기상청의 기상자료 개방포털에는 전국의 관측 지점들에서 종관기상관측장비(automated surface observing system; ASOS)를 이용하여 수집된 일별 평균기온, 일교차, 일 강수량, 평균 풍속, 평균 상대

습도, 평균 증기압, 평균 현지기압, 평균 해면기압, 합계 일조시간 자료를 제공하고 있다. 본 연구에서는 이 포털에서 제공한 2014년 1월 1일부터 2014년 12월 31일까지의 관측소별 자료를 16개 시도로 묶어 주별 평균값을 구하였다. 분석에서는 통상적으로 보건 및 질병에 영향을 줄 것으로 예상되는 기온, 습도, 기압, 일조와 관련된 8개 변수를 선택하였다. 일반적으로 질병은 이전 주의 기상 상황에도 영향을 받을 수 있기 때문에 분석에서는 해당 주의 자료뿐만 아니라 이전 주의 자료도 설명변수에 포함 시켰다.

2.3. 자료병합

기상자료와 심평원 자료를 주별(2주차에서 52주차 까지)로 나누어 기상자료는 각 변수의 주간 평균을 계산하였고 심평원 자료는 각 환자별로 51개 주에서 질병감염상태에 대한 세 범주로 구분하였다. 한 주의 시작은 일요일로 하고 주별 분석이기 때문에 한 주가 7일이 되지 않는 1월의 첫 번째 주와 12월의 마지막 주는 분석에 제외하였다.

기상 요인이 질병 발생에 미치는 영향을 확인하기 위해서 환자의 거주지역에 대한 정보가 필요하지만 심평원 데이터에는 환자의 거주지역에 대한 정보가 없다. 하지만 심평원 데이터는 명세서를 기반으로 하여 내원한 병원의 지역을 확인할 수 있으며 환자의 거주지역은 내원한 병원의 위치와 밀접한 관련이 있을 것으로 예상되어 각 주에 대해 다음의 규칙을 이용하여 거주지역을 추정한다.

1. 내원 기록이 1개인 경우 내원한 병원의 지역을 거주지역으로 추정한다.
2. 내원 기록이 없는 경우 가장 근접한 날짜의 내원 지역을 거주지역으로 추정한다.
3. 내원 기록이 2개 이상인 경우, 일 기준으로 빠른 날짜에 내원한 병원의 지역을 해당 주의 지역으로 추정한다.

기상 데이터와 심평원 데이터를 각각 전처리하고 최종 데이터를 만드는 과정은 다음과 같다.

- 심평원 데이터: 연결 변수를 이용해 명세서 일반내역, 상병, 요양기관 현황 데이터를 합친 하나의 심평원 데이터를 만든다.
- 심평원 데이터: 각 환자에 대해 2주부터 52주까지 위의 거주지역 배분규칙에 따라 거주지역을 추정한다.
- 심평원 데이터: 해당하는 주에서의 발병 여부 변수를 만든다. 발병 여부 변수는 질병에 걸리지 않음, 관심 질병 발생, 그 외의 질병 발생으로 구분한다.
- 통합 데이터: 각 환자에 대해 시·도별 51개 주의 주별 평균기상 데이터와 심평원 데이터와 합친다. 이때, 설명변수에는 이전 주의 기상 데이터도 포함되어 있다.
- 통합 데이터: 결측값이거나 오류가 있는 관측치의 경우 삭제한다.
- 통합 데이터: 기상 데이터의 경우 각 기상 변수마다 표준화한다. 통계모형에서는 문제가 되지 않으나 신경망 모형에서는 학습률(learning rate)가 설명변수의 척도에 영향을 크게 주고받기 때문에 일반적으로 설명변수를 0과 1사이의 값으로 표준화를 먼저 실시한다.

