코로나 바이러스 확진자 데이터 기반 시뮬레이션 모델 학습 방법 제안

장미*, 이복주**, 강봉구***, 서경민****
*한국기술교육대학교 에너지신소재응용화학공학부 **한국기술교육대학교 컴퓨터공학부
***한국생산기술연구원 ****한국기술교육대학교 융합학과
dntdmami@koreatech.ac.kr, bokju618@koreatech.ac.kr, bgkang@kitech.re.kr, kmseo@koreatech.ac.kr

Suggestion of Corona Virus Infection Data-based Simulation Model Update Method

Mi Jang*, Bok-Ju Lee**, Bong-Gu Kang***, Kyung-Min Seo****
*School of Energy Materials and Chemical Engineering, Korea University of Technology and Education
**School of Computer Science and Engineering, Korea University of Technology and Education
*** Korea Institute of Industrial Technology
****Dept. of Future Technology, Korea University of Technology and Education

요 약

코로나감염-19, 사스, 메르스 등 바이러스성 질병이 전세계적으로 확산되어 많은 인구가 감염되어 왔다. 바이러스성 질병의 확산 예측 및 종결을 위해 실제 감염자 데이터를 기반으로 한 시뮬레이션 연구는 반드시 필요하다. 본 연구는 지역 내 클러스터 감염 시뮬레이션을 위한 바이러스 감염 모델을 제안한다. 제안하는 모델은 여러 개의 셀로 구성되어 있으며, 각 셀은 군집을 표현하고 있다. 본 논문에서 제안한 모델은 실제 데이터를 기반으로 하여 정확도가 높으며, 이를 바탕으로 향후 지역의 특성을 반영한 전파 시뮬레이션 혹은 지역 간의 전파를 예상하는 시뮬레이션의 기초로 사용될 수 있다.

1. 서론

전세계적으로 코로나감염-19 의 유행이 지속적으로 확산되고 있다. 중국 이외의 지역에서는 코로나감염-19 로 인한 사망률이 낮아 국민 건강에 미치는 영향이 크지 않을 것이라고 판단했으나, 대구 · 경북 지역에서 31 번째 확진자 발생 이후 예상하지 못한 클러스터 감염이 발생하면서 상황이 급변하였다. 그 결과 코로나감염-19 유행의 종결 시점을 정확하게 예측하기 어려워졌으며, 특정 집단에서 감염자 수가 폭등하는 현재와 같은 상황에서는 추이를 예측하는 것이 불가능하다[1].

코로나감염-19 발생 이전에도 사스, 메르스 등과 같은 바이러스성 질환의 전파에 대해 많은 데이터 기반의 분석이 이루어졌다. 데이터 기반의 모델의 경우 실제로 제공된 데이터가 바탕이 되므로 정확한 분석이 가능하다는 장점이 있다. 본 연구는 시뮬레이션 모델을 구성을 위한 이전 과정으로서 실제 발생한 일별 감염자 수 데이터를 바탕으로 바이러스 전파 모델을 설계하였다. 뿐만 아니라 클러스터 감염에 대한 표현이 가능하도록 각 행정구역을 셀로 나타내고, 각 셀의 내부에 군집 표현하였다.

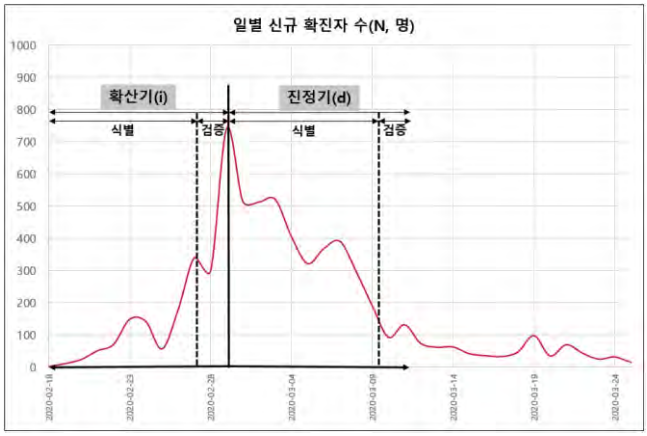
본 연구에서 제시한 모델을 통해 지역 내 클러스터 감염의 추이를 예측할 수 있게 된다. 현재 모델에서는 각 셀을 일반화하여 표현하였으나, 추후 진행될 연구에서는 각 셀에 인구 밀도, 단체 시설의 개수와 같은 지역적 특성을 반영하여 더 현실적인 전파 모델의 구현이 가능해질 것이다. 또는 지역 간의 연결을 통해 넓은 범위에 대한 분석이 가능할 것으로 보인다.

