

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0170544

(43) 공개일자 2023년12월19일

(51) 국제특허분류(Int. Cl.)

G16H 50/80 (2018.01) G06N 20/00 (2019.01)

G06N 3/04 (2023.01) G06N 3/08 (2023.01)

G16H 50/50 (2018.01) G16H 50/70 (2018.01)

(52) CPC특허분류

G16H 50/80 (2018.01)

G06N 20/00 (2021.08)

(21) 출원번호 10-2022-0135062

(22) 출원일자 2022년10월19일

심사청구일자 2022년10월19일

(30) 우선권주장

1020220070968 2022년06월10일 대한민국(KR)

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

황인준

서울특별시 성동구 매봉길 15(옥수동, 래미안 옥수 리버젠)

노윤아

서울특별시 도봉구 도봉로150길 42(방학동, 삼성래미안아파트2차)

(뒷면에 계속)

(74) 대리인

특허법인주연케이알피

전체 청구항 수 : 총 10 항

(54) 발명의 명칭 메타학습을 활용한 신종 감염병 확산 예측 장치 및 방법

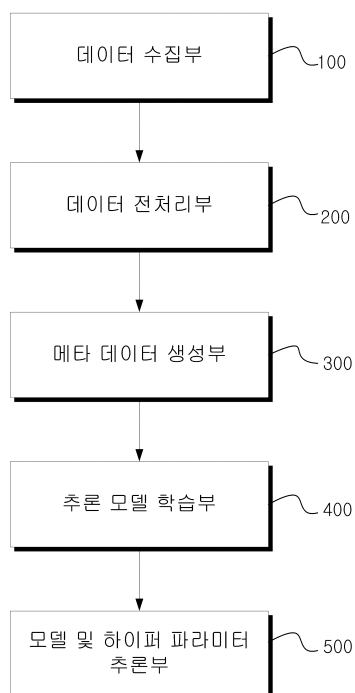
(57) 요약

본 발명은 신종 감염병의 발생량을 예측하기 위한 장치 및 방법에 관련된 것이며, 특히 다양한 시계열 데이터로부터 시간에 따른 패턴을 학습하여 새로 발생하여 과거 데이터가 부족한 감염병의 미래 발생량을 예측하기 위한 장치 및 방법에 관한 것이다. 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치는, 감염병에 관한 데이터

(뒷면에 계속)

대표도 - 도1

10



셋과 상기 감염병과 관련된 분야 이외의 상이한 분야와 관련된 복수의 시계열 데이터 셋을 수집하는 데이터 수집부; 상기 데이터 수집부에서 수집한 상기 감염병에 관한 데이터 셋과 상기 시계열 데이터 셋에 대하여 전처리를 수행하는 데이터 전처리부; 상기 시계열 데이터 셋으로부터 특징 정보를 추출하고, 랜덤 서치를 통한 하이퍼 파라미터 탐색을 하여, 탐색된 하이퍼 파라미터로 예측한 결과를 비교하여 모델을 선정하는 과정을 시계열 데이터 마다 반복하여 메타 데이터를 생성하는 메타 데이터 생성부; 상기 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 모델을 학습하는 추론 모델 학습부; 및 상기 추론 모델 학습부에서 학습된 추론 모델을 통해 감염병에 관하여 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론하는 모델 및 하이퍼 파라미터 추론부를 포함한다. 따라서, 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치 및 방법은 감염병 확산 초기에 예측 모델의 학습에 활용할 데이터가 충분하지 않은 상황에서도 신종 감염병의 확산을 정확하고 효율적으로 예측할 수 있다.

(52) CPC특허분류

G06N 3/04 (2023.01)
G06N 3/08 (2023.01)
G16H 50/50 (2018.01)
G16H 50/70 (2018.01)

정승원

서울특별시 노원구 우이천로 102(월계동, 초안아파트)

(72) 발명자

문재욱

서울특별시 광진구 아차산로78길 7(광장동, 유진스퀘어)

이 발명을 지원한 국가연구개발사업

과제고유번호	1465036120
과제번호	HG19C0682020022
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	감염병위기대응기술개발(방역연계 범부처 감염병 연구개발사업단)
연구과제명	감염병 예측 정확도 향상을 위한 딥러닝 모델 고도화 연구
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

감염병에 관한 데이터 셋과 상기 감염병과 관련된 분야 이외의 상이한 분야와 관련된 복수의 시계열 데이터 셋을 수집하는 데이터 수집부;

상기 데이터 수집부에서 수집한 상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 전처리를 수행하는 데이터 전처리부;

상기 복수의 시계열 데이터 셋으로부터 특징 정보를 추출하고, 랜덤 서치를 통한 하이퍼 파라미터 탐색을 하여, 탐색된 하이퍼 파라미터로 예측한 결과를 비교하여 모델을 선정하는 과정을 시계열 데이터마다 반복하여 메타 데이터를 생성하는 메타 데이터 생성부;

상기 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 모델을 학습하는 추론 모델 학습부; 및

상기 추론 모델 학습부에서 학습된 추론 모델을 통해 감염병에 관하여 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론하는 모델 및 하이퍼 파라미터 추론부를 포함하는 신종 감염병 확산 예측 장치.

청구항 2

제1 항에 있어서,

상기 데이터 전처리부에서 상기 데이터 수집부에서 수집한 상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 최댓값은 1, 최소값은 0으로 정규화 하는 신종 감염병 확산 예측 장치.

청구항 3

제1 항에 있어서,

상기 데이터 수집부에서 수집하는 상기 감염병에 관한 데이터 셋은 코로나 19 데이터 셋이며, 상기 복수의 시계열 데이터 셋은 M4데이터 셋으로 구성되는 신종 감염병 확산 예측 장치.

