Journal of Korea Multimedia Society Vol. 21, No. 2, February 2018(pp. 271-280)
https://doi.org/10.9717/kmms.2018.21.2.271

의약품 처방 데이터 기반의 지역별 예상 환자수 및 위험도 예측

장정현[†], 김영재[††], 최종혁[†††], 김창수[††††], 나스리디노프 아지즈[†††††]

A Prediction of Number of Patients and Risk of Disease in Each Region Based on Pharmaceutical Prescription Data

Jeong Hyeon Chang[†], Young Jae Kim[††], Jong Hyeok Choi[†††],
Chang Su Kim[††††], Nasridinov Aziz[†††††]

ABSTRACT

Recently, big data has been growing rapidly due to the development of IT technology. Especially in the medical field, big data is utilized to provide services such as patient-customized medical care, disease management and disease prediction. In Korea, 'National Health Alarm Service' is provided by National Health Insurance Corporation. However, the prediction model has a problem of short-term prediction within 3 days and unreliability of social data used in prediction model. In order to solve these problems, this paper proposes a disease prediction model using medicine prescription data generated from actual patients. This model predicts the total number of patients and the risk of disease in each region and uses the ARIMA model for long-term predictions.

Key words: Medicine Prescription Data, Disease Prediction, Region, ARIMA Model, Long-Term Predictions

1. 서 론

최근 IT 기술과 다양한 분야의 융합은 각각의 분야로부터 다양한 빅데이터의 생성을 유발하였으며, 이러한 빅데이터들의 활용하여 새로운 가치를 창출하는 다양한 서비스들이 전 세계적으로 등장하고 있다. 이러한 흐름에 대응하기 위해 우리나라의 경우 공공기관과 정부가 보유한 여러 분야의 데이터를 개방과 공유하는 정책을 통해 누구나 학술적, 상업적으로 자유롭게 이용할 수 있도록 서비스를 제공하고 있으며 이에 대한 개발과 활용을 촉진하고 있다[1]. 특히나 최근 의료분야 데이터가 공개됨에 따라 의료 데이터의 활용 사례가 증가 하고 있으며, 정부에서는 의료분야의 빅데이터 시범 사업으로 국민건강 주의 예보 시범 서비스, 빅데이터 기반 의약품 안전성 조기경보 서비스, 보건의료 빅데이터 활용 시범 사업 등을 진행하고 있다[2]. 특히 국민건강 알람서비스의 경우 범국민적으로 현재 유행하는 질병과 이에 대한

※ Corresponding Author : Nasridinov Aziz, Address: (28644) S1-4 225, Chungbuk Natinal University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk, Korea, TEL : +82-43-261-3597, FAX : +82-43-273-2265, E-mail : aziz@chungbuk.ac.kr
Receipt date : Jan. 2018, Approval date : Jan. 24, 2018
[†] Dept. of Software, ChungBuk National University (E-mail : practice1356@gmail.com)
[††] Dept. of Software, ChungBuk National University (E-mail : juk1413@naver.com)
[†††] Dept. of Software, ChungBuk National University (E-mail : leopard@chungbuk.ac.kr)
[††††] Dept. of Computer Engineering, PaiChai University (E-mail : ddoja@pcu.ac.kr)
[†††††] Dept. of Software, ChungBuk National University
※ This work is supported by Seondo project of the Ministry of Education in Korea

위험 정도를 제공하고 있다. 하지만 해당 서비스의 경우 감기, 눈병, 식중독, 천식, 피부염에 대하여 오늘, 내일, 모레와 같이 3일 이내의 단기적인 예측 결과를 제공하고 있으며, 진료 동향에 대해서는 15일 이내의 예측 결과만 제공 하고 있다. 또한 해당 질병 예측 모델에는 소셜 정보(SNS, 뉴스, 블로그)가 포함되는데 문맥이 아닌 단순 단어 포함 여부에 따라 수치를 측정하기 때문에 신뢰도의 문제가 있으며 모델의 수식이 선형식에 기초하고 있기 때문에 질병의 계절적인 특성을 고려치 못하고 있다[3,4].

본 논문에서는 상기 언급한 문제점들을 해결하기 위하여, 국민건강 심사 평가원에서 관리하고 제공하는 의약품 처방 데이터 및 보건의료 빅데이터 서비스와 통계청에서 제공하는 인구 통계정보를 기반으로 연월별 시별 질병별 예상 환자수를 산출해내는 예측 모델을 제안하고자 한다. 제안하고자하는 예측 모델은 유행성이 높고 질환 발생 대상자가 일반적인 감기, 눈병, 중이염, 비염, 장염 등으로 기존 국민건강 알림서비스와 달리 실제 환자를 대상으로 얻어진 의약품 처방 데이터를 활용하여 예측 결과의 신뢰성을 확보하고자 한다. 또한 질병에 가장 주요하게 작용하는 의약품 정보를 얻기 위하여 수집한 의약품 처방 데이터와 보건 의료 빅데이터 서비스에서 제공하는 질병별 실제 환자수 정보를 대상으로 상관분석을 실시하여 가장 높은 상관관계를 가지는 의약품을 해당 질병의 주요 처방 의약품으로 선별하여 활용하였다. 이후 질병별 실제 환자수와 의약품 처방 총 건수를 바탕으로 선형 회귀 분석을 통해 질병별 예상 환자수를 도출하고 각 지역별 의약품 처방 건수와 인구수를 통해 산출한 가중치 값을 통해 지역별 예상 환자수를 산출한다. 이렇게 생성된 1차 예측 모델의 결과는 최종적으로 시계열 형태를 보이기 때문에 이후 단계에서는 시계열 분석 모델인 ARIMA 모델을 적용하여 장기 예측을 실시함으로써, 기존 국민건강 알람 서비스의 단기 예측의 문제점을 보완한다.

