



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0160974
(43) 공개일자 2022년12월06일

(51) 국제특허분류(Int. Cl.)
G06N 20/20 (2019.01) G01N 33/00 (2006.01)
G06N 3/04 (2006.01) G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 20/20 (2021.08)
G01N 33/0004 (2013.01)
(21) 출원번호 10-2021-0069288
(22) 출원일자 2021년05월28일
심사청구일자 2021년05월28일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
대한민국 (관리부서 : 환경부 국립환경과학원장)
인천 서구 환경로 42, 종합환경연구단지 (경서동)
(72) 발명자
권장우
부산광역시 남구 오륙도로 85, 109동 2503호(용호동, 오륙도 에스케이뷰 아파트)
이선우
서울특별시 강서구 우장산로 82, 101동 102호(화곡동, 우장산길성그랑프리텔)
(뒷면에 계속)
(74) 대리인
양성보

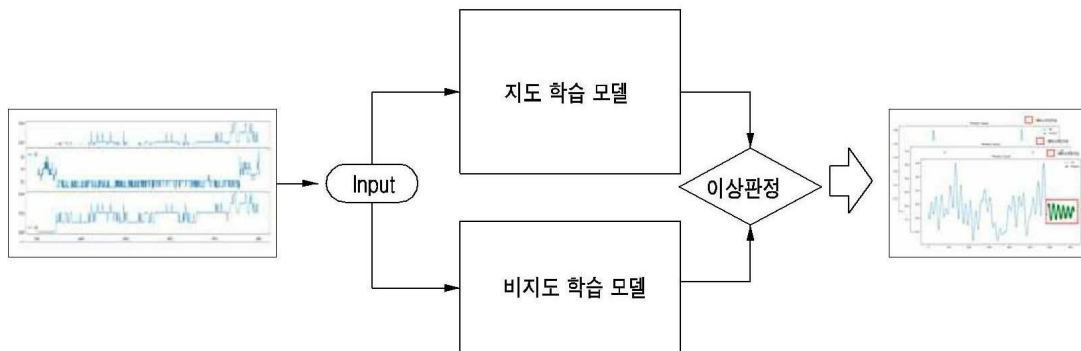
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 지도 학습 모델 및 비지도 학습 모델의 이상불 구조를 이용한 대기질 데이터의 이상 판정 방법 및 시스템

(57) 요약

지도 학습 모델 및 비지도 학습 모델의 이상불 구조를 이용한 대기질 데이터의 이상 판정 방법 및 시스템이 개시된다. 일 실시예에 따른 대기질 데이터의 이상 판정 방법은, 복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습 모델과 비지도 학습 모델이 이상불된 학습 모델에 입력받는 단계; 및 상기 지도 학습 모델과 비지도 학습 모델이 이상불된 학습 모델을 이용하여 상기 대기질 데이터의 이상 여부를 판정하는 단계를 포함할 수 있다.

대표도



(52) CPC특허분류

G06N 3/0454 (2013.01)

G06N 3/049 (2013.01)

G06N 3/08 (2013.01)

(72) 발명자

최정무

전라남도 광양시 금호로 73, 11동 401호(금호동, 사랑아파트)

이문형

인천광역시 부평구 경원대로 1269, 104동 104호(산곡동, 현대아파트)

양호준

인천광역시 미추홀구 인주대로366번길 23, 11동 102호(주안동, 남광로알아파트)

신혜정

서울특별시 영등포구 당산로42길 13, 104동 1305호(당산동5가, 당산동1차 효성아파트)

박지훈

인천광역시 서구 원당대로685번길 23, 204동 901호(마전동, 당하3차풍림아이원아파트)

이지선

서울특별시 서초구 방배중앙로 207-10, 104동 803호(방배동, 아크로리버)

이 발명을 지원한 국가연구개발사업

과제고유번호	1000637901
과제번호	63790-1
부처명	환경부
과제관리(전문)기관명	국립환경과학원
연구사업명	학술(기술)연구용역사업
연구과제명	인공지능을 활용한 대기오염측정망 자료확정 알고리즘 개발
기 여 율	1/2
과제수행기관명	국립환경과학원
연구기간	2020.09.23 ~ 2021.03.22

이 발명을 지원한 국가연구개발사업

과제고유번호	1711125779
과제번호	2017-0-01642-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	대학 IT연구센터 육성·지원
연구과제명	[Ezbaro][정부] 인공지능을 활용한 콘텐츠 창작 기술(5차년도)
기 여 율	1/2
과제수행기관명	인하대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명세서

청구범위

청구항 1

대기질 데이터의 이상 판정 방법에 있어서,

복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델에 입력받는 단계; 및

상기 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델을 이용하여 상기 대기질 데이터의 이상 여부를 판정하는 단계

를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 2

제1항에 있어서,

상기 지도 학습 모델은, 심층 합성곱을 포함하는 DeepLab V3+모델을 사용하고,

상기 입력받는 단계는,

복수 개의 성분에 대한 시간별 측정값인 1차원 데이터로 구성된 대기질 데이터를 상기 지도 학습 모델에 입력받는 단계

를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 3

제2항에 있어서,

상기 지도 학습 모델은,

레즈넷(ResNet34) 기반의 모델이 수정된 특징 추출기가 구성되고, 상기 대기질 데이터가 상기 레즈넷 기반의 모델에 통과됨에 따라 제1 특징맵이 출력되고, 상기 출력된 특징맵이 ASSPP(Atrous Separable Spatial Pyramid Pooling)에 통과됨에 따라 성분별 특징맵이 생성되고, 상기 생성된 성분별 특징맵과 상기 제1 특징맵을 통합하도록 학습된

것을 특징으로 하는 대기질 데이터의 이상 판정 방법.

