Journal of the KIMST, Vol. 24, No. 3, pp. 264-271, 2021
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2021.24.3.264

Research Paper 센서·신호처리 부문

잡음과 스펙트럼 이동에 강인한 CNN 기반 라만 분광 알고리즘

박재현[1] · 유형근[1] · 이창식[1] · 장동의[*,1] · 박동조[1] · 남현우[2] · 박병황[2]

[1] 한국과학기술원 전기및전자공학부
[2] 국방과학연구소 제4기술연구본부

CNN based Raman Spectroscopy Algorithm That is Robust to Noise and Spectral Shift

Jae-Hyeon Park[1] · Hyeong-Geun Yu[1] · Chang Sik Lee[1] · Dong Eui Chang[*,1] · Dong-Jo Park[1] · Hyunwoo Nam[2] · Byeong Hwang Park[2]

[1] *Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Korea*
[2] *The 4th Research and Development Institute, Agency for Defense Development, Korea*

(Received 29 January 2021 / Revised 7 April 2021 / Accepted 30 April 2021)

Abstract

Raman spectroscopy is an equipment that is widely used for classifying chemicals in chemical defense operations. However, the classification performance of Raman spectrum may deteriorate due to dark current noise, background noise, spectral shift by vibration of equipment, spectral shift by pressure change, etc. In this paper, we compare the classification accuracy of various machine learning algorithms including k-nearest neighbor, decision tree, linear discriminant analysis, linear support vector machine, nonlinear support vector machine, and convolutional neural network under noisy and spectral shifted conditions. Experimental results show that convolutional neural network maintains a high classification accuracy of over 95 % despite noise and spectral shift. This implies that convolutional neural network can be an ideal classification algorithm in a real combat situation where there is a lot of noise and spectral shift.

Key Words : Raman Spectroscopy(라만 분광기), Convolutional Neural Network(합성곱 신경망), Machine Learning(기계학습), Spectral Shift Robustness(스펙트럼 이동 강인성), Noise Robustness(잡음 강인성)

1. 서 론

라만 산란(Raman scattering)이란 빛이 물질에 입사되었을 때 입사광과 물질의 상호작용으로 입사광과 다른 파수(wavenumber)들의 빛이 산란광으로 전 방향으로 방출되는 현상이다. 입사광과 라만 산란광과의 파수 차이를 라만 전이(Raman shift)라고 하며 물질마

* Corresponding author, E-mail: dechang@kaist.ac.kr

다 라만 전이의 산란광 세기가 다르다[1]. 라만 분광기(Raman spectroscopy)는 레이저를 표적 물질에 입사하여 나오는 산란광을 필터링하여 라만 전이를 포함하는 파수 대역의 라만 산란광만을 검출기로 보내서 빛의 세기를 검출한다[2]. 라만 전이에 따른 라만 산란광의 세기를 라만 스펙트럼(Raman spectrum)이라고 하며 물질의 고유 특징이기에 라만 스펙트럼을 분석해서 물질을 탐지하고 분류할 수 있다[3].

라만 분광기는 대기 성분 분석[4], 환경 분석[5], 지질 분석[6], 화학실험실의 물질 분류[7] 등에 사용되었으며 최근에는 지표면에 존재하는 생물 물질에 의한 형광 간섭을 회피할 수 있는 단자외선 UV 레이저 기술, 광학기술, 검출기술의 발달로 원거리 비접촉식 지표면 라만 분광기가 연구되어 군사 작전에 사용되고 있다[8,9]. 원거리 비접촉식 지표면 라만 분광기는 기존 접촉식 분석 장비의 단점인 장비의 오염이 없기에 화생방 작전 하에서 살포된 오염물질을 탐지하고 구별하는데 기존 접촉식 분석 장비들을 대체하고 있다[10,11].

라만 분광기의 물질 분류를 방해하는 요인으로는 잡음(noise), 스펙트럼 이동(spectral shift)이 있다. 잡음은 물질의 고유 라만 스펙트럼 신호에 더해지는 불규칙한 확률 신호로 잡음의 원인으로는 산란광 이외의 빛, 검출기의 암전류 등이 있으며[12], 잡음을 제거하기 위한 알고리즘 연구가 진행되고 있다[13]. 스펙트럼 이동은 측정된 라만 스펙트럼이 물질의 고유 라만 스펙트럼보다 파수를 축으로 평행이동된 상태가 되는 것이다. 스펙트럼 이동은 물질 주변의 압력 변화[14]나 화생방 차의 떨림 등으로 라만 분광기에 가해지는 진동으로 내부 광학기계가 조금씩 틀어져서 생긴다. 이 경우는 관리자가 직접 주기적으로 파수를 보정하는 것으로 해결한다.

