(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0178500
(43) 공개일자 2024년12월31일

(51) 국제특허분류(Int. Cl.)
G06F 11/30 (2006.01) *G06N 3/04* (2023.01)
G06N 3/0464 (2023.01) *G06N 3/048* (2023.01)
(52) CPC특허분류
G06F 11/3072 (2013.01)
G06F 11/3058 (2013.01)
(21) 출원번호 10-2023-0080423
(22) 출원일자 2023년06월22일
심사청구일자 2023년06월22일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
정성원
서울특별시 양천구 오목로 300, 204동 604호
이강우
경기도 화성시 영통로50번길 27, 110동 202호
김윤영
서울특별시 마포구 광성로6안길 14, 401호
(74) 대리인
정부연

전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 장치 및 방법**

(57) 요 약

본 발명은 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 장치 및 방법에 관한 것으로, 상기 장치는 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 상기 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성하는 학습 데이터 수집부; 상기 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축하는 모델 구축부; 및 상기 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지하는 이상 탐지부;를 포함한다.

대 표 도 - 도3

(52) CPC특허분류
G06N 3/0464 (2023.01)
G06N 3/048 (2023.01)
G06N 3/049 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711195337
과제번호 2021-0-00180-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신·방송 기술개발사업(데이터경제를위한블록체인기술개발(R&D))
연구과제명 다양한 산업 분야 활용성 증대를 위한 분산 저장된 대규모 데이터 고속 분석 기술개발
기 여 율 1/1
과제수행기관명 한국전자통신연구원
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

복수의 센서들로부터 측정된 센서 데이터를 전처리하고 상기 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성하는 학습 데이터 수집부;

상기 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축하는 모델 구축부; 및

상기 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지하는 이상 탐지부;를 포함하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 2

제1항에 있어서, 상기 학습 데이터 수집부는

상기 그래프 모델링 과정에서 상기 복수의 센서들 각각을 노드(node)로 표현하고 상기 노드 간의 상관계수를 엣지(edge)로 표현하여 상기 인접행렬의 성분값으로 표현하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 3

제1항에 있어서, 상기 학습 데이터 수집부는

상기 센서 데이터를 특정 구간 범위 내의 데이터로 변환하여 정규화하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 4

제3항에 있어서, 상기 학습 데이터 수집부는

상기 정규화된 센서 데이터를 기초로 TLCC(Time Lagged Cross Correlation)를 통해 상기 센서별 상관계수를 산출하여 상기 인접행렬을 생성하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 5

제4항에 있어서, 상기 학습 데이터 수집부는

상기 센서별 상관계수에 기 설정된 임계치를 적용하여 상기 인접행렬을 최적화하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 6

제1항에 있어서, 상기 모델 구축부는

상기 학습 데이터의 센서 데이터와 상기 인접행렬 간의 행렬곱 연산을 수행한 후 상기 GCN의 가중치 레이어(weight layer)를 통과시켜 상기 이상 탐지 모델의 입력 데이터를 생성하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 7

제6항에 있어서, 상기 모델 구축부는

상기 가중치 레이어의 출력에 대해 활성화함수를 적용한 다음 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 데이터를 생성하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 8

제1항에 있어서, 상기 이상 탐지부는

상기 이상 탐지 모델에 상기 주어진 시계열 데이터를 입력으로 제공하여 이상치 점수를 산출하고 상기 이상치 점수를 기초로 상기 이상 여부를 결정하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 9

제8항에 있어서, 상기 이상 탐지부는

상기 이상치 점수를 특정 시간 구간 별로 산출하고 상기 이상치 점수를 기초로 동적 오류 임계치(Dynamic Error Threshold)에 따라 상기 이상 여부의 결정을 위한 임계치를 결정하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 10

제9항에 있어서, 상기 이상 탐지부는

상기 특정 시간 구간 별로 상기 이상치 점수 벡터를 산출하고 상기 이상치 점수 벡터의 평균 및 분산에 관한 임계치 후보 값을 산출하며 기 정의된 목적 함수의 값을 최대화하는 임계치 후보 값을 상기 이상 여부의 결정을 위한 임계치로서 결정하는 것을 특징으로 하는 GCN 기반의 시계열 데이터 이상 탐지 장치.

청구항 11

학습 데이터 수집부를 통해, 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 상기 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성하는 단계;

모델 구축부를 통해, 상기 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축하는 단계; 및

이상 탐지부를 통해, 상기 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지하는 단계;를 포함하는 GCN 기반의 시계열 데이터 이상 탐지 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 데이터 이상 탐지 기술에 관한 것으로, 보다 상세하게는 시계열 데이터의 특징을 반영한 TLCC를 적용하여 정확한 센서별 상관관계를 분석하고, 상관관계 표현력이 뛰어난 GCN을 활용하여 데이터 이상을 탐지할 수 있는 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 시계열 데이터의 이상 탐지는 정상 범위를 크게 초과하는 데이터를 탐지하는 것에 해당할 수 있다. 네트워크 관리/보안, 산업현장 등에서는 각 센서가 수집하는 시계열 데이터를 통해 장비를 모니터링 할 수 있다. 따라서, 시계열 데이터를 효율적으로 분석하여 장비 이상을 조기에 탐지하는 것은 더 큰 피해를 방지하고 생산성 향상에 기여할 수 있다는 점에서 매우 중요한 과제에 해당할 수 있다.

[0004] 또한, 산업현장에서의 시계열 이상 탐지 활용처는 크게는 거대 공장의 제조설비부터 작게는 자동차의 엔진 등 그 범위가 갈수록 증가하고 있어 신속하고 정확한 이상 탐지를 위해 많은 연구가 진행되고 있다.

[0005] 기존 시계열 데이터 이상 탐지 모델에서는 센서별 상관관계를 반영하지 않고 예측기반 모델과 재구성기반 모델을 통해 시계열 데이터를 분석하여 왔다. 따라서, 기존의 방법은 정상 데이터임에도 불구하고 이상으로 탐지하는 불필요 알람이 자주 발생하는 문제점을 가질 수 있다. 이러한 문제점을 보완하기 위해 시계열 데이터 이상 탐지에서 센서별 상관관계를 반영하는 것이 필요할 수 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국공개특허 제10-2017-0084445호 (2017.07.20)

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 실시예는 시계열 데이터의 특징을 반영한 TLCC를 적용하여 정확한 센서별 상관관계를 분석하고, 상관관계 표현력이 뛰어난 GCN을 활용하여 데이터 이상을 탐지할 수 있는 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0010] 실시예들 중에서, GCN 기반의 시계열 데이터 이상 탐지 장치는 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 상기 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성하는 학습 데이터 수집부; 상기 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축하는 모델 구축부; 및 상기 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지하는 이상 탐지부;를 포함한다.