2. 모델 설계

2.1 시나리오 분석

모델 설계에 앞서 바이러스 전파 시나리오 분석을 위해 선정한 지역은 가장 많은 감염자가 발생한 대구광역시이다. 그림 1은 실제 대구에서 발생한 일별 신규 확진자의 수를 나타낸 표이다. 질병관리본부에서 발표한 바에 따르면 대구광역시에서 2 월 18 일 처음 1 명의 확진자가 발생하였다. 확진자 발생 첫날부터 40 일이 지난 3 월 28 일까지의 일별 신규 확진자 추이를 살펴보면 12 일째에 확진 판정을 받은 사람이 741

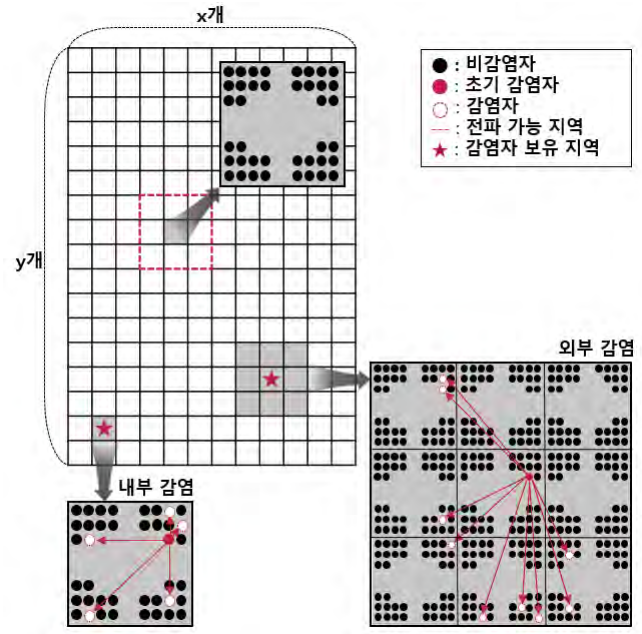
명으로 가장 높았다. 12 일째를 기준으로 이전에는 신규 확진자의 수가 점점 증가하여 바이러스가 감염이 확산되는 추세였으며, 그 이후에는 점차 확진자의 수가 줄어들어 바이러스 감염이 진정되는 경향을 확인할 수 있다. 본 논문에서는 상반되는 경향을 보이는 두 시기를 각각 확산기와 진정기로 구분하였다.



(그림 1) 대구 광역시 일별 확진자 수 현황

2020 년 2 월 행정안전부에서 발표한 KOSIS 에 따르면 대구광역시는 총 204 개의 행정구역(읍, 면, 동)으로 이루어져 있으며, 약 243 만 2 천명이 거주 중이다. 모든 구성원을 하나의 모델로 간주할 경우, 모델의 개수가 굉장히 많아지므로 구현하는데 한계가 있다. 이러한 점을 반영하여 모든 구성원을 각각 모델로 나타내는 대신 행정구역을 모델 단위로 설정하였다. 연구 대상을 셀 단위로 분할하여 분석하는 방법을 셀룰러 오토마타(이하 CA)라고 한다. 모든 셀의 상태를 각각 설정할 수 있으며, 각 셀들의 관계를 나타내는 규칙에 의해 동작이 이루어지는 동적 계산 시스템이다. CA 를 사용할 경우 수치적으로 분석하기 어려운 대상을 단순화하여 직관적으로 계산할 수 있다 [2].

대구 지역을 CA 로 표현하기 위해 1 개의 행정구역을 1 개의 셀로 설정하였으며 각 셀에는 동일한 인구가 거주 중이라고 가정하였다. 즉, 지역 내 전체 인구의 수가 M 이고, 행정구역의 수가 D 일 때, 각 셀에 표현되는 인구의 수는 D/M 이다. 초기 단계에서 전체 셀 중에 임의의 하나의 셀에 한 명의 감염자를 발생시킨다. 그리고 그 감염자는 감염자가 위치한 셀 내부와 그 셀을 중심으로 인접한 8 개의 셀에 거주 중인 구성원을 확률적으로 감염시킬 수 있다. 그림 2 과 같이 감염자와 같은 셀 안에 포함된 인구를 감염시키는 경우를 내부 감염, 인접 셀에 포함된 인구를 감염시키는 경우를 외부 감염이라고 지칭하였다.



