한국CAD/CAM학회 논문집 Vol. 17, No. 2, pp. 97-110. April 2012
Transactions of the Society of CAD/CAM Engineers DOI http://dx.doi.org/10.7315/CADCAM.2012.097

〈응용논문〉 ISSN 1226-0606

시스템 결함원인분석을 위한 데이터 로그 전처리 기법 연구

이양지[1] · 김덕영[1†] · 황민순[2] · 정영수[2]
[1]울산과학기술대학교 디자인 및 인간공학부, [2]현대중공업 통신운영부

A Study on Data Pre-filtering Methods for Fault Diagnosis

Yang Ji Lee[1], Duck Young Kim[1†], Min Soon Hwang[2], and Young Soo Cheong[2]
[1]School of Design and Human Engineering, UNIST.
[2]Communication Network Management Dept., Hyundai Heavy Industries

Received 9 February 2012 ; received in revised from 27 February 2012; accepted 28 February 2012

ABSTRACT

High performance sensors and modern data logging technology with real-time telemetry facilitate system fault diagnosis in a very precise manner. Fault detection, isolation and identification in fault diagnosis systems are typical steps to analyze the root cause of failures. This systematic failure analysis provides not only useful clues to rectify the abnormal behaviors of a system, but also key information to redesign the current system for retrofit. The main barriers to effective failure analysis are: (i) the gathered data (event) logs are too large in general, and further (ii) they usually contain noise and redundant data that make precise analysis difficult. This paper therefore applies suitable pre-processing techniques to data reduction and feature extraction, and then converts the reduced data log into a new format of event sequence information. Finally the event sequence information is decoded to investigate the correlation between specific event patterns and various system faults. The efficiency of the developed pre-filtering procedure is examined with a terminal box data log of a marine diesel engine.

Key Words: Data log, Fault diagmosis, Marine diesel engine, Pre-filtering

1. 서 론

최근 시스템이 더욱더 복잡해지고 빠르게 변하였으며, 이러한 시스템의 신뢰성과 안정성을 유지하는 일이 무엇보다도 중요한 일로 대두되고 있다. 이를 위해서는 그 시스템에서 발생하는 고장에 대해서 보다 빠르게 대응할 수 있어야 하며, 고장으로 인한 손실을 최대한 줄여야 할 필요가 있다[1].

†Corresponding Author, dykim@unist.ac.kr

이를 위한 고장 진단 시스템은 크게 모델을 기반으로 하는 방법과 모델을 기반으로 하지 않고 수집된 데이터를 분석하는 방법으로 분류할 수 있다[2]. 수학적 모델의 해석을 중심으로 하는 모델기반의 방법에는 상태추정법(State estimation)과 파라미터 추정법(Parameter estimation) 등이 있으며, 모델기반이 아닌 수집된 데이터를 이용하여 분석하는 방법에는 한계치 검사 기법(Limit measurement method), 전문가 시스템(Expert system) 그리고 신경망 응용 방식이 일반적으로 사용되고 있다. 이런 다양한 종류의 고장 진단 시스템은 일반적으로

시스템의 결함 유무를 판별하는 결함검출(Fault detection), 검출된 결함이 어떤 종류인지를 판단하는 결함분리(Fault isolation) 그리고 결함의 Size 및 Type 등을 판별하는 결함평가(Fault identification)의 단계적 프로세스로 구성되어 있다. 이 세가지 단계를 통해서 시스템의 고장이 발생하였을 경우, 신속하게 결함을 파악하여 격리조치하고, 그 결함의 종류, 특성 및 경중을 평가하여, 추후 시스템의 원활한 운용(Operation) 및 재설계(Redesign)를 위한 중요한 정보를 제공하게 된다.

현재 자동차, 철도차량, 비행기 그리고 선박 시스템에 대한 고장 분석을 위한 시스템이 지속적으로 개발되고 있다. 선박의 경우 안정성을 평가하기 위해 실제 해상상태에서 손상선박의 거동을 정확하게 예측할 뿐만 아니라 파랑 중에서의 선박의 구조적인 결함여부를 나타내는 구조안정성 평가 등을 수행할 수 있는 많은 시뮬레이션 기반 시스템이 구축되어 있다[3]. 또한 선박의 사고사례나 관련 데이터들을 이용하여 정의된 시나리오 및 변화하는 환경조건에 반응하는 선박의 고유 특성에 근거하여 안정성을 평가하는 많은 연구가 진행되고 있다[4].

그러나 이런 구조적 결함을 분석하기 위한 시뮬레이션 기반 외에 선박 엔진의 이상 유무를 판단할 수 있는 시스템을 탑재한 선박은 거의 전무하며, 단순하게 고장인지 아닌지의 판별과 엔진 자체의 고장을 판단하는 시스템은 개발되어 있으나 고장의 전조 증상을 미리 판단한다든지 인공 지능적으로 고장 확률, 부품의 교체 시기 등을 알려줄 수 있는 시스템은 발견하기 힘들다[5]. 고장 유무를 판별하는 기존의 고장시스템에 고장에 대한 조기 대응이나 예방정비계획을 수립할 수 있도록 하는 시스템을 추가하여 선박의 해상사고를 미연에 방지할 수 있다.

이와 같이 선박 엔진의 정확한 진단을 위해서는 고장을 진단하기 이전에 수집된 데이터에 대한 정확한 이해가 필요하다. 최근에는 고성능 센서 및 Data logging 기술을 적용하여 시스템 운용 상태를 실시간으로 저장 및 모니터링 할 수 있다. 예를 들면, 자동차의 경우 OBD(On Board Diagnostics)를 통해 실시간으로 데이터를 저장 하고 있으며, 자동차뿐만 아니라 철도차량, 비행기, 그리고 선박 등에서도 주행 운항 상태 데이터를 실시간으로 수집/관리하고 있다. 하지만 데이터 저장 장치를 통해 수집된 방대한 양의 데이터를 바로 고장 분석을 위해 사용하기에는 어려움이 있다. 일반적으로 선박 엔진 데이터뿐만 아니라 모든 시스템에서 실시간으로 수집된 데이터는 단기간이 아닌 적어도 1개월 길게는 수년 동안 수집된 데이터이기 때문에 데이터의 양이 방대하며, 각 시스템을 구성하는 모든 센서에서 수집된 데이터이기 때문에 각 데이터의 단위 및 표준크기가 다양하다. 그리고 실시간으로 들어오는 데이터가 데이터 저장 시스템에 저장되는 형태로 변환(Transformation)될 때 결측데이터(Missing value)를 포함하며, 주변 환경의 영향을 많이 받기 때문에 잡음(Noise) 및 중복정보(Redundancy)를 포함하고 있다. 즉 실측데이터는 그 자체를 바로 활용하는데 어려움이 있으며, 데이터의 불완전(Incomplete), 잡음(Noise) 그리고 불일치(Inconsistent)를 해결하기 위해 실제 데이터에서 필요한 정보만을 추출하는 전처리 과정이 요구된다.

따라서 본 연구에서는 기존의 선박 엔진 고장 진단시스템의 효율적인 고장 분석을 위해 시스템에 수집된 센서 데이터를 이벤트 로그 정보로 축소/변환하는 프로토콜을 개발하여, 시스템의 데이터 로그에서 중요한 정보만을 안전하게 추출하기 위한 전처리 기법을 제안한다.

2. 관련 연구

2.1 선박 엔진 고장 진단 시스템

IT 기술의 급격한 발전에 따라 선박을 운용하기 위한 각종 시스템들이 점차 자동화되어 운용중인 장비의 데이터를 실시간으로 감시할 수 있게 되었으며, 다양한 고장진단기법에 관한 연구가 수행되었다. 기존의 선박 감시 시스템은 미리 정해놓은 설정치와 비교해서 이에 도달하거나 혹은 미치지 못한 경우 경보를 발생하는 한계치 검사 기법이 대부분이다[6]. 즉 정상상태에서의 모델과 실시간으로 측정되는 모델을 비교하며, 고장을 검출하는 이중화 비교방식과 통계적 분석을 통해 부하들의 관계를 분석하는 방법을 이용한다.

2.1.1 이중화 비교 방식에 의한 고장 진단

대부분의 선박 엔진 고장 진단 시스템은 미리 정의해 놓은 정상상태의 데이터와 실시간으로 들어오는 데이터를 비교하여 고장을 감지하는 기법

이다. 즉 전문가가 미리 각 센서데이터에 대한 정상상태(Normal state) 범위를 정해놓고, 실시간으로 데이터를 수집하면서 데이터가 정상상태(Normal state)의 범위를 벗어났을 경우 알람을 발생한다. 이중화 비교 방식을 이용하는 데이터는 대부분 진동 센서의 데이터나 실시간으로 입력되는 Signal의 패턴을 이용한다.

