

## 深層学習を用いた共通特徴量抽出による蛍光指紋識別 Excitation-Emission Matrix Classification by Common Feature Extraction using Deep Learning

林田 純弥<sup>†</sup> 柿下 容弓<sup>†</sup> 服部 英春<sup>†</sup>  
Junya Hayashida Yasuki Kakishita Hideharu Hattori

### 1. はじめに

蛍光指紋(Excitation-Emission Matrix, EEM)とは、分光蛍光光度計を用いて、波長を変えながら試料に励起光を照射し、試料から放出される蛍光スペクトルを記録した等高線状データである。蛍光指紋は、試料の品質の解析等に用いられる。

現行の蛍光指紋識別手法には、Lasso 回帰[1]と呼ばれる線形回帰手法が主に用いられる。また、近年、深層学習モデルによる蛍光指紋解析手法が提案されている[2][3]。

Lasso 回帰は、蛍光指紋を構成する波長帯の内、少数の波長帯のみを識別に使用するように正則化をかけた、識別モデルを構築する手法である。以降、識別に使用される波長帯を有効波長帯と呼ぶ。識別モデル構築後は、有効波長帯の蛍光強度のみを測定することで、識別が可能である。Lasso 回帰は、線形回帰モデルであるため、CPU 環境で高速に実行可能であるものの、非線形回帰モデルや深層学習モデルと比較すると表現力が乏しい。そのため、十分な識別精度を得るために手動での前処理(特徴量の設計や、波長帯の選択等)が必要となる場合がある。

一方、深層学習モデルは、複数の層に渡って線形/非線形写像を繰り返すことで、線形回帰モデルでは捉えられない複雑な特徴量を自動学習することが可能となり、高精度化が期待できる。しかし、深層学習モデルは、識別タスクに基づいて、大量のパラメータを最適化する必要があり、計算コストが高い。また、一般的な深層学習モデルでは、蛍光指紋全体から、様々なパラメータを複数回掛け合わせることで特徴量を抽出するため、有効波長帯を限定することが難しいという課題も存在する。

そこで、本稿では、深層学習を用いた共通特徴量抽出器と、Lasso 回帰を連携した蛍光指紋識別手法を提案する。図 1 に提案する蛍光指紋識別手法の概要図を記載する。

共通特徴量抽出器(Common feature extractor)とは、類似タスク(識別対象データの分布が共通特徴量抽出器の学習データと類似した識別タスク)に対し、共通的に利用可能な特徴量抽出器である。図 1 に示すように、2 クラス識別タスクの情報として、正例(Positive)と負例(Negative)の蛍光指紋集合であるコンテキスト集合(Context set)を与えることで、正例と負例を識別するために必要な特徴量(共通特徴量)を、識別対象の蛍光指紋(Target EEM)から抽出する。このように、共通特徴量抽出器は、コンテキスト集合に応じて、蛍光指紋から適応的に特徴量を抽出するモデルである。そのため、学習済みの共通特徴量抽出器を用いることで、類似タスクに対しても、追加の学習を介さずに、類似タスクのコンテキスト集合から適応的に特徴量を抽出することが可

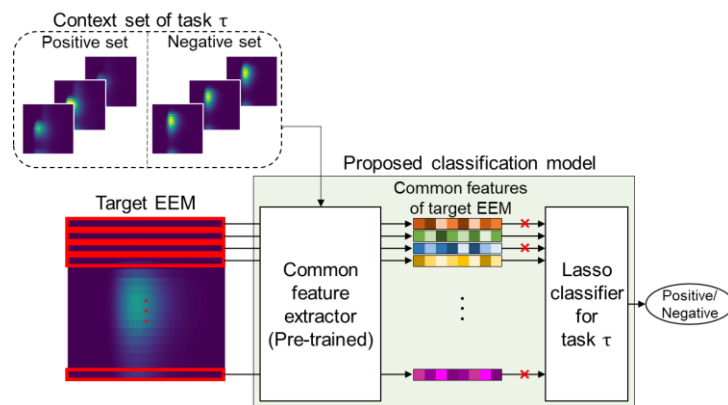


図 1 共通特徴量抽出器を用いた蛍光指紋識別手法

能である。これにより、深層学習モデル利用における学習コストの低減を実現する。

対象となる蛍光指紋から共通特徴量を抽出した後、Lasso 識別器(Lasso classifier)に入力することで、識別を行う。Lasso 識別器は、識別タスク毎に Lasso 回帰を用いて構築する。Lasso 識別器の構築は深層学習とは異なり、CPU のみでも短時間で実行することが可能である。

更に、励起波長帯単位(図 1 赤枠単位)で特徴量を抽出するように共通特徴量抽出器を設計することで、Lasso 識別器で使用する有効波長帯を励起波長帯単位で限定可能とした。以降、励起波長帯単位の有効波長帯を、有効励起波長帯と呼ぶ。

このように、提案手法は、新規の識別タスクにおいても、学習を介さずに深層学習モデルによる特徴量抽出を実現する共通特徴量抽出器と、低い計算コストで識別を行う Lasso 識別器を用いることで、高精度な識別と学習コスト低減を両立する。更に、励起波長帯単位で共通特徴量を抽出することで、Lasso 識別器で使用する励起波長帯を限定する。運用時は有効励起波長帯のみを測定した蛍光指紋で識別が可能のため、蛍光指紋の測定時間短縮が可能となる。