청구항 4

제2항에 있어서,

상기 입력받는 단계는,

상기 대기질 데이터에 조각별 집계근사법을 적용하여 상기 대기질 데이터에 포함된 노이즈를 제거하고, 상기 조각별 집계근사법이 적용됨에 따라 사라진 노이즈 데이터에 대하여 선형보간법(Linear interpolation)을 사용하여 데이터를 보완하는 단계

를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 5

제4항에 있어서,

상기 입력받는 단계는,

상기 대기질 데이터가 측정된 시간에 따라 구간을 형성하고, 상기 형성된 구간에 포함된 대기질 데이터의 평균값을 상기 구간의 대표값으로 근사하여 노이즈를 제거하는 단계

를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 6

제1항에 있어서,

상기 판정하는 단계는,

상기 지도 학습 모델을 통해 상기 대기질 데이터에 대한 시간별 베이스라인 판정 결과를 출력하는 단계를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 7

제1항에 있어서,

상기 비지도 학습 모델은, 생성자와 판별자로 구성된 적대적 생성 신경망(GAN)을 사용하고,

상기 입력받는 단계는,

시계열 기반의 대기질 데이터를 상기 비지도 학습 모델에 입력받는 단계

를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 8

제7항에 있어서,

상기 비지도 학습 모델은,

특징을 추출하는 인코더와 특징을 되살리는 디코더로 구성된 구성되고, 다변수 시계열 데이터의 이상 패턴을 감지하는 BeatGan 모델을 사용하여 시계열 데이터의 비정상 패턴을 감지하도록 학습된

것을 특징으로 하는 대기질 데이터의 이상 판정 방법.

청구항 9

제8항에 있어서,

상기 비지도 학습 모델은,

정상 패턴을 가지고 있는 시계열 데이터로 학습을 진행하여 BeatGan 모델이 정상 데이터에 관하여 특징 정보를 이해할 수 있도록 구성된 것을 특징으로 하는 대기질 데이터의 이상 판정 방법.

청구항 10

제1항에 있어서,

상기 판정하는 단계는,

상기 비지도 학습 모델을 통해 상기 대기질 데이터에 포함된 비정상 데이터에 대해 이상치를 표출하는 단계를 포함하는 대기질 데이터의 이상 판정 방법.

청구항 11

이상 판정 시스템에 있어서,

복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델에 입력받는 입력부; 및

상기 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델을 이용하여 상기 대기질 데이터의 이상 여부를 판정하는 판정부

를 포함하는 이상 판정 시스템.

발명의 설명

기술 분야

[0001] 아래의 설명은 지도 학습 모델 및 비지도 학습 모델의 앙상블 구조를 이용하여 대기질 데이터에 대한 이상을 판정하는 기술에 관한 것이다.

배경 기술

[0003]최근 지구 온난화 및 이상기후에 대한 관심이 높아지면서, 전국의 대기오염측정망 측정소에 대한 숫자가 늘어나고 있는 추세이다. 대기오염측정망 측정소로부터 관측된 데이터 셋에는 기록 장치의 결함이나 자연재해 등의 이유로 결측치나 이상치가 포함될 수가 있다.

[0004]이러한 이상치나 결측값이 많아질 경우, 정보량이 정보의 질을 담보하지 못한다는 점에서, 정보의 질의 문제 또한 정보량의 문제만큼 데이터 분석에 있어서 중요한 역할을 한다. 그 이유는 첫째, 정보량이 많더라도 개개 정보의 내용을 신뢰할 수 없으면 그 정보를 이용한 통계적 추론의 신뢰성 또한 보장할 수 없기 때문이다. 둘째, 대부분 통계자료가 가지는 근원적 특성인 표본으로서의 조건을 충족하지 못하기 때문이다. 어떤 현상을 설명하기 위한 통계적 방법은 모집단을 대표할 수 있는 표본(sample)을 구성하고 그로부터 통계적 추론을 수행하는 것을 요체로 한다. 그러나 표본으로부터 조사한 자료가 불완전하여 자료의 질을 담보하지 못할 경우 자료의 대표성을 담보하기 어렵다.

[0005]시간과 공간의 자료를 포함하는 시공간 자료의 결측치는 시공간 분석의 주요한 장애물이다. 측정소 데이터의 고신뢰성을 위하여 측정망 데이터에서 이상치 탐지와 결측치 처리는 주요 관심사 중 하나이다. 결측치가 자료를 공간적, 시간적으로 단절시켜 공간적 현상의 변화를 분석하는데 어려움을 주기 때문이다. 이상 및 결측치 발생에 의한 불완전한 자료는 자료 분석 시 모델링에서의 편향된 모수 추정등의 요소에서 문제가 발생 문제가 발생하여 잘못된 결과를 초래할 수 있기 때문에, 적절한 결측치 처리는 분석을 할 때 중요한 요소라고 할 수 있다. 또한 연구 대상 공간에 속한 일부 지역의 속성값이 결측되어 있다면, 그 자료를 이용한 대상공간 전체에 대한 통계적 추론은 완전한 자료를 이용할 때에 비해서 부정확할 가능성이 높다. 또한 결측이 발생한 어떤 시점의 공간표본이 다른 시점의 공간표본들로부터 얻은 값들과 다른 특이성을 가진다면, 공간분석의 결과가 왜곡될 수 있기 때문에 높은 시공간 자료는 신뢰성을 요구한다.

발명의 내용

해결하려는 과제

[0007]인공지능을 이용하여 대기오염 측정망 자료의 이상 데이터를 검출하는 방법 및 시스템을 제공할 수 있다.