진동 센서의 데이터를 이용하는 연구에는 정상 데이터를 먼저 수집하고 엔진 고유 진동 주파수를 기준으로 하여 각 채널로부터 획득되는 엔진 진동 주파수를 비교하여 엔진의 고장 유무를 판별하는 기법[7], 선박엔진 구조물의 각 부위에 진동을 측정하고 수신되는 진동신호를 고속푸리에변환(FFT)된 진동데이터로 변환시켜 주파수 영역에서 선박엔진의 고장을 감지하는 기법이 있다. 즉 선박엔진에서 감지된 진동데이터 값이 진동데이터 임계값의 일정 범위에 포함되면 정상신호로 처리하고, 임계값의 일정 범위에 포함되지 않으면 비정상신호로 처리한다.

Signal 패턴을 이용하는 방식은 엔진의 고장을 나타낼 수 있는 파형이 있다고 가정하고, 엔진 고장을 알리는 실측 데이터가 입력되었을 때 두 파형을 비교하는 방식이다[7]. 간단하게 두 곡선 파형으로부터 샘플링 간격을 아주 작게 하여 샘플링된 지점의 주파수 값을 각각 대조하여 이상 Signal을 추출한다.

2.1.2 통계적 방식을 이용한 고장 진단

통계적 방식을 이용한 고장 진단 시스템도 이전 단락에서 설명된 전문가 시스템을 통한 비교방식의 형태를 이용한다. 숙련된 전문가의 지식을 활용하여 감시데이터의 상호 연관성을 검토하며, 검토된 감시데이터 항목 사이의 관계를 통계적 분석 기법을 이용하여 정량화하여 이상 데이터와 비교하여 고장을 분석한다.

선박 엔진 시스템을 구성하는 다양한 부하의 변동에 따라 계측항목별 상관관계를 검토해 상관관계가 높은 항목과 낮은 항목으로 분류할 수 있다. 정상상태에 계측항목은 부하와 높은 상관관계를 가지고, 고장이 발생하였을 경우 기계의 특성이 정상이 아니기 때문에 상관관계는 낮게 된다[2]. 이러한 점을 이용하여 정상상태에서 부하의 상태에 따른 상관계수(Correlation coefficient)의 크기를 측정하여 고장을 분석한다.

또한 신경회로망을 이용하여 고장을 분석하는 방법은 데이터의 정상적인 것과 정량적인 것을 동시에 처리하는 방법으로 부하와의 상대적인 중요도를 이용한다. 데이터를 신경회로망을 통하지 않고 구분하게 되면 그 구분은 단순히 현재 그 데이터의 상태만을 나타낼 뿐 다른 데이터와의 연관성은 없게 된다. 신경회로망을 이용하여 데이터의 학습을 통해 다른 데이터와의 상대적 중요도를 측정하여 데이터 값의 범위 이상으로 나오는 데이터를 이상 데이터로 측정한다. 또한 엔진은 어느 한 데이터가 표준 운전범위를 벗어났다고 하여 계통고장이라고 단정할 수 없으므로 표준운전범위를 벗어난 데이터를 중심으로 상호 영향을 미치는 관련 데이터를 체계적으로 조사하여 계통고장가능성을 진단하는 것이 이상적이다[8].

2.2 데이터 전처리

선박 엔진 데이터뿐만 아니라 대부분의 시스템을 구성하는 센서 데이터는 그 양이 매우 방대하여 이를 저장하기가 매우 어렵고, 동시에 외부환경으로 인해 발생되는 잡음(Noise), 장기간 동안 같은 데이터가 반복적으로 저장되는 중복정보(Redundancy), 그리고 변환(Transformation)될 때 발생하는 결측데이터(Missing data)를 처리하기 위해서는 매우 많은 시간이 소요된다. 시간적인 측면과 저장 공간의 효율성을 위해 대용량의 고차원 원본 데이터를 특정 이벤트 로그로 변환하여 데이터를 축소하는 다양한 연구가 진행되고 있으며, 크게 조건부 Table을 이용한 Clustering 의 방법과 데이터의 Signal을 이용하여 데이터의 중요한 특징만을 추출하는 방식이 있다.

2.2.1 데이터 축소를 위한 Clustering 알고리즘

Clustering 방법은 주어진 데이터를 의미 있는 집단(Subgroup)들로 분류하며, 데이터 분석, 시각화, 압축 및 전처리와 관련된 많은 분야에서 널리 이용되고 있다. Clustering을 통해 데이터는 비슷한 항목들마다 별개의 Cluster를 형성하며, 이 Cluster를 형성하기 위한 척도는 주로 데이터간의 유사도(Similarity) 또는 거리(Distance)를 이용하여 형성된다[9].

실제 시스템에서는 주로 전문가가 데이터의 특징, 즉 선박 엔진 시스템일 경우 각 선박에 수집되는 센서 데이터의 정상범위를 정해놓거나 선박에

고장이 발생하였을 경우의 센서 데이터 값을 미리 조건부 Table로 정의해 놓는다. 예를 들어 선박의 고장이 발생하였을 경우의 센서 데이터 값의 조건부 Table을 이용한다면, 선박 운용 중에 새로 들어오는 데이터를 미리 정의되어 있는 조건부 Table과 비교하여 고장이 발생하였을 경우의 센서 데이터 값과 유사한 데이터만을 받아들여 데이터의 차원을 축소한다. 즉 이벤트를 특정 센서 노드에서 수집된 데이터가 사용자가 미리 제시한 조건들을 만족시키는 상황으로 정의하고, 시스템 운용 중 수집되는 Sensor 테이블과 센서 네트워크 밖에서 만들어진 정적 테이블간의 조인 연산을 통해 이벤트를 검출한다[10-12]. 이와 유사한 방법으로 전문가가 미리 시스템의 조건부 Table을 정의하는 것이 아니라 과거에 수집된 많은 양의 데이터를 이용하여 고장에 관련된 Episode를 만들어놓고, 실제 들어오고 있는 패턴 중 이와 가장 유사한 Episode만을 추출하는 Episode fragment 방식이 있다[13].

이와 같이 고차원의 데이터를 직접 바로 사용하는 것이 아닌 고장 패턴과 유사한 데이터만 Clustering 하거나 정상상태(Normal state)를 벗어난 데이터만 Clustering하여 사용함으로써 데이터를 저장하는 공간을 줄일 수 있을 뿐만 아니라 고장을 분석하는 데이터 처리 과정의 시간적 낭비를 줄일 수 있다.

2.2.2 Signal data를 이용한 데이터 특징 추출 기법

실시간 모니터링 시스템에서는 일반적으로 실시간으로 들어오는 방대한 양의 데이터 Signal을 관찰하면서 임계치(Threshold)를 넘는 값만 저장하여 데이터의 양을 축소하고 통신 비용을 줄이며, 데이터 저장용량의 효율성을 높인다.

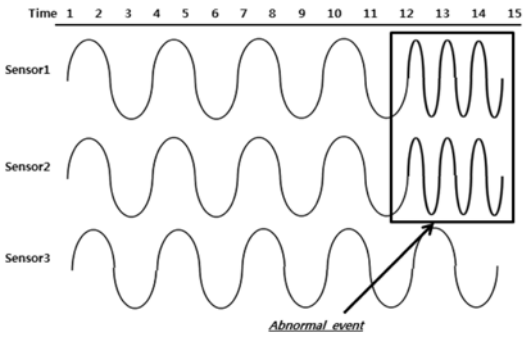


Fig. 1 Real-time data acquisition[14]

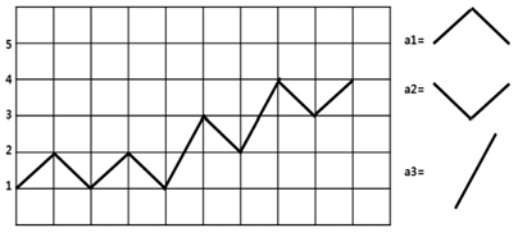


Fig. 2 Rule discovery from time series[15]

Fig. 1과 같이 실시간으로 측정되는 다양한 종류의 센서에서 각 센서의 모든 데이터를 사용하지 않고, 실시간으로 들어오는 Signal을 모니터링 하면서 Time(12-15) 구간과 같이 이상 구간만을 이벤트로 추출하여 저장하는 방식을 사용하여 데이터 처리 시간과 저장 공간의 소비를 줄인다[14].