이러한 잡음과 스펙트럼 이동을 일일이 보정하지 않고 라만 분광기로 바로 물질을 분류할 수 있다면 간편할 것이다. 기계학습 알고리즘은 클래스별 데이터의 특징을 학습해서 이후에 새로운 데이터가 들어와도 해당 클래스로 분류할 수 있는 알고리즘으로, 다양한 잡음과 스펙트럼 이동이 있는 데이터를 그대로 기계학습 알고리즘에 학습시킨다면 이후에 들어온 새로운 데이터에 잡음이나 스펙트럼 이동이 있어도 제대로 구별할 것이다. 기존의 기계학습 알고리즘들로는 K-근접 이웃(K-nearest neighbor), 선형판별분석(Linear discriminant analysis), SVM(Support Vector Machine), 결정 트리(Decision tree)[15,16]가 있다. 최근에는 인간의 뇌를 모방한 기계학습 알고리즘인 신경망 알고리즘을 사용한 연구가 주를 이루고 있다[17,18].

본 논문에서는 기존 기계학습 알고리즘과 신경망 알고리즘 중에서 이미지 분류에 뛰어나다고 알려진 CNN(Convolutional Neural Network) 알고리즘[19]을 사용한다. 화학물질 42종의 순수 라만 스펙트럼 데이터를 얻은 후 임의로 다양한 수준의 잡음과 스펙트럼 이동을 부여해서 데이터를 생성하고, 생성된 데이터를 기계학습 알고리즘으로 학습한다. 이후 잡음과 스펙트럼 이동이 있는 새로운 라만 스펙트럼 데이터를 학습한 알고리즘으로 분류하여 분류 정확도를 분석하고 잡음과 스펙트럼 이동에 강인한 라만 분광 알고리즘을 찾아서 제안한다.

2. 실험 방법 및 데이터 획득

실험에 사용할 화학물질들의 라만 스펙트럼 측정을 위해 국방과학연구소에서 제작한 라만 분광기를 사용하였다. 라만 분광기는 KrF 엑시머 레이저의 248.68 nm 단자외선을 사용하였고, 라만 전이의 분광범위는 0~3442 cm^{-1}이었다. 라만 스펙트럼은 총 1024개의 라만 산란 파수에서 산란광의 세기를 측정하였으며 각 파수간의 차이는 10 cm^{-1} 정도였다. 각 라만 산란 파수에서의 산란광의 세기는 검출기인 CCD(Charge Coupled Device)의 픽셀의 전자 축적 개수로 결정된다. 따라서 하나의 라만 스펙트럼은 1024차원의 벡터로 이루어져 있다. 분광기가 하나의 라만 스펙트럼을 찍는 속도는 10 Hz이었다. 여기서 사용한 분광기의 자세한 성능은 국방과학연구소에서 쓴 논문[20]에서 찾아볼 수 있다.

라만 스펙트럼을 측정할 화학물질은 국립환경과학원에서 지정한 산업화학물질 및 유독물질 42종을 선정하였다. UV를 흡수해서 측정되는 산란광에의 배경 영향을 최소화하는 물질로 이루어진 cover slip(Duran group, D263) 위에 화학물질 1 μl를 피펫으로 한 방울 떨어뜨리고 라만 분광기로 라만 스펙트럼을 연속해서 100번 측정하였다. 휘발성 화학물질은 3 μl를 피펫으로 한 방울 떨어뜨리고 측정하였다. 정리하자면 42종의 화학물질의 배경 영향을 최소화한 순수 라만 스펙트럼을 각각 100번씩 측정하였다.