[0008]지도 학습 모델 및 비지도 학습 모델의 앙상블 구조를 이용하여 대기질 데이터에 대한 이상 판정을 수행하는 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0010]대기질 데이터의 이상 판정 방법은, 복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델에 입력받는 단계; 및 상기 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델을 이용하여 상기 대기질 데이터의 이상 여부를 판정하는 단계를 포함할 수 있다.

[0011]상기 지도 학습 모델은, 심층 합성곱을 포함하는 DeepLab V3+모델을 사용하고, 상기 입력받는 단계는, 복수 개의 성분에 대한 시간별 측정값인 1차원 데이터로 구성된 대기질 데이터를 상기 지도 학습 모델에 입력받는 단계를 포함할 수 있다.

[0012]상기 지도 학습 모델은, 레즈넷(ResNet34) 기반의 모델이 수정된 특징 추출기가 구성되고, 상기 대기질 데이터가 상기 레즈넷 기반의 모델에 통과됨에 따라 제1 특징맵이 출력되고, 상기 출력된 특징맵이 ASSPP(Atrous

Separable Spatial Pyramid Pooling)에 통과됨에 따라 성분별 특징맵이 생성되고, 상기 생성된 성분별 특징맵과 상기 제1 특징맵을 통합하도록 학습된 것일 수 있다.

- [0013] 상기 입력받는 단계는, 상기 대기질 데이터에 조각별 집계근사법을 적용하여 상기 대기질 데이터에 포함된 노이즈를 제거하고, 상기 조각별 집계근사법이 적용됨에 따라 사라진 노이즈 데이터에 대하여 선형보간법(Linear interpolation)을 사용하여 데이터를 보완하는 단계를 포함할 수 있다.
- [0014] 상기 입력받는 단계는, 상기 대기질 데이터가 측정된 시간에 따라 구간을 형성하고, 상기 형성된 구간에 포함된 대기질 데이터의 평균값을 상기 구간의 대표값으로 근사하여 노이즈를 제거하는 단계를 포함할 수 있다.
- [0015] 상기 판정하는 단계는, 상기 지도 학습 모델을 통해 상기 대기질 데이터에 대한 시간별 베이스라인 판정 결과를 출력하는 단계를 포함할 수 있다.
- [0016] 상기 비지도 학습 모델은, 생성자와 판별자로 구성된 적대적 생성 신경망(GAN)을 사용하고, 상기 입력받는 단계는, 시계열 기반의 대기질 데이터를 상기 비지도 학습 모델에 입력받는 단계를 포함할 수 있다.
- [0017] 상기 비지도 학습 모델은, 특징을 추출하는 인코더와 특징을 되살리는 디코더로 구성된 구성되고, 다변수 시계열 데이터의 이상 패턴을 감지하는 BeatGan 모델을 사용하여 시계열 데이터의 비정상 패턴을 감지하도록 학습된 것일 수 있다.
- [0018] 상기 비지도 학습 모델은, 정상 패턴을 가지고 있는 시계열 데이터로 학습을 진행하여 BeatGan 모델이 정상 데이터에 관하여 특징 정보를 이해할 수 있도록 구성된 것일 수 있다.
- [0019] 상기 판정하는 단계는, 상기 비지도 학습 모델을 통해 상기 대기질 데이터에 포함된 비정상 데이터에 대해 이상치를 표출하는 단계를 포함할 수 있다.
- [0020] 이상 판정 시스템은, 복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델에 입력받는 입력부; 및 상기 지도 학습 모델과 비지도 학습 모델이 앙상블된 학습 모델을 이용하여 상기 대기질 데이터의 이상 여부를 판정하는 판정부를 포함할 수 있다.

발명의 효과

- [0022] 지도 학습을 활용하여 조각별 집계근사법 적용과 대기질 데이터의 이상치 제거를 통해 성능을 향상시키고, 비지도 학습을 활용하여 대기질 데이터에 대한 이상 판단의 정확도를 향상시킬 수 있다.