이와 같이 Signal의 실시간 모니터링을 통해 특정 Signal만 추출하는 방법 외에도 Signal의 패턴에 따라 Signal을 특정 이벤트 형태로 변환하는 Rule discovery 방법이 있다. 이 방법은 먼저 실시간으로 들어오는 Signal을 사용자가 설정한 Size 크기의 Subsequence로 나누게 되면, 다양한 형태의 Signal로 분류된다. 분류된 서로 다른 패턴을 가지는 Signal을 각각 다른 이벤트 로그로 정의한다. 즉 형태가 비슷한 Signal들을 Clustering하여 특정 이벤트 로그로 변환하는 방법으로 Time series 형태의 Signal을 Discrete한 Data 형태로 변환할 수 있도록 해준다.

예를 들어 Fig. 2와 같이 연속적인 Signal이 입력될 때 Signal은 아래와 같이 Discrete한 Data의 형태로 표현 가능하다.

[Original time series = (1,2,1,2,1,2,3,2,3,4,3,4)]

Window size가 3일 경우 원본 데이터를 크기가 3인 Sub-data로 Sampling 한다. 이를 통해 서로 다른 Signal 패턴을 가지며, Fig. 2의 a1, a2, a3과 같이 특정 Signal의 패턴 ‘a1’, ‘a2’, ‘a3’를 얻는다. 즉 Window size로 나누어진 Signal은 이벤트 로그 ‘a1’, ‘a2’, ‘a3’로 정의된다. 이 특정 이벤트를 미리 정해두고, 실시간으로 들어오는 Signal과 정의된 이벤트 Signal과의 유사성을 측정하여 Signal의 형태가 같은 것끼리 Clustering 한다. 위와 같은 Original time Series data를 이벤트화하면 아래와 같은 이벤트 로그로 변환된다.

[Event log = (a1, a2, a1, a2, a3, a1, a2, a3, a1, a2)]

위와 같은 효율적인 고차원 데이터의 전처리에 관한 많은 연구를 기반으로 본 연구에서는 선박 엔진 시스템을 중심으로 시스템 고장분석의 효율성을 높이고자 시스템의 저장장치에 수집된 다양한 종류의 센서 데이터를 중요한 이벤트 로그 정보로 변환하는 프로토콜을 개발하여 데이터를 축소할 수 있는 다양한 전처리 방법을 제안한다.

3. 데이터 전처리의 필요성

고장 분석을 위한 데이터 수집에 있어서 고려해야 할 중요 사항은 크게 Sensor, Measurement, Reading의 3가지 관점에서 살펴볼 수 있다[13].

3.1 Sensor

시스템의 고장 분석을 위해 수집되는 데이터는 시스템을 구성하고 있는 다양한 센서로부터 수집된 데이터이다. 이러한 센서 데이터들은 한 종류의 통일된 데이터가 아닌 다양한 종류로 구성되어 있다. 자동차의 경우 자동차 엔진의 고장을 분석하기 위해서는 공기 유량 센서, 온도 센서, 산소 센서, TPS(Throttle Position Sensor), CKP(Crankshaft Position Sensor), CMP(Camshaft Position Sensor), 배터리 전원 등의 여러 종류의 센서 데이터를 이용해야 하며, 선박 엔진의 경우에도 온도 센서, 압력 센서, 회전속도 그리고 가속도 센서 등의 다양한 종류의 센서 데이터를 이용해야 한다. 이렇게 다양한 종류의 센서로부터 수집된 데이터들은 센서의 종류에 따라 단위가 모두 다르고, 각 센서의 정상상태(Normal state)의 범위도 다양하다. 즉 동시간에 다양한 종류의 센서로부터 수집되는 데이터는 기준이 없이 수집된 데이터이기 때문에 센서 각각의 기준을 모두 비교하기 위한 시간과 메모리의 소비가 크다. 시간적이고 공간적인 효율성을 위해서는 동시간에 수집된 정상상태(Normal state)의 범위가 다른 데이터를 한꺼번에 처리하기 위한 기준과 각 센서 데이터에서 상관관계를 통해 중요한 정보만을 추출해 낼 수 있는 방법이 요구된다.

3.2 Measurement

실제 연속적으로 측정되는 센서 데이터는 고장 진단을 위해서 Numerical value로 변환되어야 한다. 보통 시스템에 발생하는 고장은 단기의 데이터 이상으로 인해 일어나는 것이 아니라 장기간에 걸친 여러 데이터들의 상관관계나 다른 조건과의 영향으로 일어나는 경우가 대부분이며, 고장 분석을 위해 수집되는 데이터는 짧게는 1개월 길게는 수년에 걸쳐 운용된 데이터이기 때문에 데이터의 양은 매우 방대할 것이다. 또한 실제 측정되는 데이터는 매우 좁은 시간 간격으로 측정되기 때문에 각 데이터의 변동 범위의 차이가 매우 작은 경우가 많으며 중복 정보(Redundancy)의 발생도 빈번하다. 이러한 변동 범위의 차이가 적거나 중복 정보(Redundancy)를 직접 사용할 수도 있지만 방대한 양의 데이터일 경우는 데이터 처리의 시간적 측면에서 비효율적이다.

이와 같이 데이터를 처리하지 않고 직접 고장 분석에 사용한다면 데이터를 검색하는 시간이 많이 들고, 데이터의 저장 공간 측면에서도 비효율적인 문제가 발생한다. 이러한 문제로 인해 데이터의 양을 적게 수집하여 사용하면 효율적일 수 있겠지만 고장진단의 정확도적인 측면에서는 원본 데이터 양이 많을수록 정확도가 높기 때문에 기존의 많은 양의 원본 데이터를 유지하고 이 데이터를 활용하여 고장 분석에 사용할 수 있도록 접근해야 한다.

3.3 Reading

실제 운용되는 시스템에서 데이터 저장 장치로 옮겨지게 될 때 데이터가 정확하게 Transform이 되었다고 할 수는 없다. 데이터가 전송되는 과정에 외부 환경의 영향으로 인해 잡음(Noise)에 의한 측정오류의 형태로서 값의 왜곡이 일어날 수도 있고, 데이터의 누락 및 비정상적인 값이 저장될 수도 있다. 이러한 정확하지 않은 데이터를 이용하여 고장 진단에 직접 이용하게 된다면, 고장 진단의 결과에 대한 신뢰도가 떨어지게 될 것이며, 시스템의 고장 진단의 목적인 원활한 운용을 위한 정확한 피드백을 할 수 없게 된다. 즉 위와 같이 센서의 작동 실패, 각 센서에 대한 데이터 기입 표기 문제, 데이터 전송문제, 기술적인 속성문제 그리고 속성값의 부정확성과 같은 원인으로 데이터의 잡음(Noise)이 발생한다. 정확하고 신뢰성이 중요시되는 고장 진단 시스템을 위해서는 정확한 형태의 원본 데이터가 가장 기본이다. 그렇기 때문에 데이터 전처리 과정을 통해 일관성 있는 통합된 데이터 형태로의 변환이 필요하다. 데이터 변환 시 축소된 데이터는 원래 데이터와 같은 분석

결과를 얻을 수 있어야 하며, 데이터 Computing 시간을 고려하고, 방대한 로그 데이터의 경우 시간 단위로의 데이터 축소 과정을 거쳐야 한다.

이러한 이유로 본 논문에서는 Raw 데이터를 특정 이벤트 로그로 변환하기 위해 Simple mean 형태로 Binning하는 방법과 수치값을 속성값으로 변환하여 데이터를 이산화(Discretization) 방법을 제시하여, 선박엔진 데이터뿐만 아니라 시스템으로 수집된 데이터를 이용한 고장 분석의 효율성과 정확성을 높일 수 있도록 한다.

4. 이벤트 로그 정보 추출

4.1 이벤트 로그 정보 추출

선박 엔진의 경우 순간적인 기계고장의 경우보다는 서서히 기계적인 소모에 의해서 고장이 발생하는 경우가 더 빈번하다. 따라서 고장을 분석하기 위해서는 순간의 데이터의 분석으로 끝나는 것이 아니라 고장이 발생하기 전까지의 데이터의 경향(Trend)를 분석해야 한다. 이를 위해 장기간의 데이터가 수집되어야 하므로 데이터의 양이 상당하다.