3. 분석모형

분석할 자료는 어떤 환자에 대해 51 주간 세 상태(질병에 걸리지 않음, 관심 질병 발생, 그 외의 질병 발생) 중 어떤 상태였는지를 반응변수, 해당 주의 기상 데이터와 환자의 인구생태학적 정보를 설명변수로 구성한 것이다. 이는 전체적으로 개별 환자 내에서는 종속성이 있고 환자들 간에는 독립이라고 가정할

수 있는 경시적 자료(longitudinal data)로 볼 수 있다. 통계학적으로 보면, 어떤 i번째 환자의 j번째 시점에서의 질병 상태를 $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$라고 표시하자. 여기서 $(Y_{ij1}, Y_{ij2}, Y_{ij3})$는 질병에 걸리지 않은 경우 $(1,0,0)$, 관심 질병에 걸린 경우 $(0,1,0)$, 다른 질병에 걸린 경우 $(0,0,1)$의 값을 가지고 다음과 같이 범주가 세개인 다항분포를 따른다.

$$(Y_{ij1}, Y_{ij2}, Y_{ij3}) \sim M(1; \theta_{ij1}, \theta_{ij2}, \theta_{ij3}).$$

모수 θ_{ijk}는 i번째 환자가 j 시점에서 k번째 상태에 있을 확률로 이 환자의 연령과 성별, 거주지역의 기상자료에 영향을 받을 것이라고 가정한다.

각 환자들의 주별 반응변수들 간 상관성이 존재할 수 있기 때문에 통계적 모형으로 일반화 선형혼합모형(generalized linear mixed model; GLMM)과 GEE를 이용할 수 있다. 일반화 선형혼합모형은 설명변수에 고정효과와 변량효과 모두를 포함한다. 여기에서 변량효과는 동일 환자의 자료들 간에 미치는 그룹의 영향력을 간접적으로 모형화할 수 있다. 한편, GEE는 동일 환자의 자료들 간 상관관계를 모형적합식에 직접 설정하고 있다. 다항반응변수에 대한 GEE에서는 기준범주로짓모형(baseline category logit model)처럼

$$\log\left(\frac{\theta_{ijk}}{\theta_{ij3}}\right) = \beta_{0k} + \boldsymbol{\beta}_k \boldsymbol{x}_{ij}, \quad k = 1, 2$$

가정하고 회귀계수 $\boldsymbol{\beta}$는 다음과 같은 추정방정식을 통해 추정한다.

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\rho}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\theta}_i) = \boldsymbol{0},$$

여기서 $\boldsymbol{D}_i = \partial\boldsymbol{\theta}_i/\partial\boldsymbol{\beta}$이고 $\boldsymbol{V}_i$는 $\boldsymbol{Y}_i$의 가상공분산(working covariance)이며 현재 분석에서 51주 동안 세 범주에 대한 분석이기 때문에 $51(3-1) \times 51(3-1)$ 행렬로 구성된다. 이와 관련된 자세한 GEE 분석은 Touloumis 등 (2013)을 참조할 수 있고 R에서는 multgee라는 패키지를 제공하고 있다.

우리가 분석해야 할 데이터가 매우 커 R로는 분석이 불가능했으며 실제 분석에서는 SAS를 이용하였다. SAS 9.4/M4버전에서 명목형 다항분포에 대한 'PROC GEE'를 사용할 수 있었으며 다항분포인 경우에는 가상관행렬(working correlation matrix)은 독립 형태만 가능했다. 이러한 이유로 본 연구에서는 GEE와 RNN을 비교하는 연구를 진행하였다. GEE 모형과 RNN 모형에 대한 내용은 기존 문헌에 방대하게 나와 있어 별도로 내용을 설명하지 않고 해당모형에서 변수를 어떻게 설정했는지를 중심으로 설명하고자 한다 (GEE는 Diggle 등 (2002), Hardin과 Hilbe (2003)). 아래의 SAS 프로그램은 실제분석에서 사용된 코드로 'Hira'는 데이터세트 이름, Id는 환자의 '명세서 연결코드', Y는 질병상태를 $(1,2,3)$으로 표시한 것이다. 설명변수에는 해당 주의 8개 기상자료(Weather)와 이전 주의 기상자료(pWeather), 지역, 성별 그리고 연령이 포함되었다. 연령의 경우 수치자료이지만 5세 단위로 구분되어 있고 연령대에 따라 질병 감염에 증감 패턴이 있다고 보기 어렵기 때문에 요인으로 처리하였다.

```
PROC GEE DATA = Hira;
    CLASS Id Y Area Gender Age Week;
    MODEL Y = Weather1 - Weater8 pWeather1 - pWeater8 Area Gender Age
        / DIST = MULTINOMIAL LINK = GLOGIT;
    REPEATED SUBJECT = ID / WITHIN = Week;
    OUTPUT OUT = Result PROB=Prob;
RUN;
```