청구항 4

제1 항에 있어서,

상기 메타 데이터 생성부에서 수행되는 랜덤 서치는 통계 기반 예측 모델을 대상으로 하며,

상기 통계 기반 예측 모델은 ARIMA(Autoregressive Integrated Moving Average), SARIMA(Seasonal ARIMA), Prophet, Holt-Winters, Theta, STL(Seasonal-Trend decomposition using Local regression Forecasting)중 적어도 하나 이상으로 구성되는 신종 감염병 확산 예측 장치.

청구항 5

제4 항에 있어서,

상기 메타 데이터 생성부에서 생성된 메타 데이터를 가지고 학습하는 기계학습 모델로 랜덤 포레스트와 신경망을 사용하는 신종 감염병 확산 예측 장치.

청구항 6

제1 항에 있어서,

상기 모델 및 하이퍼 파라미터 추론부에서 추론된 모델과 하이퍼 파라미터를 활용해 일일 감염병 확진자 수를 예측하는 신종 감염병 확산 예측 장치.

청구항 7

신종 감염병 확산 예측 장치에 의해 수행되는 예측 방법에 있어서,

감염병에 관한 데이터 셋과 상기 감염병과 관련된 분야 이외의 상이한 분야와 관련된 복수의 시계열 데이터 셋을 수집하는 단계;

수집한 상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 전처리를 수행하는 단계;

상기 복수의 시계열 데이터 셋으로부터 특징 정보를 추출하고, 랜덤 서치를 통한 하이퍼 파라미터 탐색을 하여, 탐색된 하이퍼 파라미터로 예측한 결과를 비교하여 모델을 선정하는 과정을 시계열 데이터마다 반복하여 메타 데이터를 생성하는 단계;

상기 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 모델을 학습하는 단계; 및

학습된 추론 모델을 통해 감염병에 관하여 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론하는 단계를 포함하는 신종 감염병 확산 예측 방법.

청구항 8

제7 항에 있어서,

상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 최댓값은 1, 최소값은 0으로 정규화하는 단계를 더 포함하고,

수집하는 상기 감염병에 관한 데이터 셋은 코로나 19 데이터 셋이며, 상기 복수의 시계열 데이터 셋은 M4데이터 셋으로 구성되는 신종 감염병 확산 예측 방법.

청구항 9

제7 항에 있어서,

상기 메타 데이터를 생성하는 과정에서 랜덤 서치는 통계 기반 예측 모델을 대상으로 하는 단계를 더 포함하고,

상기 통계 기반 예측 모델은 ARIMA(Autoregressive Integrated Moving Average), SARIMA(Seasonal ARIMA), Prophet, Holt-Winters, Theta, STL(Seasonal-Trend decomposition using Local regression Forecasting)중 적어도 하나 이상으로 구성되는 신종 감염병 확산 예측 방법.

청구항 10

제9 항에 있어서,

생성된 메타 데이터를 가지고 학습하는 기계학습 모델로는 랜덤 포레스트와 신경망을 사용하는 단계를 더 포함하는 신종 감염병 확산 예측 방법.

발명의 설명

기술 분야

[0001]

본 발명은 신종 감염병의 발생량을 예측하기 위한 장치 및 방법에 관련된 것이며, 특히 다양한 시계열 데이터로부터 시간에 따른 패턴을 학습함으로써 새로 발생하여 과거 데이터가 부족한 신종 감염병의 확산을 예측하기 위한 장치 및 방법에 관한 것이다.

배경 기술

[0002]

최근 코로나19로 인한 인명 피해와 경제적 손해로 인해 세계 각국은 다양한 확산 방지 대책을 시행하고 있다. 감염병의 경우, 감염자 수를 조기에 예측할 수 있다면 상황에 따른 효과적인 대책을 강구할 수 있어 확산 방지 효과를 증진시킬 수 있다. 코로나19의 경우, 확산 초기부터 국내 확진자 수를 예측하기 위한 다양한 연구가 진행되었다. 세계적으로 확산이 2년 넘게 지속되어 온 현재에는 발생 데이터가 많이 확보되었으나, 확산 초기에는, 충분한 양의 데이터를 수집하는 것이 어려워 예측 모델 연구에 어려움이 많았다. 일반적으로 데이터

부족 문제는 정확한 예측 모델 구성을 어렵게 하며 예측 성능의 저하를 초래한다. 따라서, 신종 감염병 확산 초기에 데이터가 부족하더라도 예측 성능 저하 우려가 없는 예측 모델을 필요로 하게 되었다.

발명의 내용

해결하려는 과제

[0003] 본 발명의 목적은 확산 초기에 예측 모델의 학습에 사용될 데이터가 충분하지 않은 상황에서 신종 감염병의 확산을 정확하고 효율적으로 예측할 수 있는 예측 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0004] 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치는, 감염병에 관한 데이터 셋과 상기 감염병과 관련된 분야 이외의 상이한 분야와 관련된 복수의 시계열 데이터 셋을 수집하는 데이터 수집부; 상기 데이터 수집부에서 수집한 상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 전처리를 수행하는 데이터 전처리부; 상기 복수의 시계열 데이터 셋으로부터 특징 정보를 추출하고, 랜덤 서치를 통한 하이퍼 파라미터 탐색을 하여, 탐색된 하이퍼 파라미터로 예측한 결과를 비교하여 모델을 선정하는 과정을 시계열 데이터마다 반복하여 메타 데이터를 생성하는 메타 데이터 생성부; 상기 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 모델을 학습하는 추론 모델 학습부; 및 상기 추론 모델 학습부에서 학습된 추론 모델을 통해 감염병에 관하여 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론하는 모델 및 하이퍼 파라미터 추론부를 포함한다.

[0005] 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치에 의해 수행되는 신종 감염병 확산 예측 방법은, 감염병에 관한 데이터 셋과 상기 감염병과 관련된 분야 이외의 상이한 분야와 관련된 복수의 시계열 데이터 셋을 수집하는 단계; 수집한 상기 감염병에 관한 데이터 셋과 상기 복수의 시계열 데이터 셋에 대하여 전처리를 수행하는 단계; 상기 복수의 시계열 데이터 셋으로부터 특징 정보를 추출하고, 랜덤 서치를 통한 하이퍼 파라미터 탐색을 하여, 탐색된 하이퍼 파라미터로 예측한 결과를 비교하여 모델을 선정하는 과정을 시계열 데이터마다 반복하여 메타 데이터를 생성하는 단계; 상기 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 모델을 학습하는 단계; 및 학습된 추론 모델을 통해 감염병에 관하여 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론하는 단계를 포함한다.