이를 위한 본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구와 관련된 기존의 연구들에 대해 알아본다. 제 3장에서는 본 논문을 통해 제안하고자하는 예측 모델에 대해 설명한다. 제 4장에서는 본 논문에서 제안된 예측 모델을 실험을 통해 평가한다. 마지막 제 5장에서는 본 논문의 다시금 정리하며 마친다.

2. 관련 연구

“복합 이벤트 처리 기술을 이용한 의료기관 빅데이터 응용 시스템 개발”[5]에서는 오픈소스로 제공되는 CEP 기반 기술을 바탕으로 대량의 데이터를 실시간 처리하고, 개인정보를 분석하여 개인을 대상으로 실시간 모니터링 및 조기 경보에 관한 방법을 제안한다. 하지만 본 논문의 경우 유행성이 높고 질환 대상자가 일반적인 질병을 대상으로 지역별 예상 환자수와 위험도를 예측하여 제공한다는 점에서 적용 대상이 범국민적인 차별성을 갖고 있다.

“의료정보 빅데이터 분석을 통한 개인 맞춤형 유의질병 및 병원정보 제공 앱 서비스 개발”[6] 에서는 건강보험심사평가원에서 제공하는 건강보험청구데이터를 기반으로 개인 맞춤형 서비스에 대한 연구를 진행하였으며 유사그룹 기반 데이터 분석을 통해 현재 서비스에 등록된 환자와 유사한 진료 및 처방 패턴이 비슷한 환자의 정보를 고려하여 추가적으로 발병할 가능성이 높은 질병에 대한 정보를 안내한다. 그러나 유사그룹 기반의 질병 예측 방법은 등록된 환자와 유사한 진료 패턴이 없거나 분석 데이터가 부족한 경우 예측이 불가능하거나 예측 결과의 신뢰성을 보장하지 못한다.

“소셜 빅데이터를 활용한 인플루엔자 일일 예측모형”[7]에서는 인플루엔자 질병에 대한 계절형 ARIMA 모형, 계절형 ARX모형, 계절형 ARIMA 오차 회귀모형 등의 여러 예측 모델의 결과를 비교하였다. 3가지 예측 모델에 대해서 환경요인만을 적용한 경우, 소셜요인만을 적용한 경우, 환경요인과 소셜요인을 모두 적용한 경우로 총 3가지의 상황에 대한 예측 결과를 도출하였다. 그러나 해당 연구의 분석 절차에서 사용된 인플루엔자 진료 자료는 서울시의 자료만을 사용하였기에 서울시 이외 지역에 해당 예측모형을 적용하지 못한다는 문제점을 가지고 있다.

“기상 기후 및 질병 빅데이터 기반의 질병 예측 및 건강 정보 어플리케이션 구현”[8]에서는 기상 기후와 질병간의 상관분석을 실시하고 이를 통해 지역별 질병 정보, 의약품, 의료기관 정보를 제공하는 서비스를 구축하며 연구를 진행하였다. 해당 연구의 경우 기상 기후의 데이터와 이와 관련된 질병만을 고려하여 예측한다는 점과 예측 결과와 실제 환자수와의 오차율 검증에 대한 언급이 없어 연구 결과에 대한 신뢰성을 보장하지 못한다는 문제점을 가지고 있다.

3. 제안 방법

3.1 활용 데이터 및 수집 방법 명세

본 논문에서는 질병별 연월별 지역별 총 예상 환자수를 예측할 수 있는 예측 모델을 생성하기 위해 국민건강심사평가원의 '의약품 처방 데이터' 및 '보건의료 빅데이터 서비스'의 데이터와 통계청에서 제공되는 '인구 통계정보' 데이터를 활용하였다.

우선 국민건강심사평가원의 '의약품 처방 데이터'의 경우 처방된 연월 정보와 함께 의약품이 처방된 시·군 또는 구의 행정구역 코드, 국제 의약품 분류 코드인 ATC 코드, 의약품 총 처방 건수 등의 정보를 분석에 사용하였으며, '보건의료 빅데이터 서비스'의 경우 질병별, 연월별로 전국 기준 총 환자수의 정보를 분석에 사용하였다. 마지막으로 통계청의 '인구 통계 정보'의 경우 '의약품 처방 데이터'에서 처방 정보가 제공되는 행정구역들의 환자수 예측에 필요한 총 인구 수 정보를 반영하기 위해 사용하였다.

또한 이들 데이터는 현재 REST API 또는 Excel 형태로 제공되고 있기 때문에 본 논문에서는 이들의 수집 및 데이터베이스화를 자동화할 수 있는 Python 기반의 자동 수집 도구를 제작하여 해당 데이터들의 수집 및 데이터베이스화를 수행하였으며 이후 데이터베이스에 저장된 데이터들을 분석 자료로써 사용하였다.

3.2 분석 절차 개요

본 논문에서 제안하는 예측 모델의 분석 절차는 Fig. 1과 같이 진행되며, 각 단계별 상세 내용은 다음과 같다.