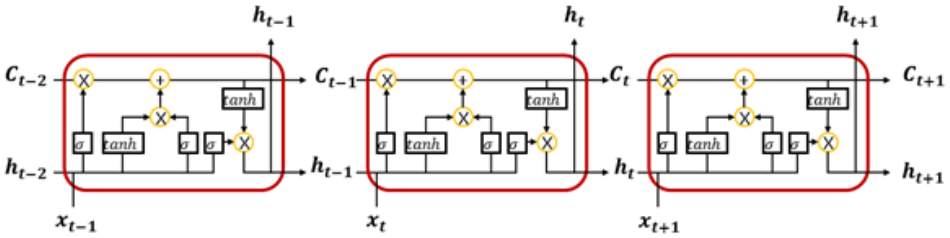

Figure 3.1. Structure of RNN based on LSTM. RNN = recurrent neural network; LSTM = long short term memory.

현재 자료와 같이 분석자료가 시간순으로 순차적으로 만들어져 있고 자료들간 관련성이 존재하는 경우 사용되는 대표적인 기계학습 방법으로는 RNN이 있다. RNN은 Hopfield (1982)에 의해 고안된 알고리즘이며 본 연구에서는 Hochreiter와 Schmidhuber (1997)이 제안한 long short term memory (LSTM) 기반 RNN을 이용하여 시간의 순서에 따른 주요 질병의 발병 빈도수를 예측하였다.

LSTM 기반 RNN 모형은 모형적합에 있어 과거의 정보를 활용한다. Figure 3.1에서 보는 것과 같이 입력값 x_t가 유닛셀(unit cell)에 입력되었을 때, 망각 게이트, 입력게이트, 출력게이트로 이루어진 3개의 게이트를 복합적으로 거친 후 h_t라는 값을 출력한다. 이 출력값은 다음 시점의 셀의 입력값으로 사용되며 이러한 과정을 거쳐 반복적으로 정보를 업데이트한다. 이 과정을 통해 셀 스테이트(C_t)는 유닛셀에서 동일한 게이트를 공유하면서 중요한 정보를 다음 단계로 전달한다. 그림에서 σ로 표시된 것은 시그모이드 함수를 의미한다.

LSTM 기반 RNN을 수행하기 위해서는 활성화 함수(activate function), 학습률(learning rate), 최적화 알고리즘(optimizer), 손실함수(loss function), 은닉층(hidden layers), 출력층 포함여부 등을 정해야 한다. 본 연구에 적용된 RNN에서는 51주 자료이기 때문에 순차적으로 읽는 단계는 51개이며 반응변수의 형태에 맞게 손실함수는 교차엔트로피(cross-entropy)를 적용했다. 다양한 세팅 하에서의 실험을 진행하여 은닉층은 8개, 학습률은 0.0002으로 선택하였으며 Adam Optimizer를 이용하였다.

분석에서는 연산의 효율성을 고려하여 데이터를 한꺼번에 입력하지 않고 적절한 크기로 나누어 학습을 시키는데 이를 위한 배치(batch)의 크기와 epoch을 정해야 하는데 훈련 데이터의 배치는 433이며, Epoch은 세 질환 모두 2,000번으로 했다. 2,000번의 epoch이 실행되고 추정된 계수값을 이용해 평가 데이터에 적용할 때에도 배치를 이용하여 해당 관측값에서의 세 가지 범주에 대한 확률값을 얻을 수 있다. 분석은 64bit Windows 10운영체제 하에서 Tensorflow 1.8을 NVIDIA GeForce GTX 1080 Ti가 장착된 GPU컴퓨터에서 실행되었다.

4. 분석결과와 모형 비교

전체 데이터 중 70%를 무작위로 선택하여 훈련자료로 사용하였고 나머지 30%는 모형비교를 위한 검정자료로 사용하였다. Table 4.1은 호흡기 질환 자료에 대한 GEE의 분석결과로 대부분의 공변량은 유의한 것으로 나타났으며 지역, 성별, 연령의 요인은 기준 수준과 차이가 있는 것으로 나타났다. 순환질환과 피부질환에서도 모수 추정값에 차이가 있을 뿐 대부분 유의한 것으로 분석되었다.