발명의 효과

[0006] 본 출원의 실시예에 따른 신종 감염병 확산 예측 장치 및 방법은 시계열 데이터를 활용해 메타 데이터를 생성하여 최적의 예측 모델 및 하이퍼 파라미터를 추론하여 예측을 수행한다. 따라서, 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치 및 방법은 감염병 확산 초기에 예측 모델의 학습에 활용할 데이터가 충분하지 않은 상황에서도 정확하고 효율적인 예측이 가능하다.

도면의 간단한 설명

[0007] 도 1은 본 출원의 실시예에 따른 신종 감염병 확산 예측 장치를 보여주는 블록도이다.
 도 2는 본 출원의 일 실시예에 따른 도 1의 메타 데이터 생성부의 메타 데이터 생성 과정을 상세히 보여주는 도면이다.
 도 3은 본 출원의 일 실시예에 따른 도 1의 추론 모델 학습부의 구성을 더 상세히 보여주는 도면이다.
 도 4는 본 출원의 실시예에 따른 도 1의 모델 및 하이퍼 파라미터 추론부에서 최적의 모델 및 하이퍼 파라미터를 추론하는 과정을 상세히 보여주는 블록도이다.
 도 5는 본 출원의 실시예에 따른 신종 감염병 확산 예측 방법의 순서도를 나타낸 도면이다.
 도 6은 본 출원의 실시예에 따라 복수의 시계열 데이터로부터 메타 데이터가 생성되는 모습을 예시적으로 보여주는 도면이다.
 도 7a 내지 7b는 본 출원의 실시예에 따라 신종 감염병 확산 예측을 수행한 결과를 코로나 19를 예시로 하여 보여주는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0008] 이하에서는, 본 발명의 기술적 사상을 본 출원의 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있을 정도로 상세히 설명하기 위하여, 첨부된 도면을 참조하여 본 출원의 실시 예들이 자세히 설명될 것이다.
- [0009] 도 1은 본 출원의 실시예에 따른 신종 감염병 확산 예측 장치를 보여주는 블록도이다.
- [0010] 도 1을 참조하면, 신종 감염병 확산 예측 장치(10)는 데이터 수집부(100), 데이터 전처리부(200), 메타 데이터 생성부(300), 추론 모델 학습부(400) 및 모델 및 하이퍼 파라미터 추론부(500)를 포함한다.
- [0011] 데이터 수집부(100)는 신종 감염병 확산 예측에 필요한 각종 시계열 데이터를 수집한다. 예를 들어, 데이터 수집부(100)는 M4 데이터셋과 신종 감염병 데이터를 수집할 수 있다. M4 데이터셋은 금융, 산업 등 다양한 분야에서 수집한 총 10만개의 연도별, 분기별, 주별, 일별, 시간별 데이터를 제공한다. 신종 감염병 데이터는 예측모델이 학습을 마치고 예측할 데이터로써 수집된다. 예를 들어, 데이터 수집부(100)는 국내 코로나19 발생자 수 데이터를 Our World in Data에서 수집할 수 있다.
- [0012] 데이터 수집부(100)에서 수집하는 시계열 데이터로 반드시 M4 데이터셋으로 한정되지는 않는다. 예를 들어, 감염병 예측을 위한 학습에 활용할 데이터로 다른 감염병에 관한 데이터셋을 수집할 수도 있다.
- [0013] 데이터 전처리부(200)는 데이터 수집부(100)에서 수집한 데이터를 예측 모델에 사용하기 전에 보다 더 안정적인 학습 및 추론을 위해 전처리를 수행한다. 전처리를 수행하는 이유는 시계열 데이터별로 데이터의 범위가 크게 차이가 날 경우 모델의 학습에 방해가 되기 때문이다.
- [0014] 예를 들어, 데이터 전처리부(200)는 감염병 데이터와 M4 데이터 모두 최댓값은 1, 최소값은 0으로 정규화하는 최대-최소 정규화를 수행한다. 정규화로 구성된 전처리가 끝난 데이터는 예측 모델의 입력변수로 활용된다.
- [0015] 메타 데이터 생성부(300)는 데이터 전처리부(200)를 거친 다양한 시계열 데이터의 특징을 추출하고 해당 데이터에 적합한 모델과 하이퍼 파라미터 정보를 담고 있는 메타 데이터를 생성하는 것을 목적으로 한다. 메타 데이터 생성부(300)에서는 시계열 데이터로부터 특징을 추출하면서, 최적의 예측 모델 및 하이퍼 파라미터 탐색이 수행된다.
- [0016] 본 출원의 일 실시예에 있어서, 메타 데이터 생성부(300)에서 하이퍼 파라미터 탐색의 기법으로 랜덤 서치가 수행되고, 탐색 대상은 통계 기반 예측 모델일 수 있다. 메타 데이터 생성부(300)는 이하의 도 2에서 좀 더 자세히 설명될 것이다.
- [0017] 추론 모델 학습부(400)는 메타 데이터 생성부(300)에서 생성된 메타 데이터를 활용하여 시계열 데이터 예측에 적합한 모델과 하이퍼 파라미터를 추론하는 기계학습 모델을 학습한다.
- [0018] 본 출원의 일 실시예에 있어서, 기계학습 모델로는 랜덤 포레스트(Random Forest)와 신경망(Neural Network)을 사용할 수 있다. 추론 모델 학습부(400)는 이하의 도 3에서 좀 더 자세히 설명될 것이다.
- [0019] 모델 및 하이퍼 파라미터 추론부(500)는 추론 모델 학습부(400)에서 학습된 추론 모델을 통해, 새로운 시계열 데이터 예측에 적합한 모델과 하이퍼 파라미터를 추론한다. 모델 및 하이퍼 파라미터 추론부(500)는 추론 결과로 나온 최적의 모델과 최적의 하이퍼 파라미터에 따라 미래 확진자 수 예측을 진행한다. 모델 및 하이퍼 파라미터 추론부(500)는 이하의 도 4에서 좀 더 자세히 설명될 것이다.
- [0020] 상술한 바와 같이, 본 출원의 실시 예에 따른 신종 감염병 확산 예측 장치(10)는 시계열 데이터를 활용해 메타 데이터를 생성하여 최적의 예측 모델 및 하이퍼 파라미터를 추론하여 예측을 수행한다. 신종 감염병 확산 예측 장치(10)는 다양한 시계열 데이터로부터 각 시계열 데이터의 특징 정보와 최적의 예측 모델 및 하이퍼 파라미터 정보를 담고 있는 메타 데이터를 확보하고, 이를 참조하여 데이터 수가 부족한 경우에도 예측 정확도가 높은 예측 모델 및 하이퍼 파라미터를 추론한다. 따라서, 본 출원의 실시예에 따른 신종 감염병 확산 예측 장치 및 방법은 감염병 확산 초기에 예측 모델의 학습에 활용할 데이터가 충분하지 않은 상황에서도 신종 감염병의 확산을 정확하고 효율적으로 예측할 수 있다.
- [0021] 도 2는 본 출원의 일 실시예에 따른 도 1의 메타 데이터 생성부에서의 메타 데이터 생성 과정을 상세히 보여주는 도면이다.
- [0022] 도 2를 참조하면 메타 데이터를 생성하기 위해 일 실시예에 따라, M4 데이터셋의 각 시계열 데이터마다 다음과 같은 과정을 수행한다. 우선, 시계열 데이터에서 분산, 계절성, 엔트로피 등 40개의 특징 정보를 추출한다. 그