試料に含まれる化合物を識別する実験において、一部の化合物情報のみで学習した共通特徴量抽出器を用いて、各成分の Lasso 識別器を構築し、識別を行った結果、従来手法(Lasso 回帰)と比較して、提案手法は高精度かつ少ない有効励起波長帯数での識別を達成した。

### 2. 蛍光指紋について

蛍光指紋とは、分光蛍光光度計を用いて測定される、励起波長(Excitation)、蛍光波長(Emission)、蛍光強度(Intensity)の 3 軸で構成される等高線状のデータである。本研究で用いる蛍光指紋の可視化例を図 2 に示す。図 2(a)は測定後の蛍光指紋の可視化例、図 2(b)は図 2(a)に対して前処理を適用した結果であり、横軸を蛍光波長、縦軸を励起波長とし、蛍光強度を色で表現したヒートマップ画像形式

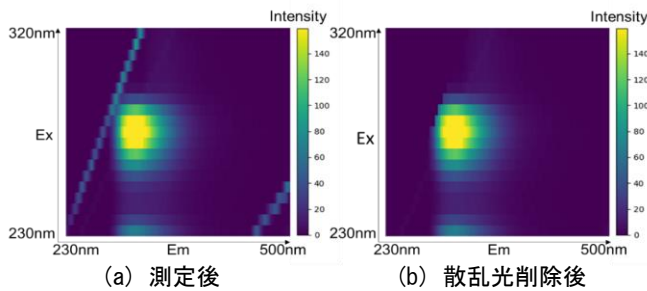


図 2 蛍光指紋の可視化例

となっている。蛍光指紋の形状やピークを解析することで、試料の性質の推定が行われる。

1 つの励起波長帯における、蛍光波長帯毎の蛍光強度のスペクトルを蛍光スペクトル、1 つの蛍光波長帯における励起波長帯毎の蛍光強度のスペクトルを励起スペクトルと呼ぶ。試料に対し、照射する励起波長帯を変化させながら蛍光スペクトルを測定する、あるいは受光する蛍光波長帯を変化させながら励起スペクトル単位での測定を繰り返すことで、蛍光指紋が得られる。そのため、測定する励起波長帯あるいは蛍光波長帯を限定することが、蛍光指紋の測定時間短縮に繋がる。提案手法においては、蛍光スペクトル単位(励起波長帯単位)で深層学習を用いた特徴抽出を行うことで、有効励起波長帯の限定を可能とする。

図 2(a)において、蛍光指紋は、励起波長 $\geq$ 蛍光波長となる非蛍光領域、励起光が試料表面で反射しそのまま検出された、励起波長=蛍光波長 $\pm \beta$ の散乱光、および回折格子によって生じる散乱光の 2、3、 $\dots$ 、 $n$ 次光(励起波長 $=n \times$ 蛍光波長 $\pm \beta$ および  $1/n \times$ 蛍光波長 $\pm \beta$ )を含む。ただし、 $\beta$ は散乱光の帯域幅を表すハイパーパラメータである。非蛍光領域および散乱光は、試料が放出する蛍光とは異なり、試料特性に関係しないノイズを含むため、図 2(b)のように、解析前に削除することが一般的である。今回の実験における前処理については 5.1 節にて説明する。

### 3. 関連手法

#### 3.1 Lasso 回帰

Lasso 回帰[1]とは、線形回帰手法の一つであり、目的関数に正則化項を追加することで、幾つかの係数が 0 に設定される、スパース性を持つことを特徴とする。0 と推定された係数に対応する説明変数は、推定に用いる必要が無いため、推定時には該当する説明変数を測定不要となる。つまり、Lasso 回帰は係数推定と、説明変数の選択を同時に行う回帰手法である。

従来手法である Lasso 回帰を用いた蛍光指紋の識別手法においては、訓練用の蛍光指紋に対し、各蛍光強度を説明変数、試料が属するクラスを目的変数として、変数選択を行う。運用時には、評価対象の試料に対し、選択された有効波長帯の蛍光強度のみを測定し、定量や識別を行うことで、測定時間の短縮が可能となる。

#### 3.2 深層学習

蛍光指紋に対して、深層学習を用いた回帰/識別手法が提案されている。例えば Itakura ら[2]は、CNN(Convolutional Neural Network)[4]を用いて柑橘類の蛍光指紋から、成熟度

の推定を行う手法を提案しており、Rutherford[3]らは、CNNを用いて、蛍光指紋から空気中の粒子状物質の発生源(排気ガスや煙草等)を特定する手法を提案している。しかし、CNNのような深層学習モデルの場合、複数回に渡って線形/非線形変換を繰り返すことで蛍光指紋全体から特徴量を抽出するため、有効波長帯を限定することが難しい。また、タスクに応じて CNN のパラメータを学習する必要があり、学習には高性能の計算機が必要となる場合があるため、計算コストが高いという課題がある。

### 4. 提案手法

#### 4.1 概要

有効励起波長帯を選択可能かつ、類似タスクに共通利用可能な特徴量抽出器(共通特徴量抽出器)、および抽出した特徴量を用いた蛍光指紋識別手法を提案する。提案手法は大きく二つの要素から構成されている。一つは共通特徴量抽出器、もう一つは Lasso 識別器である。

