|  |  |  |  |
| --- | --- | --- | --- |
| **이상치 탐지 기준 선정 (Reference)**  제출자 : 숭실대학교 AI 데이터 연구단 박준혁 |  |  |  |

텍스트이(가) 표시된 사진

자동 생성된 설명

[Minimum volume ellipsoid classification model for contamination event detection in water distribution systems (sciencedirectassets.com)](https://pdf.sciencedirectassets.com/271872/1-s2.0-S1364815214X00046/1-s2.0-S1364815214000899/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEHkaCXVzLWVhc3QtMSJGMEQCIAYEkPI04pi%2F9ZNrtBA6hwPJCARIWTVzcvJ4%2FGxZ4hkEAiAaBkJ340wEZj9nfz3%2Fz%2FP%2FJVe5x3lM40bWLzCcJVv5FyqDBAiC%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F8BEAQaDDA1OTAwMzU0Njg2NSIMYNoMd4k7MWCy1wE0KtcDEo4p9okncHlIGI9zDotbfzc28FDUDuepOjN4%2B2FycBEjoXmFTQ4phO5eEX7ZvllrGNgqDIZMJSx%2F4GFIzK5C%2BSRuronXL9oEo6S0IMe5bi9n%2BaSZAXp6uMzWC5jEzHvBSD8tzw1nbSdL%2FfjNj9daNa2FASst%2BIYuHCZk7YClcFZa9xsYUMdxMCvDBdD2CeSoM3oeegidg234eXZ%2Bk9O74eZgSwwjwdSo1LE3PxUiNsp6OenGhkJOgnqkm0MYusjZSBL2%2BTmiJGirFrlnDktdtnoWfV2a0w2O6AH%2FrjKWYcR7e7wvXKUErKbRSxEtGhpOKTGSF%2FtoYfqmTSJq40OB7q8VQcYRN%2FSpyZyBf4lMumINBs%2BERUktmZuj%2Fled3gLG6psYc2HGYB55PuwWzHh9MRPYaS3w07etVzfEeEHVC1x8sIsPKw8E9oJauYLhAGYq49Y33tQdAx1PFg2Gu9LMyhFvmg98oydCJootH97ESHnD1v78%2BptScTBti1OisbjooAtaf3K4LJdYTSUuD4SV9E%2BstAv0KDThGw6ppK66%2BejiraPEha%2BnSi9RD190NGSyVI5RkdeQey%2FMbWjLbucX8KRXDPzNbNgLwx8VSb4%2FC7EHcPywSsN5MKTT044GOqYB3GS%2BfWYmxP9XgVWI3CSRZZVkBPVyyoPxiCQQp%2BvgacWaEvDt7kshClS580G50PHHjri5RTAxUWgXz4Qn3oHggHOl1smuQxgspQRrBLXOih4PyJnsx6uUiMh%2Bs%2BD9TFDE1uUqyfko8KRXFBHY1HqJI97WzSKt9xic5g9PFdAQSl7ZYW6HgxCExAW49pUX9LSUnLaZqPzla09d5tk1UePPo4JdipP9sg%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20220105T014223Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYUGIURTOV%2F20220105%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=f7aba1897e79035de5d0fd34897125ed4aed0d9ee6436a13578be983ca12b6f6&hash=a824dbba493f66ecc02ea12a19fc4f82352add0386c70a40ad1e3bd920f163d3&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1364815214000899&tid=spdf-8b60ca9d-319d-4951-9e9b-f8569ccf4995&sid=53d8af832d98d043d81b9e35fa730ff19d94gxrqa&type=client)

**Abstract**

본 논문은 물 분배 시스템 상에서 수질 오염 징후를 감지하는 이벤트 탐지 시스템의 특성을 이야기하고자 함. 개발된 모델은 이상치를 측정하는 최소 부피 타원체(MBE) 분류기와 상황 분류를 위해 MBE 이진 출력을 사용하는 차후 시계열 분석으로 이루어져 있다. 본 모델은 지속적으로 업데이트 되는 연속적인 데이터를 사용하며, MVE는 수질에 관한 파라미터에 대한 자동적인 분석을 가능하게 한다. 다변량 분석은 수질 파라미터 간의 관계를 탐색하고 고통 패턴의 변화를 탐지한다. 수질 오염 상황에 대한 어떠한 가정 없이 모델을 구축하면 모델의 신뢰성과 일반성 부문에서 효과를 입증할 수 있다.