후, ARIMA(Autoregressive Integrated Moving Average), SARIMA(Seasonal ARIMA), Prophet, Holt-Winters, Theta, STLF(Seasonal-Trend decomposition using Local regression Forecasting) 6개의 통계 기반 예측 모델을 대상으로 하이퍼 파라미터 탐색 기법인 랜덤 서치를 수행하여 모델마다 최적의 하이퍼 파라미터를 탐색한다.

- [0023] 예측 모델을 구체적으로 살펴보면, ARIMA는 이전 관측값이 이후 관측값에 영향을 주는 AR(Autoregressive) 모델, 이전 항에서의 오차 혹은 변동 값을 이용해 현재 항의 상태를 추론하는 MA(Moving Average) 모델을 일반화한 모델이며, SARIMA는 ARIMA에 계절적 패턴을 추가로 반영한 모델이다. Prophet은 ARIMA에서 발전되어 추세와 계절적 패턴뿐만 아니라 예외적인 휴가철과 같은 상황도 반영할 수 있는 모델이다. Holt-winters는 추세나 계절적 특성에 급격하게 변화하는 시점의 개수를 추가로 반영하여 예측하는 모델이다. Theta는 원본 시계열로부터 변형된 2개의 시계열을 생성한 후 생성된 각 시계열의 예측 결과를 결합하여 최종 예측값을 도출하는 모델이다. STLF는 STL 분해를 사용하여 시계열 데이터에서 계절성을 제거하고, 분해된 성분으로 예측을 진행하고, 예측된 결과를 다시 계절화 하여 최종 예측을 진행하는 모델이다.
- [0024] 각 모델의 하이퍼 파라미터에 대한 설명은 다음과 같다. ARIMA 모델은 AR 모형의 시간 지연을 의미하는 p , MA 모형의 시간 지연을 의미하는 q , 차분(Difference) 횟수를 의미하는 d 를 하이퍼 파라미터로 가지고 있다. SARIMA는 ARIMA의 하이퍼 파라미터에 seasonal order와 trend를 추가로 사용한다. Prophet 모델은 seasonality prior scale, change point prior scale, change point range, yearly seasonality, weekly seasonality, daily seasonality, seasonality mode를 하이퍼 파라미터로 가지고 있으며, Holt-Winters 모델은 seasonality periods, trend, damped, seasonal을 하이퍼 파라미터로 설정하고 있다. Theta 모델은 계절적 패턴이 반복되는 길이를 나타내는 m 을 하이퍼 파라미터로 사용한다. STLF는 Theta의 하이퍼 파라미터에 추가로 예측 모델을 지정하는 method 하이퍼 파라미터를 가지고 있다. ARIMA 모델을 예로 들면 p, d, q 세개의 하이퍼 파라미터를 가지고, $p + d + q \leq 10$ 이 되도록 그 안에서 p, d, q 의 값을 조합하여 랜덤 서치를 수행한다.
- [0025] 랜덤 서치를 통한 하이퍼 파라미터 탐색을 마친 후, 최적의 하이퍼 파라미터로 예측한 각 모델의 결과를 비교함으로써 가장 우수한 예측 성능을 내는 최적의 모델을 선정한다. 예를 들어, 각 모델에 대해 탐색된 최적의 하이퍼 파라미터를 가지고 모델 탐색을 위한 시계열 데이터에서 예측했을 때, 평균 제곱근 편차(RMSE, Root Mean Square Error)와 PCC(Pearson Correlation Coefficient) 평가지표를 기준으로 예측 성능이 가장 우수한 예측 모델이 최적의 모델로 선정된다.
- [0026] 이러한 과정을 많은 수의 N 개 시계열 데이터에 반복하여 각 시계열 데이터마다 40개의 특징 정보와 최적의 모델 및 최적의 하이퍼 파라미터 정보를 담은 메타 데이터를 얻을 수 있다.
- [0027] 도 3은 본 출원의 일 실시예에 따른 도 1의 추론 모델 학습부의 구성을 더 상세히 보여주는 도면이다.
- [0028] 도 3을 참조해보면, 추론 모델 학습부(400)는 메타 데이터 생성부(300)에서 생성한 메타 데이터를 활용하여 시계열 데이터 예측에 적합한 모델과 하이퍼 파라미터를 추론하는 기계학습 모델을 학습한다. 일 실시예로, 기계학습 모델로는 랜덤 포레스트(Random Forest)와 신경망(Neural Network)을 사용할 수 있다. 랜덤 포레스트는 무작위로 복원 추출한 데이터 집합을 여러 개의 의사결정 나무(Decision Tree)에 입력하여 나온 결과들을 가중 투표하여 최종 분류값을 결정하는 모델이다. 랜덤 포레스트는 앙상블 알고리즘 기반으로, 메타 데이터의 수가 적더라도 좋은 분류 성능을 보이기 때문에 최적의 모델을 예측하는 데 활용한다. 한편, 신경망은 입력층, 은닉층, 출력층에서 가중치 값들을 스스로 학습하여 최종 출력 값을 예측하는 구조다. 신경망은 다른 기계학습 모델보다 범주형과 수치형의 하이퍼 파라미터 값을 각각 예측해야 하는, 다중 예측에 적합하기 때문에 최적의 하이퍼 파라미터를 예측하는 데 활용한다.
- [0029] 구체적으로, 시계열 데이터에 대해 각 시계열 당 최적의 예측 모델과 하이퍼 파라미터 정보를 사전에 랜덤 포레스트와 신경망 모델이 각각 학습한다. 그 후, 새로운 감염병 데이터에 대해 40가지의 특징 정보를 추출하고, 이를 학습된 랜덤 포레스트에 입력하면 최적의 모델을 추천받게 된다. 또한, 40가지의 특징 정보와 추천받은 최적의 모델 이름을 신경망에 입력하면 최적의 하이퍼 파라미터를 추천받게 된다. 이렇게 추천받은 최적 모델과 하이퍼 파라미터 정보를 가지고 신종 감염병 데이터의 확산 추세를 예측한다.
- [0030] 도 4는 본 출원의 일 실시예에 따른 도 1의 모델 및 하이퍼 파라미터 추론부에서 최적의 모델 및 하이퍼 파라미터를 추론하는 과정을 상세히 보여주는 블록도이다.
- [0031] 본 출원의 실시예에 따르면, 감염병 데이터에 대해서는 국내 코로나19 일일 확진자 수 데이터를 대상으로 하고

