

# 다변수 시계열 데이터 군집화 알고리즘의 사이버 물리 시스템 오브 시스템즈 실패 분석을 위한 연구 조사\*

김한수<sup>o</sup> 현상원 배두환  
 한국과학기술원 전산학부  
 {hansukim, swwhyun, bae}@se.kaist.ac.kr

## A Survey on Multivariate Time-Series Clustering Techniques for Analyzing Cyber-Physical System-of-Systems Failures

Hansu Kim<sup>o</sup> Sangwon Hyun Doo-Hwan Bae  
 School of Computing, KAIST

### 요 약

사이버 물리 시스템 오브 시스템즈(Cyber-Physical System-of-Systems, CPSoS)는 독립적으로 작동하는 여러 사이버 물리 시스템(Cyber-Physical System)들로 이루어진 크고 복잡한 시스템이다. 이러한 복잡 시스템의 실패 분석을 위해서는 각 구성 시스템에서 기록되는 관측 변수에 대한 포괄적 분석이 필요하다. 이러한 분석을 위해서 사용할 수 있는 기법이 다변수 시계열 데이터(Multivariate Time-Series, MTS) 군집화 알고리즘이다. MTS 군집화 알고리즘은 다양한 연구 분야에서 복잡한 시계열 데이터를 분석하고 공통되는 패턴을 추출하기 위해 많이 사용되어왔다. 본 연구에서는 시스템 데이터를 활용하는 MTS 군집화 알고리즘들의 CPSoS 실패 분석에 대한 적용 가능성을 중점으로 분석하였다. 하지만 MTS 군집화 알고리즘 연구 조사에서 기술된 알고리즘들이 CPSoS의 주요 특징 중 Dynamic Reconfiguration에 대한 고려가 충분히 되지 않았음을 확인하였고, 해당 특징을 고려한 동적 차원 대상 MTS 군집화 기법의 필요성을 분석하였다.

### 1. 서 론

사이버 물리 시스템 오브 시스템즈(Cyber-Physical System-of-Systems, 이하 CPSoS)는 여러 사이버-물리 시스템(Cyber-Physical System, 이하 CPS)으로 이루어진 복잡 시스템이다[1]. CPSoS는 개별 시스템이 달성하기 힘든 상위수준의 목표를 구성 시스템(Constituent System, 이하 CS) 간의 상호작용을 통하여 달성할 수 있다. 이러한 CPSoS의 대표적인 예시로는 차량 군집 주행 시스템(Platooning system)과 스마트 시티 등이 있다.

CPSoS의 각 CS는 주변 환경을 관측하고, 다른 CS와 상호작용하며 의사결정을 하게 된다. 이 과정에서 환경적 변수, 네트워크 지연 등을 포함한 불확실성으로 인해 생기는 시스템의 실패가 발생할 수 있다. 하지만 크고 복잡한 시스템인 CPSoS에서 모든 실패 사례에 대한 매뉴얼한 분석을 하는 것은 큰 비용이 필요하다. 따라서 CPSoS의 실패에 대한 효과적인 분석을 위한 기법이 요구된다.

CPSoS와 같은 복잡 시스템의 실패 분석 방법으로 기존의 연구들은 시계열 데이터 군집화(Time-series clustering)를 통한 방법을 많이 사용해왔다[2-3]. 시계열 데이터 군집화는 시계열 데이터에서 유용한 정보를 추출하는 데 사용되는 방법으로, 데이터 셋에 있는 여러 시계열 데이터에서 자주 나타나는 패턴을 추출하는 것을 목표로 한다[4]. 이렇게 추출된 패턴은 실패 예측(Failure prediction), 실패 원인 분석(Fault localization) 등의 다양한 분석에 사용될 수 있다.

CPSoS에서 생기는 다양한 실패 사례는 단일 CS의 실패뿐만

아니라 여러 CS의 상호작용에 의한 실패를 포함하여 다양한 원인에 의해 발생한다[5]. 이러한 복잡 시스템의 실패 사례의 분석을 위해 단일 변수로 이루어진 시계열 데이터를 따로 분석하는 것이 아닌 각각의 CS에서 관측된 변수를 모두 통합한 다중 변수 시계열 데이터(Multivariate Time-series, 이하 MTS)를 이용해 군집화된 데이터로부터 knowledge를 추출하여 제공할 수 있다. 따라서 본 연구에서는 MTS 군집화 알고리즘에 관련된 여러 연구와 알고리즘이 적용된 대상 데이터를 조사하여 해당 알고리즘이 CPSoS의 여러 특징에 부합하여 적용 가능한지를 분석하였다.