[0011] 상기 학습 데이터 수집부는 상기 그래프 모델링 과정에서 상기 복수의 센서들 각각을 노드(node)로 표현하고 상기 노드 간의 상관계수를 엣지(edge)로 표현하여 상기 인접행렬의 성분값으로 표현할 수 있다.

[0012] 상기 학습 데이터 수집부는 상기 센서 데이터를 특정 구간 범위 내의 데이터로 변환하여 정규화할 수 있다.

[0013] 상기 학습 데이터 수집부는 상기 정규화된 센서 데이터를 기초로 TLCC(Time Lagged Cross Correlation)를 통해 상기 센서별 상관계수를 산출하여 상기 인접행렬을 생성할 수 있다.

[0014] 상기 학습 데이터 수집부는 상기 센서별 상관계수에 기 설정된 임계치를 적용하여 상기 인접행렬을 최적화할 수 있다.

[0015] 상기 모델 구축부는 상기 학습 데이터의 센서 데이터와 상기 인접행렬 간의 행렬곱 연산을 수행한 후 상기 GCN의 가중치 레이어(weight layer)를 통과시켜 상기 이상 탐지 모델의 입력 데이터를 생성할 수 있다.

[0016] 상기 모델 구축부는 상기 가중치 레이어의 출력에 대해 활성화함수를 적용한 다음 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 데이터를 생성할 수 있다.

[0017] 상기 이상 탐지부는 상기 이상 탐지 모델에 상기 주어진 시계열 데이터를 입력으로 제공하여 이상치 점수를 산

출하고 상기 이상치 점수를 기초로 상기 이상 여부를 결정할 수 있다.

[0018] 상기 이상 탐지부는 상기 이상치 점수를 특정 시간 구간 별로 산출하고 상기 이상치 점수를 기초로 동적 오류 임계치(Dynamic Error Threshold)에 따라 상기 이상 여부의 결정을 위한 임계치를 결정할 수 있다.

[0019] 상기 이상 탐지부는 상기 특정 시간 구간 별로 상기 이상치 점수 벡터를 산출하고 상기 이상치 점수 벡터의 평균 및 분산에 관한 임계치 후보 값을 산출하며 기 정의된 목적 함수의 값을 최대화하는 임계치 후보 값을 상기 이상 여부의 결정을 위한 임계치로서 결정할 수 있다.

[0020] 실시예들 중에서, GCN 기반의 시계열 데이터 이상 탐지 방법은 학습 데이터 수집부를 통해, 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 상기 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성하는 단계; 모델 구축부를 통해, 상기 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축하는 단계; 및 이상 탐지부를 통해, 상기 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지하는 단계;를 포함한다.

발명의 효과

[0022] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0023] 본 발명의 일 실시예에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 장치 및 방법은 시계열 데이터의 특징을 반영한 TLCC를 적용하여 정확한 센서별 상관관계를 분석하고, 상관관계 표현력이 뛰어난 GCN을 활용하여 데이터 이상을 탐지할 수 있다.

도면의 간단한 설명

[0025] 도 1은 본 발명에 따른 이상 탐지 시스템을 설명하는 도면이다.

도 2는 도 1의 이상 탐지 장치의 시스템 구성을 설명하는 도면이다.

도 3은 도 1의 이상 탐지 장치의 기능적 구성을 설명하는 도면이다.

도 4는 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 설명하는 순서도이다.

도 5는 본 발명에 따른 데이터 이상 탐지를 위한 학습 구조의 일 실시예를 설명하는 도면이다.

도 6은 본 발명에 따른 TLCC를 적용한 센서별 상관관계 히트맵의 일 실시예를 설명하는 도면이다.

도 7은 본 발명에 따른 동적 오류 임계치 계산 및 이상탐지의 일 실시예를 설명하는 도면이다.

도 8은 본 발명에 따른 이상 탐지 방법에 관한 실험 결과를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0026] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0027] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0028] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2

구성요소도 제1 구성요소로 명명될 수 있다.

[0029] 어떤 구성요소가 다른 구성요소에 "연결되어"있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어"있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0030] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다"또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0031] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0032] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0033] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.

[0035] 도 1은 본 발명에 따른 이상 탐지 시스템을 설명하는 도면이다.

[0036] 도 1을 참조하면, 이상 탐지 시스템(100)은 센서 단말(110), 이상 탐지 장치(130) 및 데이터베이스(150)를 포함할 수 있다.

[0037] 센서 단말(110)은 사용자에 의해 운용되는 단말 장치에 해당할 수 있으며, 적어도 센서를 포함하여 구성될 수 있다. 즉, 센서 단말(110)은 적어도 하나의 센서를 통해 센서 데이터를 측정하고 저장할 수 있다. 본 발명의 실시예에서 사용자는 하나 이상의 사용자로 이해될 수 있으며, 하나 이상의 사용자들 각각은 하나 이상의 센서 단말(110)에 대응될 수 있다. 즉, 도 1에서는 하나의 센서 단말(110)로 표현되어 있으나, 필요에 따라 센서 단말(110)은 복수개로 구현될 수 있다. 이 경우, 각 센서 단말(110)은 동일 사용자에 의해 운용되거나 또는 서로 다른 사용자에 의해 운용될 수 있다.

[0038] 또한, 센서 단말(110)은 본 발명에 따른 이상 탐지 시스템(100)을 구성하는 하나의 장치로서 구현될 수 있으며, 이상 탐지 시스템(100)은 GCN 기반의 시계열 데이터 이상 탐지 목적에 따라 다양한 형태로 변형되어 구현될 수 있다.

[0039] 또한, 센서 단말(110)은 이상 탐지 장치(130)와 연결되어 동작 가능한 다양한 단말로 구현될 수 있으며, 예를 들어 적어도 센서를 포함하여 동작하는 스마트폰, 노트북 또는 컴퓨터 등의 다양한 디바이스에 해당될 수도 있다.

[0040] 한편, 센서 단말(110)은 이상 탐지 장치(130)와 네트워크를 통해 연결될 수 있고, 복수의 센서 단말(110)들은 이상 탐지 장치(130)와 동시에 연결될 수도 있다.

[0041] 이상 탐지 장치(130)는 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 수행하는 컴퓨터 또는 프로그램에 해당하는 서버로 구현될 수 있다. 또한, 이상 탐지 장치(130)는 센서 단말(110)과 유선 네트워크 또는 블루투스, WiFi, LTE 등과 같은 무선 네트워크로 연결될 수 있고, 네트워크를 통해 센서 단말(110)과 데이터를 송·수신할 수 있다. 또한, 이상 탐지 장치(130)는 독립된 외부 시

스템(도 1에 미도시함)과 연결되어 동작하도록 구현될 수 있다.

