J. Comput. Struct. Eng. Inst. Korea, 34(1)
pp.25~33, February, 2021
https://doi.org/10.7734/COSEIK.2021.34.1.25

pISSN 1229-3059 eISSN 2287-2302
Computational Structural Engineering Institute of Korea

진동 데이터의 시간영역 특징 추출에 기반한 고장 분류 모델

김 승 일[1]·노 유 정[2†]·강 영 진[3]·박 선 화[4]·안 병 하[5]

[1]부산대학교 기계공학부 대학원생, [2]부산대학교 기계공학부 부교수, [3]부산대학교 기계기술연구원 연구원,
[4]LG전자 H&A연구소 선임연구원, [5]LG전자 H&A연구소 연구위원

Fault Classification Model Based on Time Domain Feature Extraction of Vibration Data

Seung-il Kim[1], Yoojeong Noh[2†], Young-jin Kang[3], Sunhwa Park[4] and Byungha Ahn[5]

[1]Graduate Student, School of Mechanical Engineering, Pusan National University, Busan, 46241, Korea
[2]Associate Professor, School of Mechanical Engineering, Pusan National University, Busan, 46241, Korea
[3]Research Institute of Mechanical Engineering, Pusan National University, Busan, 46241, Korea
[4]Senior Researcher, H&A Research Center, LG Electronics, Changwon, 51554, Korea
[5]Chief Researcher, H&A Research Center, LG Electronics, Changwon, 51554, Korea

Abstract

With the development of machine learning techniques, various types of data such as vibration, temperature, and flow rate can be used to detect and diagnose abnormalities in machine conditions. In particular, in the field of the state monitoring of rotating machines, the fault diagnosis of machines using vibration data has long been carried out, and the methods are also very diverse. In this study, an experiment was conducted to collect vibration data from normal and abnormal compressors by installing accelerometers directly on rotary compressors used in household air conditioners. Data segmentation was performed to solve the data shortage problem, and the main features for the fault classification model were extracted through the chi-square test after statistical and physical features were extracted from the vibration data in the time domain. The support vector machine (SVM) model was developed to classify the normal or abnormal conditions of compressors and improve the classification accuracy through the hyperparameter optimization of the SVM.

Keywords *: fault diagnosis, twin rotary compressor, health index, intersection area, data augmentation, support vector machine (SVM)*

1. 서 론

일상생활 속에서 누구나 쉽게 접할 수 있는 기계 시스템 가운데 하나인 에어컨(air-conditioner)은 필수 가전제품으로 자리매김하였다. 이에 따라 에어컨의 수요도 증가하는 반면, 다양한 형태의 고장도 급격히 증가하고 있다. 고장은 현장에서 바로 수리가 되는 단순한 고장이 있는 반면, 전체 시스템을 교체해야 하는 치명적인 고장이 발생하기도 한다. 따라서 학계와 산업계에서는 에어컨의 고장을 사전에 감지하는 연구가 오래전부터 진행되었다. 기계 시스템의 고장진단과 관련하여 압축기에서 발생하는 근원적인 소음 메커니즘을 파악하고 설계를 개선(Son *et al*., 2017)하거나 압축기 소음원을 규명하기 위한 실험을 통해 주요 인자간의 관계를 파악한 연구(Son *et al*., 2015)도 수행되었다. 압축기의 소음과 관련된 연구 이외에도 진동분석을 통한 기계의 상태를 진단하는 연구가 진행되어 왔다. 회전기계에서 발생하는 진동 신호에서 노이즈 성분을 제거하여 강건한 고장진단 모델의 설계(Ko *et al*., 2018), 압축기로부터 수집된 진동 신호를 웨이블릿 변환을 통해 특징을 추출한 압축기의 상태 진단(Lim *et al*., 2003), 압축기에서 발생하는 진동과 관에서 유체의 맥동 때문에 발생하는 진동의 고조파 성분을 분석(Lee *et al*., 2012)하거나, 압축기의 압력맥동 때문에 발생하는 진동분석을 기반으로 소음 저감에 대한 연구(Sano and Mitsui, 1984)가 수행되었다. 위와 같은 연구들의 공통점은 소음과 진동을 실험적 환경에서 측정하고 그 신호에 대한

† Corresponding author:
Tel: +82-51-510-2308; E-mail: yoonoh@pusan.ac.kr
Received November 18 2020; Revised December 10 2020;
Accepted December 11 2020

분석을 통해 소음 혹은 진동의 근원을 규명 또는 기계의 상태를 진단하는 연구이다. 최근에는 컴퓨터 성능의 비약적인 발전에 힘입어 머신러닝과 관련된 연구가 활발히 진행되고 있는데, 압축기로부터 측정된 진동 신호를 별도의 전처리 없이 딥러닝 모델의 입력 데이터로 사용하여 분류 모델을 설계(Yang *et al*., 2019)하고, 진동 신호를 웨이블릿 변환을 거쳐 이미지로 변환하여 딥러닝 알고리즘을 통해 정상과 비정상을 분류하는 연구(Verstraete *et al*., 2017)가 수행되었다. 이러한 연구들은 진동 신호를 사용하여 기계 시스템의 상태를 진단하고 분류함으로써 진동 신호를 기반으로 한 고장 진단의 필요성을 확인하였다.

기존에 진행된 압축기와 관련된 연구들은 하나의 실험 조건을 대상으로 데이터를 취득하였지만(Son et al., 2015), 본 연구에서는 실제 에어컨 압축기의 다양한 운전 조건을 모사하기 위해서 정상 트윈 로터리 압축기와 압축부의 마모가 진행된 비정상 압축기를 대상으로 회전주파수, 팽창밸브의 개도 각 및 실외기 팬의 회전속도를 변화시켜가면서 다양한 실험 조건을 대상으로 진동 데이터를 취득하였다. 그리고 강건한 머신러닝 모델을 학습하기 위해 한정된 데이터 수를 늘리고 효과적으로 학습하기 위하여 정규화(normalization) 및 데이터 증대(augmentation)를 수행하였고, 전처리 전후에 따른 분류 난이도의 특성을 정량화하여 비교·분석하였다. 전처리된 데이터를 SVM 모델을 이용하여서 고장 분류 모델을 생성하였고, 강건하고 정확한 모델 생성을 위해 하이퍼 파라미터(hyper parameter)를 최적화하였다.