선박으로부터 장기간에 걸쳐 시스템에 수집된 데이터는 Table 1과 같이, 다양한 종류의 센서(m Sensors)와 데이터 수집속도(Hz) 및 수집 기간에 따른 센서 측정값 들의 Array 형태로 구성되어 있으며, 수집된 데이터는 모두 다양한 범위(Range)를 가지고 있으며, 정상상태(Normal state) 범위(Range)도 모두 다르다. 또한 순간의 사고로 인한 선박의 고장이 아닌 서서히 고장이 발생하는 경우 고장이 발생하기 이전의 수집된 데이터들은 각각 다른 데이터 값과 데이터 경향(Trend)을 가지고 있기 때문에 각 고장에 대한 원인을 파악하기 위해서는 수집된 데이터 전체를 이용해야 할 필요가 있다.

이렇듯 일정하지 않은 방대한 양의 데이터들을 직접 시스템의 고장 분석에 사용하는 것은 데이터를 분석하는 시간적인 측면과 데이터를 저장하기 위한 공간적인 측면에서 효율성이 떨어진다. 실제 수집된 Raw 데이터를 변환하는 방법에는 데이터로부터 잡음(Noise)를 제거하기 위해 데이터 추세에 벗어나는 데이터를 추세에 맞게 변환하는 방법인 ‘Smoothing’, 특정 구간에 분포하는 값으로 스케일을 변환시키는 ‘Generalization’, 최소값이나 최대값 또는 Z-score을 통한 ‘Normalization’ 그리고 데이터 통합을 위해 새로운 속성이나 특징을 만들어 주어진 여러 데이터 분포를 대표할 수 있는 특징을 활용하는 ‘Feature construction’ 방법이 있다.

선박 엔진 데이터의 경우 실시간으로 선박에서 수집되는 데이터 중 전문가 시스템에 의해 미리 정해진 고장 패턴과 매칭하여 관련 데이터만 수집하여 데이터베이스 용량을 감소시킨다[7]. 본 연구에서는 시스템의 고장 분석의 정확성을 높이기 위해 선박 엔진의 모든 센서 데이터로부터 수집된 센서 데이터를 이용하여 고장을 분석한다. 수집된 모든 데이터를 이용하기 위해서는 기존의 raw 데이터를 다른 형태로 변환하는 과정이 필요하다. 이를 위해 PCA(Principal Component Analysis), Auto/Cross-Correlation, Entropy기반 Discretization을 통해 데이터를 분석하고 축소된 데이터를 Simple

Table 1 Raw data format of engine sensors

Time	Sensor1	Sensor2	Sensor3	…	Sensor(m)
1	5.2	10	0.2	…	1
2	2.5	21.2	0.4	…	3.2
…	Failure				
1002	2	35	0.25	…	5.2
1003	…	…	…	…	…
…	…	…	…	…	…
3015	Failure				
3016	1.2	25	0.82	…	5.1
…	…	…	…	…	…
10023	Failure				
…	…	…	…	…	…

mean을 통한 Binning 방법, Entropy 방법을 통해 이벤트 시퀀스 정보로 변환하여 데이터 기반의 선박 에진 고장 분석의 효율성을 높인다.

4.1.1 PCA와 Correlation을 이용한 데이터 전처리

PCA(Principal Component Analysis)는 Raw 데이터에 많은 양의 변수가 있을 때 그 변수의 양보다 작은 수의 주성분으로 전체 변동 중 상당 부분을 설명할 수 있으며, 축소된 주성분을 통해 자료를 해석하여 Raw 데이터에 나타나지 않은 새로운 관계들을 찾을 수 있다.

Fig. 3과 같이 기존의 Raw data를 가장 잘 표현하고 있는 직교상의 데이터 벡터들을 찾아서 데이터 압축을 하고, 속성들을 선택하고 다시 조합시켜서 다른 작은 집합을 만든다. 즉 주성분으로 새로운 차원을 만들어 낸 후, 기존의 데이터를 사영하여 새로운 차원의 데이터를 만들 수 있다. PCA는 계산하는 과정이 간단하고 정렬되지 않은 속성, 빈약한 데이터나 일률적인 데이터 처리가 가능하다. 즉 정보 손실을 최소화하면서 저차원 공간에서 데이터 해석을 가능하게 한다.

시스템에 수집된 센서 데이터의 경우 PCA를 통해 시스템을 구성하는 모든 센서 중에 주성분 값과 관련된 몇 개의 주요 센서를 선별하여, 데이터의 차원을 축소하여, 시스템의 고장 분석을 용이하게 한다.

PCA를 통해 주요인 센서 데이터만을 뽑아내는 방법과 달리 Correlation 분석은 시스템을 구성하고 있는 센서간의 상관관계를 측정하여, 서로 관련이 있는 센서들을 추출해 낼 수 있다. Auto-correlation 기법은 하나의 센서에서의 상관관계를 분석한 것으로 Fig. 4와 같이 하나의 센서의 signal에서 주기성을 확인하여 상관관계를 분석하여, 센서의 Signal이 주기성을 가지고 발생하는지의 여부를 확인한다. Cross-correlation은 Fig. 5와 같이

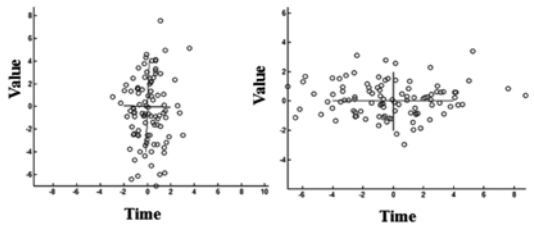


Fig. 3 Raw data vs. PCA result

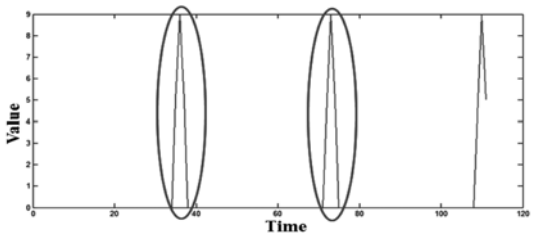


Fig. 4 Auto-correlation

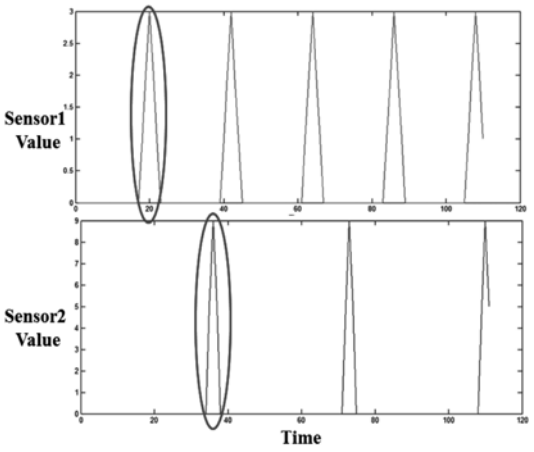


Fig. 5 Cross-correlation

자기상관관계인 Auto-correlation과 다르게 서로 다른 종류의 센서의 상관관계를 분석한다. 예를 들어 Fig. 5에서 Sensor1에서 Signal이 발생한 후에 Sensor2의 특정 Signal이 발생한다. 이와 같이 Sensor1에서 발생하는 특정 Signal이 Sensor2의 이상 Signal과 상관관계가 있는지의 여부를 확인하여, 다양한 센서들의 상관관계를 분석하여 시스템을 구성하는 센서들 중 서로 관련이 있는 센서를 파악하고, 센서 데이터 간의 상호연관 패턴을 분석한다.