図 3 は提案手法の概要を示した図であり、図 3(a)は共通特徴量抽出器の学習、図 3(b)は、ある識別タスク  $\epsilon$  に対する Lasso 識別器の構築、図 3(c)はタスク  $\epsilon$  に対する蛍光指紋の識別について記しており、提案手法は図 3(a)、(b)、(c)の順番で構築および運用される。

共通特徴量抽出器は、様々な 2 クラス識別タスクの蛍光指紋を用いて学習を行うことで、類似タスクに対しても、追加の学習を介さずに、適応的に特徴量を抽出することを目的とする。図 3(a)は、学習に用いる複数の 2 クラス識別タスクの一つである識別タスク  $\tau$  の蛍光指紋集合を入力した際の共通特徴量抽出器の学習過程を示した図である。まず、タスク  $\tau$  の情報として、コンテキスト集合を与え、コンテキスト特徴量(Context feature)を抽出する。その後、ターゲット集合(Target set)を入力し、コンテキスト特徴量に応じて、適応的に特徴量を抽出する。得られた特徴量が、正例と負例の間で異なり、かつ、正例内および負例内で類似した特徴を持つように学習を行う。

図 3(b)は、学習済みの共通特徴量抽出器を用いた、識別タスク  $\epsilon$  に対する Lasso 識別器の構築を記した図である。識別タスク  $\epsilon$  は共通特徴量抽出器の学習に用いたタスクとは別のタスクを想定している。共通特徴量抽出器に対しては新たな学習を行わず、Lasso 識別器の学習のみを行うことで、少ない計算コストでタスクに応じた識別器を作成する。また、共通特徴量を励起波長単位で抽出することで、Lasso 識別器において使用する特徴量を励起波長単位で取捨選択できるように設計した。識別に使用する励起波長帯(有効励起波長帯)が少ないほど、蛍光指紋の測定時間削減に繋がる。

図 3(c) は、構築した Lasso 識別器と、その際に抽出したコンテキスト特徴量を用いた蛍光指紋識別について示した図である。識別の際、Lasso 識別器の構築過程で特定した有効励起波長帯のみを測定することで測定時間を短縮する。

4.2 節において、共通特徴量抽出器の学習方法、4.3 節において、共通特徴量を用いた Lasso 識別器の構築方法、4.4 節において、共通特徴量器および Lasso 識別器を用いた識別方法について各々説明を行う。