도면의 간단한 설명

- [0024] 도 1은 일 실시예에 따른 이상 판정 시스템의 동작을 설명하기 위한 도면이다.
- 도 2는 일 실시예에 있어서, DeepLab V3+모델 구조를 설명하기 위한 도면이다.
- 도 3은 일 실시예에 있어서, 베이스라인 이상 판정 모델 구조를 설명하기 위한 도면이다.
- 도 4는 일 실시예에 있어서, 조각별 집계근사법의 적용 전과 적용 후를 비교한 도면이다.
- 도 5는 일 실시예에 있어서, GAN 모델의 학습 구조를 설명하기 위한 도면이다.
- 도 6은 일 실시예에 있어서, BeatGan 모델의 학습 구조를 설명하기 위한 도면이다.
- 도 7은 일 실시예에 있어서, 비정상 데이터에 대해 높은 이상값을 표출하는 것을 확인할 수 있다.
- 도 8은 일 실시예에 있어서, 학습 데이터를 생성하는 것을 설명하기 위한 도면이다.
- 도 9 및 도 10은 일 실시예에 있어서, 비정상 데이터 및 정상 데이터에 대한 실험 결과를 나타낸 그래프이다.
- 도 11은 일 실시예에 따른 이상 판정 시스템의 구성을 설명하기 위한 블록도이다.
- 도 12는 일 실시예에 따른 이상 판정 시스템에서 대기질 데이터의 이상 판정 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.
- [0027] 실시예에서는 대기오염 측정망 자료(예를 들면, 국립환경 과학원 대기오염 측정망 자료)의 신뢰도 향상을 위하여 이상 데이터를 검출하는 방법 및 시스템에 대하여 설명하기로 한다. 이를 위해 기존의 측정망 데이터를 분석 후 오폭기되어 있는 레이블링 데이터를 수정하여 인공지능 데이터의 학습에 맞도록 개선함으로써 이상 데이터의 검출을 위한 성능을 향상시킬 수 있다.
- [0028] 도 1은 일 실시예에 따른 이상 판정 시스템의 동작을 설명하기 위한 도면이다.
- [0029] 도 1은 지도 학습 모델과 비지도 학습 모델을 앙상블화 시킨 구조를 시각화한 프로세스로서, 대기질 데이터를 입력 데이터로 하여 지도 학습과 비지도 학습을 모두 사용하여 이상 판정이 진행될 수 있다.
- [0030] 불명확한 레이블링은 학습 모델의 정확도를 저하시키는 요인이 되는데, 이러한 현상이 발생하는 원인은 레이블링을 수행하는 담당자마다 판단기준이 상이하며 측정소마다 사용되는 측정 기기가 다르고 측정 기기별로 이상 데이터를 판단하는 기준이 다르기 때문이다. 이러한 문제를 해결하기 위하여 실시예에서는 각 이상 증상별 일정한 판단 기준을 설정한 뒤 레이블링을 다시 진행하여 학습용 데이터 셋을 생성하기로 한다. 학습용 데이터 셋을 이용하여 학습 모델이 선정되고, 이를 대기질 데이터에 적용하여 학습시킴에 따라 학습 모델이 구축될 수 있다.
- [0031] 이상 판정 시스템은 인공지능을 이용하여 대기질 데이터의 이상 여부를 판단할 수 있다. 인공지능에서 학습 데이터와 이를 정답으로 표기한 레이블링이 된(Labeled) 학습 데이터를 기반으로 지도 학습과 정답이 없고 학습 데이터만을 이용하여 판단하는 비지도 학습을 포함하는 두 가지 접근법을 적용한 앙상블 모델을 통해 대기질 데이터에 대한 이상 여부를 판단할 수 있다. 이와 같이, 기존의 지도 학습의 레이블링이 측정소별, 사용자별로 기준이 각기 다르기 때문에 모델이 학습하기에 어려움이 많아 레이블없이 할 수 있는 비지도 학습 모델과 지도 학습 모델을 모두 적용하여 성능을 극대화시킬 수 있다.
- [0032] 이상 판정 시스템은 베이스라인 이상의 경우, 기존 데이터의 레이블링으로 학습된 지도 모델의 결과와 비정상 데이터를 제외한 정상 데이터만으로 학습이 진행된 비지도 모델의 결과를 통합하여 입력 데이터의 베이스라인 이상 여부를 판단할 수 있다.
- [0033] 도 2는 일 실시예에 있어서, DeepLab V3+모델 구조를 설명하기 위한 도면이다.
- [0034] DeepLab V3+모델은 딥러닝(Deep Learning) 분야 중 전경 분할(Semantic Segmentation) 연구분야에서 사용되는 모델이다. DeepLab V3+ 모델에 경우 세그멘테이션(Segmentation)을 해결하기 위한 모델로서, 이미지가 입력값으로 들어오게 되면, 심층 합성곱(Deep Convolutional Neural Network; DCNN)을 통과함에 따라 복수 개의 크기의 특징맵(Feature Map)이 생성될 수 있다. DeepLab V3+모델은 심층 합성곱(DCNN) 결과를 통해 출력된 복수 개의 크기의 특징맵들과 기 생성된 연산초기 특징맵을 통합한 뒤, 각 특징맵들의 합성곱 연산을 통해 다시 확장하여 구성될 수 있다. 복원된 이미지의 픽셀값을 통해 분류하고자 하는 베이스라인 이상 여부가 출력될 수 있다.
- [0035] DeepLab V3+ 에서, 대기질 데이터의 경우 PM10, PM2.5, NO, NO2, NOx, O3, CO, SO2 8개의 성분에 대한 시간별 측정값인 1차원 데이터를 입력으로 가지므로, 1차원 합성곱(1 Dimension Convolution Neural Network)이 활용될 수 있다.
- [0036] 실시예에서 데이터의 경향성을 파악하고 해당 경향성에서 벗어나는 데이터값이 베이스라인 이상으로 판정될 수 있다. 예를 들면, 각 원소의 경향성을 관측하기 위해서는 최소한 이전 데이터를 기준으로 한 1개월 데이터와 판정을 위한 1개월 데이터를 더한 2개월(720시간) 분량의 데이터가 입력 데이터로 설정될 수 있다.
- [0037] 도 3을 참고하면, 베이스라인 이상 판정 모델 구조를 설명하기 위한 도면이다. 베이스라인 이상 판정 모델 구조는 입력된 데이터의 특징을 고차원의 특징으로 추출할 수 있도록 특징 추출기(Feature Extractor)는 레즈넷(ResNet34)기반의 모델을 수정하여 구성될 수 있다. 레즈넷을 통과함에 따라 출력된 최종 특징맵은 SO2, CO, O3 3개의 ASSPP(Atrous Separable Spatial Pyramid Pooling)을 거쳐, 성분별 특징맵이 생성될 수 있다. ASSPP에서는 다양한 크기의 특징을 추출하기 위해 합성곱 필터의 크기를 다르게 적용하여 출력된 특징을 추출한 뒤, 이러한 특징들을 다시 통합할 수 있다. 최종적으로 생성된 각각의 성분별 특징들과 초기에 나온 특징을 통합하고, 통합된 결과를 원본 시간과 동일한 크기로 확장시켜 구성할 수 있다. 최종적으로 입력 720시간(약 2달)의 입력과 출력에 대한, 각성분별로 '시간별 베이스라인 판정결과'가 출력될 수 있다.