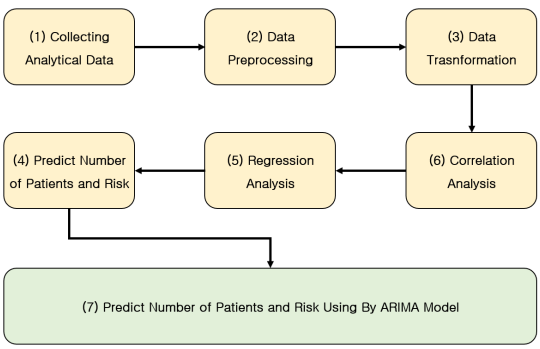


Fig. 1. Flowchart for predict number of patients and risk on each city.

(1) 데이터 수집 단계로써 자동 수집 도구를 통해 '의약품 처방 데이터', '보건의료 빅데이터 서비스', '인구 통계정보' 데이터를 수집한다.

(2) 데이터 전처리 단계로써 앞선 단계에서 수집된 데이터들의 정보 누락이나 불필요하게 중복된 데이터를 제거하는 등 분석을 위한 데이터 전처리를 수행한다.

(3) 데이터 변환 단계로써 기 구축된 행정구역 코드 테이블과 인구 통계 정보의 행정구역 코드, 의약품 처방 데이터의 군, 구 행정구역 코드를 활용하여 같이 처방 연월별, 시 또는 군별, 해당 의약품의 총 처방 건수와 해당 시 또는 군의 총 인구 수를 계산한 데이터로 변환한다.

(4) 상관분석 단계로써 시 또는 군별 의약품 총 처방 건수를 연월별로 모두 합산한 결과와 보건의료 빅데이터 서비스의 해당 의약품이 처방되는 질병의 연월별 실제 총 환자수를 대상으로 상관분석을 수행하여 실제 해당 질병에 주요하게 작용되는 의약품 정보를 선별한다.

(5) 회귀분석 단계로써 앞선 상관분석 단계에서 사용된 연월별 의약품 총 처방 건수와 연월별 실제 총 환자수 데이터를 바탕으로 단순 선형회귀분석을 실시, 연월별 의약품 총 처방 건수를 바탕으로 예상 총 환자수를 산출하는 선형회귀식을 얻는다. 이후 해당 식에 연월별 의약품 총 처방건수를 대입하여 예상 총 환자수를 산출한다.

(6) 지역별 예상 총 환자수 및 위험도 산출 단계로써 3번째 단계에서 변환된 데이터에서 총 처방 건수를 해당 지역의 총 인구로 나눈 가중치 값을 계산하고, 연월별 예상 총 환자수의 1%에 해당하는 환자수를 해당 가중치와 곱하고 이를 소수 첫째자리에서 반올림하여 지역별 예상 총 환자수를 산출한다. 이후 연월별 지역별 예상 총 환자수를 사분위수로 구분하여 경미, 보통, 주의, 위험 등으로 해당 질병에 대하여 지역별 위험도를 구분한다.

(7) 이전 단계에서 회귀모형을 통해 예측된 시계열 형태의 결과를 장기 예측 가능하도록 연월별, 지역별, 예상 총 환자수를 시계열 예측 모형인 ARIMA 모델에 적용시켜 장기 예측을 실시한다.

3.3 제안 수식

본 절에서는 본 논문을 통해 제안되는 의약품 처

방 데이터 기반의 연월별 질병별 지역별 예상 환자수 및 위험도 예측 모델의 회귀분석 단계 및 지역별 예상 총 환자수 및 위험도 산출 단계에서 사용되는 다양한 수식들을 보인다.

제안되는 예측 모델은 질병별 연월별 의약품 총 처방 건수와 해당 질병의 실제 총 환자수를 대상으로 회귀분석을 통해 도출된 선형회귀식을 바탕으로 총 환자수를 예상해야할 필요가 있으며, 이를 위한 계산식은 식 (1)과 같다.

$$predictPatients = A \times MonthlyTotalMedicineQuantity + B \quad (1)$$

식 (1)을 통해 얻어진 연월별 예상 총 환자수는 이후 각 지역별 예상 환자수를 계산하기 위해 활용된다. 이때 보다 정확한 지역별 예상 환자수를 고려하기 위해서는 지역별 인구 대비 의약품 처방 건수를 고려한 가중치 값을 활용해야할 필요가 있다. 이러한 가중치 값은 지역별 의약품 처방 건수의 편차를 고려하여 얻을 수 있으며 이를 위한 계산식은 식 (2)와 같다.

$$WeightOfPerCity = (MonthlyTotalMedicineQuantityOnCity \div CityPopulation) \times 100 \quad (2)$$

식 (1)의 계산식은 모든 지역의 해당 질병에 대한 의약품 처방 건수를 고려하여 연월별 예상 총 환자수를 도출하였다. 이를 활용하여 지역별 예상 총 환자 수를 구하기 위해서는 연월별 예상 총 환자 수를 식 (2)를 통해 얻은 지역별 가중치 값에 따라 분배해야 한다. 가중치 값에 따른 분배를 위해서는 해당 연월의 지역별 가중치 값이 전체 가중치 값에 차지하는 비율을 알아야하며, 이에 따른 전체 가중치 값을 구하는 계산식은 식 (3)과 같다.

$$MonthlyTotalWeight = \sum WeightOfPerCity \quad (3)$$

각 지역별 예상 총 환자수를 구하기 위하여 지역별 인구 대비 의약품 처방 건수의 편차를 고려한 가중치 값을 식 (2)를 활용하여 구하였다. 연월별 예상 총 환자 수를 지역별 가중치 값에 따른 분배를 위해 식 (3)을 활용하여 해당 연월의 지역별 가중치 합을 구하였고 이에 따라 각 지역별 가중치 값의 비율을 구하는 것이 가능해졌다. 해당 비율에 따라 연월별 예상 총 환자 수를 분배하기 위해서는 연월별 예상 총 환자 수에서 1%를 차지하는 환자 수를 알아야하며, 이에 따른 계산식은 식 (4)와 같다.