본 논문의 이후 구성은 다음과 같다. 2장에서 CPSoS의 주요 특징과 해당 특징에 관련된 실패 사례를 서술한다. 3장에서는 조사된 MTS 군집화 알고리즘과 대상 데이터를 서술하고, 마지막으로 4장에서는 조사 결과를 바탕으로 결론 및 향후 연구를 기술한다.

### 2. CPSoS의 주요 특징

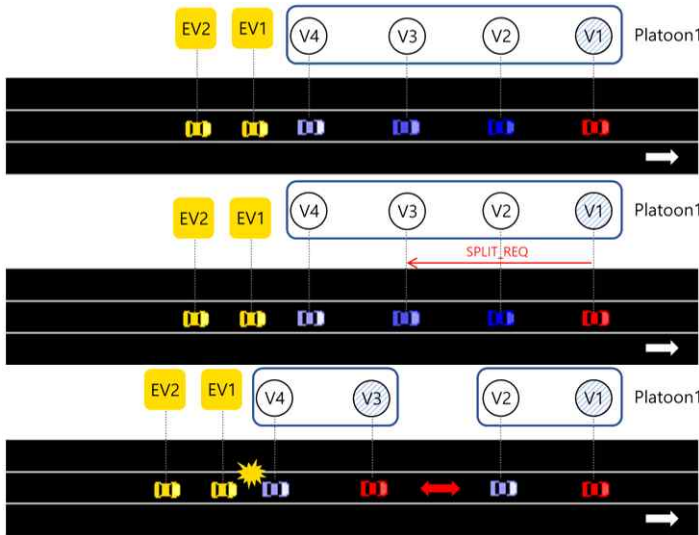
Nielson의 [6]의 연구 조사에서 서술된 시스템 오브 시스템즈(System-of-Systems, 이하 SoS)의 8가지 특징 중 실패 분석과정에서 가장 큰 영향을 끼치는 두 가지 특징으로는 Diversity와 Dynamic reconfiguration이 있다. Diversity의 경우 서로 다른(Heterogeneous) 시스템이 SoS를 구성할 수 있음을 말한다. 예를 들면, 차량 군집 주행 시스템에서 군집을 이루는 차량의 종류와 각 차량이 가진 센서의 수, 그리고 다른 형태의 자율 주행 알고리즘을 가질 수 있다. Dynamic reconfiguration의 경우 SoS의 구조와 조합이 변화할 수 있음을 말한다. 이는 SoS를 구성하던 CS의 추가 혹은 이탈할 수 있음을 의미한다. 위의 군집 주행 예시에서, 함께 군집을 이뤄 주행 중에 차량의 상태와 목표에 따라 새로운 차량이나 군집이 추가되기도 하고, 군집 주행 중이던 차량이 군집에서 이탈할 수 있다.

\* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2015-0-00250, (SW 스타랩) 모델 기반의 초대형 복잡 시스템 분석 및 검증 SW 개발)과 대학ICT연구센터지원사업의 연구결과로 수행되었음. (IITP-2022-2020-0-01795)

[표 1] MTS clustering관련 연구 및 CPSoS 실패 분석에 대한 적용 가능성 분석

First author	Target domain	Similarity calculation method	Diversity	Dynamic reconfiguration
P. D' Urso[7]	Financial data	DTW	-	-
D. Hallac[8]	Automobile data	TICC	-	-
B. Sürmeli[9]	Smart factory	HMM	X	X
V. Kapp[10]	Smart manufacturing system	TICC	X	X
G. Liu[2]	Photovoltaic array system	DTW	X	X
T. Schmidt[3]	ALFA	DTW	X	X
	SWaT	DTW	O	X

특히, CPSoS에서의 Dynamic reconfiguration은 각 CS의 동작 메커니즘에 변화를 주게 된다. 차량 군집 주행 시스템 예시에서의 경우, 차량 간의 의사소통을 통해 차량 사이의 거리를 짧게 유지함으로써 연료 효율성을 증가시킨다. 따라서 군집에서 차가 이탈하는 경우 이탈하는 차와 나머지 군집 사이의 거리를 다시 증가시킬 필요가 있다. 이렇게 구성 시스템의 동작이 변화하는 경우에도 시스템이 실패 없이 작동하는지 확인하기 위해 Dynamic reconfiguration을 고려해 주어야 한다.