2. 실험 환경 및 데이터 취득

가정용 에어컨에는 비교적 가격이 저렴하여 경제적이고 구조가 단순한 로터리 압축기(rotary compressor)가 사용되는데, 편심 회전하는 회전자와 실린더, 베인(vane)으로 구성되어 흡입구에서 흡입된 냉매 증기를 회전자가 1회전하는 동안 냉매를 압축하여 토출하는 방식으로 구동된다. 로터리 압축기는 회전자가 한 개인 싱글 로터리 압축기와 회전자가 두개인 트윈 로터리 압축기로 구분 지을 수 있는데, 비교적 진동이 적고 회전축이 1회전하는 동안 180°의 위상차를 가지므로 압축 효율이 좋은 트윈 로터리 압축기의 보급이 증가하고 있다. 압축기는 구동 중 냉매의 부족 혹은 유출, 실린더 내부 부품의 마모, 윤활유 부족 등 다양한 원인에 의해 고장이 발생하는데 본 연구에서는 실린더 내부 부품의 마모로 인한 고장을 모사하여 고장 데이터를 수집하였다.

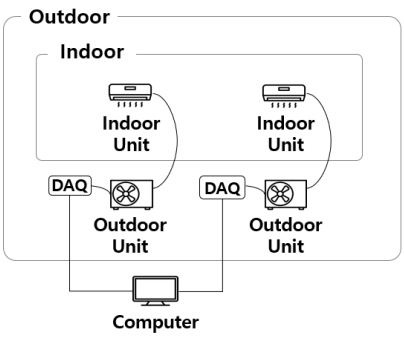


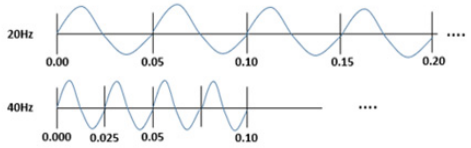
Fig. 1 Home environment laboratory

압축기의 진동 데이터를 취득하기 전에 계절에 따른 온도 조건을 설정하기 위하여 대형 챔버(chamber)가 설치된 '가정환경 실험실'에서 진행되었으며, 여름환경과 겨울환경을 모사하여 실험을 진행하였다. '가정환경 실험실'의 구조는 Fig. 1과 같으며 내부에는 실내 환경과 실외 환경을 구분 짓는 작은 방이 있으며, 실내에는 실내기를 설치하였고 실외에는 실외기와 측정 장비를 설치하였다. 챔버를 통해 실내와 실외의 온도를 설정할 수 있었으며, 여름환경을 모사하기 위해서 실내와 실외 온도를 각각 27°C와 35°C로 설정하였다. 또한, 겨울환경을 모사하기 위하여 실내와 실외 온도를 각각 7°C와 6°C로 설정하였다. 그 이외에 통제 변수로는 압축기의 회전 주파수, 팽창밸브(electronic expansion valve, EEV)의 개도 각 및 실외기 팬의 회전 속도가 있으며, 각각의 통제 변수를 조절하기 위하여 회사 측에서 자체 개발된 소프트웨어를 이용하였다. 실험 대상으로 사용된 압축기의 작동 가능 범위에 맞춰 회전주파수는 저주파에서 고주파수까지 총 8개의 조건에 대해 실험을 순차적으로 진행하였다. 또한, 팽창밸브의 개도 및 실외기 팬의 회전속도 역시 소프트웨어를 통해 통제되었으며, 각각 6개와 3개의 조건에 대해 실험을 진행하였다. 위와 같은 실험조건을 바탕으로 압축기가 작동되며 발생하는 진동 데이터는 트윈 로터리 압축기의 외벽에 설치된 가속도계를 통해 측정하였다. 진동 신호의 측정은 PCB 353B15를 사용하여 약 6.2초 동안 측정하였으며, DAQ를 통해 컴퓨터로 데이터를 전송하였다. 앞서 설명한 288개의 실험조건에 대하여 각각 실험을 수행하여 총 288개의 실험 데이터를 취득하였다. 자세한 고장원인과 실험조건은 회사의 기밀사항에 해당되어서 자세한 내용은 논문에 수록되지 않았다.

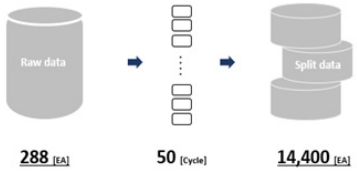
3. 데이터 전처리

3.1 데이터 증대

일반적으로 머신러닝 모델의 학습에 사용되는 데이터의 수는 모델의 성능에 많은 영향을 끼치게 되는데 데이터 수가 많을수록 더 정확한 모델을 만들 수 있다고 알려져 있지만(Cortes *et al*., 1995; Stockwell and Peterson, 2002), 데이터가 부족하거나



(a) Data segmentation for each cycle



(b) Before and after data segmentation

Fig. 2 Data augmentation

불균형을 이룰 경우 모델의 성능은 저하되는 경우가 많다. 이러한 문제를 해결하기 가짜 데이터를 생성하는 Generative Adversarial Network(GAN) 알고리즘을 사용하여 데이터 부족 문제를 해결(Wang *et al*., 2018)하거나, 소수(minority) 클래스를 오버샘플링(over-sampling)하여 다수(majority) 클래스의 수만큼 데이터를 증가시키는 Synthetic Minority Over-sampling Technique (SMOTE) 알고리즘을 사용하여 데이터 불균형을 해소하는 연구(Chawla *et al*., 2002) 등이 진행되었다. 하지만 오버샘플링 기법은 동일한 데이터를 반복 학습하므로 과적합(overfitting)될 수 있는 단점이 있기 때문에 사용에 주의가 필요하다(Son *et al*., 2019).