배경 영향이 없는 순수 라만 스펙트럼은 물질 고유 스펙트럼과 백색 잡음(white noise)의 합으로 나타낼 수 있으며, 각 라만 전이마다 산란광의 평균 세기와 백색

잡음이 정도가 다르다. 동일한 화학물질의 라만 스펙트럼을 여러 번 찍어서 구한 각 라만 전이에서의 산란광 세기의 평균과 잡음의 표준편차로 해당 물질의 라만 스펙트럼 측정값의 성질을 정의할 수 있다. 즉, 화학물질의 라만 스펙트럼 평균 세기가 $\boldsymbol{s}[i](i = 1, ..., 1024)$이고 잡음의 표준편차가 $\boldsymbol{\sigma}[i](i = 1, ..., 1024)$라고 하자. 그러면 해당 화학물질을 라만 분광기로 측정한 스펙트럼 $\boldsymbol{x}[i](i = 1, ..., 1024)$은 다음의 식으로 나타내어진다.

$$\boldsymbol{x}[i] = \boldsymbol{s}[i] + N(0, \boldsymbol{\sigma}[i]) \quad (1)$$

여기서 $N(0, \boldsymbol{\sigma}[i])$은 평균이 0이고 표준변차가 $\boldsymbol{\sigma}[i]$인 가우시안 분포이다. Fig. 1에 임의로 선정한 두 화학물질 Camphor(1,7,7-Trimethylbicyclo [2.2.1]heptan-2-one)와 Cedrene((1S,2R,5S,7R)-2,6,6,8-tetramethyltricyclo [5.3.1.01,5]undec-8-ene)의 분광기로 찍은 순수 라만 스펙트럼 100개의 평균(mean)과 잡음의 표준편차(std)를 그래프로 Fig. 1에 나타내었다.

42종의 화학물질의 라만 스펙트럼을 100번씩 찍은 데이터를 물질마다 50개는 training set, 50개는 test set으로 분류하였다. 그리고 화학물질마다 training set의 라만 스펙트럼 50개의 평균과 잡음의 표준편차를 구해서 저장하고, test set의 라만 스펙트럼 50개의 평균과 잡음의 표준편차를 구해서 저장하였다. Training set은 기계학습 알고리즘이 학습하는 데이터이고, test set은 알고리즘의 정확도를 검증하는 데이터이다. 기계학습 알고리즘을 잡음과 스펙트럼 이동에 강인하게 만들기 위해서는 강한 잡음과 강한 스펙트럼 이동이 있는 라만 스펙트럼 데이터의 수가 많이 필요하다. 따라서 이미 구한 test set과 training set에 인위적으로 잡음과 스펙트럼 이동을 부여해서 많은 수의 데이터를 획득하였다.

잡음의 세기를 나타내는 잡음 곱계수(noise multiplication factor) α와 스펙트럼 이동의 정도를 나타내는 이동계수 (shift factor) β를 정의하였다. 잡음과 스펙트럼 이동을 주어서 새롭게 생성하는 데이터를 다음과 같이 생성하였다.

$$\boldsymbol{x}[i] = \boldsymbol{s}[i+\gamma] + N(0, \alpha\boldsymbol{\sigma}[i+\gamma]) \quad (2)$$

$$\gamma = U(-\frac{\beta}{10}, \frac{\beta}{10}) \quad (3)$$

입사광의 세기가 커지면 라만 노이즈도 커지므로 $\boldsymbol{s}[i]$와 $\boldsymbol{\sigma}[i]$에 같은 γ를 사용하였다. 여기서 $U(-\frac{\beta}{10}, \frac{\beta}{10})$는 $-\frac{\beta}{10}$와 $\frac{\beta}{10}$사이의 임의의 정수를 뽑는 균등분포이다. 10으로 나누어준 이유는 라만 스펙트럼의 i점과 $i+1$점의 파수 차이가 10 cm^{-1} 였기 때문이다.