[그림 1] Dynamic Reconfiguration 과정에서의 충돌

위 그림은 차량 군집 주행 시스템에서 Dynamic reconfiguration 과정에서 발생하는 실패인 차량 간의 충돌사태를 보여주고 있다. 이 차량 군집 시스템은 서로 다른 제조사에서 만들어진 차량인 V1-V4로 이루어져 있으며, 해당 차량들은 다양한 종류의 자율 주행 알고리즘을 가지는 것을 가정한다. 그림의 상단에 있는 초기 상태에는 V1-V4 차량이 하나의 군집을 이루고 있고, 빗금 친 V1의 경우 Leader 차량이며, 나머지 세 차량 V2-V4는 Follower 차량이다. 나머지 두 노란색의 차량 EV1과 EV2의 경우 군집의 차량과 의사소통을 하지 않는 환경 차량을 나타낸다.

군집 주행을 하던 4대의 차량에서 V3 차량이 SPLIT 명령에 따라 V1을 Leader로 하는 군집과 V3를 Leader로 하는 두 군집으로 분할된다. 두 그룹 사이의 거리를 증가시키기 위해 뒤쪽

그룹인 V3와 V4는 속도를 감소하게 된다. 그 결과로 V4 차량이 뒤따라오던 환경 차량 EV1과 충돌하게 되는 예시이다.

위 사례는 Dynamic reconfiguration에 의한 실패 사례로, 충돌이 일어난 V4 차량에서 관측된 변수만으로는 충돌 원인에 대한 정보를 충분히 포함하고 있지 않다. 하지만 군집 주행하던 4대의 차량의 데이터의 경우 V1과 V3 사이의 의사소통을 포함한 모든 정보를 포함하고 있다. 따라서 해당 사례에서 나타난 충돌을 분석하기 위해서는 V4뿐만 아니라, 충돌 당시에 군집 시스템을 함께 이루고 있던 V3, 그리고 SPLIT 명령 이전에 군집을 이루고 있던 V1과 V2 역시 고려해 주어야 한다.

하지만 서로 다른 제조사에서 만들어진 차량은 서로 다른 센서의 종류와 수를 가질 수 있다. 또한, 군집 주행 시스템을 이루는 차량의 수가 동작 중에 변하게 되면, 실험 데이터 역시 그에 따라 변수의 수가 시간에 따라 변하게 되고, 그 결과, 동적 차원을 가진 MTS가 생성된다. 따라서 Heterogenous한 시스템과 Dynamic reconfiguration을 가진 CPSoS의 주행 기록 데이터 셋은 서로 다른 수와 타입의 변수를 가질 수 있을 뿐만 아니라, 각 데이터는 시간에 따라 차원 역시 변할 수 있다. 따라서 이러한 MTS 데이터 셋을 군집화하는 방법이 요구된다.

### 3. 조사 방법 및 결과

MTS 군집화에 대한 조사는 다음 조건에 따라 이루어졌다. CPSoS의 경우 생성된 MTS의 길이가 다를 수 있기 때문에 서로 다른 길이의 MTS를 비교할 수 있어야 한다. 또한 이 조사의 목적에 따라 분석 대상 데이터에 대한 설명이 상세하게 기술되어 있어야 하며 되도록 많은 CPS혹은 SoS를 대상으로 분석한 연구에 초점을 맞추었다. 특히, 시스템을 대상으로 한 연구의 경우 CPSoS의 적용 가능성을 확인하는 데 필요한 SoS의 특징인 Diversity와 Dynamic reconfiguration이 고려되었다. 또한 MTS간의 유사도를 계산하는 Similarity method 역시 여러 알고리즘이 포함되도록 하였다.