있으므로, 코로나19 데이터에서 40개의 특징 정보를 추출하고 학습된 랜덤 포레스트와 신경망에 입력데이터로 사용한다. 모델 및 하이퍼 파라미터 추론부(500)에서 그 결과로 나온 최적의 모델과 하이퍼 파라미터에 따라 신종 감염병 코로나19의 일일 확진자 수 예측을 진행한다.

- [0032] 도 5는 본 출원의 실시예에 따른 신종 감염병 확산 예측 방법의 순서도를 나타낸 도면이다.
- [0033] S110 단계에서, 데이터 수집부(100)는 감염병에 관한 데이터 셋과 감염병과 상이한 분야의 복수의 시계열 데이터 셋을 수집한다. 예를 들어, 감염병에 관한 데이터 셋은 코로나19에 관한 것일 수 있고, 복수의 시계열 데이터 셋은 M4 데이터셋 또는 다른 감염병에 관한 데이터셋을 수집할 수도 있다.
- [0034] S120 단계에서, 데이터 전처리부(200)는 데이터 수집부(100)에서 수집한 감염병에 관한 데이터 셋과 복수의 시계열 데이터셋에 대해 전처리를 수행한다. 예를 들어, 감염병에 관한 데이터 셋과 복수의 시계열 데이터셋 모두 최댓값은 1, 최소값은 0으로 정규화 하는 최대-최소 정규화를 수행할 수 있다.
- [0035] S130 단계에서, 메타 데이터 생성부(300)는 데이터 전처리부(200)를 거친 시계열 데이터를 가지고 최적의 모델 및 하이퍼 파라미터를 탐색하여 메타 데이터를 생성한다. 예를 들어, 메타 데이터 생성부(300)에서는 시계열 데이터로부터 특징을 추출하면서, 랜덤 서치를 통해 최적의 예측 모델 및 하이퍼 파라미터를 탐색할 수 있다.
- [0036] S140 단계에서, 추론 모델 학습부(400)는 메타 데이터 생성부(300)에서 생성된 메타 데이터를 활용하여 시계열 데이터 예측 모델과 하이퍼 파라미터를 추론하는 예측 모델을 학습한다. 예를 들어, 학습되는 기계학습 모델로는 랜덤 포레스트와 신경망을 사용할 수 있다.
- [0037] S150 단계에서, 모델 및 하이퍼 파라미터 추론부(500)는 추론 모델 학습부(400)에서 학습된 추론 모델을 통해 새로운 시계열 데이터 예측에 활용할 모델과 하이퍼 파라미터를 추론한다.
- [0038] S160 단계에서, 모델 및 하이퍼 파라미터 추론부(500)는 추론된 모델과 하이퍼 파라미터로 새로운 시계열 데이터 예측을 통해 신종 감염병 확산 예측 동작을 수행한다. 예를 들어, 모델 및 하이퍼 파라미터 추론부(500)에서 예측을 수행하는 대상은 코로나19 확진자 수 일 수 있다.
- [0039] 도 6은 본 출원의 실시예에 따라 복수의 시계열 데이터로부터 메타 데이터가 생성되는 모습을 예시적으로 보여주는 도면이다.
- [0040] 도 6을 참조하면, 도 2의 메타 데이터 생성에 있어서, 5개의 시계열 데이터가 있다고 가정하였을 때 생성된 메타 데이터의 예시적인 모습을 확인할 수 있다. 특정 시계열 데이터에 따라, 해당 시계열에서 가장 성능이 좋은 모델을 선정하고, 각각 선정된 해당 모델에서 가장 예측 성능이 좋을 것으로 보이는 하이퍼 파라미터를 탐색한다.
- [0041] 도 7a 내지 도 7b는 본 출원의 실시예에 따라 신종 감염병 확산 예측을 수행한 결과를 코로나 19를 예시로 하여 보여주는 도면이다.
- [0042] 도 7a는 과거 2개월의 확진자 수를 기반으로 예측한 결과이고, 도 7b는 과거 6개월의 확진자 수를 기반으로 예측한 결과를 그래프로 나타낸 것이다.
- [0043] 도 7a 내지 7b를 참조하면, 코로나19와 같이 데이터가 부족한 신종 감염병 예측 시 과거 데이터가 2개월과 같이 매우 적게 주어졌을 때와 과거 데이터가 6개월과 같이 어느 정도 주어졌을 때 발명한 기법과 기존의 예측 기법의 성능 차이를 시각화한 결과를 볼 수 있다. 구체적으로, 도 7a 내지 7b 각각 과거 2개월과 6개월의 코로나19 시계열 데이터가 주어졌을 때, 본 출원의 실시예를 통해 추론된 모델인 SARIMA에서 추론된 하이퍼 파라미터로 예측한 결과와 랜덤 서치로 SARIMA의 최적의 하이퍼 파라미터를 탐색하여 예측한 결과를 비교한 것이다. 도 7b를 참조하면, 6개월과 같이 과거 데이터가 충분히 있는 경우에는 랜덤 서치를 수행하여 탐색한 하이퍼 파라미터로 예측한 결과가 본 출원의 실시예를 통해 추론한 모델과 유사한 예측 성능을 보여주었다. 하지만, 도 7a를 참조하면, 2개월과 같이 주어진 과거 데이터의 수가 부족한 경우에는 2021년 11월 이후 발생한 오미크론 변이 바이러스의 대유행으로 추세가 급격히 변화하는 후반 부분에서 다소 변동 폭이 큰 예측 결과를 보이며 제안한 기법보다 정확도가 떨어지는 것을 확인할 수 있었다.
- [0044] 반면, 본 출원의 실시예에 따라 추론된 모델을 사용하여 예측했을 때는 과거 데이터의 수가 부족한 상황에서도 확진자 수가 불규칙하게 변화하는 초반의 추세와 오미크론 대 확산의 추세 및 정점을 더욱 정확히 예측하였다. 따라서 신종 감염병 발생 초기와 같이 활용할 수 있는 과거 데이터의 수가 부족할 때 제안한 기법을 통해 신종 감염병 예측의 정확도를 높일 수 있다.