$$OnePercentPatients = predictPatients \div 100 \quad (4)$$

상기 제안된 계산식을 통해 연월별 예상 총 환자수와 지역별 인구 대비 의약품 사용량을 고려한 가중치 값과 가중치 비율에 따른 분배를 위한 가중치 합, 예상 총 환자수의 1%에 해당하는 환자수를 구하였다. 이후 연월별 지역별 가중치 값의 비율에 따른 예상 총 환자수를 구하는 계산식은 식 (5)와 같다.

$$predictPatientsOfCity = ((WeightOfPerCity \div MonthlyTotalWeight) \times 100) \times OnePercentPatients \quad (5)$$

4. 실험 결과 및 고찰

4.1 실험 개요

본 논문에서 제안하는 예측 모델의 정확성을 평가하기 위해서 본 논문에서는 실제 환자수와 예측 모델을 통해 예측된 환자수를 비교하는 실험을 진행하였으며, 질병에 대한 예측은 감기, 장염, 눈병, 비염, 중이염과 같이 계절성, 그리고 유행성을 띄는 질병에 대하여 진행하였다. 이를 위한 약품 처방 데이터의 선별은 2012년도부터 2017년도까지 사용된 의약품들과 실제 환자수를 대상으로 수행된 상관분석을 통해 각 질병별 주요 처방 의약품을 선별하였으며, 이후의 분석은 이들 약품에 대한 처방 데이터만을 수집하여 진행하였다. 최종적으로 분석에 사용된 의약품 처방 데이터는 2012년 1월부터 2017년 6월까지 얻어진 감기 644,894건, 장염 562,138건, 눈병 298,583건, 비염 647,497건, 중이염 1,692,726건이며 이들은 환자수의 예측 및 예측 모델의 검증을 위한 비교 실험에서 사용되었다.

4.2 연월별 예상 총 환자수 예측 결과

본 절에서는 제안된 분석 절차와 수식을 통해 앞서 선택된 질병들에 대한 연월별 예상 총 환자수 예측을 수행하고 이들 결과와 실제 환자수 사이의 차를 보인다.

Table 1의 총 약품 건수(Total Medicine)와 실제 환자수(Real Patients)는 2016년도의 월별 감기 관련 의약품의 총 처방 건수와 실제 환자수를 나타낸다. 이들을 이용하여 환자수 예측을 수행하기 위해서는 우선 연월별 의약품 총 처방 건수와 감기 환자수를

이용한 선형회귀식을 통해 우선 회귀 계수를 탐색해야할 필요가 있으며 2016년 감기의 회귀 계수 A와 B는 각각 0.003146, 1739019라는 값이 얻어지게 된다. 이후 이들 회귀 계수를 식 (1)에 반영하여 예상 환자수를 예측하게 되면 Table 1의 예측 환자수(Predict Patients)와 같은 결과를 얻을 수 있으며, 예측 환자수와 실제 환자수 사이의 오차율(Error Rate)는 평균 14%로 나타남을 알 수 있다.

다음의 Fig. 2는 감기와 동일한 방법을 통해 감기 이외의 질병인 (a) 눈병, (b) 장염, (c) 비염, (d) 중이염에 대한 월별 예상 총 환자수를 예측한 후, 각 질병에 대한 예측 환자수와 실제 환자수를 함께 나타낸 그림이다. 감기의 경우 앞선 실험을 통해 얻어진 실제 환자수와 예측 환자수 사이의 평균 오차율은 14%였으나, 눈병의 경우는 23.8%, 장염의 경우 16%, 비염의 경우 17%, 중이염의 경우 8.1%의 평균 오차율을 보임을 실험을 통해 확인할 수 있다.

4.3 시별 예상 총 환자수 및 위험도 산출

본 절에서는 시별 예상 총 환자수를 산출하기 위해 4.1에서 계산된 연월별 예상 총 감기 환자수를 이용하며, 서울, 대전, 대구, 부산, 인천 그리고 청주를 대상으로 환자수 예측 과정을 설명한다.

Table 2는 2016년 1월 서울, 대전, 대구, 부산, 인천, 청주의 감기관련 의약품 처방건수(Medicine)와 각 시별 인구(Population)를 기록한 표이다.

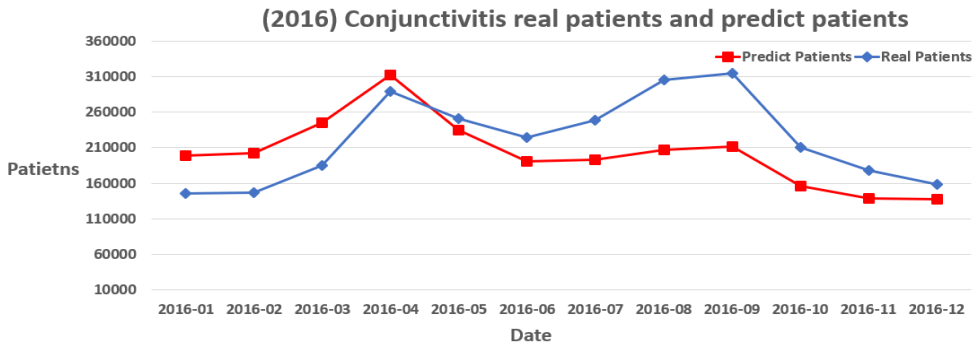
시별 인구 수 대비 의약품 처방 건수의 편차를 반영하기 위해서는 가중치 계산식인 식 (2)를 적용하여 시·군별 가중치 값을 구해야하며 그 결과는 Table 2의 도시별 가중치(Weight of per city)와 같다. 이후 시·군별 감기 질병에 대한 예상 환자수를 구하기 위해 식 (3)을 통해 각 시·군별 가중치 합과 1%의 환자수를 계산하는 식 (4)의 결과를 식 (5)에 대입하면 각 시별 예상 환자수를 구할 수 있으며, 이러한 각 시별 예상 환자수와 해당 환자수를 대상으로 사분위수를 통해 위험도를 나타낸 결과는 Table 3과 같다.