본 연구에서는 데이터 수집 실험을 통해 동일한 수의 정상 및 비정상 데이터를 취득할 수 있었기 때문에 데이터 불균형 문제는 발생하지 않았지만, 강건한 머신러닝 모델의 설계를 위해서는 실험을 통해 수집된 288개의 데이터는 부족하므로 데이터 증대가 필요하다. 따라서 Fig. 2(a)와 같이 운전 주파수의 역수에 해당하는 주기를 계산하여 데이터 분할의 기준을 설정하였고, Fig. 2(b)에 나타낸 그림의 내용과 같이 각각의 주기마다 데이터를 50분할하여 288개의 데이터를 14,400개의 데이터로 증가시켰다. 분할에 의한 데이터 증대는 한 주기에 해당하는 데이터를 일정하게 분할하여 분할 기준이 명확하다는 점과 기계적 결함과 같이 회전 기계에서 주기적으로 신호가 발생하는 경우에는 정상 및 비정상의 특징을 명확하게 확인이 가능하다는 장점이 있다.

3.2 특징 추출

고장 유무를 판단하는데 필요한 특징은 시간 영역과 주파수 영역 모두 추출 가능하지만, 본 연구의 목적은 실시간으로 측정되는 데이터의 고장 유무를 판단하는데 활용하는 것을 목표로 하므로 시간 영역에서의 특징만 추출하였다. 하지만, 시간

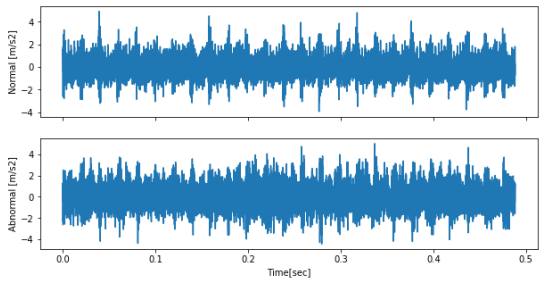


Fig. 3 Normal and abnormal vibration data

영역의 진동 신호만으로 정상과 비정상을 판단하는 것은 Fig. 3에서 확인할 수 있듯이 쉬운 일이 아니다. 따라서 판단의 근거가 되는 신호의 특징들을 모델이 학습할 수 있도록 특징 추출(feature extraction)이 필요하다. 특징 추출과 관련하여 많은 연구가 진행되었는데, 데이터의 평균(mean), 분산(variance), 왜도(skewness) 및 첨도(kurtosis) 등의 통계적 특징들을 비롯하여 신호의 실효치를 나타내는 RMS(root mean square)와 신호의 Peak를 표현하는 Peak to Peak 등의 물리적 특징들을 추출하여 고장을 대표하는 특징에 해당된다(Kim *et al*., 2010; Saxena *et al*., 2013; Caesarendra and Tjahjowidodo, 2017).

한편, 시간영역에서의 통계적 특징은 측정된 진동 데이터와 측정된 다른 진동데이터 간의 차이를 나타낸다. 정상적으로 작동되는 회전기계의 상태에 이상이 생기면 진동 데이터의 확률밀도함수(probability density function, PDF)도 변하므로 통계적 특징 역시 변하게 된다. 특히 진동 데이터의 PDF가 한 쪽으로 치우친 정도를 나타내는 왜도와 PDF의 Peak 값을 나타내어 진동 데이터의 뾰족한 정도를 나타내는 첨도는 가장 크게 변하는 특징이다.

본 연구에서는 Table 1와 같이 통계적 특징(statistical features) 5개와 물리적 특징(physical features) 4개에 대해 특징 추출을 진행하였다. Peak to Peak은 시간에 독립적이며 진동 신호의 피크 크기를 나타내고, 진동 신호를 표현할 때 가장 흔하게 사용되며, 진동 신호의 에너지와 직접적으로 관련이 있는 RMS는 급격한 신호 변동이 있는 고장을 나타내는데 주요 특징으로 사용된다. Impulse Factor 는 일반적으로 베어링 고장진단에서 결함을 나타내는 지표로 사용되며 충돌과 관련된 고장을 나타내는 특징으로 자주 사용되며, Shape Factor는 회전기계에서 불균형, 정렬불량(misalignment)을 대표하는 특징으로 사용된다.

3.3 데이터 정규화 및 특징선택

추출된 특징은 모두 다른 수치적 범위를 가지게 되는데, 그대로 학습에 사용하게 되면 수치가 큰 값들에 편향되어 학습될

Table 1 Statistical & physical features

Statistical features	Formula	Physical features	Formula
Absolute mean	$\frac{1}{N}\sum_{k=1}^{N}\lvert X_k\rvert$	Peak to Peak	$Max(X_k)-Min(X_k)$
Mean	$\frac{1}{N}\sum_{k=1}^{N}X_k$	RMS	$\left(\frac{1}{N}\sum_{k=1}^{N}X_k^2\right)^{1/2}$
Variance	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{X_k-m}{\sigma}\right)^2$	Impulse Factor	$\frac{Max(\lvert X_k\rvert)}{\frac{1}{N}\sum_{k=1}^{N}\lvert X_k\rvert}$
Skewness	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{X_k-m}{\sigma}\right)^3$	Shape Factor	$\frac{\mathrm{RMS}}{\left(\frac{1}{N}\sum_{k=1}^{N}\lvert X_k\rvert\right)^{1/2}}$
Kurtosis	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{X_k-m}{\sigma}\right)^4$		

수 있기 때문에 학습이 제대로 이루어지지 않을 수 있다. 따라서 추출된 특징들을 일정한 범위로 정규화 시키기 위해 최대-최소 정규화(min-max normalization) 방법을 적용하였다. 최대-최소 정규화는 전체 데이터를 0~1 사이의 값으로 정규화 하는 방법으로서 다음과 같이 표현된다.

$$x_{scale}=\frac{x-x_{min}}{x_{max}-x_{min}} \tag{1}$$

여기서, x_{min}은 데이터의 최솟값, x_{min}은 최댓값이다.