[0045] 한편, 상술한 설명은 예시적인 것이며, 본 출원의 기술적 사상은 이에 한정되지 않음이 이해될 것이다. 예를 들어, 동일한 알고리즘의 다른 모델에서 효율성과 정확도에서 높다면 이를 활용해 예측을 수행될 수도 있다.

[0046] 이상에서는 본 발명에 따른 바람직한 실시 예들에 대하여 도시하고 또한 설명하였다. 그러나 본 발명은 상술한 실시예에 한정되지 아니하며, 특허 청구의 범위에서 첨부하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변형 실시가 가능할 것이다.

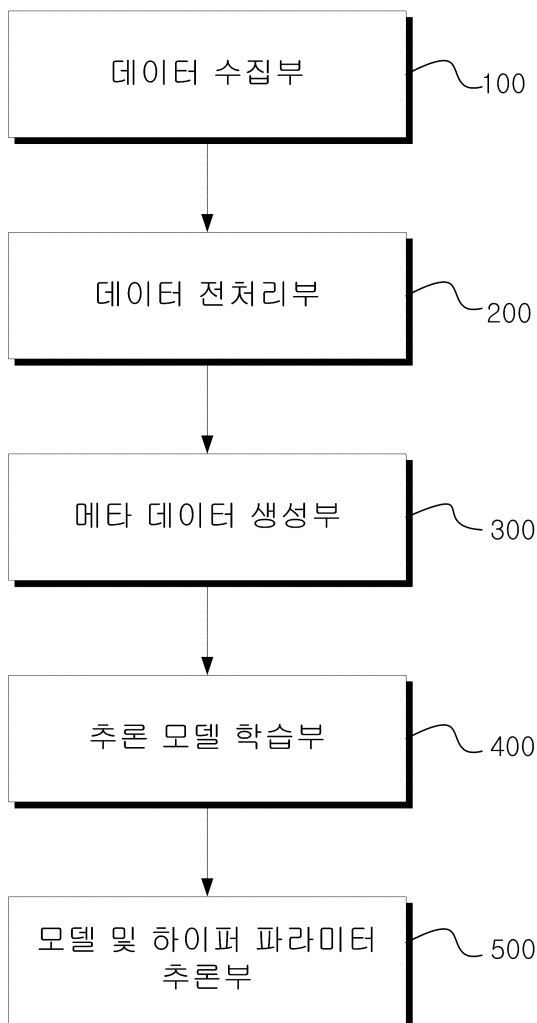
부호의 설명

- [0047]
- 100: 데이터 수집부
 - 200: 데이터 전처리부
 - 300: 메타 데이터 생성부
 - 400: 추론 모델 학습부
 - 500: 모델 및 하이퍼 파라미터 추론부