- [0038] 도 4는 일 실시예에 있어서, 조각별 집계근사법의 적용 전과 적용 후를 비교한 도면이다.
- [0039] 조각별 집계근사법(Piecewise Aggregate Approximation)이 적용되어 대기오염망 자료(대기질 데이터)가 확정될 수 있다. 매 시각 레이블링을 하는 경우, 사용자마다 오탐지하는 부분이 있을 수 있고, 데이터의 노이즈가 많다는 점에 착안하여, 기존 데이터가 관측시간에 따라 구간이 형성된 뒤, 각 구간별 데이터 평균값이 각 구간의 대표값으로 근사되어 노이즈가 최소화될 수 있다. 도 4(a)는 조각별 집계 근사법의 적용 전을 나타낸 것이고, 도 4(b)는 조각별 집계 근사법의 적용 후를 나타낸 것이다. 도 4와 같이 데이터의 특징을 유지하며, 데이터의 노이즈를 줄일 수 있음을 확인할 수 있다.
- [0040] 도 4에서 녹색 그래프는 측정소의 실제 측정값을 의미하고, 회색 그래프는 측정소를 기준으로 가장 인접한 복수개의 측정소의 시간대별 최소값과 최대값을 의미한다. 조각별 집계 근사법의 적용 전과 조각별 집계 근사법의 적용 후의 데이터를 비교 시, 그래프의 크게 튀는 값들이 감소하고, 그래프에서 나타나지 않던 작은 그래프의 변화 값이 더 부각되게 된다. 그 결과, 조각별 집계 근사법 적용 전에 비하여 그래프의 경향성이 잘 나타나게 된다. 노이즈가 감소된 데이터는 선형보간법(Linear interpolation)을 사용하여, 사라진 노이즈 데이터를 보완할 수 있다. 이에, 기존의 데이터에 대하여 노이즈가 줄어들고 데이터의 흐름에 대한 특성은 대체적으로 보존되는 결과를 보인다. 또는, 예를 들면, 대기질 데이터에 대한 이상치 제거 전 정규화 과정이 수행될 수 있다. 이때, 이상치 제거는 사분위수 범위 제거(Interquartile Range; IQR) 방식을 이용하여 제거될 수 있다.
- [0041] 도 5는 일 실시예에 있어서, GAN 모델의 학습 구조를 설명하기 위한 도면이다.
- [0042] 비지도 학습 모델로서, 생성자(Generator)와 판별자(Discriminator)로 불리는 모델 2가지를 한번에 학습시키는 모델인 적대적 생성 신경망(Generative Adversarial Network; GAN)이 적용될 수 있다.
- [0043] 기존의 대기오염망 측정소에서 자료 확정을 하시는 담당자들의 레이블링을 하는 근거가 각기 다르기 때문에 레이블의 신뢰성이 많이 떨어지기 때문에 학습 데이터만을 이용하여 학습하고자 비지도 학습이 수행될 수 있다. 적대적 생성 신경망은 특정 범주에 대한 예측 형태를 학습하는 모델이다. 도 5를 참고하면, 대기질 데이터를 접목시킨 적대적 생성 신경망 모델의 학습 구조를 나타낸 것이다.
- [0044] 가짜 기상 데이터 제조가는 분류 전문가를 속이는 것을 목적으로 하여 최대한 비슷한 기상 데이터를 생성하고, 분류 전문가는 가짜 기상 데이터 제조가에게 속지 않는 것을 목적으로 하므로 서로 적대적인 관계 속에서 상호발전하며 학습이 진행되는 것이 큰 특징이다. 실시예에서는 적대적 생성 신경망 모델 중에서 다변수 시계열 데이터의 이례적 패턴을 감지할 수 있는 'BeatGan'이라고 하는 모델을 사용할 수 있다.
- [0045] 도 6을 참고하면, BeatGan은 특징을 추출하는 인코더(Encoder)와 특징을 되살리는 디코더(Decoder)로 구성되며, 여기에 적대적 생성 신경망의 학습 방식을 접목한 모델이다. 시계열 데이터의 이례적인 패턴을 감지할 수 있는 것이 큰 특징이다. 정상 패턴을 가지고 있는 시계열 데이터로만 학습을 진행하여 BeatGan 모델이 정상 데이터에 관하여 중요한 특징 정보를 이해할 수 있도록 할 수 있다. 도 7을 참고하면, 비정상 데이터에 대해 높은 이상값을 표출하는 것을 확인할 수 있다.
- [0046] 도 8을 참고하면, 학습 데이터를 생성하는 것을 설명하기 위한 도면이다. 비지도 학습 모델을 이용하여 대기질 데이터 적용이 결정될 수 있다. 기존의 대기질 자료 확정 담당자에 의해 제작된 데이터 레이블이 활용되어 학습 데이터가 구성될 수 있다. 도 8의 왼쪽 도면은 전문가들에 의한 데이터 레이블을 나타낸 것이고, 도 8의 오른쪽 도면은 해당 구간의 데이터를 나타낸 것이다.
- [0047] 도 9 및 도 10은 일 실시예에 있어서, 비정상 데이터 및 정상 데이터에 대한 실험 결과를 나타낸 그래프이다.
- [0048] 도 9는 비정상 데이터에 대한 실험 결과이다. 도 9를 참고하면, 베이스라인이 갑자기 높아지는 현상에 대해서는 높은 재현율을 보이는 결과를 보이지만, 낮아지는 현상에 대해서는 재현율이 떨어지는 결과를 보인다. 도 10은 정상 데이터에 대한 실험 결과이다. 도 10을 참고하면, 진폭이 큰 패턴을 가지고 있는 정상 데이터의 경우 이상치로 탐지하는 현상 발생하는 결과를 보인다. 이는 입력이 굉장히 진폭이 크기 때문에 오차합수가 커져서 모두 다 잘못된 값이라고 하는 거짓 양성(False Positive)라고 하기 때문이다. 이에, 임계값 등이나 전/후처리 등의 수정 과정이 추가적으로 수행될 수 있다.
- [0049] 도 11은 일 실시예에 따른 이상 판정 시스템의 구성을 설명하기 위한 블록도이고, 도 12는 일 실시예에 따른 이상 판정 시스템에서 대기질 데이터의 이상 판정 방법을 설명하기 위한 흐름도이다.
- [0050] 이상 판정 시스템(100)의 프로세서는 입력부(1110) 및 판정부(1120)를 포함할 수 있다. 이러한 프로세서의 구성요소들은 이상 판정 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는