[0042] 데이터베이스(150)는 이상 탐지 장치(130)의 동작 과정에서 필요한 다양한 정보들을 저장하는 저장장치에 해당할 수 있다. 예를 들어, 데이터베이스(150)는 각 센서 단말(110)로부터 수집된 센서 데이터를 유형에 따라 분류하여 저장하거나 또는 이상 탐지 모델의 구축을 위한 학습 알고리즘에 관한 정보를 저장할 수 있으며, 반드시 이에 한정되지 않고, 이상 탐지 장치(130)가 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 수행하는 과정에서 다양한 형태로 수집 또는 가공된 정보들을 저장할 수 있다.

[0043] 또한, 도 1에서, 데이터베이스(150)는 이상 탐지 장치(130)와 독립적인 장치로서 도시되어 있으나, 반드시 이에 한정되지 않고, 논리적인 저장장치로서 이상 탐지 장치(130)에 포함되어 구현될 수 있음은 물론이다.

[0045] 도 2는 도 1의 이상 탐지 장치의 시스템 구성을 설명하는 도면이다.

[0046] 도 2를 참조하면, 이상 탐지 장치(130)는 프로세서(210), 메모리(230), 사용자 입출력부(250) 및 네트워크 입출력부(270)를 포함할 수 있다.

[0047] 프로세서(210)는 본 발명의 실시예에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 프로시저를 실행할 수 있고, 이러한 과정에서 읽혀지거나 작성되는 메모리(230)를 관리할 수 있으며, 메모리(230)에 있는 휘발성 메모리와 비휘발성 메모리 간의 동기화 시간을 스케줄 할 수 있다. 프로세서(210)는 이상 탐지 장치(130)의 동작 전반을 제어할 수 있고, 메모리(230), 사용자 입출력부(250) 및 네트워크 입출력부(270)와 전기적으로 연결되어 이들 간의 데이터 흐름을 제어할 수 있다. 프로세서(210)는 이상 탐지 장치(130)의 CPU(Central Processing Unit) 또는 GPU(Graphics Processing Unit)로 구현될 수 있다.

[0048] 메모리(230)는 SSD(Solid State Disk) 또는 HDD(Hard Disk Drive)와 같은 비휘발성 메모리로 구현되어 이상 탐지 장치(130)에 필요한 데이터 전반을 저장하는데 사용되는 보조기억장치를 포함할 수 있고, RAM(Random Access Memory)과 같은 휘발성 메모리로 구현된 주기억장치를 포함할 수 있다. 또한, 메모리(230)는 전기적으로 연결된 프로세서(210)에 의해 실행됨으로써 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 실행하는 명령어들의 집합을 저장할 수 있다.

[0049] 사용자 입출력부(250)는 사용자 입력을 수신하기 위한 환경 및 사용자에게 특정 정보를 출력하기 위한 환경을 포함하고, 예를 들어, 터치 패드, 터치 스크린, 화상 키보드 또는 포인팅 장치와 같은 어댑터를 포함하는 입력장치 및 모니터 또는 터치 스크린과 같은 어댑터를 포함하는 출력장치를 포함할 수 있다. 일 실시예에서, 사용자 입출력부(250)는 원격 접속을 통해 접속되는 컴퓨팅 장치에 해당할 수 있고, 그러한 경우, 이상 탐지 장치(130)는 독립적인 서버로서 수행될 수 있다.

[0050] 네트워크 입출력부(270)는 네트워크를 통해 센서 단말(110)과 연결되기 위한 통신 환경을 제공하고, 예를 들어, LAN(Local Area Network), MAN(Metropolitan Area Network), WAN(Wide Area Network) 및 VAN(Value Added Network) 등의 통신을 위한 어댑터를 포함할 수 있다. 또한, 네트워크 입출력부(270)는 데이터의 무선 전송을 위해 WiFi, 블루투스 등의 근거리 통신 기능이나 4G 이상의 무선 통신 기능을 제공하도록 구현될 수 있다.

[0052] 도 3은 도 1의 이상 탐지 장치의 기능적 구성을 설명하는 도면이다.

[0053] 도 3을 참조하면, 이상 탐지 장치(130)는 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 수행할 수 있다. 이를 위하여, 이상 탐지 장치(130)는 학습 데이터 수집부(310), 모델 구축부(330), 이상 탐지부(350) 및 제어부(370)를 포함할 수 있다.

[0054] 이때, 본 발명의 실시예는 상기의 구성들을 동시에 모두 포함해야 하는 것은 아니며, 각각의 실시예에 따라 상기의 구성들 중 일부를 생략하거나, 상기의 구성들 중 일부 또는 전부를 선택적으로 포함하여 구현될 수도 있다. 이하, 각 구성들의 동작을 구체적으로 설명한다.

[0055] 학습 데이터 수집부(310)는 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성할 수 있다. 즉, 학습 데이터 수집부(310)는 센서 데이터를 기초로 그래프 모델링을 통해 센서들 간의 상관관계를 표현하는 그래프 구조를 도출할 수 있다. 또한, 학습 데이터 수집부(310)는 그래프 모델링 과정에서 센서별 상관관계를 인접행렬로서 표현

할 수 있다. 학습 데이터 수집부(310)에 의해 전처리된 센서 데이터와 인접행렬은 이후 이상 탐지 모델을 위한 학습 과정에 사용될 수 있다.

[0056] 일 실시예에서, 학습 데이터 수집부(310)는 그래프 모델링 과정에서 복수의 센서들 각각을 노드(node)로 표현하고 노드 간의 상관계수를 엣지(edge)로 표현하여 인접행렬의 성분값으로 표현할 수 있다. 즉, 학습 데이터 수집부(310)는 각 센서들을 그래프의 노드에 대응시켜 표현할 수 있고, 노드 간의 엣지(또는 간선)는 센서 사이의 상관관계에 대응시켜 표현할 수 있다. 이때, 학습 데이터 수집부(310)는 센서들 간의 상관관계를 완전 그래프 형태로 표현하는 경우 모든 센서들에 대한 연산이 비효율적으로 수행되는 점을 고려하여 인접행렬을 최적화할 수 있다.

[0057] 일 실시예에서, 학습 데이터 수집부(310)는 센서 데이터를 특정 구간 범위 내의 데이터로 변환하여 정규화할 수 있다. 예를 들어, 학습 데이터 수집부(310)는 센서 데이터에 대해 'Min-Max Scaler(최소-최대 스케일러)'를 적용하여 0과 1 사이의 값으로 정규화하는 전처리 동작을 수행할 수 있다. 여기에서, Min-Max Scaler는 스케일(scale)을 조정하는 정규화 함수로서 모든 값을 0과 1 사이의 값으로 변환해 주는 함수에 해당할 수 있다. 한편, 학습 데이터 수집부(310)는 다양한 정규화 함수를 통해 센서 데이터의 정규화를 수행할 수 있음은 물론이다.