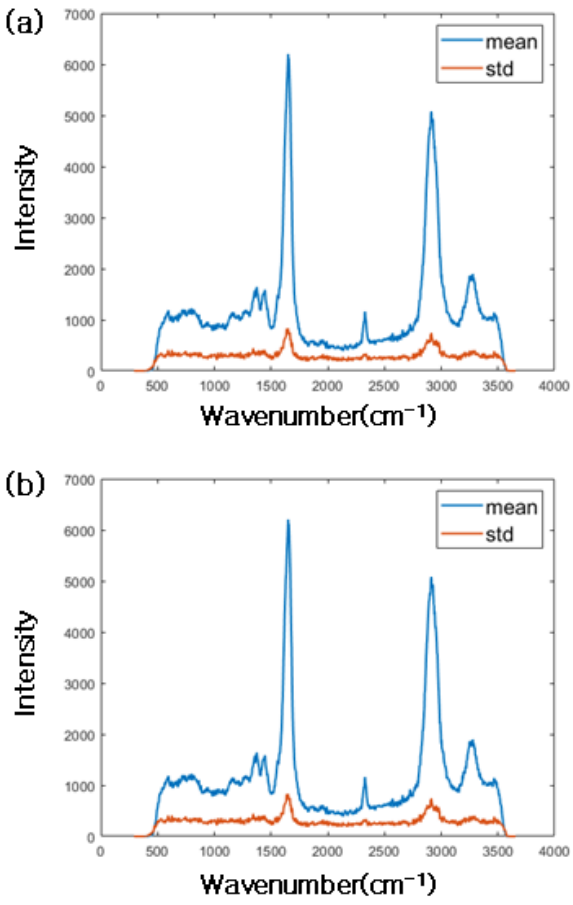

Fig. 1. Mean and noise standard deviation of 100 Raman spectra of (a) Camphor and (b) Cedrene

이후 같은 잡음 곱계수와 이동계수를 주어서 training set의 물질마다 500개의 새로운 라만 스펙트럼을 생성하고, test set의 물질마다 500개의 새로운 라만 스펙트럼을 생성하였다. Training set의 500 × 42개의 라만 스펙트럼으로 기계학습 알고리즘이 라만 스펙트럼으로 화학물질 클래스를 분류하도록 학습하고, test set의

500 × 42개의 라만 스펙트럼으로 기계학습 알고리즘의 정확도(= 제대로 분류한 test 스펙트럼 개수 / 총 test 스펙트럼 개수)를 측정하였다. 여기서 화학물질 클래스란 화학물질의 종류를 의미하며 총 42개이다. 잡음 곱계수 α와 이동계수 β를 다양하게 변화시키며 여러 기계학습 알고리즘의 라만 스펙트럼 분류 정확도를 비교하여 CNN이 가장 잡음과 스펙트럼 이동에 대한 강인성이 높음을 확인하였다.

3. 기계학습 알고리즘

라만 스펙트럼의 물질 분류에 대한 잡음과 스펙트럼 이동 강인성을 가진 기계학습 알고리즘을 찾기 위해서 K-최근접 이웃, 결정 트리, 선형판별분석, 선형 SVM, 비선형 SVM, CNN 알고리즘을 사용하였다. 알고리즘들에 대한 자세한 설명은 전공도서[16]에 있으며 여기에 간단히 설명한다. 알고리즘 학습을 위한 training set은 앞에서 설명한 잡음과 스펙트럼 이동이 부여된 라만 스펙트럼 데이터로, 42개의 클래스로 분류되어 있고 클래스마다 500개씩 라만 스펙트럼이 있다.

K-최근접 이웃은 새로운 다변수 벡터가 입력값으로 들어왔을 때 해당 벡터에서 가장 가까운 k개의 점을 training set에서 찾고 찾은 점 중에서 가장 많이 들어있는 클래스를 선택하는 알고리즘이다. 같은 수의 클래스가 들어있는 경우 등에는 가장 가까운 거리로 판별한다든지 다양한 방법이 만들어져있다.

결정 트리는 특정 기준(질문)에 따라 데이터를 구분하는 모델이다. 예를 들어서 라만 스펙트럼 데이터인 1024 × 1 벡터로 이루어진 training set의 경우 '벡터의 1번째 요소가 특정 값 이상인가?'라는 질문에 따라 2개의 node로 나뉘고, 그 node들이 또 '벡터의 2번째 요소가 특정 값 이상인가?'라는 질문으로 총 4개의 node로 나뉘게 된다. Node를 나누는 기준은 엔트로피를 최소화하는 방향으로 정해지며 이 기준을 정하는 것이 학습이다. 엔트로피를 구하는 공식은 아래와 같다.

$$Entropy = -\sum_i (p_i)\log_2(p_i) \tag{4}$$

여기서 p_i는 한 node 안에 존재하는 데이터 가운데 클래스 i에 속하는 데이터의 비율이다.

선형판별분석은 주성분 분석과 비슷한 차원 축소 방법의 하나이다. 이 경우에는 데이터의 클래스를 one-hot vector로 나타내는데 이 논문에서 사용하는 데이터의 경우 42 × 1 벡터이고 해당 클래스에 해당하는 성분만 1이고 나머지 성분은 0이다. 라만 스펙트럼이 x값, 클래스 one-hot vector가 y값이라면 선형판별분석은 training set에서 클래스 간의 분산은 최대한 크게 가져가고, 클래스 내부의 분산은 최대한 작게 가져가는 축들을 찾아서 y와 x와의 선형 관계를 찾는 방법이다.