서로 다른 기능들(different functions)의 표현들일 수 있다. 프로세서 및 프로세서의 구성요소들은 도 2의 대기질 데이터의 이상 판정 방법이 포함하는 단계들(1210 내지 1220)을 수행하도록 이상 판정 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

[0051] 프로세서는 대기질 데이터의 이상 판정 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 이상 판정 시스템에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 이상 판정 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서가 포함하는 입력부(1110) 및 판정부(1120) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(1210 내지 1220)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

[0052] 단계(1210)에서 입력부(1110)는 복수 개의 측정소로부터 측정된 대기질 데이터를 지도 학습과 비지도 학습이 앙상블된 학습 모델에 입력받을 수 있다. 일례로, 복수 개의 측정소로부터 측정된 대기질 데이터가 수집될 수 있으며, 주기적으로 매일 모든 측정소의 데이터를 회수하여 전일 시간별 측정값이 선별될 수 있다. 입력부(1110)는 복수 개의 성분에 대한 시간별 측정값인 1차원 데이터로 구성된 대기질 데이터를 지도 학습 모델에 입력받을 수 있다. 입력부(1110)는 대기질 데이터에 조각별 집계근사법을 적용하여 대기질 데이터에 포함된 노이즈를 제거하고, 조각별 집계근사법이 적용됨에 따라 사라진 노이즈 데이터에 대하여 선형보간법(Linear interpolation)을 사용하여 데이터를 보완할 수 있다. 입력부(1110)는 대기질 데이터가 측정된 시간에 따라 구간을 형성하고, 형성된 구간에 포함된 대기질 데이터의 평균값을 구간의 대표값으로 근사하여 노이즈를 제거할 수 있다. 입력부(1110)는 시계열 기반의 대기질 데이터를 비지도 학습 모델에 입력받을 수 있다.

[0053] 단계(1220)에서 판정부(1120)는 지도 학습과 비지도 학습이 앙상블된 학습 모델을 이용하여 대기질 데이터의 이상 여부를 판정할 수 있다. 판정부(1120)는 지도 학습 모델을 통해 대기질 데이터에 대한 시간별 베이스라인 판정 결과를 출력할 수 있다. 판정부(1120)는 비지도 학습 모델을 통해 대기질 데이터에 포함된 비정상 데이터에 대해 이상치를 표출할 수 있다.

[0054] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0055] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0056] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체

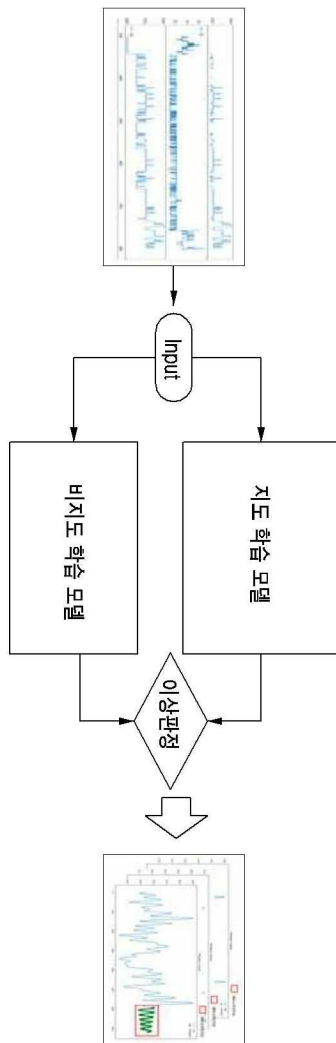
(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0057] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

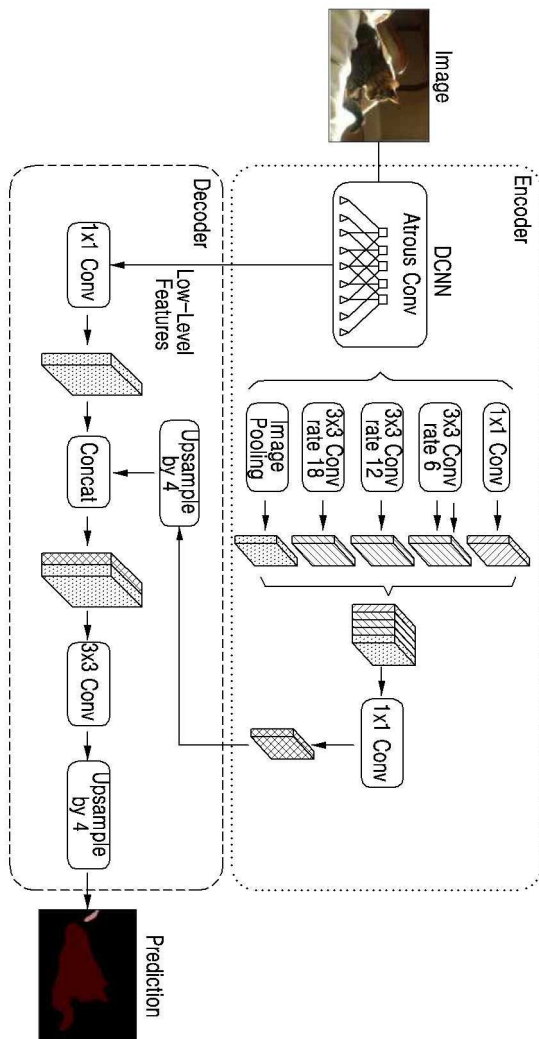
[0058] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