[0058] 일 실시예에서, 학습 데이터 수집부(310)는 정규화된 센서 데이터를 기초로 TLCC(Time Lagged Cross Correlation)를 통해 센서별 상관계수를 산출하여 인접행렬을 생성할 수 있다. 즉, 학습 데이터 수집부(310)는 센서별 상관관계 분석을 위하여 센서 사이의 상관계수를 산출할 수 있다. 여기에서, TLCC는 피어슨(Pearson) 상관계수에 시간 종속성을 반영한 상관계수에 해당할 수 있으며, 다음의 수학식 1과 같이 표현될 수 있다.

[0059] [수학식 1]

[0060] $$\rho(i,j) = \max\left(\frac{cov(i,j_\tau)}{\sigma_i \sigma_{j_\tau}}\right), \tau = 0,1,\ldots,t$$

[0062] 여기에서, i는 기준 센서이고, j_τ는 센서 i의 시작 시간 τ를 0에서부터 학습 데이터셋의 마지막 인덱스 t까지 변화했을 때의 비교신호이다. $cov(i,j_\tau)$는 기준센서 i와 j_τ의 공분산이고, σ_i와 σ_{j_τ}는 i와 j_τ의 표준편차이다. 이를 통해, 학습 데이터 수집부(310)는 TLCC를 사용하여 시간 종속성을 포함한 센서별 상관관계를 분석할 수 있다.

[0063] 일 실시예에서, 학습 데이터 수집부(310)는 센서별 상관계수에 기 설정된 임계치를 적용하여 인접행렬을 최적화할 수 있다. 상기의 수학식 1로 표현되는 센서 간의 상관관계 $\rho(i,j)$ 값을 그대로 사용하여 모델링을 수행하는 경우 완전 그래프 모델링과 마찬가지로 모든 센서에 대한 연산이 진행되어 비효율적일 수 있다. 따라서, 학습 데이터 수집부(310)는 센서별 상관계수에 기 설정된 임계치를 적용하여 인접행렬을 최적화할 수 있다. 예를 들어, 학습 데이터 수집부(310)는 다음의 수학식 2와 같이 표현되는 γ값을 적용하여 상관관계가 높은 센서만 연산에 포함되도록 최적의 인접행렬 A^{TLCC}를 생성할 수 있다.

[0064] [수학식 2]

[0065] $$A_{i,j}^{TLCC} = \begin{cases} 1 & if\ \rho(i,j) \geq \gamma \\ 0 & otherwise \end{cases}$$

[0067] 여기에서, γ값은 하이퍼 파라미터이고, 예를 들어 0.9로 설정될 수 있다. 또한, 상기의 수학식 2에 따른 인접행렬의 시각화에 대해서는 도 6을 통해 보다 자세히 설명한다.

[0068] 모델 구축부(330)는 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축할 수 있다. 모델 구축부(330)는 그래프 모델링 이후 GCN, GRU 및 예측 모델을 적용하여 이상 탐지 모델을 학습시키는 동작을 수행할 수 있다.

[0069] 일 실시예에서, 모델 구축부(330)는 학습 데이터의 센서 데이터와 인접행렬 간의 행렬곱 연산을 수행한 후 GCN

의 가중치 레이어(weight layer)를 통과시켜 이상 탐지 모델의 입력 데이터를 생성할 수 있다. 구체적으로, 모델 구축부(330)는 전처리 과정을 거친 센서 데이터와 센서별 상관관계가 반영된 인접행렬 간의 행렬곱 연산을 수행할 수 있고, 이후 1D-컨볼루션(Convolution) 연산을 수행할 수 있다. 1D-컨볼루션 연산의 결과는 다시 가중치 레이어를 통과할 수 있고 이후 이상 탐지 모델의 학습을 위한 입력으로 사용될 수 있다.

[0070] 일 실시예에서, 모델 구축부(330)는 가중치 레이어의 출력에 대해 활성화함수를 적용한 다음 GRU(Gated Recurrent Unit)에 입력하여 입력 데이터를 생성할 수 있다. 예를 들어, 모델 구축부(330)는 가중치 레이어의 결과에 활성화함수 'ReLu'를 적용하여 GCN 단계의 출력을 생성할 수 있다. 이때, GCN 단계의 출력은 다음의 수학식 3과 같이 표현될 수 있다.

[0071] [수학식 3]

[0072) $$H^1 = ReLU(A^{TLCC}H^0W + b)$$

[0074] 여기에서, H^1는 최종 임베딩벡터이고, A^{TLCC}는 TLCC를 통해 추출된 인접행렬이며, H^0은 센서 데이터이다. 또한, W와 b는 모델 파라미터이고, 예를 들어, 가중치 값과 바이어스(bias)에 해당할 수 있다.

[0075] 또한, GCN 단계를 거친 최종 임베딩벡터 H^1은 시계열 데이터 분석에 최적화된 GRU를 거친 후 이상 탐지 모델의 입력으로 생성될 수 있다. 이상 탐지 모델은 입력 데이터를 기반으로 다음 t+1의 1차원의 예측값인 $\hat{y}_t$을 출력할 수 있다. 모델 구축부(330)는 손실(Loss)을 산출하기 위하여 t+1의 예측값인 $\hat{y}_t$와 실제값인 y_t를 비교하여 그 차이를 줄여나가는 MSE(Mean Square Error) Loss를 활용할 수 있다. 즉, 모델 구축부(330)는 산출된 손실을 역전파하여 각 레이어의 가중치(weight)를 학습할 수 있다.

[0076] 한편, 모델 구축부(330)는 이상 탐지 모델로서 재구성기반 모델이 아닌 예측기반 모델만 사용할 수 있다. 예측기반 모델은 분석 모델을 최종 통과한 예측값과 정답값 간의 오차를 기반으로 이상 탐지를 진행할 수 있다. 예측기반 모델은 LSTM-NDT, DAGMM 등을 포함할 수 있다. LSTM-NDT는 시계열 데이터 이상 탐지에 중점을 둔 모델로서, LSTM을 통해 생성된 예측값과 원본 데이터와의 차이를 동적 Threshold 설정을 적용하여 이상을 탐지할 수 있다. DAGMM은 Deep Auto-Encoder와 GMM(Gaussian Mixture Model)을 사용하여 시간 종속성이 없고 하나 이상의 특징(feature)을 입력으로 받는 다변량 데이터의 이상 탐지에 중점을 둔 모델에 해당할 수 있다.

[0077] 또한, 재구성기반 모델은 입력 데이터와 동일하게 재구성한 벡터와 입력 데이터와의 차이를 통해 산출된 오차를 학습시켜 시계열 데이터의 이상을 탐지하는 모델에 해당할 수 있다. 재구성기반 모델은 LSTM-VAE와 OmniAnomaly 등을 포함할 수 있다. LSTM-VAE은 인코더(Encoder)와 디코더(Decoder) 부분에 LSTM을 적용하고, 평균과 분산을 통해 이상치 점수(Anomaly Score)를 산출하는 모델로 시간 종속성을 반영할 수 있다. OmniAnomaly는 GRU(Gated Recurrent Unit)와 VAE(Variational Auto-Encoder)를 통해 입력 데이터와 동일하게 재구성하고 입력 데이터와의 차이를 확률적 모델을 적용 및 비교하여 이상을 탐지할 수 있다.