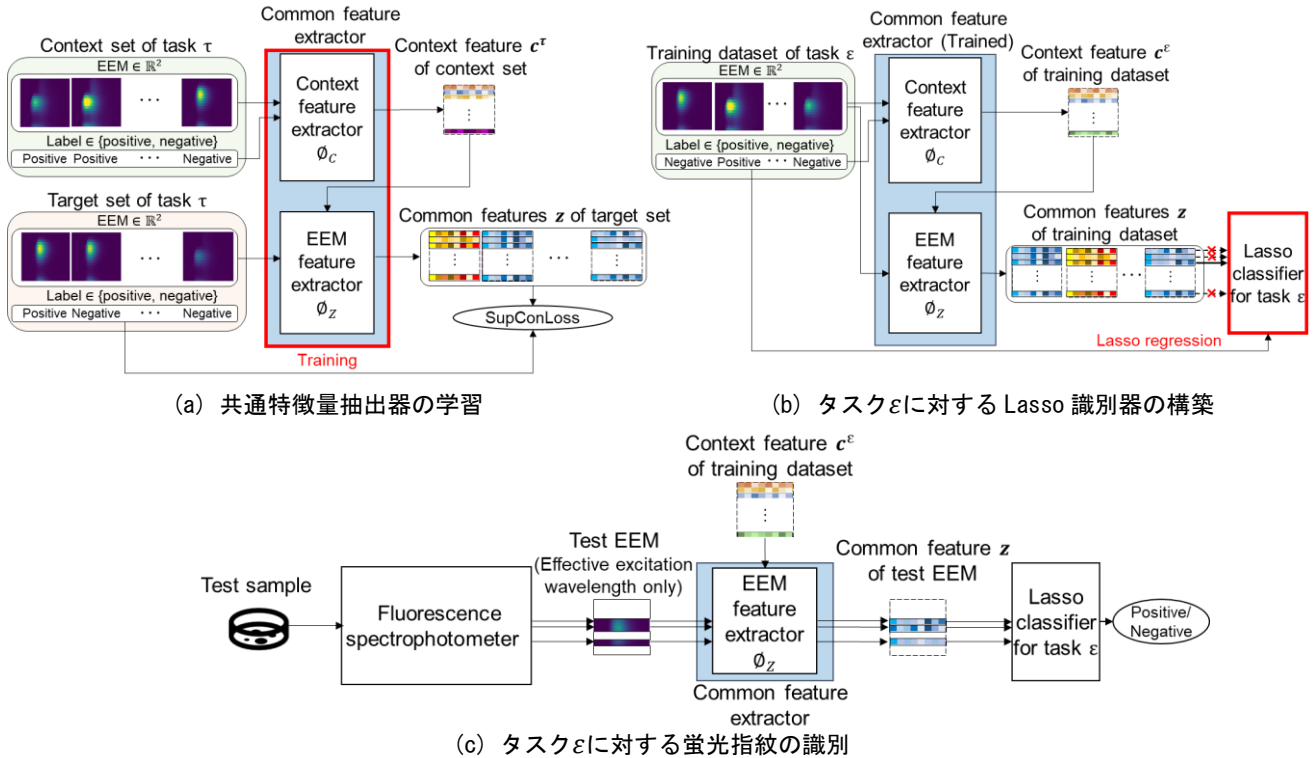


図 3 提案手法の概要図

## 4.2 共通特徴量抽出器の学習

共通特徴量抽出器の学習について説明する。図 3(a)は、ある 2 クラス識別タスク  $\tau$  の蛍光指紋集合を入力した際の、共通特徴量抽出器の学習過程を示した図である。図 3(a)に示すように、共通特徴量抽出器は、コンテキスト集合という形で、2 クラス識別タスクの情報を与え、コンテキスト集合から適応的に特徴量を抽出するモデルである。様々な 2 クラス識別タスクを用いて、共通特徴量抽出器が、それぞれのタスクから適応的に特徴量を抽出するように学習を行う。

学習には、タスク  $\tau$  の蛍光指紋と正例/負例ラベルのペアの集合であるコンテキスト集合と、特徴量抽出対象の蛍光指紋集合であるターゲット集合を用いる。まず、タスク  $\tau$  の情報として、コンテキスト集合を与え、正例蛍光指紋と負例蛍光指紋間の関係性(コンテキスト)を抽出した特徴量(コンテキスト特徴量)を抽出する。その後、コンテキスト特徴量に基づいて、ターゲット集合の各蛍光指紋から、タスク  $\tau$  の識別に必要な特徴量(共通特徴量)を抽出する。共通特徴量が正例と負例の間で異なり、かつ、正例内および負例内で類似し、Lasso 回帰のような線形回帰手法で識別可能な特徴量になるように、後述する損失関数(SupConLoss)を用いて、共通特徴量抽出器を学習する。

上記を、様々な 2 クラス識別タスクの蛍光指紋集合に対して行うことで、共通特徴量抽出器を学習する。ただし、3 クラス以上の識別データセットであっても、クラスを 2 グループに分けることで、2 クラス識別データセットとして扱うことができる。以下、ある 2 クラス識別タスク  $\tau$  における共通特徴量の抽出から損失関数の計算までを説明する。

### 4.2.1 蛍光指紋集合の準備

図 3(a)に示すように、共通特徴量抽出器は、コンテキスト特徴抽出器(Context feature extractor)  $\phi_c$  と、蛍光指紋特徴抽出器 (EEM feature extractor)  $\phi_z$  で構成される。ここで、ある 2 クラス識別タスク  $\tau$  に従って、正例/負例ラベルが付けられた蛍光指紋の集合として、要素数  $N$  のコンテキスト集合  $D_\tau = \{(X_n^\tau, y_n^\tau)\}_{n=1}^N$  と、要素数  $M$  のターゲット集合  $D_{\tau^*} = \{(X_m^*, y_m^*)\}_{m=1}^M$  を定義する。ただし、 $X_n^\tau$ 、 $X_m^*$ 、および  $y_n^\tau$ 、 $y_m^*$  は、コンテキスト集合、ターゲット集合としてサンプリングした蛍光指紋  $X \in \mathbb{R}^2$  および正例/負例ラベル  $y \in \{posi, nega\}$  を示す。コンテキスト集合  $D_\tau$  は、コンテキスト特徴量の抽出に用いる情報であり、ターゲット集合  $D_{\tau^*}$  は、共通特徴量抽出対象となる蛍光指紋集合である。

### 4.2.2 コンテキスト特徴量の抽出

まず、タスク  $\tau$  の情報として、コンテキスト集合  $D_\tau$  をコンテキスト特徴抽出器  $\phi_c$  に入力し、 $D_\tau$  に含まれる正例蛍光指紋と負例蛍光指紋間の関係性を表したコンテキスト特徴量  $c^\tau = (c_1^\tau, c_2^\tau, \dots, c_{E_x}^\tau)$  を得る。 $c_i^\tau$  は励起波長帯  $i$  における  $d$  次元の特徴量ベクトルであり、 $c^\tau$  は、各励起波長帯  $i \in \{1, 2, \dots, E_x\}$  に対応した  $d$  次元ベクトル  $c_i^\tau$  を連結して得られる  $E_x \times d$  次元ベクトルである。 $D_\tau = \{(X_n^\tau, y_n^\tau)\}_{n=1}^N$  と  $c^\tau$  の関係は以下で表される。

$$(c_1^\tau, c_2^\tau, \dots, c_{E_x}^\tau) = \phi_c(\{(X_n^\tau, y_n^\tau)\}_{n=1}^N) \quad (1)$$

$\phi_c$  は、コンテキスト集合を入力として受け付け、要素間の関係性を表すコンテキスト特徴を、励起波長帯毎に抽出する。集合を入力とするため、要素の入れ替えに対して同一の結果を出力する必要がある。これを実現するため、Attention[5]の Encoder layer を用いた。Encoder layer は、集合を入力として受け付け、集合の要素間のコンテキストを抽出し、各要素に 1 対 1 対応した集合を出力する。得られた集合の各要素を平均し、多層パーセプトロンを用いて適