4.1.2 이벤트 로그 정보 변환

다양한 값을 가지는 데이터의 효율적인 분석을 위해서 이벤트 로그 정보로의 변환이 필요하다. 이를 위해 슬라이딩 윈도우 개념을 이용하여 고정된 크기로 데이터를 묶어 이벤트화 한다. 슬라이딩 윈도우로 묶여진 데이터를 이벤트화하기 위해서는 일정한 기준이 필요하지만 Fig. 6과 같이 기존의 센서 데이터는 종류가 모두 다르고 데이터의 크기, 시스템이 정상적으로 운용될 때의 데이터의 정상상태범위(Normal state range)가 모두 다르기 때

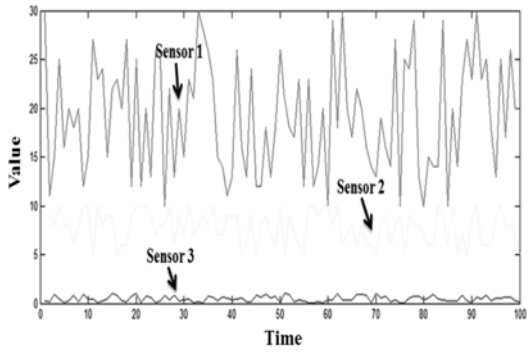


Fig. 6 Sensor signal

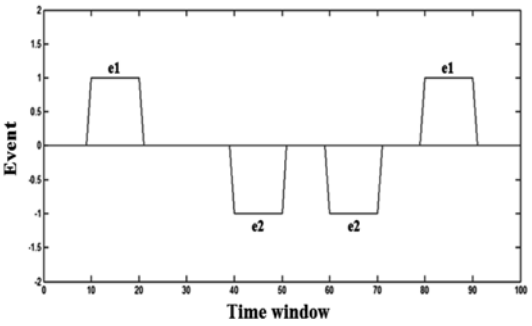


Fig. 7 Event log data

문에 모든 센서데이터들의 불일치를 해결할 방법이 필요하다. 본 연구에서는 이를 위해 다양한 범위의 데이터를 대표값으로 변환시키는 'Binning' 방법과 수치형 속성의 도메인을 여러 개의 분절된 부분으로 나누는 Entropy 기반 이산화(Discretization)을 통해 Raw 데이터를 특정 이벤트 로그로 변환한다.

'bin'이란 현재 주어진 데이터 집합에서 인스턴스가 최소한 한번 이상 나타난 좌표의 값으로 정의한다[16]. 즉 'binning'이란 Raw 데이터에서 일정한 크기의 데이터를 'Bin'이라는 공간으로 분할하는 것을 의미한다.

Binning 방법도 이산화(Discretization) 기법에 속하며, 'Equal width binning', 'Equal depth binning' 그리고 'Class dependent binning' 방법이 있다. 'Equal width binning'은 비교사, 정적, 전역적 이산화 방법으로 각각 속성의 도메인 구간을 일정한 간격, 일정한 인스턴스 수를 갖는 구간으로 분할한다[17]. 몇 개의 구간으로 변환되어야 할지는 사용자의 주관으로 결정한다. 'Equal depth binning'은 데이터에서 비슷한 것끼리 묶는 방식으로 Distance를 이용하는 방식을 주로 사용한다. 'Class dependent binning'은 Equal width로 분할된 도메인 구간에서 각 데이터의 가중치(Weight)를 이용하여 대표값으로 데이터를 변환시킨다.

위와 같은 Binning 방법 중 'Equal depth binning'의 경우 시간관계를 고려하지 않기 때문에 시스템의 운용시간에 따라 차례대로 저장되는 센서 데이터의 경우 'Equal width binning' 방법을 이용한다. 'Bin'의 수는 보통 도메인 내의 데이터 상한값(Upper bound)와 하한값(Lower bound)를 구한수 상한값과 하한값 사이의 일정 구간 k로 나눈다[17]. 그러나 기존의 'Equal width binning' 방법에 Window 개념을 도입하면 사용자가 지정한 크기의 Window에 대한 이산화가 가능하다. 실용적인 시스템의 고장 데이터 분석을 위한 이벤트 로그를 구하기 위해서는 동일한 Window 크기로 분할한다. 분할된 데이터의 대표값은 각 센서 데이터의 Mean value를 이용하여, 이벤트 로그로 변환한다. 수집된 데이터에서 Failure가 일어나기 전까지의 각 센서 데이터의 Mean value를 기준으로 실제 Raw 데이터에서의 각 센서의 시간별 데이터의 이벤트를 결정한다. Mean value를 시스템의 정상상태(Normal state)로 설정하고, 정상상태(Normal state)의 범위에 포함될 경우 '0', 정상상태(Normal state) 이상일 경우를 '이벤트1(e1)', Normal state 이하일 경우를 '이벤트2(e2)'로 변환한다면 기존의 불일치 데이터를 3가지의 이벤트로 변환할 수 있다. 즉 Fig. 6의 불안정한 불일치 데이터 Sensor1은 Fig. 7과 같이 '0', 'e1' 그리고 'e2'라는 세 가지 경우의 이벤트로 간단하게 변환할 수 있다. 이와 같이 기존의 데이터를 이벤트화된 데이터로 변환하여 이용하면 데이터의 Computing 시간이 줄어들고, 데이터의 저장공간을 줄여 보다 효율적이고 빠르게 데이터 분석을 할 수 있다. 슬라이딩 윈도의 사이즈를 크게 할수록 Raw 데이터를 큰 도메인 형태로 분할 가능하여 이벤트 로그 양이 더 축소되며, 정상상태(Normal state)의 범위를 크게 할수록 '0'에 속하는 이벤트가 많기 때문에 이벤트 로그의 양은 더 축소된다.

또 다른 이산화(Discretization) 방법 중의 하나인 엔트로피(Entropy)는 확률 정보(Probability)를 불확실성 정보(Uncertainty)로 변환해 주는 일종의 변환 정보 함수이다[18]. 즉 엔트로피는 불확실성의 척도로 높을수록 불확실한 정보이며, 엔트로피가 낮을수록 확실성은 증가하게 된다. 모든 목적속성

(Failure 여부)의 분포가 균일할수록 큰 값을, 어떤 한 목적속성의 비율이 다른 목적속성의 비율보다 높아질수록 작은 값을 갖는다. 즉 엔트로피는 어떤 속성의 값들이 목적속성 값과 연관되어 있는 정도를 측정할 수 있다.

실제 시스템을 구성하는 센서로부터 수집된 측정 데이터는 시스템의 운용상태와 주변환경의 영향에 의해 값이 근소한 범위 또는 큰 범위의 다양한 형태를 가지고 있다. 이런 데이터를 직접 사용하는 것도 가능하지만 기존의 데이터는 잡음(Noise) 뿐만 아니라 결측데이터(Missing value)와 불일치(Inconsistent)가 존재한다. 기존의 연구에서 결측데이터를 처리하기 위해서 결측데이터에 'unknown'이라는 별도의 표기를 하거나, 대표값(Mean, Variance)로 대체한다. 이와 달리 본 연구에서는 다양한 값을 가지는 데이터의 엔트로피를 이용한 이산화(Discretization) 과정을 통해 데이터 값의 범위를 축소하고, 축소된 데이터를 통해 이벤트 로그로 변환한다.

센서데이터에서 목적속성(Failure 여부)와의 관계를 이용하여 데이터의 엔트로피 값을 측정하고, 분할하는 임계치(Threshold)를 대표값으로 설정한다. 예를 들어 Table 2에서와 같이 '66.56', '70.5'와 같은 각 임계치(Threshold)에서의 엔트로피를 구해 적정할 임계치(Threshold)를 대표값으로 이용해 A-F의 5가지 이벤트로 데이터를 구별할 수 있다.

위와 같이 Entropy를 통해 대표값으로 대체된 데이터들은 Simple mean을 이용한 Binning 방식과 같이 특정 이벤트로그로의 변환이 필요하다.

Fig. 8의 기존의 데이터(a)는 (b)의 그림과 같이 연속형 범주에 속하는 수치형 데이터는 대표 속성값으로 이산화(Discretization)되어 변환된다. (b)와 같이 이산화(Discretization)된 데이터는 (b)의 그림과 같이 여러 구간으로 나누어지며, 이 구간을 서로 다른 이벤트로 정의할 수 있다. 이와 같은 경우 Simple mean을 이용한 Binning 방식에서 필요한

Table 2 Entropy based discretization

						Value								
Sensor value	64	65	68	69	70	71	72	72	75	75	80	81	83	85
Failure	1	0	1	1	1	0	1	0	1	1	0	1	1	0
			F(66.5)			E(70.5)			D(73.5)		C(77.5) B(80.5)		A(84.5)	

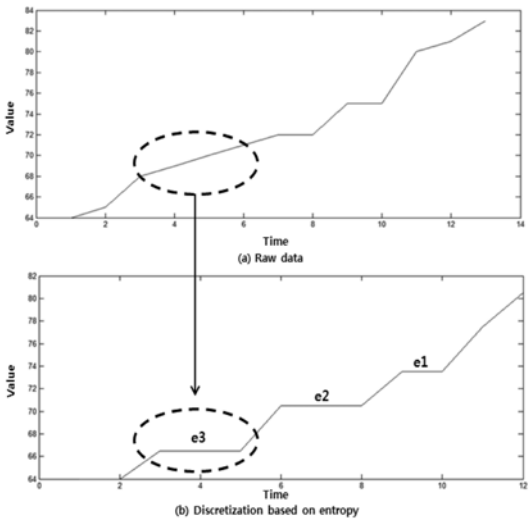


Fig. 8 Event log generation by discretization

Window size는 필요 없이 자동으로 동일한 속성값을 가지는 연속 데이터를 하나의 이벤트로 묶어진다. 이와 같이 여러 구간으로 나누어진 이벤트를 통해 특정 이벤트 로그로 변환된 데이터를 이용하여 시스템 고장 분석에 사용하여 Simple mean 기반 Binning 방법과 같이 시간적, 공간적 효율성을 높인다.