[0078] 다만, 기존의 모델들은 모두 센서별 상관관계를 분석하지 않고 각 센서의 이상 유무만을 분석하기 때문에 이를 통해 이상 탐지를 수행하는 경우 불필요한 허위 알람이 발생한다는 문제점이 존재할 수 있다. 또한, MTAD-GAT는 센서별 상관관계를 고려하고, 예측기반 모델과 재구성기반 모델을 합친 모델에 해당할 수 있다. 해당 모델은 센서들을 완전 그래프로 모델링을 하고, 시간 종속성 및 센서별 상관관계를 파악하기 위해 두 가지의 GAT(Graph Attention Networks)를 적용할 수 있다. 이후, MTAD-GAT는 예측기반 모델과 재구성기반 모델을 모두 활용하여 이상을 탐지할 수 있다. 하지만, 해당 방식은 정확한 센서별 상관관계를 파악하는데 불필요한 연산량이 증가하여 많은 분석시간이 소요된다는 단점을 가질 수 있다.

[0079] 이상 탐지부(350)는 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지할 수 있다. 예를 들어, 이상 탐지부(350)는 주어진 센서 데이터를 기초로 TLCC 인접행렬을 통한 그래프 모델링과 GCN 및 GRU를 통해 최종 생성된 입력 데이터를 이상 탐지 모델에 입력하여 센서별 예측값을 산출하고 센서별 예측값과 실제값 간의 차이를 기초로 이상치 점수(Anomaly Score)를 산출하여 이상 여부를 결정할 수 있다.

[0080] 일 실시예에서, 이상 탐지부(350)는 이상 탐지 모델에 주어진 시계열 데이터를 입력으로 제공하여 이상치 점수를 산출하고 이상치 점수를 기초로 이상 여부를 결정할 수 있다. 예를 들어, 이상 탐지부(350)는 다음의 수학식 4를 통해 이상치 점수를 산출할 수 있다.

[0081] [수학식 4]

[0082] $$Anomaly\ Score = \sum_{i=1}^{k} s_i = \sum_{i=1}^{k} (\hat{y}_i - y_i)^2$$

[0084] 여기에서, i는 센서의 인덱스이고, k는 전체 센서의 개수이다. 이상 탐지부(350)는 센서별 예측값 $\hat{y}$과 실제값 y의 차이를 제곱한 후 합한 이상치 점수를 시간 구간(time step) 별로 산출할 수 있다. 이상 탐지부(350)는 이상치 점수를 기 설정된 임계치와 비교하여 이상 여부를 결정할 수 있다.

[0085] 일 실시예에서, 이상 탐지부(350)는 이상치 점수를 특정 시간 구간 별로 산출하고 이상치 점수를 기초로 동적 오류 임계치(Dynamic Error Threshold)에 따라 이상 여부의 결정을 위한 임계치를 결정할 수 있다. 이상 탐지부(350)는 동적 오류 임계치를 통해 이상 탐지를 위한 적절한 임계치를 결정할 수 있으며, 임계치보다 크거나 같은 값을 이상으로 판단할 수 있다. 여기에서, 동적 오류 임계치는 수시로 변화하는 주변 환경과 시간별로 수집되는 시계열 데이터들을 고려하여 임계치를 선택하기 위한 빠르고 비지도적인 방법에 해당할 수 있다. LSTM 기반 모델에서는 급격한 센서 데이터의 변화를 예측하기 어려워 정상임에도 불구하고 예측값과 결과값을 통해 산출된 이상치 점수가 Sharp Spike 형태로 나타날 수 있다. 따라서, 이러한 오차를 줄이기 위해 별도 라벨이나 파라미터 없이 비지도적인 방법을 적용하여 임계치를 선택할 수 있다.

[0086] 일 실시예에서, 이상 탐지부(350)는 특정 시간 구간 별로 이상치 점수 벡터를 산출하고 이상치 점수 벡터의 평균 및 분산에 관한 임계치 후보 값을 산출하며 기 정의된 목적 함수의 값을 최대화하는 임계치 후보 값을 이상 여부의 결정을 위한 임계치로서 결정할 수 있다. 보다 구체적으로, 다음의 수학식 5는 임계치 후보 ϵ의 계산식이고, 수학식 6은 목적함수이며, 수학식 7은 수학식 6을 산출하기 위한 변수의 값에 해당할 수 있다.

[0087] [수학식 5]

[0088] $$\epsilon = \mu(E) + Z\sigma(E)$$

[0089] [수학식 6]

[0090] $$f(\epsilon) = \frac{\Delta\mu(E)/\mu(E) + \Delta\sigma(E)/\sigma(E)}{|E_a| + |E_{seq}|^2}$$

[0091] [수학식 7]

[0092] $$\Delta\mu(E) = \mu(E) - \mu(\{e \in E | e < \epsilon\})$$

[0093] $$\Delta\sigma(E) = \sigma(E) - \sigma(\{e \in E | e < \epsilon\})$$

[0094] $$E_a = \{e \in E | e > \epsilon\}$$

[0096] 여기에서, μ는 평균이고, σ는 분산이며, E는 이상치 점수의 벡터이고, Z는 2.5부터 11.5까지 0.5의 간격을 가진 실수 19개의 집합이며, $\{e \in E | e < \epsilon\}$는 이상치 점수 중 ϵ보다 작은 값들의 집합이고, E_a는 ϵ보다 큰 값들의 집합, E_{seq}는 E_a 중 연속된 time step을 가진 값의 집합이다.

[0097] 또한, 이상 탐지부(350)에 의해 수행되는 동적 오류 임계치(Dynamic Error Threshold)를 이용한 임계치 설정과 이상 탐지 프로세스는 다음과 같이 진행될 수 있다. 먼저, 전체 훈련 데이터에 대한 상기의 수학식 4의 이상치 점수 벡터 E_{train}을 구한 후 $\mu(E_{train})$, $\sigma(E_{train})$이 산출될 수 있다. 그 후, 상기의 수학식 5을 이용하여 Z에 따른 19개의 임계치 후보 값 ϵ가 산출될 수 있다. ϵ에 따른 상기의 수학식 7의 변수 값을 이용하여 상기의 수학식 6의 계산 값을 가장 크게 만드는 ϵ가 이상 탐지를 위한 임계치로 설정될 수 있다. 테스트 데이터에 대한 상

기의 수학식 4의 이상치 점수 벡터 E_{test} 의 각 요소에 대해 ϵ_{best} 보다 크거나 같으면 이상, 작으면 정상으로 탐지될 수 있다. 이에 대해서는 도 7에서 보다 자세히 설명한다.