当な非線形写像を行うことで、単一の出力  $\mathbf{c}^{\tau} = (\mathbf{c}_1^{\tau}, \mathbf{c}_2^{\tau}, \dots, \mathbf{c}_{\text{Ex}}^{\tau})$  を得る。

このように、蛍光指紋とラベルの集合から、タスク固有の特徴ベクトルを抽出することで、タスクに応じた共通特徴量をターゲット集合から抽出することをめざす。

#### 4.2.3 共通特徴量の抽出

次に、 $\Phi_c$  より得られたコンテキスト特徴量  $\mathbf{c}^{\tau} = (\mathbf{c}_1^{\tau}, \mathbf{c}_2^{\tau}, \dots, \mathbf{c}_{\text{Ex}}^{\tau})$  と、特徴量抽出対象である蛍光指紋  $\mathbf{X}$  から、 $\Phi_Z$  を用いて励起波長帯  $i \in \{1, 2, \dots, \text{Ex}\}$  毎に  $d$  次元の共通特徴量  $\mathbf{e}_i$  を得る。学習においては、 $D_{\tau^*}$  の蛍光指紋それぞれから共通特徴量を抽出する。励起波長帯  $i$  の蛍光スペクトルを  $\mathbf{X}_i$  とすると、 $\mathbf{c}_i^{\tau}$ 、 $\mathbf{X}_i$  と、 $\mathbf{e}_i$  の関係は以下で表される。

$$\mathbf{e}_i = \Phi_Z(\mathbf{X}_i, \mathbf{c}_i^{\tau}) \quad (2)$$

ここで、各励起波長帯から得られる共通特徴量を連結した  $\text{Ex} \times d$  次元ベクトルを、共通特徴量  $\mathbf{z} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{\text{Ex}})$  と定義すると、 $D_{\tau^*}$  の蛍光指紋から共通特徴量の集合  $Z_{\tau^*} = \{\mathbf{z}_m^{\tau^*}\}_{m=1}^M$  が得られる。 $\mathbf{z}_m^{\tau^*}$  は  $D_{\tau^*}$  の  $m$  番目の要素に対する共通特徴量を示す。

#### 4.2.4 損失関数

$D_{\tau^*}$  から得られた共通特徴量集合  $Z_{\tau^*} = \{\mathbf{z}_m^{\tau^*}\}_{m=1}^M$  に対し、Supervised Contrastive Learning[6]にて用いられる Supervised Contrastive Loss(SupConLoss)を損失関数とした学習を行う。Supervised Contrastive Learning とは、同じクラスに属する特徴量を類似させ、異なるラベルに属する特徴量間で異なるようにニューラルネットワークを学習する手法である。提案手法では、 $D_{\tau^*}$  から得られた共通特徴量集合  $Z_{\tau^*} = \{\mathbf{z}_m^{\tau^*}\}_{m=1}^M$  に対し、以下に示す SupConLoss を用いて、学習を行った。

$$L = \sum_{\mathbf{z}_i^{\tau^*} \in Z_{\tau^*}} -\log \left( \frac{1}{|S(Z_{\tau^*}, \mathbf{z}_i^{\tau^*})|} \sum_{\mathbf{z}_s^{\tau^*} \in S(Z_{\tau^*}, \mathbf{z}_i^{\tau^*})} \frac{\exp(\mathbf{z}_i^{\tau^*} \cdot \mathbf{z}_s^{\tau^*})}{\sum_{\mathbf{z}_a^{\tau^*} \in Z_{\tau^*} \setminus \{\mathbf{z}_i^{\tau^*}\}} \exp(\mathbf{z}_i^{\tau^*} \cdot \mathbf{z}_a^{\tau^*})} \right) \quad (3)$$

ここで、 $S(Z_{\tau^*}, \mathbf{z}_i^{\tau^*})$  は、共通特徴量集合  $Z_{\tau^*}$  の内、 $\mathbf{z}_i^{\tau^*}$  に対応するラベル(正例か負例)と同じラベルを持つ要素( $\mathbf{z}_i^{\tau^*}$  自身は除く)を抽出する処理を示す。同じラベルを持つ共通特徴量同士の内積が大きくなるほど、また、異なるラベルを持つ共通特徴量同士の内積が小さくなるほど、 $L$  は小さくなる。つまり、共通特徴量が、正例と負例の間で異なり、かつ、正例内および負例内で類似した特徴を持つように学習を進める損失関数となっている。様々な 2 クラス識別タスクを用いて、式(3)による共通特徴量抽出器の学習を行う。

#### 4.3 Lasso 識別器の構築

次に、学習後の共通特徴量抽出器を用いて、Lasso 回帰による識別器(Lasso 識別器)を構築する。図 3(b)は、ある識別タスク  $\varepsilon$  に対する、Lasso 識別器の構築を示した図である。識別タスク  $\varepsilon$  の訓練用データセット  $D_{\varepsilon} = \{(\mathbf{X}_n^{\varepsilon}, \mathbf{y}_n^{\varepsilon})\}_{n=1}^{N_{\varepsilon}}$  をコンテキスト集合かつターゲット集合とし、学習済みの共通特徴量抽出器から各訓練データの共通特徴量を励起波長帯単位で抽出する。得られた励起波長帯単位の共通特徴量を説明変数とした Lasso 回帰を行うことで、有効励起波長帯を限定しながら Lasso 識別器を作成する。図 3(c)に示す識別過程においては、試料に対し、有効励起波長帯の蛍光スペクトルのみを測定して得られる蛍光指紋から識別を行うことが可能となるため、蛍光指紋の測定時間短縮に繋がる。