Table 1. Monthly cold disease number of predict and real patients and error rate

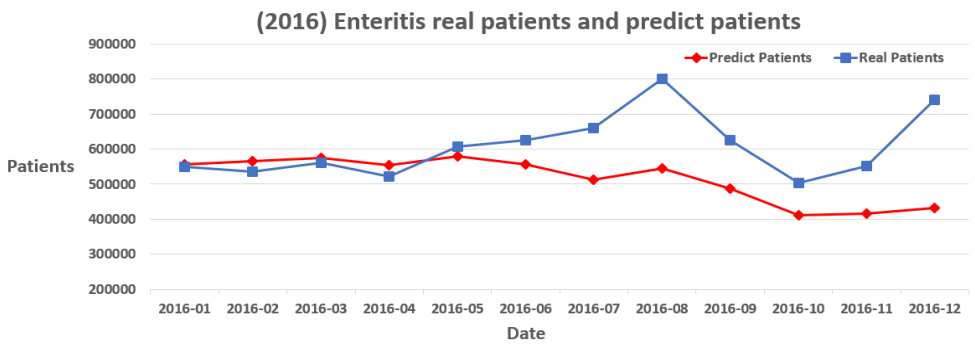
Date	Total Medicine	Real Patients	Predict Patients	Error Rate (%)
2016-01	878,531,055	3,895,605	4,502,878	15.6
2016-02	1,036,675,844	4,263,397	5,000,401	17.3
2016-03	903,167,203	4,187,911	4,580,383	9.4
2016-04	822,996,794	3,846,729	4,328,167	12.5
2016-05	695,927,239	3,352,488	3,928,406	17.2
2016-06	484,666,354	2,761,174	3,263,779	18.2
2016-07	346,265,399	2,566,717	2,828,370	10.2
2016-08	332,537,917	2,443,171	2,785,183	14
2016-09	429,176,434	3,089,361	3,089,208	0.01
2016-10	414,058,037	3,544,716	3,041,646	14.2
2016-11	488,148,185	3,883,994	3,274,733	15.7
2016-12	680,132,895	5,112,142	3,878,717	24.1

Table 2. The calculation result of weight value per city

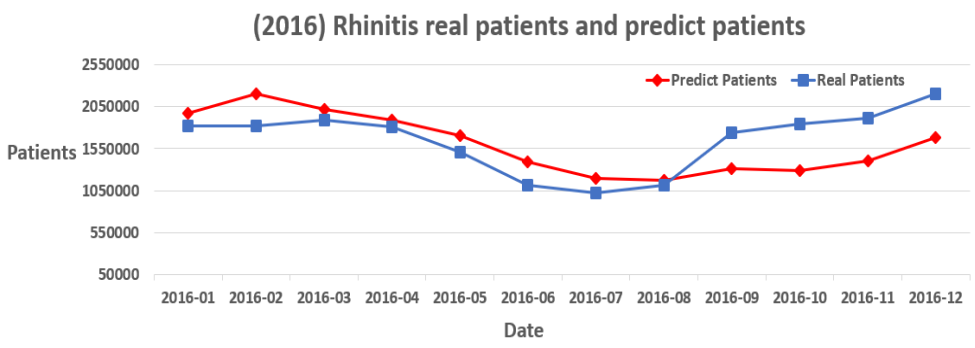
Date	CityName	Medicine	Population	Weight of per city
2016-01	Seoul	174,254,264	9,444,796	1845
2016-01	DaeJeon	29,648,772	1,499,520	1977
2016-01	DaeGu	45,751,802	2,402,744	1904
2016-01	BuSan	62,485,801	3,359,946	1860
2016-01	InCheon	47,373,860	2,783,565	1702
2016-01	CheongJu	14,306,624	808,703	1769



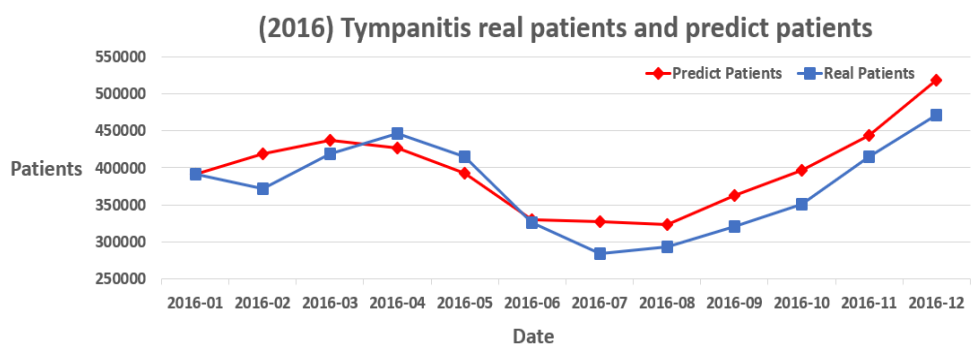
(a)



(b)



(c)



(d)

Fig. 2. Prediction Result (a) Conjunctivitis Disease (b) Enteritis Disease (c) Rhinitis Disease (d) Tympanitis Disease.