SVM은 두 데이터 집합을 나누는 초평면인 결정 경계(decision boundary)를 찾는 방법으로 결정 경계와 데이터의 거리가 최대화되도록 학습하는 이진 분류기이다. 초평면 H는 다음과 같이 정의된다.

$$H:\ W^T x + b = 0 \tag{5}$$

이때 W^T는 결정 경계의 법선 벡터이다. 학습된 SVM에 새로운 데이터 벡터 x가 입력되면 아래의

$$pred = sign(W^T x + b) \tag{6}$$

를 계산해서 sign함수의 값이 0보다 큰지 작은지로 클래스를 구분한다. 하지만 데이터의 모든 점을 하나의 초평면으로 구분하지 못 하는 경우가 존재한다. 이때 사용하는 것이 비선형 SVM이다. 비선형 SVM은 커널 함수를 통해 데이터의 차원을 고차원으로 올려서 초평면을 찾고 분류를 수행한다. 커널 함수로는 다항식 커널, 시그모이드 커널 등이 있다. 이 논문에서는 3차 다항식 커널을 사용한다. SVM은 이진 분류밖에 할 수 없지만, SVM을 응용해서 여러 클래스를 분류하는 방법이 존재한다. 하나-나머지 방법에 따르면 클래스 1과 나머지, 클래스 2와 나머지 등을 분류하는 분류기를 클래수 개수만큼 만들고, 새로운 입력 데이터 x에 대한 이항 분류 값이 가장 큰 모델의 결과를 예측값으로 삼는다.

CNN은 신경망의 일종이다. 신경망이란 입력 데이와 출력 데이터 사이의 관계를 생명체의 신경 구조를 모방한 알고리즘을 이용해서 학습한다. CNN은 신경망 중에서도 hidden layer에 합성곱(convolution)을 사용한 알고리즘을 뜻하며 이미지 분류에 탁월한 성능을 발휘한다. CNN은 데이터 클래스를 one-hot vector로 나타내서 라만 스펙트럼과 one-hot vector 사이의 관계함수를 학습한다. 이 논문에서 사용한 신경망 구조는

Table 1과 같다. 신경망의 training set 학습에 사용한 최적화 방법은 손실 함수로 cross-entropy를 사용하였고, Adam optimizer로 최적화하였다.

Table 1. CNN architecture used in this paper

Layer	구조
Input layer	1024개의 노드
Hidden layer 1	1. 21개의 8 X 1인 합성곱 layer 2. Batch normalization + ReLU 3. maxPooling 2, stride 2
Hidden layer 2	1. 11개의 16 X 1인 합성곱 layer 2. Batch normalization + ReLU 3. maxPooling 2, stride 2
Hidden layer 3	1. 32개의 5 X 1인 합성곱 layer 2. Batch normalization + ReLU 3. maxPooling 2, stride 2
Hidden layer 4	1. 1024 node의 fully connected layer 2. Batch normalization + ReLU
Output layer	1. 1024 node의 fully connected layer 2. Softmax

모든 라만 스펙트럼은 기계학습 알고리즘에 입력하기 전에 최대값으로 나눠서 normalization을 진행하였다. 이것은 너무 큰 값이 들어가면 몇몇 기계학습 알고리즘이 최적화 과정에서 발산할 수 있기 때문이다.

이 논문에서는 라만 스펙트럼 데이터의 잡음의 정도를 나타내는 잡음 곱계수와 스펙트럼 이동의 정도를 나타내는 이동계수를 다양하게 변화시키며 잡음과 스펙트럼이 이동이 들어간 데이터를 생성했다. 그리고 이 데이터 중에서 training set으로 위에서 설명한 기계학습 알고리즘을 학습시키고 test set으로 화학물질 클래스를 분류해서 기계학습 알고리즘 간의 정확도를 비교하고 잡음과 스펙트럼 이동에 강인한 알고리즘을 찾았다.

4. 학습 결과

먼저 잡음이 강할 때 기계학습 알고리즘이 제대로 분류할 수 있는지에 대한 잡음 강인성을 확인하였다. 스펙트럼 이동이 없는 상태에서 잡음 곱계수 α를 1, 2, 3으로 변화시키며 라만 스펙트럼에 잡음을 부여하고 만든 데이터로 학습하고 검증해서 기계학습 알고리즘의 분류 성능을 확인하였으며 Table 2에 정확도를 나타내었다.