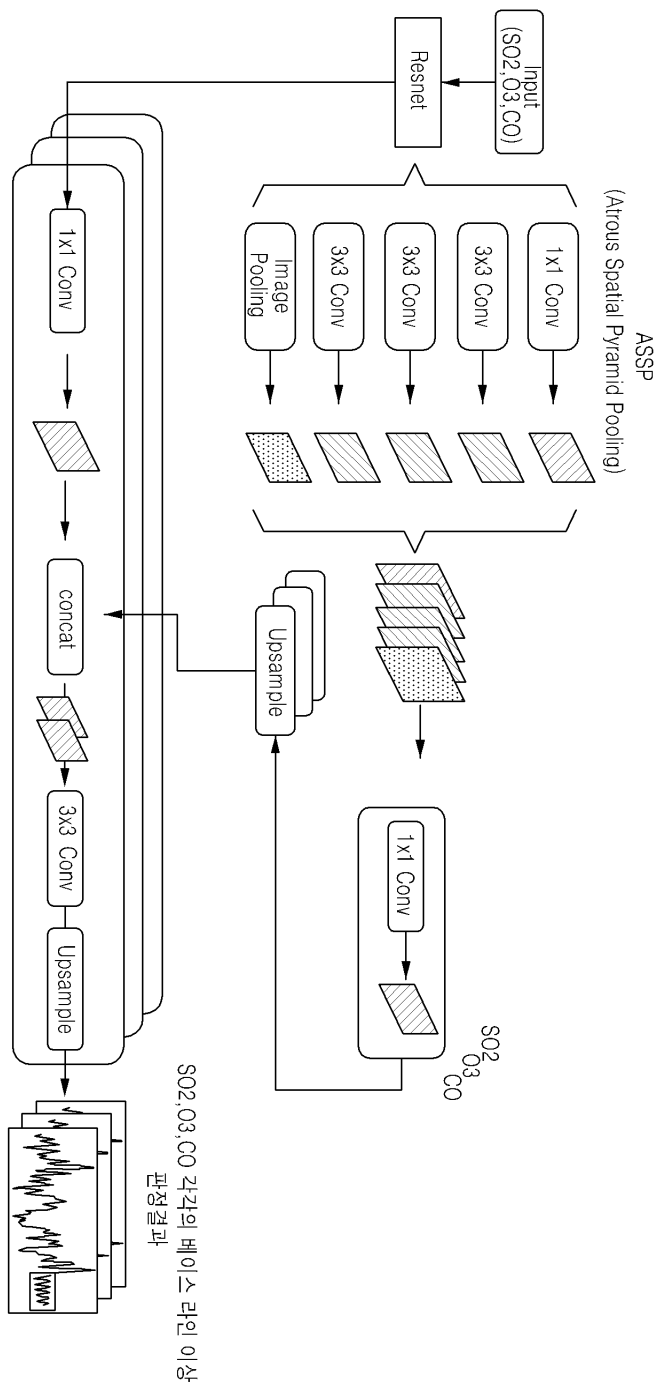
도면1



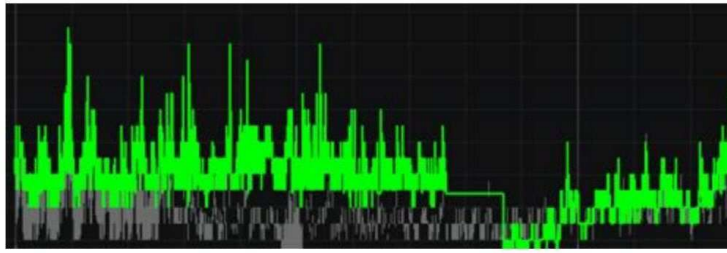
도면2



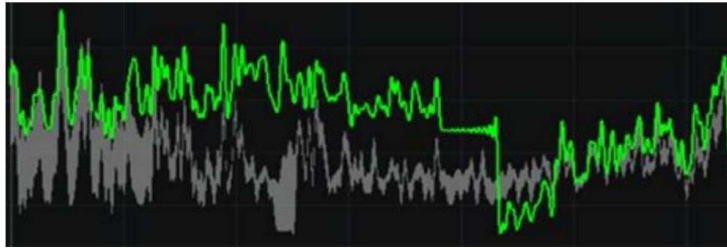
도면3



도면4

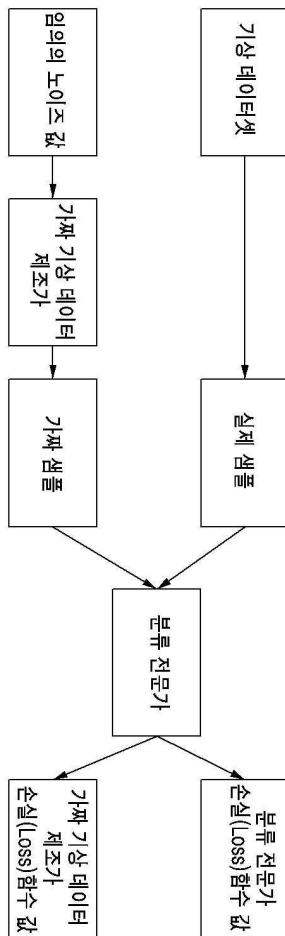


도 4(a) 조각별 집계 근사법 적용 전

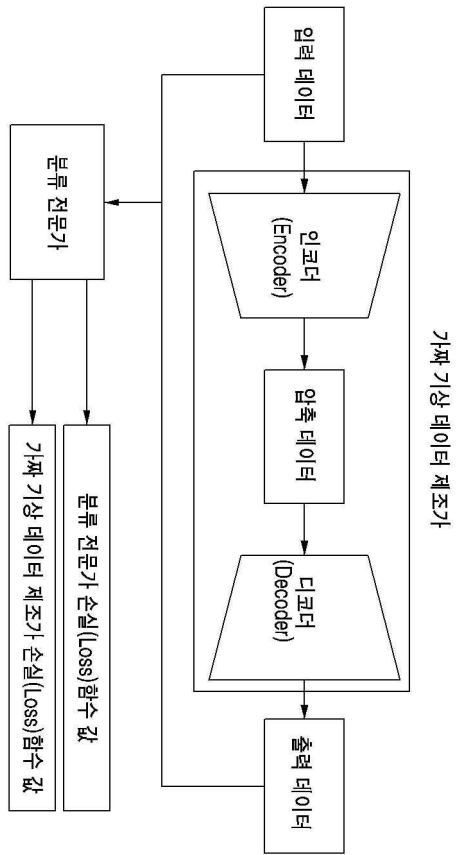


도 4(b) 조각별 집계 근사법 적용 후