4.2 선박 엔진 데이터 분석

4.2.1 데이터 전처리

앞서 설명한 데이터 전처리 방법을 실제 테스트 중인 선박 엔진 데이터에 적용하였다. 온도센서와 습도센서 등 총 10개의 센서로 구성되어 있으며, 각 센서마다 10000개의 데이터를 포함하고 있으며, 총 10만개의 데이터를 이용하여 실험하였다.

10개의 센서 데이터를 PCA(Principal Component Analysis)를 하였을 때, 구해진 PC값 중 PC1과 PC2로 약 91% 데이터 설명이 가능하다.

Fig. 9와 같이 PCA 결과를 통해 기존의 센서 데이터에서 각기 다른 10개의 센서 데이터는 Component1(PC1)과 Component2(PC2)와의 관계를 파악할 수 있다. 'S7'의 경우 Component1(PC1)과 가장 관련 있고, 'S8'의 경우 Component2(PC2)와 가장 영향력이 크다. 10개의 센서 중 가장 영향력이 있는 센서는 'S2'와 'S9'로 주성분 Component1(PC1), Component2(PC2)에 고루 영향을 미치는 요소이다.

또한 Correlation을 분석을 통해서 실제 선박엔

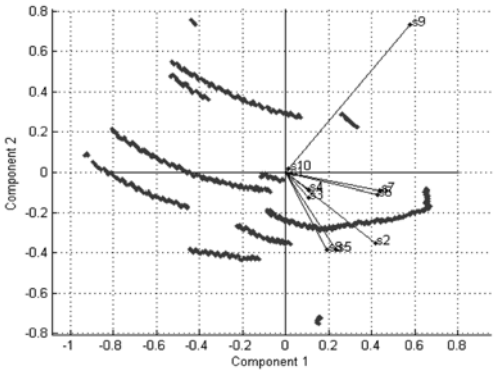


Fig. 9 PCA analysis

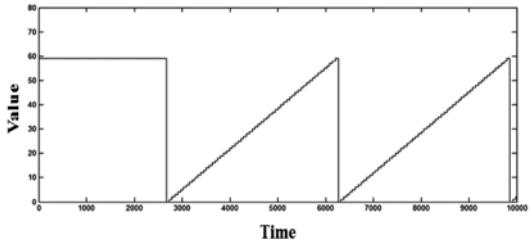


Fig. 10 'S9' (Sensor9) signal

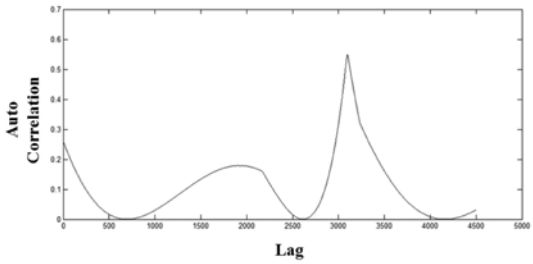


Fig. 11 S9's auto-correlation result

진 센서 데이터를 구성하는 10개의 센서들의 상관관계를 분석한다. 먼저 Auto-correlation을 이용해 각 센서의 상관관계를 분석하면 10개의 센서 중 총 4개의 센서가 주기성을 가진 상관관계가 있음을 알 수 있었다. 실험 결과 총 10개의 센서 중 Sensor 9(Engine Run Hour) 데이터가 가장 명확한 결과를 도출해 낼 수 있었으며, 기존의 Fig. 10와 같은 센서 데이터를 Auto-correlation을 실행 하였을 경우 Fig. 11과 같이 약 0.6의 상관계수를 얻을 수 있었다. Fig. 11에서 Signal의 Peak time은 3600으로 Fig. 10에서 보는 바와 같이 3600의 Time lag을 가지는 주기적인 Signal임을 알 수 있다.

Cross-correlation의 경우에는 총 10개의 센서를 이용하여 총 45개 쌍의 상관관계를 분석하였다. 그

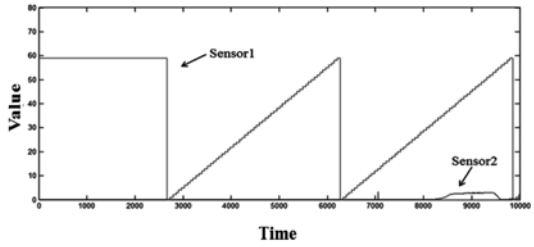


Fig. 12 'S1' (Sensor1), 'S2' signal

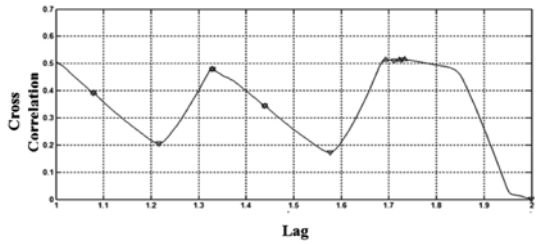


Fig. 13 Cross-correlation result between 'S1' and 'S2'

결과 총 22쌍이 상관관계가 있었으며, Sensor 1 (L.O 온도 센서)와 Sensor 9(베어링 온도 센서)가 가장 큰 상관관계를 가짐을 할 수 있다. Fig. 12와 같은 두 센서의 데이터의 Cross-correlation의 결과 Fig. 13과 같이 약 0.52의 상관계수를 얻을 수 있었으며, Time lag은 3600으로 측정되었다. 즉 Sensor1(L.O 온도 센서)와 Sensor9(베어링 온도 센서는 상관관계가 있음을 알 수 있다.

4.2.2 이벤트 로그 변환

실제 선박 엔진 데이터에서 이벤트 로그 정보로 변환하기 위해 총 56개의 센서를 이용하였으며, 시스템의 고장이 30번 발생하였을 때 수집된 데이터를 이용하였다. 수집된 데이터는 각 센서마다 41834개의 데이터로 구성되어 있으며, 전체 데이터 수는 56(센서의 수) × 41834개이다. Simple mean을 이용한 Binning 방법을 이용할 경우 각 센서 데이터의 정상상태(Normal state)를 10%의 범위로 두고, 슬라이딩 윈도우를 변경시켜가면서 이벤트 로그 정보를 추출할 때 Fig. 14와 같이 데이터의 양을 축소하였다.Window size가 클 경우 데이터 양이 축소되어 데이터를 빠르게 분석할 수 있지만 데이터 분석의 정확도 측면에서는 효율적이라고 말할 수 없다. Window size를 작게 할 경우 세분화하여 분석 가능하지만 Overfitting 현상이 발생하여, 데이터가 한쪽으로 편중되는 현상이 일어나서 이 경우 또한 정확도 측면에서 효율적이라고

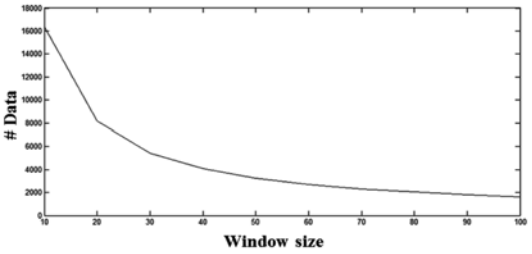


Fig. 14 Sliding window-based data reduction

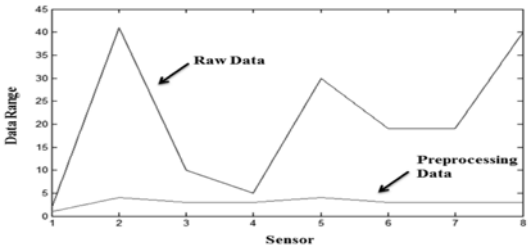


Fig. 15 Raw data discretization

할 수 없다. 따라서 적절한 슬라이딩 Window size의 결정이 효과적인 이벤트 시퀀스 분석을 결정하게 되며, 이에 대한 향후 연구가 필요하다.