분석에 사용된 질병에 대해 거의 대부분의 환자에서 세 범주 중 질병에 걸리지 않음의 확률이 가장 큰 것으로 추정되었다. 이는 모형적합의 문제이가 보다는 질병에 걸리지 않음의 데이터가 상대적으로 많은 불균형자료(imbalanced data) 문제로 인한 것으로 보인다. 이는 개별환자에 대해 단순히 확률값을 기

Table 4.1. GEE estimates for respiratory diseases

모수	질병 미 발생				해당 질병 발생			
	추정치	표준오차	z-값	p-값	추정치	표준오차	z-값	p-값
절편	0.198	0.035	5.62	<0.0001	−0.375	0.106	−3.55	0.000
해당 주 평균기온	−0.064	0.011	−5.72	<0.0001	0.531	0.030	17.53	<0.0001
해당 주 일교차	0.108	0.004	24.43	<0.0001	−0.189	0.012	−15.59	<0.0001
해당 주 강수량	−0.075	0.003	−23.79	<0.0001	−0.041	0.010	−3.98	<0.0001
해당 주 평균풍속	−0.020	0.005	−4.11	<0.0001	0.041	0.013	3.16	0.002
해당 주 평균상대습도	0.120	0.004	27.19	<0.0001	0.047	0.012	3.89	0.000
해당 주 평균증기압	−0.129	0.011	−11.56	<0.0001	−1.134	0.032	−35.05	<.0001
해당 주 평균현지기압	−1.705	0.275	−6.20	<0.0001	0.412	0.713	0.58	0.563
해당 주 평균 해면기압	1.094	0.184	5.96	<0.0001	−0.080	0.476	−0.17	0.867
해당 주 합계일조시간	0.042	0.003	13.90	<0.0001	−0.055	0.009	−6.48	<0.0001
전 주 평균기온	−0.311	0.012	−27.08	<0.0001	−0.334	0.031	−10.85	<0.0001
전 주 일교차	0.017	0.004	3.79	0.0002	0.140	0.012	11.71	<0.0001
전 주 강수량	−0.017	0.003	−5.43	<0.0001	−0.067	0.010	−6.65	<0.0001
전 주 평균풍속	0.148	0.005	30.94	<0.0001	−0.044	0.013	−3.33	0.001
전 주 평균상대습도	0.046	0.005	9.94	<0.0001	−0.024	0.012	−1.95	0.051
전 주 평균증기압	0.345	0.011	31.86	<0.0001	0.094	0.031	3.00	0.003
전 주 평균현지기압	2.425	0.282	8.60	<0.0001	−5.039	0.730	−6.90	<0.0001
전 주 평균 해면기압	−1.720	0.188	−9.13	<0.0001	3.504	0.488	7.18	<0.0001
전 주 합계일조시간	0.005	0.003	1.70	0.0887	−0.094	0.008	−11.54	<0.0001
지역(서울)	0.334	0.014	24.84	<0.0001	−0.876	0.034	−25.79	<0.0001
지역(부산)	0.208	0.012	17.86	<0.0001	−0.506	0.026	−19.65	<0.0001
지역(인천)	0.204	0.014	14.34	<0.0001	−0.798	0.035	−22.55	<0.0001
지역(대구)	0.188	0.011	17.13	<0.0001	−0.534	0.022	−24.23	<0.0001
지역(광주)	0.192	0.013	14.89	<0.0001	−0.846	0.030	−28.66	<0.0001
지역(대전)	0.119	0.013	9.29	<0.0001	−0.618	0.028	−21.82	<0.0001
지역(울산)	0.238	0.012	19.13	<0.0001	−0.117	0.023	−5.02	<0.0001
지역(경기)	0.264	0.011	24.58	<0.0001	−0.523	0.023	−23.11	<0.0001
지역(강원)	0.522	0.039	13.36	<0.0001	−2.176	0.120	−18.21	<0.0001
지역(충북)	0.265	0.029	9.08	<0.0001	−1.763	0.087	−20.31	<0.0001
지역(충남)	0.011	0.011	0.98	0.3266	−0.413	0.021	−19.68	<0.0001
지역(전북)	0.024	0.019	1.30	0.1945	−1.148	0.052	−22.25	<0.0001
지역(전남)	−0.039	0.013	−2.96	0.0031	−0.634	0.032	−19.88	<0.0001
지역(경북)	0.238	0.020	11.93	<0.0001	−1.092	0.056	−19.38	<0.0001
지역(경남)	0.118	0.012	10.28	<0.0001	−0.512	0.024	−20.98	<0.0001
성별(남성)	0.174	0.002	84.83	<0.0001	0.161	0.004	41.20	<0.0001
연령(60–64)	0.127	0.005	28.34	<0.0001	0.145	0.009	16.62	<0.0001
연령(65–69)	−0.041	0.003	−13.00	<0.0001	0.165	0.006	25.85	<0.0001
연령(70–74)	−0.206	0.003	−65.01	<0.0001	0.139	0.006	21.62	<0.0001
연령(75–79)	−0.262	0.003	−76.67	<0.0001	0.083	0.007	11.97	<0.0001