위 표는 다양한 데이터를 대상으로 MTS를 군집화하는 방법을 연구한 논문들을 보여준다. MTS 군집화 알고리즘의 대상 데이터로는 금융 데이터를 포함한 여러 실생활 데이터와 CPS 데이터, Smart factory, Smart manufacturing system과 같은 시스템들을 대상으로 하는 연구들이 포함되었다. 하지만 실제 해당 기법들의 적용 과정을 조사해본 결과, Smart factory, Smart manufacturing system과 같은 복잡 시스템을 대상으로 하였는 데도 불구하고, 해당 시스템들이 Heterogeneous한 시스템 모듈로 구성되어있음을 적용한 연구는 제한되어 있었다. 또한, 조사

를 진행한 연구 중 CPSoS에서의 Dynamic reconfiguration 특성을 고려한 연구는 발견하지 못하였다.

P. D'Urso와 [7]의 연구는 stock market과 같은 다변수 금융 데이터를 군집화하기 위해서 동적 시간 워핑(Dynamic Time Warping, 이하 DTW) 거리를 바탕으로 한 퍼지(Fuzzy) 군집 알고리즘을 이용하였다. 해당 연구에서는 Trimming이라는 과정을 통해 군집화의 정확성과 이상치 군집(Outlying cluster)과 독립적 이상치(Isolated outlier)를 찾아낼 수 있도록 하였다. 이후 이상치를 제외한 MTS를 DTW를 이용하여 군집의 중심점(medroid)과의 거리를 계산하고, 이를 바탕으로 군집화를 진행하게 된다.

D. Hallac와 [8]의 연구는 자동차의 주행 데이터를 대상으로 군집화 알고리즘을 제안하였다. 이 연구에서는 Toeplitz Inverse Covariance-based Clustering(이하 TICC) 기법을 제안 하였다. 주어진 MTS를 정해진 window 사이즈에 따라 MTS를 작은 크기로 자르고, 분할된 MTS를 상관관계 네트워크(Markov Random Field)에 따라 군집화하는 방법을 소개하고 있다.

B. Sürmeli와 [9]의 연구는 레고를 통해 구현된 모형화 된 스마트 공장과 실제 산업에서 사용되는 사이버-물리 시스템을 대상으로 하여 MTS 군집화 알고리즘을 제안하였다. 해당 연구에서는 생성된 MTS를 이동 평균(Moving average)을 사용하여 노이즈를 줄이는 전처리를 한 뒤, 주성분 분석(Pinciple Component Analysis)를 통해 차원을 줄이고, 전체 MTS 데이터셋을 바탕으로 학습된 모델을 통해 이산화 과정을 거친다. 이후 이산화된 MTS를 바탕으로 Hidden Markov Model(HMM)을 생성하고, 이에 따라 군집화하게 된다.

V. Kapp와 [10]의 연구는 이전의 연구와 비슷하게 스마트 제조 시스템을 대상으로 하여 MTS의 군집화와 패턴 인식 알고리즘을 적용한 연구이다. 이 연구에서는 이전에 기술한 TICC 알고리즘을 바탕으로 한 방법과 시계열 데이터에 Sliding window, Top-down, Bottom-up과 같은 시계열 분할(Time-series Segmentation)에서 보편적인 분할 방법을 적용한 뒤, TICC와 같이 상관관계 네트워크를 통해 군집화하는 두 가지 방법을 적용하였다.

G. Liu와 [2]의 연구에서는 여러 태양광 모듈이 결합된 Multi-agent 시스템인 태양광 발전 시스템을 대상으로 시계열 데이터 군집화에 대해 기술하였다. 해당 연구에서는 DTW를 바탕으로 한 거리 계산과 K-means 알고리즘을 결합하여 시계열 데이터 K-means 알고리즘을 제안한다.

T. Schmidt와 [3]의 연구는 두 가지 대상 시스템에 대하여 이루어졌다. 첫 번째는 AirLab Failure and Anomaly(ALFA) 데이터 셋으로 무인항공기의 주행 데이터이다. 두 번째는 SWAT(Secure Water Treatment System)으로 iTrust 연구소에서 제작한 수처리 시스템 테스트베드이다. 이 SWAT 테스트베드는 수처리에 필요한 6가지의 단계를 Multi-agent 시스템으로 구축하고 있다. 해당 연구에서는 주어진 MTS 데이터를 일정한 크기로 자르고 z-정규화 과정을 거친 뒤, DTW를 이용하여 데이터 간의 거리를 계산한다. 이후 PCA를 이용하여 차원을 줄이고, K-means 알고리즘을 이용하여 군집화하는 방법을 소개하고 있다.