또한 엔트로피를 이용한 이산화(Discritization)을 적용하였을 때 Fig. 15와 같은 결과를 얻었다. 가장 큰 변화를 보였던 센서는 Sensor2(HT. Water Outlet Temperature Sensor)로 최소값 0에서 최대값 87의 범위의 총 41개의 값으로 구성된 데이터를 4개의 범위 'A', 'B', 'C', 'D'로 이산화(Discritization)하여, 데이터를 묶어 이벤트 로그 정보로 변환하였다. 이산화(Discretization)을 통해 변환되어진 데이터를 이벤트 로그로 변환하였을 경우 Simple mean을 이용한 Binning 방법보다 이벤트가 더 세분화된다. Simple mean을 이용한 Binning의 경우 '0(Normal state)', 'e1(Normal state보다 큰 경우', 'e2(Normail state보다 작은 경우)'의 세가지 경우로 나누어 지지만 엔트로피를 이용해 이산화된 데이터를 이용할 경우 이산화된 각 구간을 이벤트로 정의할 수 있기 때문에 다양한 종류의 이벤트 로그가 발생된다.

4.2.3 패턴 분석

선박 엔진 고장의 원인을 분석하기 위해 Simple mean 기반 Binning 방법을 이용하여 축소된 데이터로 데이터마이닝 기법인 FP-Tree를 이용하여 패턴을 분석하였다[19]. FP-Tree 기법은 데이터마이닝

Table 3 FP-Tree sequence mining

	1-length	2-length	3-length
1	78-2	[78 45]-2	[78 51 45]-2
2	58-2	[51 50]-2	[53 51 50]-2
3	53-4	[60 58]-2	[62 60 58]-2
4	56-6	[62 58]-2	
5	62-2	[78 51]-2	
6	45-7	[62 60]-2	
7	50-2	[53 50]-2	
8	60-6	[53 51]-4	
9	51-21	[56 45]-6	
10		[51 45]-7	
11		[60 51]-2	

*[142 116]-12 = [Pattern]-frequency

의 Association Rule의 한 종류이며, 데이터에서의 패턴을 도출하고자 할 때 사용된다. 앞서 사용한 30번의 고장이 발생하였을 경우 56개의 센서로부터 수집된 데이터를 평균 10% 범위와 Window size 100으로 하여 이벤트 로그 정보로 변환한 데이터를 이용하였다. 임계치(Threshed)를 3으로 하였을 경우 Table 3과 같이 '패턴-sequence 수'와 같이 센서 이벤트 로그에 대한 시퀀스 패턴을 빠르게 얻을 수 있었다. Table 3에는 패턴의 길이가 1개인 1-length, 2개인 2-length 그리고 3-length까지 결과가 도출되었다. 가장 많이 Frequency를 가진 패턴은 [51 45]라는 패턴으로 총 7번이 발생하였다. 즉 '51'이라는 이벤트와 '45'라는 이벤트가 동시에 7번에 발생하였다고 해석된다. [51 45]라는 패턴을 해석하면, Window size 100으로 묶여있는 '51'이라는 이벤트와 '41'이라는 이벤트는 아래와 같은 정보를 가지고 있다.

'**51**' → 'e11e21e31e51e61e71e81e91e101e111
e121e131e141e151e161e171e181e191e201e211
e221e231e241e251e281e291e301e311e321e331
e341e351e381e391e401e411e421e431e441e451
e461e491e521e551e561'
'**45**' → 'e11e31e51e61e71e81e92e102e111e121
e131e141e151e161e171e181e191e201e211e221
e231e241e251e281e291e301e311e321e331e341
e351e381e391e401e411e421e431441e451e461
e491e521e551e561'

이벤트의 표시 ‘e’는 Event라는 표시이며, 뒤의 숫자 한자리 또는 두 자리는 Sensor 번호이고, 제일 마지막 숫자 ‘1’ 또는 ‘2’는 Normal state 범위보다 ‘크다’, ‘작다’를 의미한다. ‘1’은 Normal state 보다 큰 상태이고 ‘2’는 작은 상태이다. 즉 이와 같이 해석할 때 [51 45]라는 패턴은 이벤트 ‘51’의 정보에 속한 각 Sensor가 Normal state의 범위보다 큰 상태에 있고, 그 뒤에 ‘45’라는 정보에 속한 센서들이 Normal state 보다 큰 상태가 일어날 때 고장이 일어나는 경우가 가장 큰 확률이라고 할 수 있다. 이와 같이 전처리 과정이 이루어진 데이터를 이용하여 시스템의 고장 분석에 적용한다면, 보다 빠르게 결과를 도출해 낼 수 있다.

5. 결론 및 향후 연구

본 연구에서는 데이터 기반의 선박 엔진의 효율적인 고장 분석을 위해 방대한 양의 데이터의 전처리 과정에 대해 연구하였다. 먼저 PCA를 통해 주성분 값과 관련된 특정 센서만을 추출하였고, Auto/Cross Correlation 분석을 통하여, 시스템을 구성하고 있는 다양한 종류의 센서들간의 관계를 파악하여 센서 데이터간의 분석이 가능하다. 또한 잡음(Noise)이 포함되어 있으며, 불일치(Inconsistent)의 성질을 가지는 Raw 데이터를 특정 이벤트 로그 정보로 변환시키기 위해 Simple mean 기반의 Binning 방식과 엔트로피를 이용해 이산화(Discretization)한 데이터를 이용하는 방법을 제시하였다.

이와 같은 과정을 실제 선박 엔진의 데이터에 적용한 결과 PCA를 통해 총 10개의 센서 중 주요 센서 2가지를 추출하였고, Auto/Cross Correlation을 이용해 총 10개의 센서에 대한 상관관계를 분석함으로써 독립적인 센서들이 상관관계가 있음을 확인하여 센서간의 관계를 통해 기존 Raw 데이터의 분석을 가능하게 하였다. 또한 온도 센서, 습도 센서 등의 56종류의 센서를 이벤트 로그로 변환하기 위해 각 센서의 Mean value를 대표값으로 하여, Mean value의 몇 % 범위를 정상상태(Normal state)로 정한다. 이 범위 내의 데이터일 경우 ‘0’, 범위를 초과할 경우 ‘1’이라는 이벤트를, 미만일 경우 ‘2’라는 이벤트를 지정하면, 각 센서 데이터를 ‘e센서번호, 평균범위 초과/미만’ 이라는 이벤트로 변환되며, 기존의 Raw 데이터를 3가지 경우의 이벤트 로그로 변환이 가능하다. 변화된 이벤트를 Window라는 개념으로 고정된 Window size로 묶어 기존 데이터를 축소된 이벤트로 분석 가능하도록 하였다. 이러한 전처리 과정을 통해 축소된 데이터를 데이터마이닝의 Association rule의 FP-Tree 기법을 통해 패턴을 분석할 때 기존의 데이터를 바로 이용하는 것보다 빠른 속도로 결과를 도출해낼 수 있음을 검증하였으며, 이벤트화된 패턴을 이용하여 어떤 센서들의 관계가 시스템의 고장에 영향을 끼치는지 해석하였다. 또한 센서데이터와 목적속성(Failure 여부)의 관계를 이용하여 엔트로피를 통해 이산화(Discretization)된 데이터로 변환한다. 이 변환된 데이터를 이용하여 Simple mean 기반 Binning과 같이 이벤트로 구간을 나눈다. 엔트로피를 이용한 이벤트 로그 변환은 Simple Mean을 이용한 Binning을 이용하여 이벤트 로그로 변환하는 것과 달리 보다 다양한 종류의 이벤트 로그로 변환이 가능하였다. 같은 데이터를 이용하여 이벤트 로그로 변환하였을 경우 Simple mean 기반 Binning을 이벤트를 3가지 경우로 나누고, 엔트로피를 이용하였을 경우 이벤트를 최대 200개까지 분할이 가능하였다.

위와 같이 Simple mean 기반 Binning 방법과 엔트로피 기반 이산화(Discretization)을 통해 이벤트 로그로 변환할 경우 각 장단점이 존재한다. Simple mean 기반 Binning 방법의 경우 사용자가 직접 Window size를 결정하게 되는데, Window size가 클 경우 데이터를 아주 큰 묶음으로 Sampling하기 때문에 정확도적인 측면에서 떨어지고, Window size가 작을 경우에는 Overfitting이 발생하여 데이터가 편중되는 현상이 있을 수 있다. 적절한 사이즈의 Window를 결정하는 많은 연구가 진행되고 있지만, 현재 m-Cross validation과 같은 방법을 통해 데이터를 분석하는 중에 Test data와 Training data를 이용하여 적절한 사이즈를 결정하는 방법이 이용되고 있다. 그러나 이와 같은 경우 데이터의 양이 클 경우에는 연산을 위한 시간적 소비가 크며, 데이터를 저장하기 위한 저장공간의 소비도 크다. 이를 개선할 수 있도록 적절한 Window size를 결정할 수 있는 향후 연구가 필요하다.