Table 3. The result of predict number of patients each city

Date	CityName	Predict Patients	Risk
2016-01	Seoul	29,042	Caution
2016-01	DaeJeon	31,120	Caution
2016-01	DaeGu	29,971	Caution
2016-01	BuSan	29,278	Caution
2016-01	InCheon	26,791	Normal
2016-01	CheongJu	27,846	Normal

4.4 ARIMA 모델 생성을 위한 시계열 데이터 검정

본 논문에서 제안하는 분석 절차와 수식을 통해서 선형회귀식을 도출하고, 이에 따른 질병별, 월별, 시별 예상 총 환자수와 위험도를 예측하였다. 그러나 제안된 회귀예측 모델 당월에 대한 예측만 가능한 단기 예측의 문제점을 갖고 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 예측 결과 데이터가 연월별로 산출되는 시계열형태의 데이터인 점을 착안하여 시계열 예측 모델인 ARIMA 모델을 통해 장기 예측을 실시한다.

본 절에서 예시로 사용되는 데이터는 서울, 대전, 대구, 부산의 2012년 1월부터 2017년 6월까지의 감기 관한 예측 데이터이며, ARIMA 모델의 경우 Python 언어로 제작하였으며, sklearn, numpy, pandas, statsmodels 등의 라이브러리를 활용하였다.

시계열 예측 모형인 ARIMA 모델을 활용하기 위해선 먼저 모형에 사용되는 시계열 데이터가 안정적인 시계열인지 여부를 확인해야 한다[9]. 이를 위해 시계열 데이터의 단위근 검정 방법 중 하나인 Dickey-Fuller 검정을 실시하였으며[10], 이에 대한 결과는 다음 Table 4와 같다.

Table 4를 보면 각 시에 대한 검정통계량(Test Statistic)의 값이 1%, 5%, 10%의 영역에서의 임계값(Critical Value) 보다 작은 것을 알 수 있으며 이는 해당 데이터가 안정적인 시계열임을 의미한다.

4.5 ARIMA 모델 생성을 위한 모수 설정

ARIMA 모델은 자기 회귀와 이동평균을 고려하며, 시계열의 비정상성을 설명하기 위해 관측치의 차분을 사용한다. 이에 ARIMA(p,d,q) 모델은 3가지의 모수인 p, d, q를 갖게되며 이들은 각각 AR(p), I(d),

Table 4. The result of Dickey-Fuller test on each city

CityName	Test Statistic	Critical Value(1%)	Critical Value(5%)	Critical Value(10%)
Seoul	-2.804928	-3.535217	-2.907154	-2.591103
DaeJeon	-0.529673	-3.557709	-2.916770	-2.596222
DaeGu	-0.736650	-3.552928	-2.914731	-2.595137
BuSan	-0.314138	-3.552928	-2.914731	-2.595137
InCheon	-4.278317	-3.536928	-2.907887	-2.591493
CheongJu	-1.340830	-3.557709	-2.916770	-2.596222

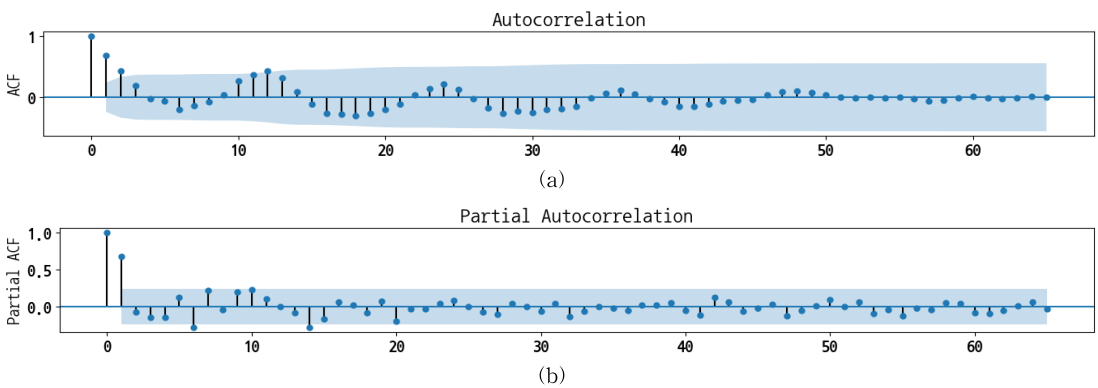
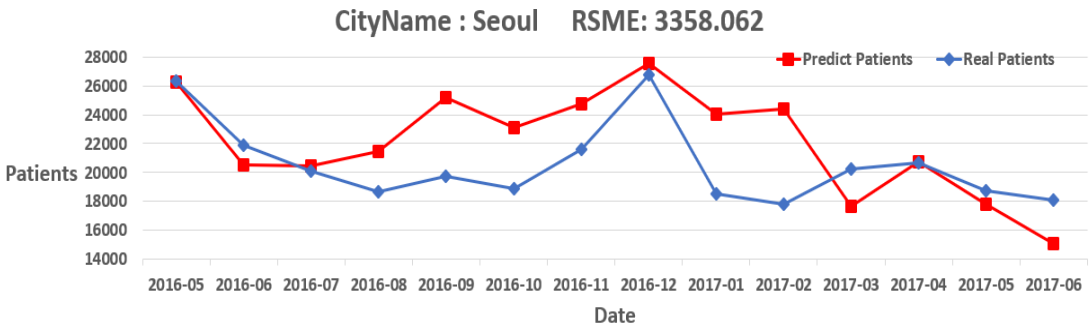
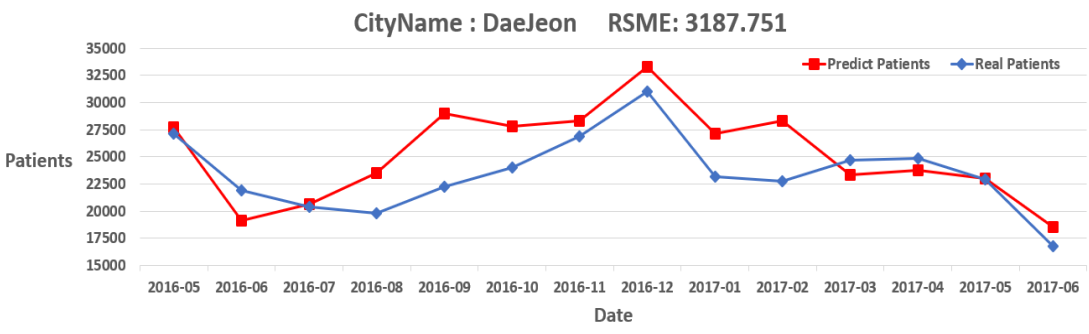