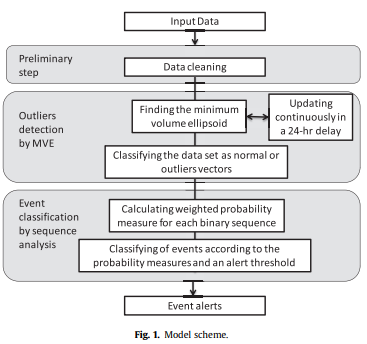
**Objectives**

제안한 연구의 목적은 : (1) 수질 시계열 데이터에 대한 다변량 분석의 적용. 다차원 분석은 시스템에 대한 여러 관점을 제공하고, 매개변수와 패턴변화 간의 관계를 밝힌다. (2) 비지도 학습 분류 방법의 적용. (3) 오염 및 패턴 변화 탐지 모델의 정확성과 민감도 개선.

모델은 두 가지의 모듈 요소로 이루어져 있다. 이상치를 탐지하는 MVE 분류기와, 이벤트 분류를 위해 MVE 이진 출력을 사용하는 차후 시계열 분석이다.

MVE는 다차원 데이터의 이상치 탐지를 위해 사용된다. 이 분류 방법은 주어진 벡터 그룹을 포함하는 최소 닫힌 사각형 표면을 찾는다. 타원의 차원은 벡터의 차원과 동일하다. 대부분의 사례에서, 타원은 주어진 벡터의 분할 집합을 포함해야 한다. 분할은 측정의 신뢰도에 따라 결정된다. 타원이 결정된 후에는 어떤 새로운 관측치가 들어오더라도 정상 혹은 이상치로 분류할 수 있다.

시간 도메인 관측치의 경우, 데이터 베이스가 시간이 갈수록 점점 커진다. 따라서 타원은 시간에 따라더 많은 벡터들을 포함하게 된다.



**Model Development**

모델의 입력 데이터는 시계열 수질 데이터로 이루어져 있다. 데이터 정제는 본격적인 분류의 전제 작업으로, 데이터 속 결측치와 노이즈를 제거하는 단계이다. 대부분의 관측치는 조금의 노이즈를 가지고 있다. 이러한 노이즈는 분류기 구축에 있어 유의미한 영향을 미친다. 이러한 경우, 노이즈의 존재는 타원체를 과도하게 확장시켜 분류기 민감도를 손상시킬 수 있다. 그러므로 분석 전에 데이터를 필터링 하는 것은 필수적으로 보인다. 본 모델에서는 평균에서 표준편차\*4 만큼 떨어지거나 양수가 아닌 값들은 제거하는 간단한 데이터 정제 방법을 사용하였다.

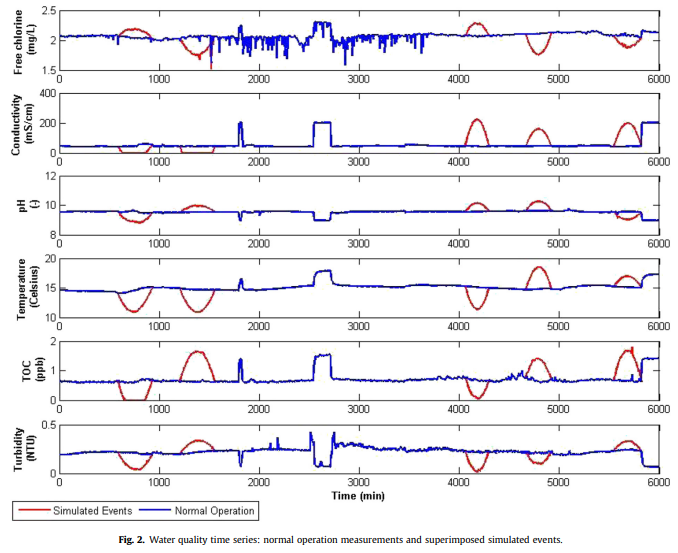
분류기는 알려진 데이터 셋 내에 있는 벡터의 95%를 포함하는 최소 타원체를 찾음으로써 구축된다. 타원체에서 경계를 이루는데 필요한 벡터 집합의 분율은 분류기의 민감도와 특이도의 상대적인 중요도를 결정한다. 분율이 높을 경우, 얻은 타원체가 더 커지고 둔감한 분류기가 된다. 반대로 분율을 줄이면, 타원이 작아져서 더 많은 이상치를 감지하게 된다. 하지만 많은 False Alarm을 만들어 낸다. 95% 분율이 일반적으로 사용되는 값이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

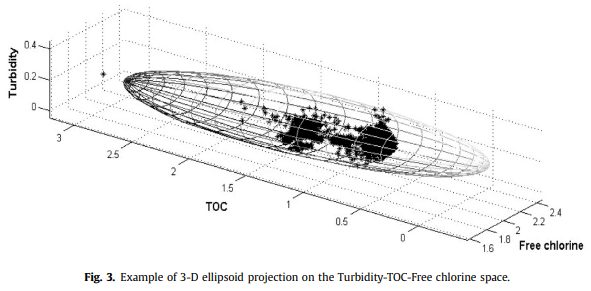
자동 생성된 설명



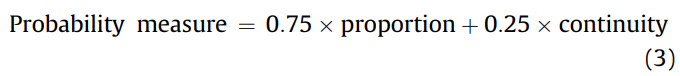
알고리즘은 모든 벡터를 포함하는 큰 부피의 타원체, 즉 실현 가능한 솔루션에서 시작한다. 반복적으로, 타원체에 포함된 샘플의 분율이 사전 정의된 값에 맞는지 검사한다. 알고리즘은 Polynoimal 범위 내 반복 횟수에서 수렴한다. 앞서 언급했듯이, 타원체의 형성은 오로지 정상 데이터만을 필요로 한다.