[0098] 제어부(370)는 이상 탐지 장치(130)의 전체적인 동작을 제어하고, 학습 데이터 수집부(310), 모델 구축부(330) 및 이상 탐지부(350) 간의 제어 흐름 또는 데이터 흐름을 관리할 수 있다.

[0100] 도 4는 본 발명에 따른 센서별 시간지연 교차 상관관계를 적용한 GCN 기반의 시계열 데이터 이상 탐지 방법을 설명하는 순서도이다.

[0101] 도 4를 참조하면, 이상 탐지 장치(130)는 학습 데이터 수집부(310)를 통해 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성할 수 있다(단계 S410). 이상 탐지 장치(130)는 모델 구축부(330)의 그래프 모델링 과정을 통해 생성된 학습 데이터를 학습하여 이상 탐지를 위한 GCN(Graph Convolution Network) 기반의 이상 탐지 모델을 구축할 수 있다(단계 S430). 이상 탐지 장치(130)는 이상 탐지부(350)를 통해 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지할 수 있다(단계 S450).

[0103] 도 5는 본 발명에 따른 데이터 이상 탐지를 위한 학습 구조의 일 실시예를 설명하는 도면이다.

[0104] 도 5를 참조하면, 이상 탐지 장치(130)는 SC-GCNAD(Sensor-specific Correlation GCN Anomaly Detection)를 통해 시계열 데이터에 관한 이상 탐지를 수행할 수 있다. SC-GCNAD는 시계열 데이터의 특징을 반영한 TLCC를 적용하여 정확한 센서별 상관관계를 분석하고, 상관관계 표현력이 뛰어난 GCN을 활용할 수 있다. 그 결과, 기존 모델 대비 F1-Score는 최대 6.37% 향상될 수 있고, 분석시간은 최대 86.72% 단축될 수 있다.

[0105] 보다 구체적으로, SC-GCNAD는 준지도 학습기반으로 정상 데이터만을 가지고 총 3단계를 거쳐 학습이 진행될 수 있다. 도 5에서, 1단계(Step 1)는 그래프 모델링(Graph Modeling)으로 TLCC(Time Lagged Cross Correlation)를 통해 센서별 상관관계를 분석한 결과로서 인접행렬을 추출하는 단계에 해당할 수 있다. 즉, 1단계에서는 센서들에 관한 그래프 구조가 도출될 수 있으며, 그래프의 노드는 각 센서를 나타내고, 엣지는 임계치가 적용된 인접행렬을 나타낼 수 있다.

[0106] 2단계(Step)는 GCN이 적용되는 단계로써 최종 손실(Loss)를 통해 가중치(Weight)가 학습되는 단계에 해당할 수 있다. 학습 과정에서는 1단계에서 도출한 그래프 구조와 가중치가 반영된 중간값이 생성되어 3단계로 전달될 수 있으며, 3단계를 통해 도출된 손실에 따라 가중치가 갱신될 수 있다. 3단계(Step)는 시계열 분석에 최적화된 GRU(Gated Recurrent Unit)와 예측 모델(Forecasting Model)을 통해 출력된 예측값과 정답값에 평균제곱오차(MSE, Mean-Square Error) Loss를 적용하여 모델을 학습시키는 단계에 해당할 수 있다. 즉, 산출된 손실에 따라 역전파를 통해 2단계 GCN의 가중치가 갱신될 수 있다.

[0107] 이상 탐지 장치(130)는 상기의 1 ~ 3단계에 의한 학습이 완료된 이상 탐지 모델에 주어진 데이터를 입력하여 이상치 점수를 산출할 수 있고, Dynamic Error Threshold를 통해 결정된 임계치를 적용하여 이상 여부를 최종 결정할 수 있다. 이때, 이상 탐지 모델에 입력되는 데이터 역시 상기의 1 ~ 3단계를 통과할 수 있으며, 해당 결과로서 최종 출력값이 생성될 수 있다.

[0109] 도 6은 본 발명에 따른 TLCC를 적용한 센서별 상관관계 히트맵의 일 실시예를 설명하는 도면이다.

[0110] 도 6을 참조하면, 이상 탐지 장치(130)는 복수의 센서들로부터 측정된 센서 데이터를 전처리하고 복수의 센서들에 관한 그래프 모델링(Graph Modeling) 과정에서 센서별 상관관계를 추출하여 인접행렬을 생성할 수 있다. 이때, 인접행렬은 센서별 상관계수에 기 설정된 임계치를 적용한 결과에 따라 최적화될 수 있다.

[0111] 도 6의 경우, 이상 탐지 장치(130)에 의해 생성된 인접행렬의 일부분이 히트맵(Heat Map)을 통해 시각화된 결과에 해당할 수 있다. 즉, 히트맵 상에서 상관관계가 높은 센서는 회색(1), 상관관계가 낮은 센서는 검은색(0)으로 표현될 수 있다. 이상 탐지 장치(130)는 해당 장비에 대한 사전정보가 부족한 상황에서 TLCC를 적용함으로써 상관관계가 높은 센서들을 분석하여 인접행렬을 추출할 수 있고, 이를 활용하여 그래프 모델링을 수행할 수 있다.

[0113] 도 7은 본 발명에 따른 동적 오류 임계치 계산 및 이상탐지의 일 실시예를 설명하는 도면이다.

[0114] 도 7을 참조하면, 이상 탐지 장치(130)는 이상 탐지 모델을 이용하여 주어진 시계열 데이터의 이상 여부를 탐지할 수 있다. 이때, 이상 탐지 장치(130)는 이상 탐지에 필요한 적절한 임계치를 선택하기 위하여 동적 오류 임계치를 적용할 수 있다. 즉, 이상 탐지 장치(130)는 동적 오류 임계치를 통해 수시로 변화하는 환경과 시간에 따른 시계열 데이터에 최적화된 임계치를 빠르게 결정할 수 있다.

[0115] 도 7에서, 여기서 t는 ϵ를 계산하기 위해 사용되는 time step의 수, a는 이상탐지의 대상이 되는 time step의 수이다. 학습 데이터의 이상치 점수 벡터 E_{train}가 첫 번째 표와 같을 때, $\mu(E_{train})$와 $\sigma(E_{train})$값은 각각 0.98과 3.06에 해당할 수 있다. 상기의 수학식 5를 이용한 Z에 따른 ϵ값들이 두 번째 표와 같이 계산될 수 있고, 상기의 수학식 6의 값을 최대화하는 ϵ_{best}가 36.17이라고 할 때, 테스트 데이터의 이상치 점수 벡터 E_{test}에 대한 이상 탐지결과는 세 번째 표와 같이 나타날 수 있다. 즉, Time step 1, 2, 3, 5, a의 이상치 점수(Anomaly Score)는 36.17보다 작기 때문에 정상으로 탐지될 수 있고, Time step 4, n, n+1, n+2의 이상치 점수는 36.17보다 크기 때문에 이상으로 탐지될 수 있다.