이산화(Discretization) 방법에서 엔트로피 기반 방법이 활발하게 사용되고 있으며, 정확도가 다른 방법들에 비해 높다[20]. 그러나 엔트로피를 통해 이산화는 가능하지만 이산화된 데이터를 이용하

여 이벤트 로그로 변환하기 위해서는 앞으로 더 많은 연구가 필요하다. Simple mean을 이용한 Binning 방법을 적용하였을 때 3가지 이벤트로 구별될 수 있었고 데이터 마이닝 기법 적용시 빠르게 패턴을 추출해 낼 수 있었지만, 엔트로피를 이용하였을 경우 이산화된 구간마다 이벤트가 부여되기 때문에 빈번한 패턴을 추출하기에 어려움이 있다. 이를 위해 이산화 된 데이터를 이용하여 신속하게 패턴을 추출할 수 있는 방법에 대한 연구가 더 요구된다. 또한 위와 같은 데이터 전처리를 통해 축소된 데이터를 시스템의 고장 분석에 적용하여, 사용자가 쉽게 고장의 원인을 파악할 수 있는 데이터 기반 고장 진단 시스템에 관해 향후 연구할 것이다.

감사의 글

본 연구는 “울산시 및 교육과학기술부의 울산과학단지 기초·원천 R&D 과제 지원 사업”의 지원을 받아 수행된 것임. 본 연구의 일부는 한국 CAD/CAM 학회 2012년 학술발표회에서 발표되었음[22].

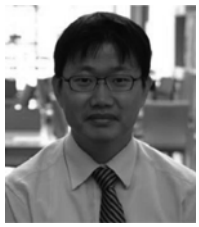
참고문헌

1. Lee, J.-H., Park, S.-W. and Seo, B.-H., 2000, Fault Diagnosis for a System Using Classified Pattern and Neural Networks, *Journal of the Korean Institute of Electrical Engineers*, 49, pp. 643-650,
2. Kim, Y.-I., Oh, H.-K. and Yu, Y.-H., 2006, The Fault Diagnosis Method of Diesel Engines Using a Statistical Analysis Method, *Journal of the Korean Society of Marine Engineering*, 30, pp. 247-252.
3. Lee, S.-S. and Lee, D.-K., 2008, Development of Integrated System for Safety Assessment of Damaged Ship, *Trans. of the SCCE*, 13, pp. 227-234.
4. Lee, K.-H., Kin, H.-S., Han, S.-W., Park, J.-H. and Oh, J., 2005, Network-based Simulation System Framework for the Safety Assessment of Ship, *Trans. of the SCCE*, 10, pp. 356-364.
5. Kim, S.-H., Kim, J.-K., Lee, D.-C. and Jang, S.-K., 2003, A Study of Torsional Vibration Monitoring System of Diesel Engine, *Journal of the Korean Society of Marine Engineering*, pp. 197-204.
6. Kim, K.-Y., Kim, Y.-I. and Yu, Y.-H., 2011, The Fault Diagnosis of Marine Diesel Engines Using Correlation Coefficient for Fault Detection, *Journal of Korea Navigation Institute*, 15, pp. 18-24.
7. Y.-M. Lee, K.-Y. Lee, S.-H. Bae, I.-S. Shin, H. Jang, J.-K. Lee, Defect Detection of Ship Engine Using Duplicated Checking of Vibration-data-distinction Method and Classification of Fault-wave, *Journal of Korean Navigation and Port Research*, 33, pp. 671-678.
8. Chun, H.-C. and Yu, Y.-H., 2007, A Data Fault Detection System for Diesel Engines Using Neural Networks, *Journal of the Korean Society of Marine Engineering*, 26, pp. 493-500.
9. Yi, S.-K., Hong, S.-E. and Park, S.-H., 2006, A Similar Price Zone Determination of Public Land Price Using a Hybrid Clustering Technique, *Journal of the Korean Geographical Society*, 41, pp. 121-135.
10. Jeon, J.-H., Yoo, J.-S. and Kim, M.-H., 2006, Path-based In-network Join Processing for Event Detection and Filtering in Sensor Networks, *Journal of the Korean Institute of Information Scientists and Engineers*, 33, pp. 620-630.
11. Yang, P. and Liu, S. S., 2006, Fault Diagnosis for Boilers in Thermal Power Plant by Data Mining. *Journal of Information and Computational Science*, 3 pp. 117-127.
12. Song, B.-H., Park, K.-W., Lee, J.-S., Lee, K.-H., Jung, M.-A. and Lee, S.-R., 2010, Efficient Processing of Multidimensional Sensor Stream Data in Digital Marine Vessel, *Journal of Communication and Network*, 35 pp. 794-800.
13. Rabatel, J., Bringay, S. and Poncelet, P., 2009, SO_MAD: SensOr Mining for Anomaly Detection in Railway Data, *In Proceedings of ICDM*, 5633, pp. 191-205.
14. Kim, J.-I., Kim, D.-I., Song, M.-J., Han, D.-Y. and Hwang, B.-H., 2010, Discovering Temporal Relation Considering the Weight of Events in Multidimensional Stream Data Environment, *Journal of the Korea Contents Association*, 10, pp. 99-110.
15. Das, G., Lin, K.-I., Mannila, H., Renganathan, G. and Smyth, P., 1998, Rule Discovery from Time Series, *Knowledge Discovery and Data Mining*, pp. 16-22.
16. Elomaa, T. and Rousu, J., 1999, General and Efficient Multisplitting of Numerical Attributes, *Machine Learning*, 36, pp. 201-244.
17. Lee, S.-H., 2003, *A Density-based Approach to Discretizing Numeric Attributes*, pp. 83-88.
18. Jun, B.-H. and Kim, J.-H., 1995, Introduction to Relative Entropy for the Discretization of

Continuous-Valued Attributes, *Journal of the institute of Electronics Engineers of Korea*, 18, pp. 1210-1213.
19. Han, J., Pei, J., Yin, Y. and Mao, R., 2004, Mining Frequent Pattern without Candidate Generation: A Frequent-Pattern Tree Approach, *Data and Knowledge Discovery*, 8, pp. 53-81.
20. Mugan, J. and Truemper, K., 2007, Discretization of Rational Data, *In Proceedings of MML (Mathematical Methods for Learning).*
21. Isermann, R., 2006, *Fault-Diagnosis Systems: An introduction from Fault Detection to Fault Tolerance*, Springer, Germany, pp. 475.
22. Lee, Y.-J., Kim, D.-Y., Hwang, M.-S. and Cheong, Y.-S., 2012, "A Study on Data Prefiltering Methods for Fault Diagnosis, *Proceedings of the Korea society of CAD/CAM Engineers Conference*, pp. 301-308.

이 양 지

2011년 한국해양대학교 제어자동화공학과 학사
2011년~현재 UNIST 디자인 및 인간공학부 석사과정
관심분야: Intelligent Failure Analysis, Signal Processing



김 덕 영

1998년 포스텍 산업공학과 학사
2000년 포스텍 산업공학과 석사
2006년 스위스 로잔 연방공과대학교(EPFL), 기계공학과 박사
2000년~2001년 고등기술연구원 생산기술센터 주임연구원
2006년~2008년 EPFL, Post-doc.
2008년~2009년 스위스 취리히 연방공과대학교(ETHZ), Post-doc.
2009년~2010년 University of Warwick, Research Fellow
2010년~현재 UNIST 디자인 및 인간공학부 조교수
관심분야: Intelligent Failure Analysis, Smart Ship, Digital Manufacturing, Remote Laser Welding

황 민 순

1998년 한양대학교 전자공학과 학사
1998년~2006년현대정보기술 (울산 현대중공업 전산실)
2006년~현재 현대중공업통합전산실 차장
관심분야: Smart Ship, e-Navigation

정 영 수

1990년 연세대학교 기계공학과 학사
2007년 울산대학교 자동차선박기술대학원 석사
1991년~2000년 현대중공업 연구소
2000년~2010년 현대중공업 e-Business 사업본부, 조선 ERP 추진
2008년~현재 현대중공업 통합전산실 부장
관심분야: Shipbuilding Logistics, Industrial IT Convergence