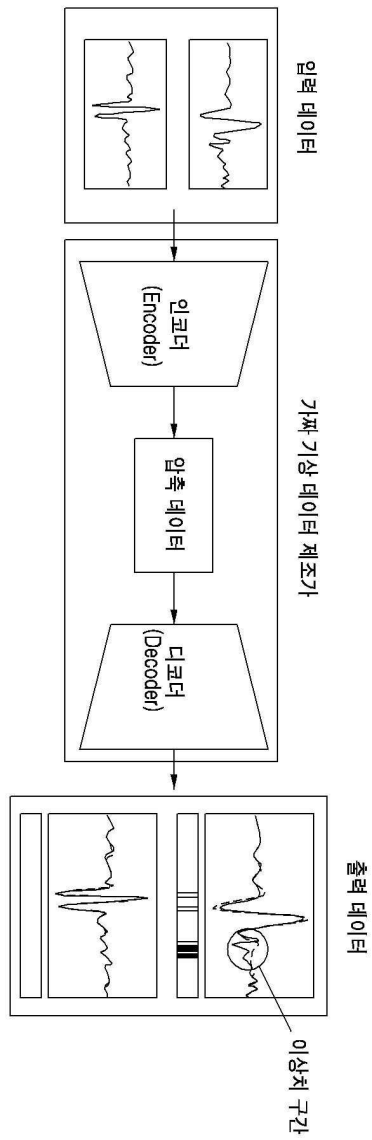
도면5



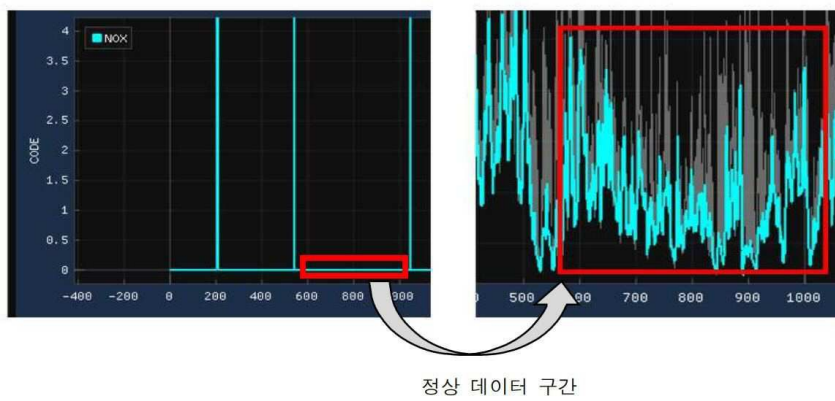
도면6



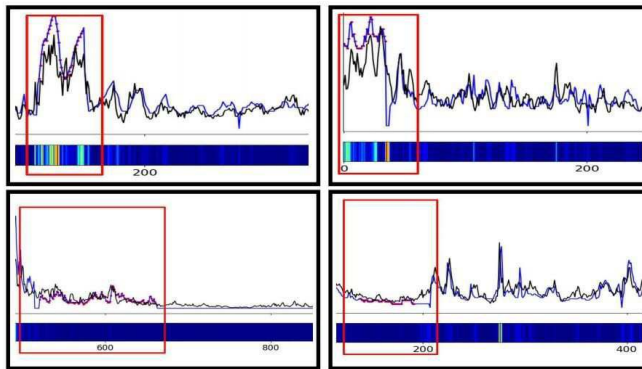
도면7



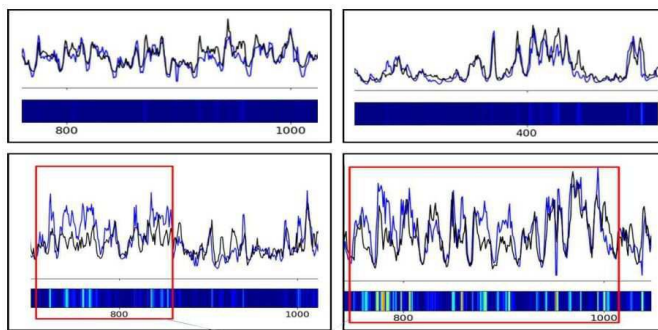
도면8



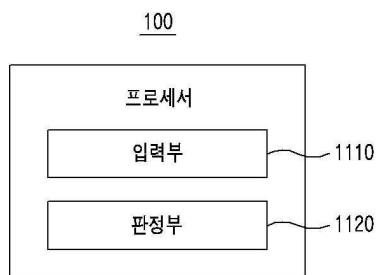
도면9



도면10



도면11



도면12