(그림 2) 셀룰러 오토마타를 이용한 각 행정구역 별 거주 인구 및 감염 모사

2.2 실 데이터 기반 모델 설계

2.2.1 확산기

확산기의 감염자는 내부 감염과 외부 감염이 가능하고 각각의 전파율은 α, β 이다. 그리고 확산기의 내부 감염을 나타내는 방정식을 $f_{i,in}$ 로, 외부 감염을 나타내는 방정식은 $f_{i,out}$ 으로 표기한다. 또한 $f_{i,in}$ 과 $f_{i,out}$ 을 통해 도출한 p 번째 날 각 셀의 신규 감염자 수를 $N_{p,x\cdot y}$, p 번째 날의 전체 신규 감염자 수를 $N_{p,tot}$ 라고 표기한다. p 번째 날 (x,y)셀의 신규 감염자 수는 (x,y)셀 내부의 (p-1)번째 날까지의 누적 감염자로 인한 내부 감염자 수와 (x,y)셀의 인근의 셀에 존재하는 감염자로 인한 외부 감염자 수의 합이다. 그리고 p 번째 날 신규 감염자 수의 합은 모든 셀에서 발생한 신규 감염자 수를 합한 값이다. 식으로 나타내면 다음과 같다.

$$\begin{aligned}\Delta N_{p,x\cdot y} = {} & f_{i,in}\left(\sum N_{p-1,x\cdot y}, \alpha\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x-1)\cdot(y-1)}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x-1)\cdot y}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x-1)\cdot(y+1)}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,x\cdot(y-1)}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,x\cdot(y+1)}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x+1)\cdot(y-1)}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x+1)\cdot y}, \beta\right) \\ & + f_{i,out}\left(\sum N_{p-1,(x+1)\cdot(y+1)}, \beta\right) \qquad \text{식(1)}\end{aligned}$$

$$N_{p,tot} = \sum_{x=1}\sum_{y=1}(\Delta N_{p,x\cdot y}) \quad \text{식(2)}$$

그리고 각 셀의 신규 감염자 수는 각 셀의 인구 수를 넘을 수 없으며, p 번째 날 전체 신규 감염자 수는 지역의 총 인구 수를 넘을 수 없으므로 다음과 같은 조건을 따르게 된다.

$$0 < \alpha, \beta \leq 1 \quad \text{식(3)}$$

$$\Delta N_{p,i\cdot j} \leq \frac{M}{D} \quad \text{식(4)}$$

$$N_{p,tot} \leq M \quad \text{식(5)}$$

2.2.2 진정기

진정기에도 확산기와 동일한 메커니즘으로 바이러스가 전파가 된다. 확산기의 전파 확률에 계수 k 와 t 를 각각 곱하여 진정기의 전파 확률을 나타낸다. 내부 감염과 외부 감염의 확률을 각각 k·α 와 t·β 로 표기하고, 감염자를 구하는 방정식은 $f_{d,in}$ 과 $f_{d,out}$ 로 표현한다.

$$\begin{aligned}\Delta N_{q,x\cdot y} = {} & f_{d,in}\left(\sum N_{q-1,x\cdot y}, \mathrm{k}\cdot\alpha\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x-1)\cdot(y-1)}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x-1)\cdot y}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x-1)\cdot(y+1)}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,x\cdot(y-1)}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,x\cdot(y+1)}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x+1)\cdot(y-1)}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x+1)\cdot y}, \mathrm{t}\cdot\beta\right) \\ & + f_{d,out}\left(\sum N_{q-1,(x+1)\cdot(y+1)}, \mathrm{t}\cdot\beta\right) \quad \text{식(6)}\end{aligned}$$