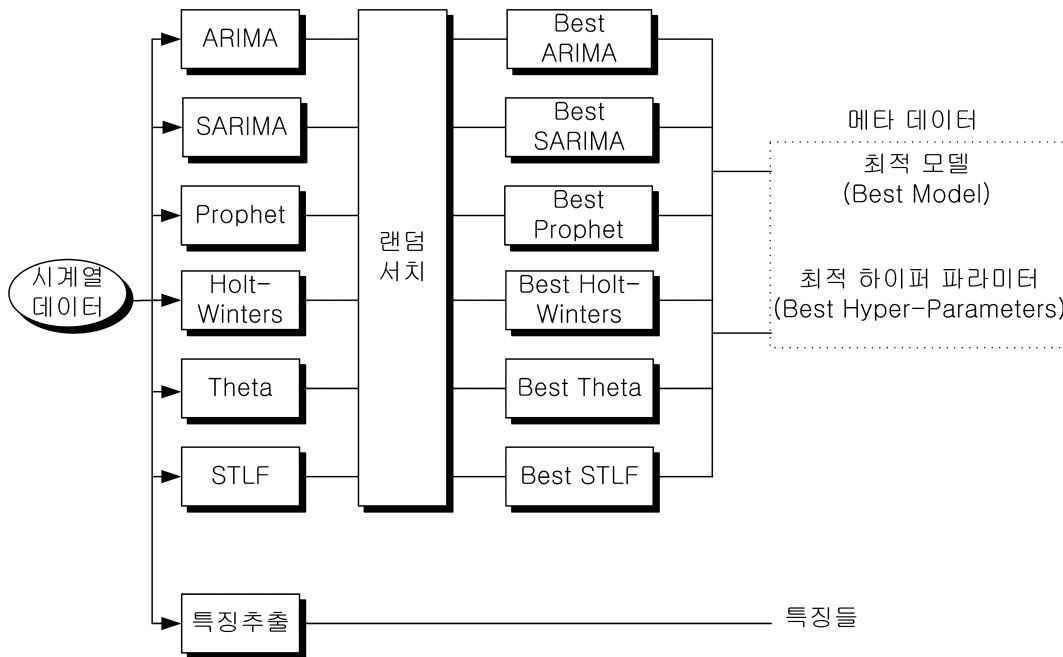
도면

도면1

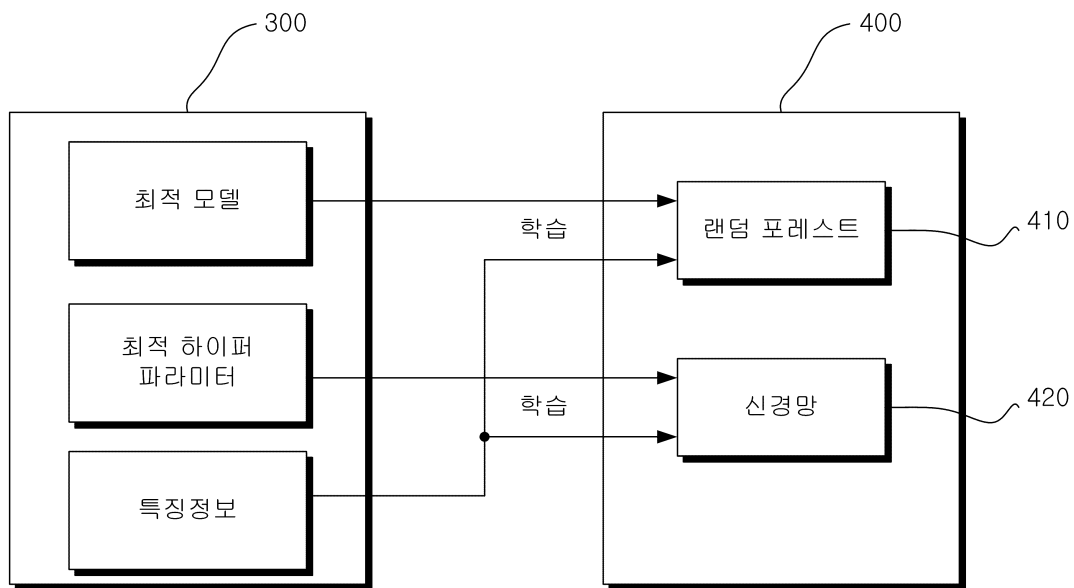
10



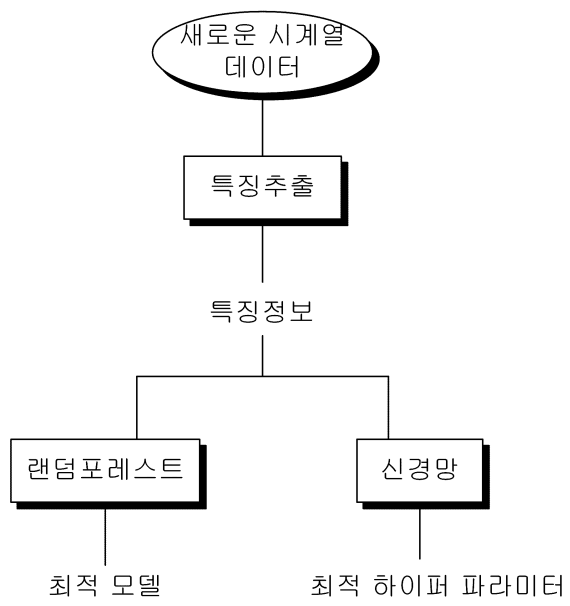
도면2