Table 2. Classification accuracy of machine learning algorithms in case of noise

알고리즘	α = 1	α = 2	α = 3
K-최근접 이웃	**0.9732**	0.8192	0.6749
결정 트리	0.8293	0.6680	0.4978
선형판별분석	**0.9890**	0.9323	0.9040
선형 SVM	0.9122	0.9095	0.8928
비선형 SVM	0.9155	0.9180	0.8810
CNN	**0.9761**	**0.9737**	**0.9740**

잡음이 작을 때는 선형판별분석, CNN, K-최근접 이웃의 정확도가 높게 나왔다. 하지만 잡음이 증가하면 다른 알고리즘들은 정확도가 낮아지지만, CNN은 여전히 높은 정확도를 보인다.

다음으로 스펙트럼 이동이 강할 때 기계학습 알고리즘이 제대로 분류할 수 있는지에 대한 스펙트럼 이동 강인성을 확인해보았다. 잡음 곱계수 α를 1로 해서 일반적인 환경잡음이 있는 상태에서 이동계수 β를 32, 64, 96으로 변화시키며 라만 스펙트럼에 스펙트럼 이동을 부여하고 만든 데이터로 기계학습 알고리즘의 분류 성능을 확인하였고 Table 3에 정확도를 나타내었다. β의 크기는 논문[14]에서 압력에 의해 변할 수 있는 스펙트럼 이동의 정도를 보고 참고하였다.

Table 3. Classification accuracy of machine learning algorithms in case of spectral shift

알고리즘	β = 32	β = 64	β = 96
K-최근접 이웃	0.9480	0.9540	0.9472
결정 트리	0.8357	0.8056	0.7794
선형판별분석	**0.9844**	**0.9755**	0.9492
선형 SVM	0.9038	0.8998	0.9020
비선형 SVM	0.9085	0.9218	0.9102
CNN	**0.9744**	**0.9693**	**0.9664**

스펙트럼 이동이 있을 때 CNN과 선형판별분석이 높은 정확도를 기록하였다. 매우 큰 스펙트럼 이동에서는 CNN이 가장 높은 정확도를 보였다.

Table 4. Classification accuracy of machine learning algorithms in case of spectral shift when α = 2

알고리즘	β = 32	β = 64	β = 96
K-최근접 이웃	0.8016	0.8072	0.7849
결정 트리	0.7092	0.6671	0.6249
선형판별분석	0.9255	0.9048	0.8842
선형 SVM	0.9038	0.8985	0.8911
비선형 SVM	0.9026	0.8999	0.8937
CNN	**0.9683**	**0.9605**	**0.9547**

Table 5. Classification accuracy of machine learning algorithms in case of spectral shift when α = 3

알고리즘	β = 32	β = 64	β = 96
K-최근접 이웃	0.6486	0.6412	0.6384
결정 트리	0.6066	0.5545	0.5227
선형판별분석	0.8914	0.8707	0.8511
선형 SVM	0.8886	0.8823	0.8720
비선형 SVM	0.8836	0.8673	0.8515
CNN	**0.9688**	**0.9638**	**0.9558**

Table 6. Classification accuracy of machine learning algorithms in case of noise when β = 32

알고리즘	α = 2	α = 3
K-최근접 이웃	0.8016	0.6486
결정 트리	0.7092	0.6066
선형판별분석	0.9255	0.8914
선형 SVM	0.9038	0.8886
비선형 SVM	0.9026	0.8836
CNN	**0.9683**	**0.9688**

이제는 잡음과 스펙트럼 이동이 모두 있을 때 기계학습 알고리즘이 제대로 분류할 수 있는지에 대해서 확인해보았다. 잡음 곱계수 α가 2, 3으로 잡음이 많을 때 각각 이동계수 β를 32, 64, 96로 해서 스펙트럼 이동을 부여하고 만든 데이터로 기계학습 알고리즘의 분류 정확도를 구하고 Table 4-8에 나타내었다.