기계 학습 모델은 데이터의 충분한 양과 양질의 데이터에 의해 성능이 좌우된다. 따라서 양질의 데이터를 학습시키기 위하여 정규화된 고차원의 특징들 중에서 유용한 특징을 선별하는 특징 선택(feature selection)이 필수적이다. 특징 선택은 일반적으로 Filter, Wrapper 그리고 Embedded method가 있는데, 본 연구에서는 가장 빠르고 효과적인 방법인 Filter 방법을 이용하였으며, 그 중에서 단변량(univariate) 및 분류(classification) 문제의 조건을 만족할 때 사용되는 카이제곱(chi-square)(Jović *et al*., 2015) 통계량을 사용하여 주요 특징을 선택하였다. 카이제곱 검정은 명목형 변수 간의 동질성(homogeneity), 연관성(association)과 독립성(independence)을 검정하는 방법으로 본

Table 2 Chi-square statistics and decision

	Feature	Statistics	P-value
1	Peak to Peak	1993.9	0.00E+00
2	RMS	1124.9	1.25E-246
3	Variance	897.3	1.68E-196
4	Skewness	493.8	2.08E-109
5	Absolute Mean	312.8	5.29E-70
6	Mean	26.0	3.35E-07
7	Impulse Factor	1.49E-27	1.00E+00
8	Shape Factor	3.16E-28	1.00E+00
9	Kurtosis	5.17E-29	1.00E+00

연구에서는 9개의 통계적 특징과 정상 및 비정상에 대한 출력변수와의 독립성 검정을 위해 사용되었다. 앞서 추출한 5개의 통계적 특징과 4개의 물리적 특징에 대하여 카이제곱통계량과 P-value를 계산하였고 유의수준 95%(α=0.05)에 대하여 귀무가설(*Ho*: 추출된 통계적 특징과 정상 및 비정상에 대한 출력변수가 서로 독립이다)의 기각 여부를 결정하였다. Table 2에 나타난 것처럼 P-value가 유의수준보다 낮은 특징들에 대해 귀무가설이 기각되므로 유의미한 6개의 변수를 선택할 수 있다. 선택된 특징들 중에서 Mean은 데이터의 주기성 특성 때문에 대분의 값이 0에 근접한 값을 가지므로 모델의 성능에 많은 영향을 끼치지 못한다고 판단되어 제외하였고, 최종적으로 상위 5개의 특징을 선택하였다.

3.4 건전성 인자 모델과 교차면적

3.1장과 3.3장에서 고장 진단의 분류 정확도를 높이기 위해서 전처리 단계에서 데이터의 정규화와 증대를 수행하였다. 두 처리의 효용성은 학습 후 분류 정확도를 통해 확인할 수 있지만, 이는 전처리를 수행할 때마다 학습을 수행해야 하는 비효율적인 방법이기 때문에 학습 전 데이터 전처리 과정이 분류 학습의 정확도에 어떻게 영향을 미칠지 분석해 볼 필요가 있다. 본 연구에서는 전처리 방법의 적용 유무에 따른 효과가 분류 학습의 난이도(정확도)에 어떻게 영향을 미치는지 정량적으로 확인하기 위해서 건전성 인자(health index, HI) 모델과 이로 부터 예측된 값들의 PDF의 교차면적(intersection area, IA)을 사용하여서 분류 난이도를 정량화하였다.

3.4.1 건전성 인자 모델

건전성 인자 모델은 데이터로부터 추출된 특징을 설명변수(explanatory variable)로 두고 제품의 상태를 수치로 표현하는 건전성 인자를 반응변수(response variable)로 두고 회귀분석을 통해 건전성 인자를 예측하는 모델이다(Jahromi *et al.,* 2009; Choi, 2020). 압축기의 건전성 인자 모델을 구축하기 위해서 설명변수는 3.3장에서 선택된 5개의 특징을 사용하고, 반응변수(건전성 인자)의 값은 정상인 경우 0, 비정상인 경우 1로 레이블링(labeling) 하였다. 회귀모델은 과대적합의 효과를 배제하기 위해서 가장 보수적인 분류 정도를 보여주는 선형함수를 사용하였다. 건전성 인자 모델을 구축한 다음, 사용된 특징 값들을 모델에 입력하게 되면 예측된 건전성 인자 값들이 계산되고 이들은 실제 건전성 인자 값과 다르게 0과 1을 중심으로 하는 분포함수를 생성하게 된다. 이때 두 PDF의 거리가 멀게 되면 분류 난이도가 낮아지는 것을 의미하고 가까워지면 분류 난이도가 높게 되는 것을 의미한다(Choi, 2020).

3.4.2 교차면적

건전성 인자의 예측 값에서 PDF의 거리는 분포 특성을 고려하지 않은 척도이므로 확률적 분포 특성을 고려하기 위해서는 두 PDF의 교차면적을 계산해서 표현할 수 있다. 교차면적은 두 PDF의 겹치는 면적을 정량화하는 방법으로 확률의 기본 정의에 의해서 0~1(0~100%)의 값을 가지며, 두 PDF가 완전 일치하면 1, 반대의 경우에는 0의 값을 가진다. 교차면적은 다음과 같이 계산된다(Kang *et al.*, 2018).