[0117] 도 8은 본 발명에 따른 이상 탐지 방법에 관한 실험 결과를 설명하는 도면이다.

[0118] 도 8을 참조하면, 해당 실험을 위하여 세 가지의 공공 데이터셋이 사용될 수 있다. SMAP(Soil Moisture Active Passive satellite)과 MSL(Mars Science Laboratory rover)은 실제 NASA의 우주선 원격 측정 데이터와 토양 수분의 이상치를 종합한 데이터셋이고, SMD(Server Machine Dataset)는 대형 인터넷 회사에서 5주일 동안 서로 다른 28개의 'Server Machine' 내에 있는 센서의 측정값을 종합한 데이터셋에 해당할 수 있다.

[0119] 도 8의 그래프는 데이터셋 별 상관관계를 반영한 모델인 MTAD-GAT과 SC-GCNAD의 분석시간을 비교한 결과에 해당할 수 있다. 여기에서, MTAD-GAT는 센서별 상관관계를 고려하고, 예측기반 모델과 재구성기반 모델을 합친 모델에 해당할 수 있다. MTAD-GAT는 센서들을 완전 그래프로 모델링하고, 시간 종속성 및 센서별 상관관계를 파악하기 위하여 두 가지의 GAT(Graph Attention Networks)를 적용할 수 있다. 이후, MTAD-GAT는 예측기반 모델과 재구성기반 모델을 모두 활용하여 이상을 탐지할 수 있다. 다만, 해당 방식은 정확한 센서별 상관관계를 파악할 수 없고, 불필요한 연산량이 증가하여 많은 분석시간이 소요된다는 단점을 가질 수 있다.

[0120] 즉, MTAD-GAT에서는 완전 그래프로 구조를 모델링하고, GAT를 적용함에 따라 연산량이 증가해 분석시간이 길어질 수 있다. 그에 반해, 본 발명에 따른 SC-GCNAD는 TLCC 기반으로 추출된 인접행렬을 통해 구조를 모델링하고, GAT 대신 GCN을 통해 학습이 진행될 수 있다. 또한, SC-GCNAD는 예측 모델만을 사용하여 불필요한 연산 부분인 재구성 모델을 제거할 수 있다.

[0121] 그 결과, SC-GCNAD는 'SMAP' 데이터셋에서는 '350.974초'로 MTAD-GAT 대비 76.30%의 분석시간을 단축할 수 있고, 'MSL' 데이터셋에서는 '129.946초'로 MTAD-GAT 대비 86.71%의 분석시간을 단축할 수 있으며, 'SMD' 데이터셋에서는 '57.412초'로 MTAD-GAT 대비 82.76%의 분석시간을 단축할 수 있다.

[0122] 본 발명에 따른 이상 탐지 방법인 SC-GCNAD는 먼저 TLCC를 적용하여 유의미한 상관관계를 산출 및 불필요한 노이즈를 제거할 수 있다. 또한, 모델 표현력이 뛰어난 GCN을 통해 효율적인 그래프 구조를 모델링하여 꼭 필요한 연산들만 추가하고 불필요한 연산을 사전에 차단할 수 있다. 이후, GRU와 예측 모델을 통해 시계열 데이터 이상 탐지를 진행할 수 있다. 그 결과, SC-GCNAD의 F1-Score는 최대 6.37%가 증가하고, 분석시간은 최대 86.72% 감소할 수 있다. 추가로 정상 데이터는 일정 범위 내에 존재한다는 것에 착안하여 훈련 데이터에 대해 Stride 기법을 적용 및 학습을 진행하면 중복되는 데이터의 학습을 최소화할 수 있으므로 최대 95.31%의 분석시간 단축 효과도 얻을 수 있다. SC-GCNAD는 높은 정확도와 짧은 분석시간으로 시계열 데이터의 이상 탐지가 중요한 네트워크 관리/보안 및 산업현장 등에서 활용될 수 있다.