Table 7. Classification accuracy of machine learning algorithms in case of noise when β = 64

알고리즘	α = 2, β = 64	α = 3, β = 64
K-최근접 이웃	0.8072	0.6412
결정 트리	0.6671	0.5545
선형판별분석	0.9048	0.8707
선형 SVM	0.8985	0.8823
비선형 SVM	0.8999	0.8673
CNN	**0.9605**	**0.9638**

Table 8. Classification accuracy of machine learning algorithms in case of noise when β = 96

알고리즘	α = 2, β = 96	α = 3, β = 96
K-최근접 이웃	0.7849	0.6384
결정 트리	0.6249	0.5227
선형판별분석	0.8842	0.8511
선형 SVM	0.8911	0.8720
비선형 SVM	0.8937	0.8515
CNN	**0.9547**	**0.9558**

잡음과 스펙트럼 이동이 모두 있을 때는 CNN만이 95 % 이상의 분류 정확도를 보여주었다. CNN이 다른 알고리즘에 비해서 잡음과 스펙트럼 이동이 있음에도 화학물질 클래스 분류에 높은 정확도를 보이는 이유는 CNN이 비선형 함수를 모사하는 능력이 가장 뛰어나기 때문이다. 화학물질 클래스 분류는 독립변수로 화학물질의 라만 스펙트럼을 받고 종속 변수로 화학물질의 종류를 one-hot vector로 받는 비선형 함수로 생각할 수 있다. 이 함수는 비선형성이 매우 강하기 때문에 생명체의 신경망을 모사한 CNN이 분류 함수

를 가장 잘 모사할 수 있다.

마지막으로 기계학습 알고리즘의 성능은 학습에 사용하는 training set에 들어있는 데이터 개수의 영향을 받으므로 잡음 곱계수 α = 3, 이동계수 β = 96일 때 training set의 개수를 변화시키며 알고리즘을 학습시키고, 학습된 알고리즘으로 test set의 데이터를 분류해서 정확도를 확인하였다. Training set 데이터의 개수를 물질당 400개, 500개, 600개로 변화시키며 데이터를 만들었다. 물질당 데이터 100개가 증가하는 것은 실제로는 전체 training set의 데이터가 4200개 증가하는 것이다. Test set 데이터의 개수는 물질당 500개로 유지하였다. 분류 결과를 Table 9에 표시하였다.

Table 9. Classification accuracy of machine learning algorithms in case of different training set data number

알고리즘	400개	500개	600개
K-최근접 이웃	0.6314	0.6384	0.6363
결정 트리	0.5388	0.5227	0.5560
선형판별분석	0.8444	0.8511	0.8472
선형 SVM	0.8694	0.8720	0.8822
비선형 SVM	0.8517	0.8515	0.8385
CNN	**0.9557**	**0.9558**	**0.9613**

Training set의 데이터를 변화시켜도 각 알고리즘의 정확도가 많이 바뀌지 않았고 CNN이 여전히 가장 높은 정확도를 보인다. 따라서 CNN이 가장 잡음 강인성과 스펙트럼 이동 강인성이 높은 기계학습 알고리즘으로 보인다. 강한 잡음과 높은 스펙트럼 이동이 화학물질 스펙트럼에 있음에도 CNN은 95 % 이상의 화학물질 분류 성능을 유지하였으므로 강인성이 매우 높은 알고리즘이라고 할 수 있다.

5. 결 론

라만 분광기는 기존 접촉식 물질 분석 장비의 단점인 장비의 오염이 없기에 화생방 작전 등에서 물질 분류에 많이 사용되는 장비이다. 하지만 암전류, 배경에 의한 잡음과 장비의 떨림, 압력변화 등에 의한 스펙트럼 이동으로 인해 분류 성능이 나빠질 수 있다. 이 논문에서는 다양한 기계학습 알고리즘을 이용해서 잡음과 스펙트럼 이동이 들어간 데이터를 학습시키고 분류 성능을 확인하였다. CNN(합성곱 신경망)이 잡음과 스펙트럼 이동이 있음에도 95 % 이상의 높은 화학물질 분류 정확도를 유지하였다. 따라서 CNN을 이용한 분류가 가장 잡음 강인성과 스펙트럼 이동 강인성이 높음을 알 수 있다. 이는 실제 전투상황같이 움직임으로 인해 떨림이 심하고 파편, 연기 등에서 나오는 빛에 의한 잡음이 높은 상황에서 CNN이 좋은 분류 성능을 보일 수 있다는 것을 보여준다.

후 기

본 연구는 국방과학연구소의 연구비 지원으로 수행되었습니다.(계약번호 UD190007GD)

References

[1] R. L. McCreery, “Raman Spectroscopy for Chemical Analysis,” John Wiley & Sons, Canada, pp. 15-34, 2000.