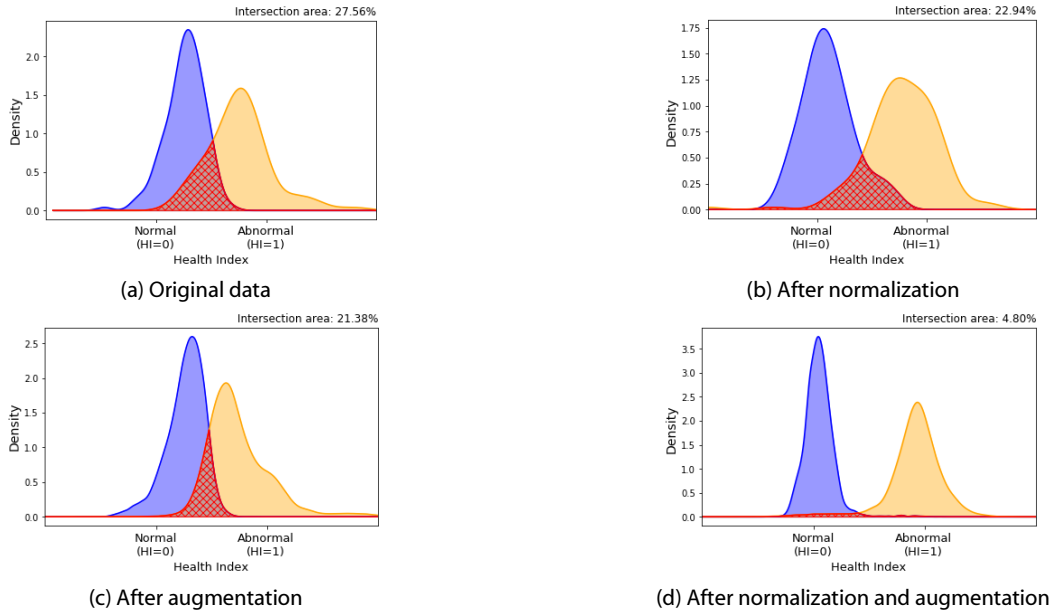
$$IA_{0,1} = \sum_{i=1}^{n} \left\{ f_{0,1}\left(\widehat{HI}_i\right) \times \left(\widehat{HI}_i - \widehat{HI}_{i-1}\right) \right\} \tag{2}$$

$$f_{0,1}\left(\widehat{HI}_i\right) = \min\left[f_0\left(\widehat{HI}_i\right), f_1\left(\widehat{HI}_i\right) \right] \tag{3}$$

여기서, f_0과 f_1은 각각 정상과 비정상 상태의 PDF를 나타내고 $\widehat{HI}$는 예측된 건전성 인자 값이다. $IA_{0,1}$는 정상-비정상 PDF의 교차면적이고, 이는 두 PDF를 균일한 n개의 간격으로 도메인을 분할한 후에 부분 교차면적의 값을 전체 도메인에서 더함으로써 계산된다.

앞서 수행된 데이터 증대 및 정규화의 효과가 데이터 분류에 미치는 영향을 확인하기 위하여 건전성 인자와 교차면적을 사용하여 확인하였으며, Fig. 4는 정상과 비정상 데이터의 예측된 건전성 인자를 이용한 PDF를 보여준다. 여기서, (a)는 데이터 증대와 정규화 전처리가 적용되지 않는 결과이고, (b)는 정규화만 수행된 결과, (c)는 증대만 수행된 결과, (d)는 정규화와 증대 모두가 수행된 후의 PDF를 보여준다. Fig. 4에서 PDF는 3.4.1장에서 계산된 정상과 비정상의 예측된 건전성 인자 값을 커널밀도추정(kernel density estimation, KDE)으로 추정하였다. KDE는 대표적인 비모수적 PDF 추정 방법으로써 PDF 함수의 형태에 종속되지 않으며 데이터에 가장 유사한 PDF를 추정하기 때문에 예측 불가능한 PDF의 형태를 표현하기 위해서 사용되었다(Kang *et al.*, 2017).

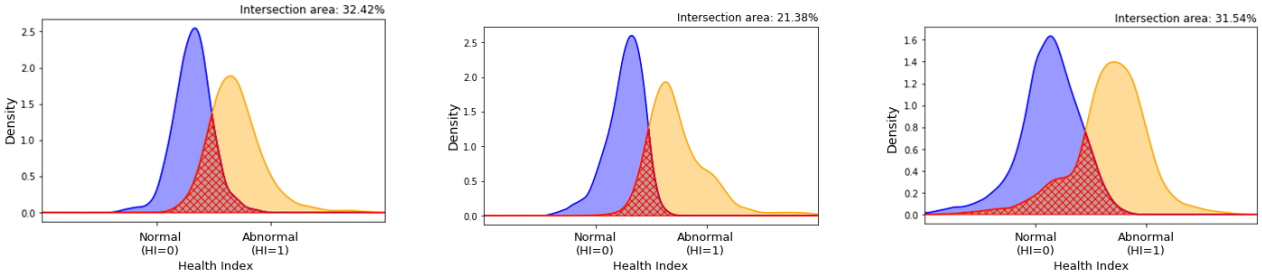
Fig. 4(a)는 데이터 증대 및 정규화를 하지 않았을 때의 결과를 나타내는데 정상과 비정상의 IA는 27.57%로 비교적 많은



(a) Original data (b) After normalization

(c) After augmentation (d) After normalization and augmentation

Fig. 4 PDFs of health index



(a) Number of splitting: 10 (b) Number of splitting: 50 (c) Number of splitting: 100

Fig. 5 HI according to number of splitting

부분이 교차하는 것을 확인할 수 있고, Fig. 4(b)는 정규화만 하였을 때의 결과를 나타내는데, IA가 22.94%로 Fig. 4(a)의 결과에 비해 IA가 줄어들었음을 확인할 수 있다. FIg. 4(c)는 데이터 증대를 수행하였을 때의 결과를 나타내는데 이 또한 Fig. 4(a)의 결과보다 IA가 줄었음을 확인할 수 있다. Fig. 4(d)는 정규화와 증대를 함께 수행하였을 때의 결과를 나타내며 앞선 결과들에 비해 IA가 4.8%로 가장 적음을 확인할 수 있다. 따라서 정규화 및 증대를 각각 독립적으로 수행하였을 경우에도 효과가 있지만 함께 사용하였을 때 분류 성능이 개선됨을 확인할 수 있다.