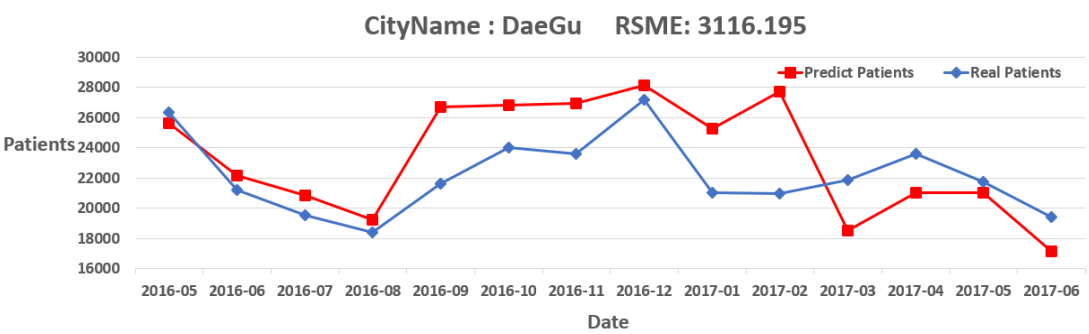
Fig. 3. (a) Autocorrelation and (b) Partial Autocorrelation graph about time series data.



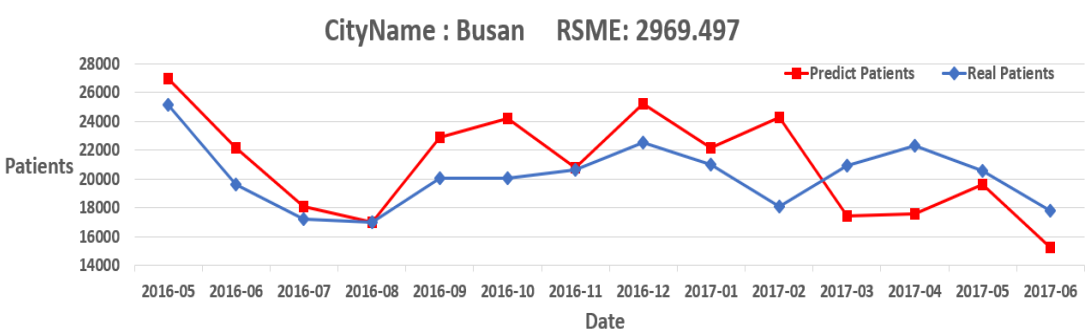
(a)



(b)



(c)



(d)

Fig. 4. A Prediction of number of cold disease patients on (a) Seoul, (b) DaeJeon, (c) DaeGu (d) BuSan using by ARIMA (2, 0, 0) Model.

MA(q)에 대한 값으로서 AR은 자기회귀 모형이며, I는 누적을 의미하고, MA의 경우 이동평균 모형이다.

이에 본 절에서는 ARIMA 모델의 각 모수의 값을 추정하기 위해서는 시계열 데이터를 대상으로 ACF, PACF 함수를 이용하였으며 ACF, PACF 함수의 결과는 Fig. 3과 같다.

Fig. 3을 보면 ACF 그래프가 천천히 0에 수렴하며, PACF의 그래프가 초기 시차 이후 급격하게 감소하는 모양을 띄므로 해당 시계열 데이터는 AR 모형에 적합하며 PACF 그래프의 Time Lag 1 이후 감소하므로 AR(2)모형이 적합하다[11]. 또한 앞서 Dickey-Fuller 검정을 통해 해당 시계열 데이터가 안정적인 시계열임을 확인하였으므로 ARIMA 모델의 모수는 $p = 2$, $d = 0$, $q = 0$로 추정되였다.

4.6 ARIMA 모델 생성 및 예측 검정

ACF, PACF 함수를 통해 ARIMA 모델 생성 시 필요한 모수를 추정하였고, 추정 모수값을 바탕으로 2012년 1월부터 2016년 4월까지의 서울, 대전, 대구, 부산의 감기 질병 예상 환자수 데이터를 ARIMA 모델에 테스트 데이터로 적용시켰다. 또한 모델의 예측력을 확인하기 위해 2016년 5월부터 2017년 6월까지의 데이터를 예측하였으며, 이에 대한 예측의 정확도를 측정하기 위해 RMSE(RootMeanSquareError)을 사용하였다[12]. 이러한 ARIMA 모델을 활용한 시계열 예측 결과와 실제 환자수는 다음의 Fig. 4와 같다.