준으로 질병 상태에 대한 판정을 한다면 대부분 질병에 걸리지 않은 것으로 판정되고 이에 따라 전체 고령자 중 해당 주상병에 걸린 환자의 수는 상당히 과소추정될 수 있다는 것을 의미한다.

이 논문의 분석 목적은 개별 환자에 대한 각 범주별 확률이 아니라 특정 연령, 지역, 성별에 따라 해당 주의 기상상태에 따라 몇 명의 환자가 발생했는지를 알아보기 위함으로 각 범주의 확률에 해당 분류 그

Table 4.2. The comparison of GEE and RNN

기준	주상병	GEE	RNN
C_1	호흡질환	99089.3	42552.4
	순환질환	97659.9	43961.8
	피부질환	91901.8	37717.4
C_2	호흡질환	952426.8	592421.6
	순환질환	967580.9	619791.4
	피부질환	923106.5	553856.5

GEE = generalized estimating equation; RNN = based recurrent neural network.

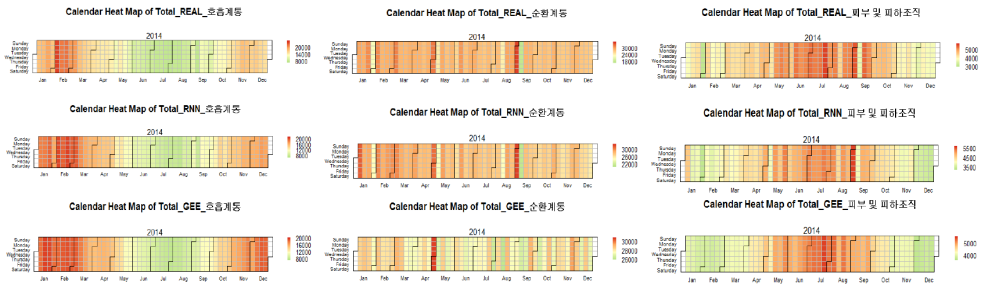

Figure 4.1. The comparison of GEE and RNN via calender heat map. GEE = generalized estimating equation; RNN = based recurrent neural network.

룹의 인원수를 곱한 기댓값을 구하는 것이다. 즉, n_{ij}를 i번째 주의 j번째 그룹(160)에 속하는 전체 환자수라 하고 $\hat{\theta}_{ijk}$를 i번째 주의 j번째 그룹의 k번째 범주($k = 1, 2, 3$)에 해당하는 확률의 추정값이라고 하면 동일 지역에 대해서는 동일 기상자료가 사용되기 때문에 j번재 그룹 내의 모든 환자들에 대한 세 범주의 확률은 동일한다. 그러므로 N_{ijk}를 i번째 주의 j번째 그룹에서 k번째 범주에 해당하는 환자의 수이고 $\hat{N}_{ijk}$는 분석모형을 통해 추정된 환자수라고 하면 $N_{ijk} = n_{ij}\hat{\theta}_{ijk}$가 된다. 만약 좀더 포괄적인 그룹 g에 대한 범주의 발생 건수를 구하고자 한다면

$$N_{igk} = \sum_{j \in g} n_{ij}\hat{\theta}_{ijk}$$

로 추정할 수 있다.