또한, 데이터 증대의 수를 10, 50, 100으로 설정하고 각각에 대한 건전성 인자를 이용하여 PDF를 Fig. 5에 정리하였다. Fig. 5(a), Fig. 5(b), Fig. 5(c)를 각각 비교한 결과, 50분할하였을 때 IA가 21.38%로 세 가지 케이스에 대하여 가장 IA가 작은 것을 확인할 수 있었으며, 이로써 분할 수를 50으로 설정하였다.

4. SVM(Support Vector Machine)

4.1 SVM 마진

본 연구에서는 분류 및 회귀(regression) 문제를 해결하기 위하여 많은 연구 및 산업에서 사용되는 지도학습 기반 모델인 SVM을 사용하여 정상과 비정상으로 이진 분류를 수행하였다. 기본적으로 SVM은 두 클래스 사이의 마진(margin)을 최대로 하는 초평면(hyper-plane)을 찾는 것을 목적으로 하고 초평면을 지지하는 관측치들을 Support Vector라고 부른다. 마진은 초평면에서 직교방향으로 Support Vector와의 거리를 의미하고 오류를 허용하지 않는 하드 마진(hard Margin)과 어느 정도 오류를 허용하는 소프트 마진(soft Margin)으로 나뉘는데, 하드 마진의 목적 함수와 조건은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} &Minimize\ \ {}_{\omega,b}\frac{1}{2}\parallel\omega\parallel^2 \\ &subject to\ y_i\left(\omega^T X_i+b\right)\ge 1 \\ &(i=1,2,...,l) \end{aligned} \tag{4}$$

여기서, ω는 초평면(hyerplane)의 법선벡터이며, b는 원점으로부터의 거리를 나타낸다.

하드마진은 데이터가 선형적으로 분류되지 않으면 초평면 생성에 문제가 있고 현실적으로 대다수의 모델은 비선형적 특성을 가지므로 소프트 마진의 사용이 일반적이다. 소프트 마진의 목적함수는 샘플이 마진을 얼마나 위반하는지를 정하는 슬랙변수(ξ)를 도입하여 다음과 같이 나타낼 수 있다.

$$\begin{aligned} &Minimize\ \ {}_{\omega,b,\xi}\frac{1}{2}\parallel\omega\parallel^2+C\sum_{i=1}^{n}\xi_i \\ &subject to\ y_i\left(\omega^T X_i+b\right)\ge 1-\xi_i,\ \xi_i\ge 0 \\ &(i=1,2,...,l) \end{aligned} \tag{5}$$

여기서, C(cost)는 슬랙변수의 패널티와 마진 너비 사이를 절충하는 정규화 하이퍼 파라미터이다. C를 사용하기에 앞서 최적화가 필요한데 경험적으로 설정하거나 Grid-search, Bayesian Optimization 등과 같은 방법으로 최적화를 수행할 수 있다.

4.2 커널(Kernel)함수

대부분 현실의 데이터는 고차원의 특징을 가지므로 선형적으로 해결하기 어렵기 때문에 커널(kernel) 함수를 사용하여 저차원에서 고차원으로 매핑(mapping)하여 분류 작업을 수행할 수 있다.

$$\Phi(X)=\Phi\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}\right)=\begin{pmatrix}x_1^2\\\sqrt{2}\,x_1x_2\\x_2^2\end{pmatrix} \tag{6}$$

식 (6)은 2차 다항식에 대한 매핑함수(Φ)이며, 2차 다항식의 매핑함수를 적용하게 되면 2차원에서 3차원의 특징 공간(feature space)으로 확장되고 그에 따라 2차원에서 분류할 수 없었던 비선형 문제를 선형으로 분류할 수 있게 된다. 매핑된 고차원의 특징공간에서 데이터 샘플들 사이의 거리 관계가 원래 공간에서의 거리 관계를 보존할 필요가 있으므로 매핑함수를 사용하여 식 (7)의 커널 함수를 정의한다.

$$K\left(X,x_i\right)=\Phi(X)^T\Phi(x') \tag{7}$$

또한, 많은 커널 함수 중에서 일반적으로 사용되는 커널의 일부를 Table 3에 나열하였다.

4.3 하이퍼 파라미터 최적화

SVM을 사용하여 데이터 분류를 위해서 최적의 하이퍼 파라미터(hyper parameter) 설정은 필수적이다. 최적화가 필요한

Table 3 Kernel functions

	Kernel Type	Formula
1	Linear	$a^T\cdot b$
2	Sigmoid	$\tanh(\gamma a^T\cdot b+r)$
3	RBF	$\exp(-\gamma\parallel a-b\parallel^2)$

하이퍼 파라미터는 C와 γ가 있고 동시에 커널도 최적의 커널을 선정하여야 한다. C는 마진의 폭을 결정하는 파라미터로써, C가 작아지면 데이터 샘플이 다른 클래스에 놓였을 때의 패널티가 낮아지므로 마진의 폭이 커지게 된다. γ는 데이터 샘플이 영향력을 행사하는 거리를 결정하는 파라미터로써 그 값이 작을수록 영향력을 행사하는 거리가 커지게 된다. 최적의 C와 γ를 찾기 위하여 Grid-search를 수행하였는데, 과적합(over-fitting)을 방지하기 위하여 5개의 fold에 대해 K-fold 교차검증(cross validation)도 함께 수행하였다. 하이퍼 파라미터 최적화하기에 앞서서 데이터를 Train set과 Test set을 각각 70:30의 비율로 분할하여 사용하였다. Train set은 하이퍼 파라미터 최적화를 위해 사용되며, Test set은 학습에 사용되지 않은 데이터로써, 최적의 하이퍼 파라미터로 학습된 모델을 검증하기 위해 사용된다. 하이퍼 파라미터를 최적화하기 위하여 Linear, RBF, Sigmoid 커널에 대해 $10^{-4} \sim 10^{4}$의 범위 내에서 Step 크기를 10으로 설정한 다음 총 10,000개의 실험점에 대해 5개 fold의 평균 정확도(accuracy)를 목표로 Grid-search를 수행하였다. 그 결과, 커널은 RBF함수를 선정하였고 C=1000, γ=1을 최적의 하이퍼 파라미터로 설정하였다.