まず、4.2 節で説明した手順で、コンテキスト特徴量  $\{\mathbf{c}_i^{\tau}\}_{i=1}^{\text{Ex}}$  および、訓練用データセットに対する共通特徴量

$\{\mathbf{z}_n^{\varepsilon}\}_{n=1}^{N_{\varepsilon}}$  を抽出する。 $\{\mathbf{c}_i^{\tau}\}_{i=1}^{\text{Ex}}$  は、識別タスク  $\varepsilon$  の評価用蛍光指紋(Test EEM)から共通特徴量を抽出する際に用いる。

$\{\mathbf{z}_n^{\varepsilon}\}_{n=1}^{N_{\varepsilon}}$  の各共通特徴量  $\mathbf{z}_n^{\varepsilon}$  は、励起波長帯単位で抽出した共通特徴量を連結した  $\text{Ex} \times d$  次元ベクトルであるため、Lasso 回帰の重み  $\mathbf{w}_{\varepsilon}$  は、それぞれの特徴量に対応した  $\text{Ex} \times d$  次元ベクトルである。

共通特徴量を説明変数、正例/負例ラベルを目的変数とした Lasso 回帰によって、重み  $\mathbf{w}_{\varepsilon}$  およびバイアス  $b$  を最適化する。目的関数は以下の式で表される。

$$\min_{\mathbf{w}_{\varepsilon}, b} \frac{1}{2} \|\mathbf{y} - (\mathbf{Z}\mathbf{w}_{\varepsilon} + b)\|_2^2 + \alpha \|\mathbf{w}_{\varepsilon}\|_1 \quad (4)$$

式(4)において、 $\mathbf{Z}$  は、各訓練データから得られた共通特徴量  $\mathbf{z}_n^{\varepsilon}$  の  $\text{Ex} \times d$  次元を行方向に並べ、 $\mathbf{z}_1^{\varepsilon}$  から  $\mathbf{z}_{N_{\varepsilon}}^{\varepsilon}$  を列方向に並べることで得られる説明変数行列であり、 $\mathbf{y}$  は、各訓練データの目的変数(正例/負例ラベル)を示すベクトルである。 $\alpha$  は正則化の強さを表し、 $\alpha$  の値が大きいく程、重み  $\mathbf{w}_{\varepsilon}$  はスパース性を持ち、識別に重要でない説明変数に対応する重みの値が 0 になる。

励起波長帯  $i \in \{1, 2, \dots, \text{Ex}\}$ 、特徴量の次元  $k \in \{1, 2, \dots, d\}$  に対応する重みを  $w_{i,k}$  とした場合、重みの集合  $\{w_{i,k}\}_{k=1}^d$  の内、値が 0 で無い重みが存在する励起波長帯  $i$  を有効励起波長帯とする。

#### 4.4 共通特徴量抽出器および Lasso 識別器による識別

学習済みの共通特徴量抽出器、識別タスク  $\varepsilon$  の Lasso 識別器を用いて、実際に試料から識別を行う方法について説明する。図 3(c)に示すように、分光蛍光光度計(Fluorescence spectrophotometer)を用いて、試料(Test sample)から有効励起波長帯の蛍光スペクトルのみを測定し、評価用蛍光指紋を得る。その後、4.3 節で得られたコンテキスト特徴量  $\{\mathbf{c}_i^{\tau}\}_{i=1}^{\text{Ex}}$  とともに試料の蛍光指紋を共通特徴量抽出器に入力し、共通特徴量を抽出する。最後に、共通特徴量を Lasso 識別器に入力することで識別を行う。このように、試料から、有効励起波長帯のみを測定することで識別が可能のため、蛍光指紋の測定時間短縮に繋がる。

### 5. 実験

#### 5.1 データセット

実験に用いるオープンデータセットとして、化合物が含まれた溶液を試料とする蛍光指紋(化合物蛍光指紋) [7]を用いた。化合物蛍光指紋は 6 種の成分の濃度(含まない場合は濃度=0)がラベルされており、このデータセットに対し、提案手法を用いて、成分の有無(濃度>0 か濃度=0 か)を識別する実験を行った。各蛍光指紋は、励起波長 230nm ~ 320nm を 5nm 間隔、蛍光波長 230nm ~ 500nm を 2nm 間隔で測定されており、励起波長帯数は 19、蛍光波長帯数は 136 である。また、1 つの試料につき同条件で 5 回測定されている。

6 種の成分はそれぞれ、Catechol、Hydroquinone、Indole、Resorcinol、Tryptophane、Tyrosine である。しかし、これらの成分の内、Resorcinol に関しては、Resorcinol を含んでいる(濃度>0である)サンプル数が、全体の約 9% と少なく、正例/負例の割合が偏っているため、データセットから除外した。