[2] M. A. Slamani, B. Fisk, T. Chyba, D. Emge and S. Waugh, “A Algorithm Benchmark Data Suite for Chemical and Biological(Chem/Bio) Defense Applications,” SPIE Defense and Security Symposium, Vol. 6969, No. 696903, 2008.

[3] R. D. Palkki, and A. D. Lanterman, “Identifying Chemicals from Their Raman Spectra Using Minimum Description Length,” Proc. of SPIE, Vol. 7698, Signal and Data Processing of Small Targets 2010, 769807, Apr. 2010.

[4] S. H. Melfi, D. N. Whiteman, and R. A. Ferrare, “Observation of Atmospheric Fronts Using Raman Lidar Moisture Measurements,” Journal of Applied Meteorology, Vol. 28, pp. 789-806, 1989.

[5] K. Xiiong, “UV Resonance Raman Spectroscopy: A Highly Sensitive, Selective and Fast Technique for Environmental Analysis,” Environmental Analytial Chemistry, Vol. 2, No. 1, 2014.

[6] F. Foucher, G. Guimbretiere, N. Bost, and F. Westall,

"Petrographical and Mineralogical Applications of Raman Mapping," Raman Spectroscopy and Applications, Khan Maaz, IntechOpen, Feb. 2017.

[7] E. V. Efremov, F. Ariese, and C. Gooijer, "Achievements in Resonance Raman Spectroscopy Review of a Technique with a Distinct Analytical Chemistry Potential," Analytica Chimica Acta 606, pp. 119-134, 2008.

[8] Y. J. Koh, "The Design and Test of the Stand-off Surface Chemical Contaminant Detection System based on Spectroscopy," Journal of the KIMST, Vol. 22, No. 3, pp. 433-440, 2019.

[9] S. K. Choi, Y. S. Jeong, J. H. Lee, and Y. C. Ha, "Deep UV Raman Spectroscopic Study for the Standoff Detection of Chemical Warfare Agents from the Agent-Contaminated Ground Surface," Journal of the KIMST, Vol. 18, No. 5, pp. 612-620, 2015.

[10] A. J. Sedlacek, M. Ray, S. Higdon, and D. Richter, "Short-Range, Non-Contact Detection of Surface Contamination Using Raman Lidar," Proc. of SPIE, Vol. 4577, 2002.

[11] P. Ponsardin, S. Higdon, T. H. Chyba, W. T. Armstrong, A. J. Sedlacek III, S. Christesen, and A. Wong, "Expanding Applications for Surface-Contaminant Sensing Using the Laser Interrogation of Surface Agents(LISA) Technique," Proc. of SPIE, 2004.

[12] J. Smulko, M. S. Wróbel, and I, Barman, "Noise in Biological Raman Spectroscopy," International Conference on Noise and Fluctuations, 2015.

[13] H. Chen, W. Xu, N. Broderick, and J. Han, "An Adaptive Denoising Method for Raman Spectroscopy based on Lifting Wavelet Transform," Journal of Raman Spectroscopy, Vol. 49, Issue. 9, June, 2018.

[14] S. A. Kirillov, "Repulsion Forces in Vibrational Spectroscopy - I. Spectral Shifts in Vibrational Spectra of Condensed Media Caused by Repulsion Forces," Spectrochimica Acta, Vol. 48A, No. 6, pp. 861-866, 1992.

[15] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of Multidimensional Data Processing Approaches for Raman and Infrared Spectrocsopy," Vol. 2, No. 1, pp. 1-38, 2015, EPJ Techniques and Instrumentation.

[16] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer Science + Business Media, LLC, 2006.

[17] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, and M. D. Dyar, "Machine Learning Tools for Mineral Recognition and Classification from Raman Spectroscopy," Journal of Raman Spectroscopy, Vol. 46, Issue 10, pp. 894-903, 2015.

[18] J. dong, M. Hong, Y, xu, and X. Zheng, "A Practical Convolutional Neural Network Model for Discriminating Raman Spectra of Human and Animal Blood," Journal of Chemometrics, Vol. 33, Issue 11, 2019.

[19] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press Cambridge, Massachusetts, pp. 326-366, 2016.

[20] Y. C. Ha, J. H. Lee, Y. J. Koh, S. K. Lee, and Y. K. Kim, "Development of an Ultraviolet Raman Spectrometer for Standoff Detection of Chemicals," Current Optics and Photonics, Vol. 1, No. 3, pp. 247-251, June, 2017.