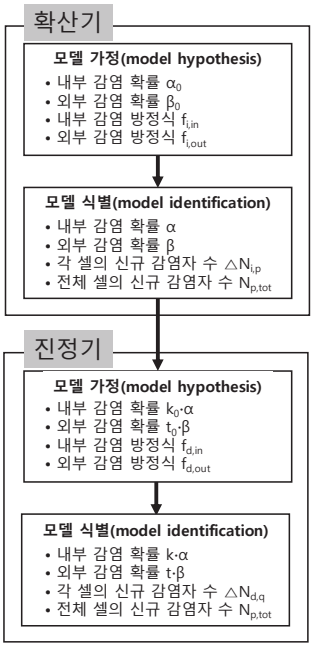
$$N_{q,tot} = \sum_{x}\sum_{y}(\Delta N_{q,x\cdot y}) \quad \text{식(7)}$$

마찬가지로 감염 확률과 감염자 수는 다음과 같은 조건을 가지게 된다.

$$0 < k\cdot\alpha, t\cdot\beta \leq 1 \quad \text{식(8)}$$

$$\Delta N_{p,i\cdot j} \leq \frac{M}{D} \quad \text{식(9)}$$

$$N_{p,tot} \leq M \quad \text{식(10)}$$



(그림 3) 바이러스 전파 시뮬레이션 식별 과정

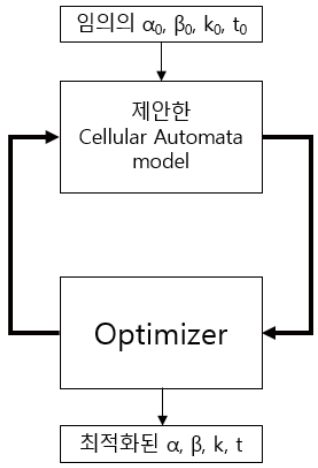
3. 모델 학습 프로세스

3.1 확산기 갱신 및 검증

그림 1 의 확산기 중 모델을 학습하여 최적의 α 와 β 를 구하기 위한 구간을 식별, α 와 β 의 타당성을 판단하기 위한 구간을 검증이라고 구분하였다. 2.2 에서 제시한 식(1)에 임의의 감염 확률 α_0, β_0 를 대입하여 그림 1 의 확산기 중 식별 구간과 유사한 결과를 내는 찾는다. 앞서 구한 α 와 β 가 타당한지 판단하기 위해 식(1)과 α, β 를 사용해 구한 일별 감염자 수와 그림 1 의 검증 구간의 감염자 수가 유사한지 비교한다. 그림 4 는 임의의 감염 확률을 적용하여 최적화된 감염 확률을 찾기위해 반복되는 과정을 도식화한 모습이다.

3.2 진정기 갱신 및 검증

3.1 에서 확산기의 최적화된 감염 확률을 구하는 방법과 동일하게 한 방법으로 진정기의 감염 확률을 도출한다. 진정기의 식별 구간을 통하여 k·α 와 t·β 를 찾는다. 찾은 k·α 와 t·β 가 최적의 전파 확률인지 진정기 검증 구간의 실제 감염자 수와 비교하여 판단한다.



(그림 4) 감염 확률 최적화 과정

4. 결론

실 데이터를 기반으로 한 시뮬레이션 모델이므로 정확도가 높다는 장점을 가진다. CA 를 사용하여 각각의 셀에 고유한 특성을 입력할 수 있다는 특징을 바탕으로 각 셀에 지역적 특성을 반영하여 좀 더 해상도가 높은 시뮬레이션 구현이 가능하다. 특히 인구밀도, 구성원들의 행정 구역 간의 이동, 단체 시설 등 클러스터 감염의 확률을 높이는 특성들을 도입할 경우, 폭발적이고 돌발적인 클러스터 감염에 대한 영향력을 미리 예측하고 대비할 수 있다. 더 나아가 현재는 하나의 지역을 분석 대상으로 설정하였으나, 여러 개의 지역을 결합한 모델을 구현하여 더 현실적인 감염 경로 및 감염자 수를 예측할 수 있다.

참고문헌

[1] 김남순, 코로나바이러스감염증-19 현황과 과제, 보건·복지 Issue & Focus, 373 호, 1-13, 2020

[2] Mohammad R. G., Kinetic of Hepatitis B Virus Infection : A Cellular Automaton Model Study, Journal of Paramedical Sciences, Vol.3, No.3, 1-8, 2012