化合物蛍光指紋は、識別対象とした成分の内、一つの成分のみを含む蛍光指紋(単成分)および、複数の成分を含む

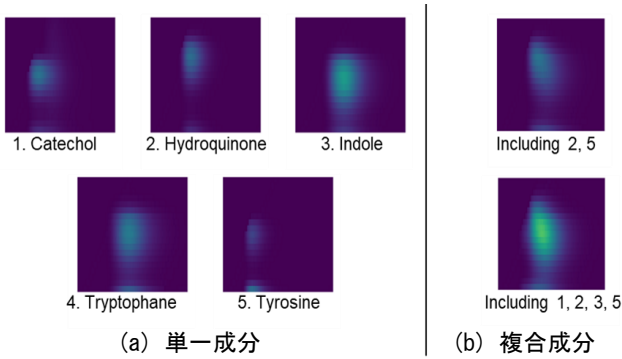


図 4 化合物蛍光指紋の可視化例

蛍光指紋(複合成分)に分けられる。単一成分および複合成分の蛍光指紋の例を図 4 に記載する。2 章で述べたように、散乱光は試料特性に関係しないノイズを多く含むため、図 2 に示すように、非蛍光領域および散乱光(1 次光, 2 次光)に相当する領域の蛍光強度を 0 にする前処理を施した。

表 1 に単一成分のデータセット構成を示す。各数値は、対象成分毎の訓練、検証、評価用データセットに含まれる試料の数を示す。各成分の試料において、訓練、検証、評価用データセットに含まれる試料の割合が 3:1:1 になるようにデータセットの分割を行った。ただし、同一の試料に対し蛍光指紋は 5 回計測されているため、表 1 の値に 5 を掛けた数だけ蛍光指紋サンプルが用意されている。訓練および検証用データセットは Lasso 識別器の構築時に用いる。後述の共通特徴量抽出器の学習においては、表 1 の緑でハイライトしたサンプル (5 回の測定データ全て) のみを用いた。

表 2 に複合成分のデータセット構成を示す。表 1 と異なり、各数値は、対象成分を含む試料数/含まない試料数を示す。成分毎に、その成分を含む試料数と含まない試料数それぞれにおいて、訓練、検証、評価用データセットに含まれる試料の割合が 3:1:1 になるようにデータセットの分割を行った。単一成分と同様に、試料毎に 5 件の蛍光指紋が用意されている。

表 1 単一成分のデータセット構成 (対象成分の試料数)

	Train	Valid	Eval
Catechol	7	2	2
Hydroquinone	9	3	3
Indole	9	3	3
Tryptophane	4	2	2
Tyrosine	11	3	3
Total	40	13	13

表 2 複合成分のデータセット構成 (対象成分を含む試料数/含まない試料数)

	Train	Valid	Eval
Catechol	79/35	26/12	26/12
Hydroquinone	73/41	24/14	25/13
Indole	77/37	25/13	26/12
Tryptophane	55/58	18/20	18/20
Tyrosine	53/61	17/21	18/20

## 5.2 共通特徴量抽出モデルの学習

5 種の成分の内、3 種の成分(Catechol、Hydroquinone、Indole)の単一成分の訓練用および検証用試料(表 1 の緑でハイライトしたサンプル)から測定した蛍光指紋(試料 25 件×5 回測定分の計 125 件)のみを用いて、共通特徴量の学習を行った。

1 回のパラメータ更新(イテレーション)毎に、正例/負例パターン  $\tau$  をバッチサイズの数だけ定義する。 $\tau$  は「Catechol、Hydroquinone、Indole の内、ランダムに 1 つの成分を選択し、その成分が含まれているか否か」というルールに基づき、選択した成分を「含む/含まない」のパターンに対し、「正例/負例」、「負例/正例」のどちらかランダムでラベルを定義する。各  $\tau$  に従って、訓練用データセットから、コンテキスト集合  $D_{\tau} = \{(\mathbf{X}_n^{\tau}, \mathbf{y}_n^{\tau})\}_{n=1}^N$  およびターゲット集合  $D_{\tau^*} = \{(\mathbf{X}_m^{\tau^*}, \mathbf{y}_m^{\tau^*})\}_{m=1}^M$  をサンプリングし、入力データとする。

上記において、バッチサイズは 8、 $N$  は 16、 $M$  は 24 とした。また、イテレーション回数は 14000 回とし、検証データに対する式(3)の損失関数の値が最も小さいモデルを評価に用いた。

## 5.3 Lasso 回帰による識別モデルの構築

5.2 節で得られた共通特徴量抽出器を用いて、単一および複合成分の訓練用蛍光指紋から、励起波長帯単位で共通特徴量を抽出し、Lasso 識別器を作成した。Lasso 識別器の訓練には、表 1、表 2 に記載の訓練用データセット(Train)を用いた。式(4)の正則化の強さ  $\alpha$  に関しては、 $\alpha=0$  から 0.001 刻みで徐々に強くし、表 1、表 2 に記載の検証用データセット(Valid)の識別精度が低下する直前の  $\alpha$  を用いた。

## 5.4 識別モデルの評価

### 5.4.1 比較手法

比較手法として、蛍光指紋の各蛍光強度を説明変数とした、蛍光指紋解析に広く用いられる Lasso 回帰の識別モデルを構築した。識別モデル構築には提案手法と同様に表 1、表 2 に記載の訓練用データセット(Train)および検証用データセット(Valid)を用いた。