[0124] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정

및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

[0126] 100: 이상 탐지 시스템

110: 사용자 단말 130: 이상 탐지 장치

150: 데이터베이스

210: 프로세서 230: 메모리

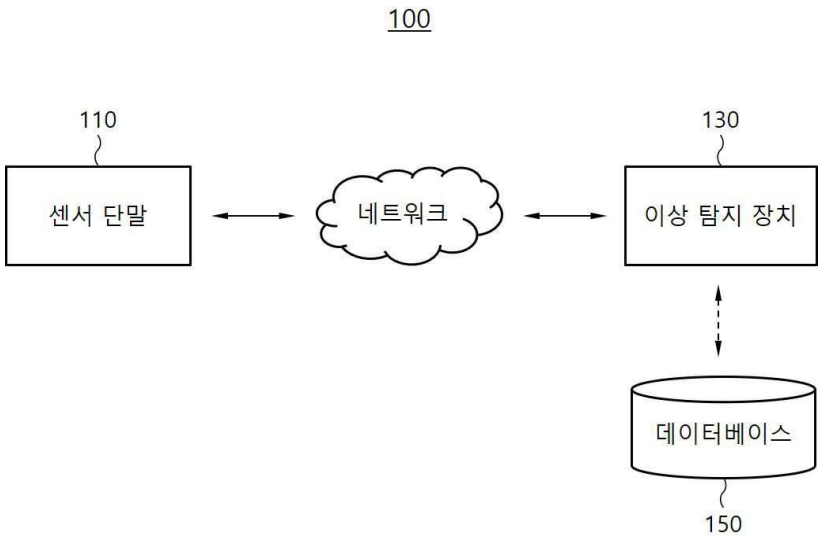
250: 사용자 입출력부 270: 네트워크 입출력부

310: 학습 데이터 수집부 330: 모델 구축부

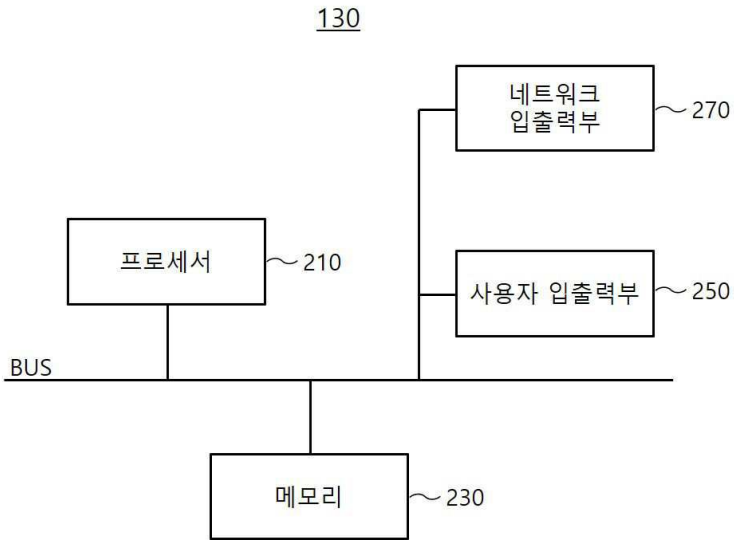
350: 이상 탐지부 370: 제어부

도면

도면1



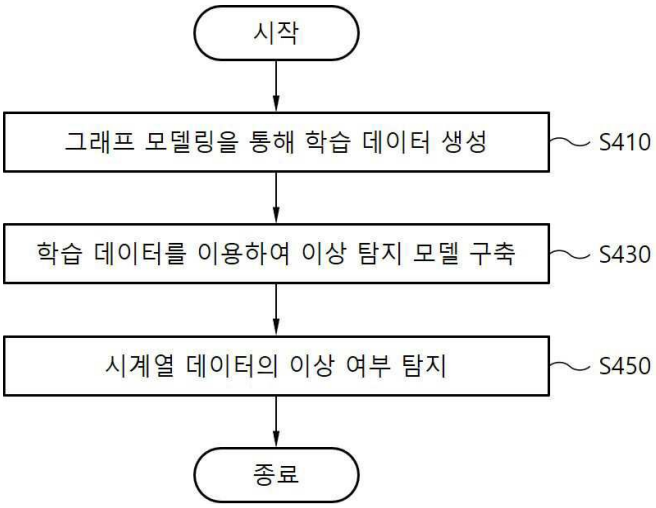
도면2



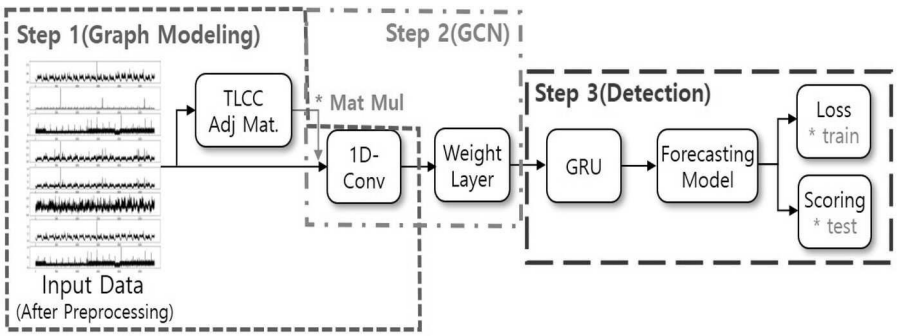
도면3


130
학습 데이터 수집부
310
모델 구축부
330
이상 탐지부
350
제어부
370

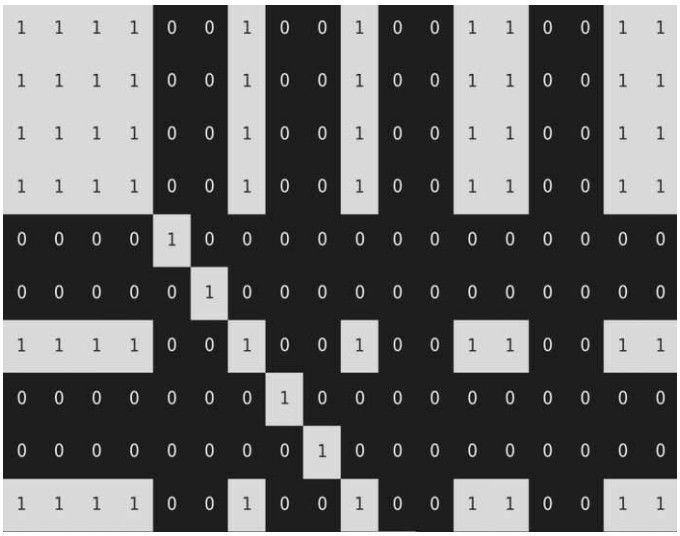
도면4


시작
그래프 모델링을 통해 학습 데이터 생성
S410
학습 데이터를 이용하여 이상 탐지 모델 구축
S430
시계열 데이터의 이상 여부 탐지
S450
종료

도면5


Step 1(Graph Modeling)
Step 2(GCN)
Step 3(Detection)
TLCC
Adj Mat.
* Mat Mul
1D-
Conv
Weight
Layer
GRU
Forecasting
Model
Loss
* train
Scoring
* test
Input Data
(After Preprocessing)

도면6



도면7

E_{train} t: # of time steps for computing threshold

Time Step	1	2	3	4	...	t
Anomaly Score	0.2	0.67	7.64	0.08	...	0.1

$\Rightarrow \mu(E_{train}) = 0.98, \quad \sigma(E_{train}) = 3.06$

Determining ϵ_{best} *using Eq.* (5) *and* (6)

Z	2.5	3	3.5	...	11	11.5
ϵ	8.63	10.16	11.69	...	34.64	36.17
$f(\epsilon)$	0.73	0.88	1.01	...	2.31	2.32

$\Rightarrow \epsilon_{best} = 36.17$

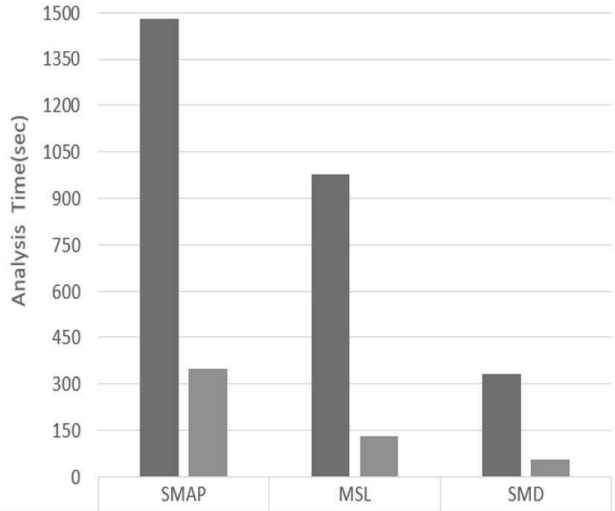
E_{test} a: # of time steps to detect anomaly

Time Step	1	2	3	4	5	...	n	n+1	n+2	...	a
Anomaly Score	0.15	0.55	1.91	48.9	30.7	...	37	89.8	89.4	...	3.16
Detection Result	N	N	N	AN	N	...	AN	AN	AN	...	N

N: Normal / AN: Anomaly

도면8

Contents	Datasets		
	SMAP	MSL	SMD
Number of Sensors	25	55	38
Training set size	135,183	58,317	708,405
Testing set size	427,617	73,729	708,420
Anomaly Rate(%)	13.13	10.27	4.16



	SMAP	MSL	SMD
MTAD-GAT[9]	1480.988	978.249	332.996
SC-GCNAD	350.974	129.946	57.412