4.4 모델 검증 및 분류 결과

최적의 하이퍼 파라미터를 선정하여 모델을 훈련하고 Test set에 대하여 분류를 진행한 결과, 모델의 정확도는 99.86%였다. 모델의 성능을 확인하기 위해 정확도 이외에 식 (8)~(11)에 표현한 정밀도(precision), 재현율(recall) 그리고 F1스코어(f1 score)를 계산하여 모델의 성능을 검증하였다.

$$Accuracy: \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$\mathrm{Precision}: \frac{TP}{TP+FP} \tag{9}$$

$$Recall: \frac{TP}{TP+FN} \tag{10}$$

Table 4 Confusion matrix for classification

Actual / Predicted	Normal	Abnormal
Normal	2181 (True Positive)	0 (False Positive)
Abnormal	6 (False Negative)	2133 (True Negative)

$$F1-score: 2\times\frac{\mathrm{Precision}\times Recall}{\mathrm{Precision}+Recall} \tag{11}$$

식 (8)~(11)에서 표현된 TP(true positive), TN(true negative), FP(false positive), FN(false negative)는 Table 4에 나타낸 혼돈행렬(confusion matrix)를 기반으로 확인할 수 있다.

Table 5는 데이터 증대 전후 및 분할 수에 따른 결과, 그리고 하이퍼 파라미터 최적화 전후에 대해 분류 정확도, 정밀도, 재현율 및 F1스코어를 계산한 결과이다. Table 5에서 확인할 수 있듯이 하이퍼 파라미터 최적화를 수행하기 전의 모든 결과는 최적화를 수행한 후의 결과에 비해 약 2~3% 정도 낮은 것을 확인할 수 있다. 이로써 하이퍼 파라미터 최적화의 수행이 분류 모델 학습에 미치는 영향이 크다는 것을 확인할 수 있었다. 그리고 데이터 증대가 분류에 미치는 영향을 확인하기 위해 분할 정도에 따른 정확도를 비교한 결과, 하이퍼 파라미터 최적화에 상관없이 증대 전과 후의 차이가 약 3~4%정도 차이를 보였고, 앞서 Fig. 5에서 비교한 결과와 같이 10분할 혹은 100분할보다 50분할을 적용하였을 때 분류 정확도 및 나머지 지표들의 결과가 비교적 우수한 것을 확인할 수 있었다. 이 결과를 통해 건전성인자에 따른 PDF의 IA의 결과와 모델의 분류 정확도가 같은 경향을 보인다는 것을 확인할 수 있었다. 그러므로 혼합행렬과 건전성인자를 이용한 분류 정확도 결과를 보면, 분류 모델의 정확도를 향상시키기 위해서는 하이퍼 파라미터 최적화와 데이터 증대가 반드시 필요하다는 사실을 알 수 있다.

5. 결 론

기존에 수행된 다수의 연구는 가상의 데이터를 활용하거나 한 가지 실험조건에서 데이터를 측정하여 고장 감지 및 진단

Table 5 Results of classification(%)

	Before hyper-parameter optimization				After hyper-parameter optimization			
	Before augmentation	After augmentation			Before augmentation	After augmentation		
		10split	50split	100split		10split	50split	100split
Accuracy	93.10	96.18	96.55	96.40	95.40	98.84	99.86	99.02
Precision	88.37	93.79	93.02	92.07	100.0	98.66	100.0	99.22
Recall	97.43	98.25	100.0	94.55	90.69	99.10	99.72	98.85
F1-score	92.68	95.97	96.38	93.29	95.12	98.87	99.97	99.03

방법을 제시하였다. 하지만 본 논문에서는 실제로 운영되는 정상 압축기와 비정상 압축기를 대상으로 실제 운전을 모사할 수 있는 다양한 실험 조건을 갖춘 후, 정상과 비정상의 분류가 가능한 SVM 모델을 제시하였다.

SVM 모델의 설계를 위하여 온도 조절이 가능한 챔버가 설치된 가정환경 실험실에서 데이터 수집을 실시하였고, 수집된 데이터를 통해 강건한 머신러닝 모델 학습을 위하여 데이터를 주기별로 분할하여 증대시켰다. 또한, 시간영역에서 통계적 특징 5개와 물리적 특징 4개를 추출하였고, 데이터 정규화를 수행한 후 카이-제곱 독립성 검정을 통해 유의미한 특징 5개를 선택하였다. 데이터 증대 및 정규화의 효과를 확인하기 위하여 건전성 인자 모델을 통한 교차면적 계산을 하였고, 데이터 증대 및 정규화를 함께 수행하였을 때 분류 난도가 효과적으로 줄어들었음을 확인하였다. 또한, Grid-search 방법을 사용하여 SVM 모델의 하이퍼파라미터 튜닝을 수행하였고 최종적으로 모델을 완성하였다.

최종적으로 모델의 분류 정확도는 99.86%로 높았으며 정밀도, 재현율 및 F1스코어 역시 높은 모델을 완성하였다. 또한, 데이터 증대를 수행하기 전과 후의 결과를 비교함으로써 데이터의 수가 모델의 성능에 미치는 영향도 확인하였다.

향후 연구에서는 기존에 완성된 SVM 모델에 에어컨 운행에 따라 냉매의 누설을 감지하고 누설량에 따라 분류가 가능한 SVM 모델을 완성할 계획이며, 주파수영역(frequency domain)에 대한 분석과 신호처리를 통해 새로운 특징 추출에 대한 연구를 진행할 계획이다.