시계열 수질 데이터는 시간이 지날수록 점점 데이터셋의 크기가 증가한다. 연구의 작동 가정은 24시간 후 측정의 본질이 명확해진다는 것이다. 그리고 정상 작동 조건으로 판별된 것은 타원체 구성에 사용될 수 있다. 그러므로, 데이터 베이스가 커지면서 24시간 딜레이로 측정된 데이터가 더해지며 점차 타원체가 점진적으로 재구성된다.

타원체 파라미터가 결정되고 나면, 새로 들어오는 관측 데이터가 정상 혹은 이상치로 분류된다. MVE 분류기의 출력은 매 시간 단위 측정치 마다 정상일 경우 0, 이상치일 경우 1로 분류되는 이진 시계열 데이터이다.

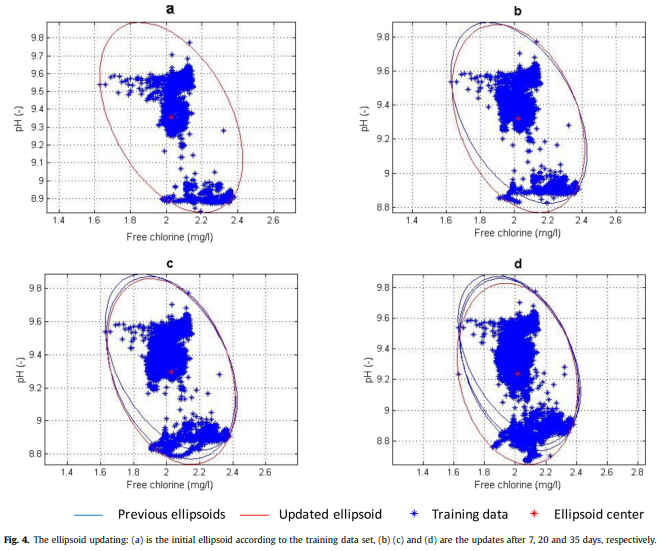


MVE 분류기의 이진 출력은 정상 혹은 이상치 분류 평가를 포함한다. 일련의 특이치가 하나의 사건보다 확실히 더 강력한 증거이기 때문에 특이치와 사건의 분류 사이에는 차이가 있다. (이상치가 더 강력함.) 따라서, 매 시간 단계의 분류는 MBE 이진 출력에서 나온 세그먼트의 시계열 분석에 기초한다. 매 시간 단계의 분류를 위해 분석된 시퀀스의 길이는 6으로, 5분마다 측정된 측정의 30분 지속 시간에 해당한다. 시퀀스 세그먼트는 두 가지 요소로 구성된 확률적 추정으로 평가된다. ‘Proportion’, ‘Continuity’.



여기서 비율은 세그먼트 외부 이상치의 비율이고 연속성은 해당 세그먼트의 연속성이다. 마지막은 세그먼트의 길이에 상대적인 분율로, 시퀀스 내에서 가장 긴 이상치 시퀀스로 표현된다. ‘Proportion’은 다른 측정값 중 시퀀스의 예외를 나타낸다. ‘Continuity’는 이상치의 신뢰성을 나타낸다. 여기서 특이치의 순차적 시퀀스는 조각난 것보다 강한 사건의 지표이다. 위 두가지 요소는 시퀀스가 이벤트의 발생을 나타낼 확률을 반영해서 하나로 합쳐진다.

확률 추정 값을 계산한 후 임계값과 비교한다. 만약 임계치를 넘으면 이벤트로 분류하고, 그렇지 않으면 정상으로 분류한다.



예를 들어, 분석된 시퀀스 세그먼트 값이 [0,1,1,1,0,1] 이라면, ‘proportion’값은 4/6이고, ‘continuity’값은 3/6으로 총 ‘probability measure’값은 0.625 (4/6 \* 0.75 + 3/6 \* 0.25)이다. 본 pm값을 Threshold 값과 비교하여 이벤트 or 정상으로 분류한다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트, 음악, 스크린샷, 피아노이(가) 표시된 사진

자동 생성된 설명