LSTM 기반 RNN 적합 정도와 GEE 적합 정도를 비교하기 위해 다음과 같은 기준을 고려해 보았다.

$$C_1 : \sum_{i=1}^{51}\sum_{j=1}^{160}\sum_{k=1}^{3} \frac{\left(N_{ijk} - \hat{N}_{ijk}\right)^2}{\hat{N}_{ijk}},$$

$$C_2 : \sum_{i=1}^{51}\sum_{j=1}^{160}\sum_{k=1}^{3} \left|N_{ijk} - \hat{N}_{ijk}\right|.$$

Table 4.2에 제시된 결과를 보면 세 가지 질병에 대하여 GEE 분석결과보다 RNN 분석결과가 더 적합을 잘하는 것으로 나타났다.

매주 동안의 전국에서 발생하는 관심 주상병의 발생건수를 예측해보기 위해 $\hat{N}_{i\cdot 2} = \sum_{j=1}^{160} n_{i2}\hat{\theta}_{ij2}$를 구하고 실제 환자수와 비교해 보았다. Figure 4.1은 총 51주 동안 관심 질병의 발생 빈도수를 비교하기 위

Table 4.3. The comparison of GEE and RNN

기준	주상병	GEE	RNN
C_3	호흡질환	5603.3	3850.3
	순환질환	8498.6	2214.1
	피부질환	1037.9	698.9
C_4	호흡질환	311785	273571
	순환질환	495225	263701
	피부질환	76232	78847

GEE = generalized estimating equation; RNN = based recurrent neural network.

한 캘런더열지도(calendar heat map)으로 열은 호흡기 질환, 순환기 질환, 피부 및 피하조직 질환을 의미한다. 첫 번째 열은 해당 질병으로 내원한 환자의 수이고 두 번째는 RNN으로 추정된 환자수, 세 번째는 GEE으로 추정된 환자수를 그린 것이다. 각 셀은 하루를 의미하는데 주별 분석이기 때문에 각 그림에서 각 열의 7개 행은 동일한 색상으로 표시된다.

호흡기 질환의 경우 RNN은 전반적으로 실제보다 조금 높게 나타났으며 GEE의 경우 RNN보다도 대체로 높고 1분기와 4분기의 예측값이 실제보다 현저히 높은 빈도를 보이고 있다. 순환기 질환의 경우 RNN의 경우 1분기 일부 주에서 실제보다 약간 높은 예측값이 있는 반면 4분기에는 대체서 낮게 예측하고 있으며 GEE의 경우 전반적으로 실제보다 적은 발생빈도를 예측하고 있다. 피부 및 피하조직 질환의 경우에도 GEE의 예측값은 실제보다 낮게 예측하 RNN의 경우 일부 주에서 차이가 있는 것으로 분석되었다.

Table 4.3은 각 질병에 대해 전국 단위로 주별 환자수와 예측된 환자수 간의 차이를 아래의 기준으로 표시한 것으로

$$C_3 : \sum_{i=1}^{51}\sum_{k=1}^{3} \frac{\left(N_{i\cdot k} - \hat{N}_{i\cdot k}\right)^2}{\hat{N}_{i\cdot k}},$$
$$C_4 : \sum_{i=1}^{51}\sum_{k=1}^{3} \left|N_{i\cdot k} - \hat{N}_{i\cdot k}\right|.$$

Table 4.2와 마찬가지로 RNN의 결과가 GEE에 비해 예측력이 높은 것으로 나타났다.

예측력에 있어 RNN이 GEE보다 정확한 것으로 나타났으나 모형 적합에 사용된 실제 모수, 즉 가중치(weight)와 편향(bias)의 개수가 RNN이 훨씬 많기 때문에 단순히 위의 기준으로 비교하는 것에는 무리가 있다. GEE 모형에서 설명변수들 간의 상호작용항을 추가하거나 적절한 변환 등을 통해 적합력을 높일 수 있으며 RNN에서는 GEE와 다르게 추정된 모수에 대한 해석이 어렵다는 문제가 있다. 그럼에도 불구하고 모형 해석보다는 단순히 예측을 목적으로 하는 사용자라면 RNN을 활용한 분석에 상당한 매력을 가질 것으로 생각된다.