감사의 글

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

References

Caesarendra, W., Tjahjowidodo, T. (2017) A Review of Feature Extraction Methods in Vibration-Based Condition Monitoring and Its Application for Degradation Trend Estimation of Low-Speed Slew Bearing, *Mach*., 5(4), p.21.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique, *J. Artif. Intell. Res*., 16, pp.321~357.

Choi, J. (2020) *PHM Practice-Case Study for Industrial Digitization: PHM Core Basic*, Korea Society for Prognostics & Health Management, Seoul, South Korea, p.29.

Cortes, C., Jackel, L.D., Chiang, W.P. (1995) *Limits on Learning Machine Accuracy Imposed by Data Quality*, In Advances in Neural Information Processing Systems, pp.239~246.

Jahromi, A., Piercy, R., Cress, S., Service, J., Fan, W. (2009) An Approach to Power Transformer Asset Management using Health Index, *IEEE Electr. Insul. Mag*., 25(2), pp.20~34.

Jović, A., Brkić, K., Bogunović, N. (2015) A Review of Feature Selection Methods with Applications, *38th International Convention on Information and Communication Technology*, Electronics and Microelectronics (MIPRO), pp.1200~1205.

Kang, Y.J., Hong, J., Lim, O.K., Noh, Y. (2017) Reliability analysis using parametric and nonparametric input modeling methods, *J. Comput. Struct. Eng. Inst. Korea*, 30(1), pp.87~94.

Kang, Y.J., Noh, Y., Lim, O.K. (2018) Kernel Density Estimation with Bounded Data, *Struct. & Multidiscip. Optim*., 57(1), pp.95~113.

Kim, Y.S., Lee, D.H., Kim, S.K. (2010) Fault Classification for Rotating Machinery using Support Vector Machines with Optimal Features Corresponding to Each Fault Type, *Trans. Korean Soc. Mech. Eng. A*, 34(11), pp.1681~1689.

Ko, J.U., Jung, J.H., Kim, M., Kong, H.B., Youn, B.D. (2018) Noise Robust Fault Diagnosis Technique to Simultaneously Learn Classification and Denoising, *In Proceedings of The Korean Soc. of Mech. Eng. (KSME)*, pp.165~167.

Lee, S.H., Ryu, S.M., Jeong, W.B. (2012) Vibration Analysis of Compressor Piping System with Fluid Pulsation, *J. Mech. Sci. & Technol*., 26(12), pp.3903~3909.

Lim, D.S., Yang, B.S., An, B.H., Tan, A., Kim, D.J. (2003) Condition Classification for Small Reciprocating Compressors Using Wavelet Transform and Artificial Neural Network, *J. Korea Soc. Power Syst. Eng*., 7(2), pp.29~35.

Sano, K., Mitsui, K. (1984) Analysis of Hermetic Rolling Piston Type Compressor Noise, and Countermeasures, *Int. Compress. Eng. Conf*., p.460.

Saxena, V., Chowdhury, N., Devendiran, S. (2013) Assessment of Gearbox Fault Detection using Vibration Signal Analysis and Acoustic Emission Technique, *J. Mech. & Civil Eng*., 7(4), pp.52~60.

Son, M.J., Jung, S.W., Hwang, E.J. (2019) A Deep Learning Based Over-Sampling Scheme for Imbalanced Data Classification, *KIPS Trans. Softw. & Data Eng*., 8(7), pp.311~316.

Son, Y., Ha, J., Lee, J. (2015) An Experimental Study on the Noise Source Identification of Rotary Compressor, *Trans. Korea Soc. Noise & Vib. Eng*., 25(11), pp.723~730.

Son, Y., Ha, J., Lee, J. (2017) The Noise Identification and Reduction of a Twin Rotary Compressor, *Trans. Korea Soc. Noise & Vib. Eng*., 27(3), pp.306~311.

Stockwell, D.R., Peterson, A.T. (2002) Effects of Sample Size on Accuracy of Species Distribution Models, *Ecol. Model*., 148(1), pp.1~13.

Verstraete, D., Ferrada, A., Droguett, E.L., Meruane, V., Modarres, M. (2017) Deep Learning Enabled Fault Diagnosis using Time-Frequency Image Analysis of Rolling Element Bearings, *Shock & Vib*., 2017.

Wang, G., Kang, W., Wu, Q., Wang, Z., Gao, J. (2018) Generative Adversarial Network (GAN) based Data Augmentation for Palmprint Recognition, *Digital Image Computing: Techniques and Applications (DICTA)*, pp.1~7.

Yang, H.B., Zhang, J.A., Chen, L.L., Zhang, H.L., Liu, S.L. (2019) Fault Diagnosis of Reciprocating Compressor based on Convolutional Neural Networks with Multisource Raw Vibration Signals, *Math. Probl. Eng*., 2019.

요 지

머신러닝 기법의 발달과 함께 기계에서 발생하는 다양한 종류(진동, 온도, 유량 등)의 데이터를 활용하여 기계의 상태를 진단하고 이상 탐지 및 비정상 분류 연구도 활발히 진행되고 있다. 특히 진동 데이터를 활용한 회전 기계의 상태 진단은 전통적인 기계 상태 모니터링 분야로 오랜 기간 동안 연구가 진행되었고, 연구 방법 또한 매우 다양하다. 본 연구에서는 가정용 에어컨에 사용되는 로터리 압축기에 가속도계를 직접 설치하여 진동 데이터를 수집하는 실험을 진행하였다. 데이터 부족 문제를 해결하기 위해 데이터 분할을 수행하였으며, 시간 영역에서의 진동 데이터로부터 통계적, 물리적 특징들을 추출한 후, Chi-square 검증을 통해 고장 분류 모델의 주요 특징을 추출하였다. SVM(Support Vector Machine) 모델은 압축기의 정상 혹은 이상 유무를 분류하기 위해 개발되었으며, 파라미터 최적화를 통해 분류 정확도를 개선하였다.

핵심용어 : 고장진단, 트윈 로터리 압축기, 건전성지표, 교차면적, 데이터 증대, 서포트벡터머신