ARIMA 모델을 활용한 시계열 예측의 결과인 Fig. 4를 보면 예측 값과 실제 값에 차이가 존재하지만 전체적인 형태에 있어서는 매우 유사한 형태로 예측을 수행하고 있음을 확인할 수 있다.

5. 결 론

본 논문에서는 '의약품 처방 데이터'와 '인구통계정보', '보건의료 빅데이터 서비스'의 데이터를 대상으로 제안된 분석 절차와 수식을 활용하여 시별 예상 총 환자수와 위험도를 예측하였다. 제안된 회귀예측 모델은 단기 예측의 문제점을 갖고 있으며 이를 해결하기 위해 시계열 예측 모델인 ARIMA 모델을 활용하여 장기 예측을 실시, 예측의 정밀도 또한 실험을 통해 검정하였다. 질병별 연월별로 각 시별 예상 총 환자수를 예측하므로, 전염성 질병에 대한 의약품보급 계획의 기준과 위험도를 통한 질병 확산 방지, 질병 발병 이후 사후 계획에 많은 도움을 줄 것으로 기대된다.

REFERENCE

[1] T. Song, "Social Big Data and Its Application: With Special Reference to MERS Information Diffusion and Risk Prediction," *Health and Welfare Policy Forum of Korea,* Vol. 227, pp. 29-49, 2015.

[2] T. Song, "Big Data Trend and Utilization Plan for Korean Health and Welfare," *Science and Technology Policy,* Vol. 192, pp. 56-73, 2013.

[3] J. Chang, Y. Kim, J. Choi, C. Kim, A. and Nasridinov, "A Study on Medicine Prescription Data Based Disease Occurrence Predictions," *Proceedings of the Korean Database Conference,* pp. 118-121, 2017.

[4] J. Yoon, S. Kim, B. Lee and B. Hwang, "A Correlation Analysis between the Social Signals of Cold Symptoms Extracted from Twitter and the Influence Factors," *Journal of Korea Multimedia Society*, Vol. 15, No. 6, pp. 667-677, 2013.

[5] M. Kim, Y. Yu, and B. Min, "Development of Bigdata Application System Using Complex Event Processing Technology for Medical Institution," *Journal of Korean Institute of Information Technology*, Vol. 14, No. 2, pp. 99-106. 2016.

[6] S. Kim and H. Hwang, "Developing a Personalized Disease and Hospital Information Application Using Medical Big Data," *Entrue Jounal of Information Technology,* Vol. 15, No. 2, pp. 7-16, 2016.

[7] E. Hwang, *Daily-based Prediction Models for Influenza Disease Using Social Big Data,* Master's Thesis of Chungbuk University, 2015.

[8] G. Kim, U. Kim, S. No, D. Lim, and J. Jeong, "Implementation of Disease Prediction and

Health Info Application Based on Climate and Disease Big Data," *Proceedings of the KITT Summer Conference,* pp. 496-497, 2017.

[9] Complete Guide To Create A Time Series Forecast With Python, https://www.analyticsvidhya. com/blog/2016/02/time-series-forecasting-codes-python, (accessed Dec., 23, 2017).

[10] Dickey-Fuller Test, https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test, (accessed Dec., 27, 2017).

[11] Time Series Model, http://env1.kangwon.ac.kr/leakage/2009/management/research/knowledge/04/Jang/%EC%8B%9C%EA%B3%84%EC%97%B4%20%EB%AA%A8%ED%98%95_Time%20Series%20Model.hwp, (accessed Dec., 30, 2017).

[12] RMSE: Root Mean Square Error, http://www.statisticshowto.com/rmse/, (accessed Feb., 10, 2018).

장 정 현

2014년 가온고등학교 졸업
2014년~현재 충북대학교 소프트웨어학과 학사과정
관심분야 : 빅데이터, 웹, 서버, 네트워크

김 영 재

2012년 서대전 고등학교 졸업
2012년~현재 충북대학교 소프트웨어학과 학사과정
관심분야 : 빅데이터, 인공지능

최 종 혁

2015년 충북대학교 컴퓨터교육과(이학사)
2017년 충북대학교 대학원 컴퓨터과학과(공학석사)
2017년~현재 충북대학교 대학원 컴퓨터과학과 박사과정

관심분야 : 데이터베이스, 빅데이터, 데이터마이닝

김 창 수

1996년 배재대학교 전자계산학과(이학사)
1998년 배재대학교 대학원 전자계산학과(이학석사)
2002년 배재대학교 대학원 컴퓨터공학과(공학박사)

2005년~2010년 청운대학교 인터넷학과
2013년~현재 배재대학교 컴퓨터공학과 조교수
관심분야 : 멀티미디어문서정보처리, 차세대 인터넷, USN, 모바일 웹서비스

나스리디노프 아지즈

2009년 동국대학교 대학원 컴퓨터공학과(공학석사)
2012년 동국대학교 대학원 컴퓨터공학과(공학박사)
2012년~2014년 숙명여자대학교 멀티미디어학과 박사 후 연구원

2014년~2015년 동국대학교 컴퓨터공학과 조교수
2015년~현재 충북대학교 소프트웨어학과 조교수
관심분야 : 데이터베이스, 빅데이터, 데이터마이닝, 분산처리 시스템