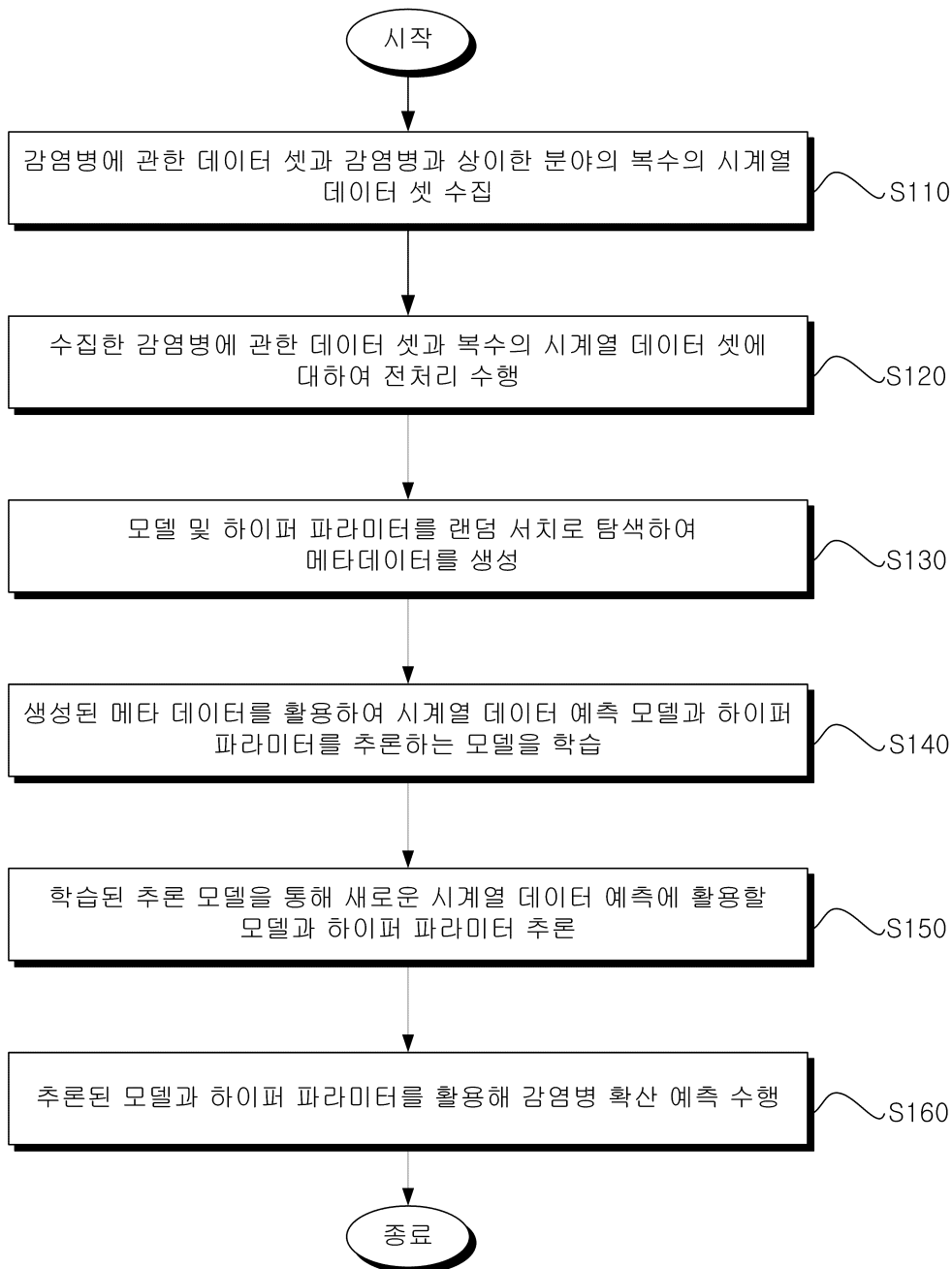
도면3



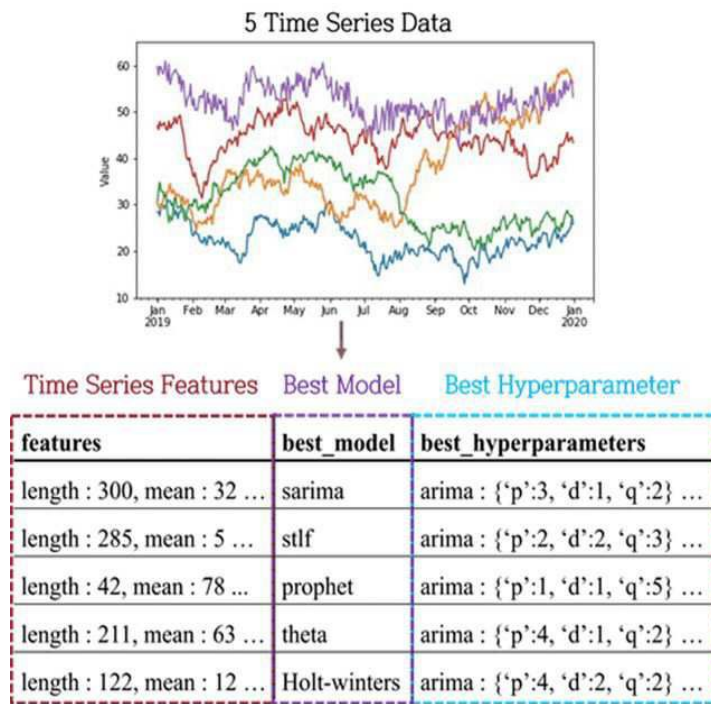
도면4



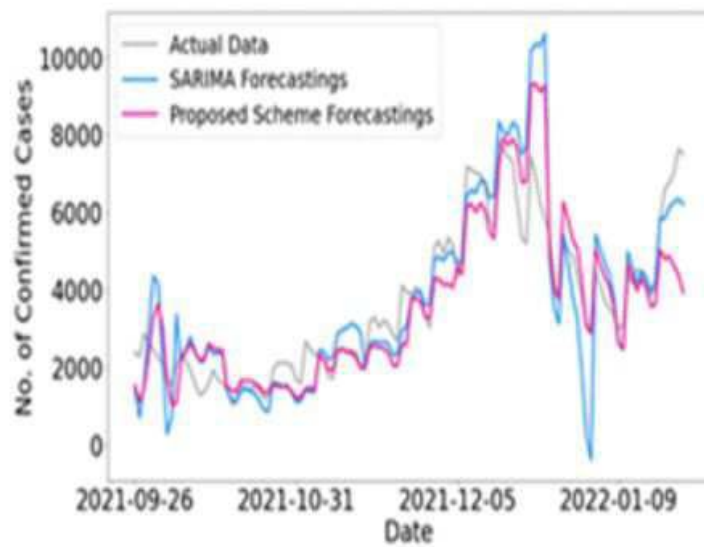
도면5



도면6



도면7a



도면7b