조사된 6개의 논문은 MTS를 군집화하기 위해서 서로 다른 방법을 사용하였다. 하지만, 군집화에 필요한 필수적인 요소인 두 시계열 데이터의 거리를 계산하는 방법으로는 DTW와 TICC, 그리고 HMM을 이용한 알고리즘으로 나눌 수 있었다. 이 기법들의 경우, 이론적으로 CPSoS의 특징인 Diversity를 고려할 수 있는 것으로 보이지만, 실제 연구 중 Heterogeneous 한 구성 시스템 혹은 시스템 모듈로 구성된 대상 시스템을 활용한 연구는 제한되었다. 또한 세 방법 모두 기법의 특성상 서로 다른 차원의 시계열 데이터 사이의 거리 계산은 불가능하였다.

따라서 CPSoS에서 동작 중에 CS의 구성에 변화가 있는 시나리오로 만들어진 데이터 셋이 존재할 때뿐만 아니라 두 데이터가 서로 다른 수의 변수를 가진 경우 두 가지 데이터의 거리를 계산하는 방법이 아직 존재하지 않았다. 이를 통해 앞서 서술한 두 가지 주요 특징을 가진 CPSoS의 실패 분석을 위한 MTS 군집화 방법이 필요함을 확인하였다.

#### 4. 결론

CPS와 같은 시스템의 실패 검출 기법으로 MTS 군집화 방법이 많이 연구되어왔다. 하지만 일부 CPSoS와 같이 동작 중에 구성 시스템이 추가되거나 이탈하여 MTS의 차원이 시간이 지남에 따라 변경되는 경우 적용할 수 있는 군집화 방법이 요구된다. 본 조사에서는 Multi-agent 시스템을 포함한 다양한 데이터 셋에 적용된 MTS 군집화 방법을 찾아 보았고, 기존에 다양한 분야에서 사용된 MTS 군집화 알고리즘에서 두 MTS사이의 유사도를 비교하는 방법이 Dynamic reconfiguration에 적용하기는 불가능함을 확인하였다.

따라서 이후 연구에서는 MTS의 차원이 시간에 지남에 따라 변화할 경우 두 MTS 사이의 유사도를 계산해 주는 방법에 관한 연구가 필요하고, 이를 바탕으로 CPSoS의 실패 분석 방법을 고안할 필요가 있다.

#### 7. 참고문헌

- [1] Engell, Sebastian, et al. "Core research and innovation areas in cyber-physical systems of systems." International Workshop on Design, Modeling, and Evaluation of Cyber Physical Systems. Springer, Cham, 2015.
- [2] Liu, Guangyu, et al. "Time series clustering and physical implication for photovoltaic array systems with unknown working conditions." Solar Energy 180 (2019): 401-411.
- [3] Schmidt, Tabea, Florian Hauer, and Alexander Pretschner. "Automated Anomaly Detection in CPS Log Files." International Conference on Computer Safety, Reliability, and Security. Springer, Cham, 2020.
- [4] Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering—a decade review." Information Systems 53 (2015): 16-38.
- [5] Hyun, Sangwon, et al. "Timed Pattern-Based Analysis of Failure-Inducing Interactions in Platooning System-of-Systems." Available at SSRN 4057041.
- [6] Nielsen, Claus Ballegaard, et al. "Systems of systems engineering: basic concepts, model-based techniques, and research directions." ACM Computing Surveys (CSUR) 48.2 (2015): 1-41.
- [7] D'Urso, Pierpaolo, Livia De Giovanni, and Riccardo Massari. "Trimmed fuzzy clustering of financial time series based on dynamic time warping." Annals of operations research 299.1 (2021): 1379-1395.
- [8] Hallac, David, et al. "Toeplitz inverse covariance-based clustering of multivariate time series data." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.
- [9] Sürmeli, Barış Gün, and M. Borahan Tümer. "Multivariate time series clustering and its application in industrial systems." Cybernetics and Systems 51.3 (2020): 315-334.
- [10] Kapp, Vadim, et al. "Pattern recognition in multivariate time series: towards an automated event detection method for smart manufacturing systems." Journal of Manufacturing and Materials Processing 4.3 (2020): 88.