### 5.4.2 評価結果

識別モデルに対する評価用データセット(表 1、2 の Eval)の評価結果を表 3 に記載する。「Lasso」は蛍光強度を説明変数とした Lasso 回帰による識別モデルを表し、「Ours」は提案手法を表す。「Accuracy」は識別精度、「Eff.」は有効励起波長帯数(The number of effective excitation wavelength bands)を表す。5.1 節に記載した通り、全励起波長帯数は 19 である。

表 3 識別精度及び有効励起波長帯数

	Lasso		Ours	
	Accuracy [%]	Eff.	Accuracy [%]	Eff.
Catechol	83.0	15	<b>96.2</b>	8
Hydroquinone	92.5	13	<b>100.0</b>	5
Indole	75.9	14	<b>96.3</b>	10
Tryptophane	92.3	15	<b>94.2</b>	10
Tyrosine	96.1	17	<b>98.0</b>	7
Average	89.2	14.8	<b>97.1</b>	8.0

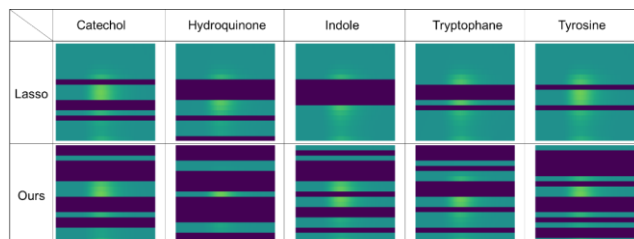


図 5 有効励起波長帯の可視化結果

表3から分かる通り、各成分において、提案手法はLasso回帰の識別精度を上回っており、Accuracyが平均7.9%向上した。学習に用いていない成分(Tryptophane、Tyrosine)においてもLasso回帰以上の識別精度を得たことから、提案手法の共通特徴量抽出器が、類似タスクに対して識別に有用な特徴量を抽出したと考える。

有効励起波長帯数においても、各成分において、Lasso回帰より提案手法の方がより少ない励起波長帯数で識別を行っていることが分かる。運用時は、有効励起波長帯の蛍光スペクトルのみの測定により識別が可能のため、提案手法を用いることで、蛍光指紋の測定時間の短縮が期待できる。

## 5.5 考察

有効励起波長帯の可視化結果を図5に記載する。図5は、蛍光指紋のサンプルから、有効励起波長帯でない蛍光スペクトルをマスクした結果である。

図5より、Lasso回帰および提案手法は、隣接した励起波長帯を選択しているケースが多いことが分かる。励起波長帯が近い場合、試料が放出する蛍光スペクトルは似た特徴量を有しやすと考えられるため、識別精度を維持したまま、隣接した波長帯を削減できる可能性が高い。

しかし、現状の提案手法は、Lasso回帰のアルゴリズムで波長帯選択および識別を行っており、各特徴量が独立した説明変数として扱われるため、隣接した波長帯の関係性を加味した波長帯選択が難しい。これに対して、提案手法におけるコンテキスト特徴抽出器 $\phi_c$ は、蛍光指紋間の関係性や、空間的な特徴を抽出する機構を有しているため、波長帯間関係性を暗に学習していると考えられる。 $\phi_c$ を拡張することで、波長帯間関係性を加味した有効励起波長帯数削減方式を今後検討する。

## 6. おわりに

本稿では、分光蛍光光度計を用いて試料から得られる蛍光指紋向けに、共通特徴量抽出器とLasso回帰を合わせた新たな識別手法を提案した。提案手法における共通特徴量抽出器は、複数のタスクを事前学習することで、類似タスク(識別対象データの分布が共通特徴量抽出器の学習データと類似した識別タスク)に対して、新たな学習を介さずに識別向けの特徴量を抽出する。これにより高精度化および深層学習の学習コスト低減を実現した。また、提案手法は、励起波長帯単位で特徴量抽出を行い、Lasso回帰による変数選択を行うことで、有効波長帯を限定しながら識別を行うことが可能である。

試料に含まれる化合物を識別する実験において、一部の化合物情報のみを用いて共通特徴量を学習し、共通特徴量

に基づく識別モデルを作成した結果、従来手法(Lasso回帰)と比較して、提案手法は高精度かつ少ない有効励起波長帯数での識別を達成した。

## 参考文献

- [1] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.58, No.1 (1996).
- [2] T.T. Wu, et al., "Estimation of Citrus Maturity with Florescence Spectroscopy Using Deep Learning", *Horticulture*, Vol.5, No.1 (2019).
- [3] J.W. Rutherford, et al., "Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification", *Atmospheric Environment*, Vol.220 (2019).
- [4] Y. LeCun, et al., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, Vol.86, No.11 (1998).
- [5] V. Ashish, et al., "Attention Is All You Need", *Advances in Neural Information Processing Systems*, Vol.30 (2017).
- [6] H. Larochelle, et al., "Supervised Contrastive Learning", *Advances in Neural Information Processing Systems*, Vol.33 (2020).
- [7] B. Rasmus, et al., "Standard error of prediction for multilinear PLS 2. Practical implementation in fluorescence spectroscopy", *Chemometrics and Intelligent Laboratory Systems*, Vol.75, No.1 (2005).