5. 결론

본 논문에서는 기상 자료와 질병 발생과의 연관성을 이용하여 질병 발생 건수를 예측해 보았다. 분석에는 기계학습 방법 중 LSTM 기반 RNN 모형과 통계학습 방법 중 GEE 모형을 사용하였으며 환자를 주별로 지역, 성별, 연령으로 나누고 관심 질병을 호흡기 질환, 순환기 질환, 피부 및 피하조직 질환에 대한 질병 발생 건수 예측하였다. 본 연구에 사용된 자료에는 환자의 거주지 정보가 없어 방문 의료기관의

소재지 정보를 활용하여 일관된 원칙대로 할당하였다 이 결과를 통해 해당 질병의 발생 건수가 인구생태학적 요인과 기상 인자와의 연관성을 가지고 있다는 점을 확인했다. 비교분석 결과 GEE 모형보다는 LSTM 기반 RNN 모형이 예측력이 높은 것으로 나타났다. 하지만 RNN을 이용한 분석에서는 인구생태학적 요인과 기상에 따른 영향이나 해석이 어렵고 GEE모형에 상호작용항 등을 추가하여 예측력을 높일 수 있다는 여지가 여전히 남아 있다. 향후 연구과제로 관심 질병을 전문가와 상의하여 더 세부적으로 또는 더 포괄적으로 지정하고 질병 발생 건수 예측이 가능하다. 향후에 미세먼지, 초미세먼지 등에 대한 데이터를 이용한 질환 발생 건수 예측이 가능할 것이라 기대된다.

References

Ballester, F., Michelozzi, P., and Iniquez, C. (2003). Weather, climate, and public health, *Journal of Epidemiology & Community Health*, **57**, 759–760.

Basu, R. and Samet, J. M. (2002). Relation between elevated ambient temperature and mortality, *A Review of the Epidemiologic Evidence*, **24**, 190–202.

Diggle, P., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd ed), Oxford University Press, Oxford.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations*, Chpman & Hall/CRC, FL.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation*, **9**, 1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 2554–2558.

Touloumis, A., Agresti, A., and Kateri, M. (2013). Generalized estimating equations for multinomial responses using a local odds ratio parameterization, *Biometrics*, **69**, 633–640.

안혜연, 정주희, 김채희, 윤진아, 김현수, 오인보, 이지호, 원경미, 이영미, 김유근 (2016). 학술논문 분석을 통한 기상민감질환 선정 및 기상 인자와의 관련성 고찰, *한국환경과학회 2016년 정기학술대회 발표논문집*, **25**, 839–851.

주영수(2008). 기후변화와 건강, *대한내과학회지*, **75**, 489–491.

순환신경망을 이용한 질병발생건수 예측

이승현[a] · 여인권[b,1]

[a]SK주식회사 C&C, [b]숙명여자대학교 통계학과

(2020년 6월 23일 접수, 2020년 7월 7일 수정, 2020년 7월 7일 채택)

요 약

본 논문에서는 건강보험심사평가원에서 제공한 약 120만명의 2014년 고령환자의료자료(HIRA-APS-2014-0053)과 기상자료를 일반화추정방정식(generalized estimating equation; GEE) 모형과 long short term memory (LSTM) 기반 순환신경망(recurrent neural network; RNN) 모형으로 분석하여 기상 조건에 따른 주요 주상병의 발생 빈도를 예측한다. 이를 위해 환자가 의료 서비스를 받은 기관의 지역을 이용하여 환자의 거주지를 추정하고 해당 지역의 주별 기상 관측소 자료와 의료자료를 병합하였다. 질병 발생 상태를 세 개의 범주(질병에 걸리지 않음, 관심 주상병 발생, 다른 질병 방생)로 나누었으며 각 범주에 속할 확률을 GEE 모형과 RNN 모형으로 추정하였다. 각 범주별 발생 건수는 해당 범주의 속할 추정확률의 합으로 계산하였으며 비교분석결과 RNN을 이용한 예측이 GEE를 이용한 예측보다 정확도가 높은 것으로 나타났다.

주요용어: 고령환자의료자료, 기상자료, GEE, RNN

[1]교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여자대학교 통계학과.
E-mail: inkwon@sookmyung.ac.kr