G-011

疑似的な蛍光指紋を用いた蛍光指紋向け表現学習手法の検討
Representation Learning using Pseudo-Excitation Emission Matrix

林田 純弥† 柿下 容弓† 長坂 晃朗†
Junya Hayashida Yasuki Kakishita Akio Nagasaka

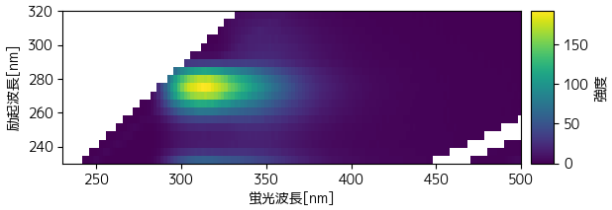

図 1 蛍光指紋の可視化例

1. はじめに

蛍光指紋(Excitation-Emission Matrix; EEM)とは、励起光照射によって試料が放出する蛍光を測定したスペクトルデータであり、試料の蛍光特性に関する情報を含む。図 1 は蛍光指紋をヒートマップ形式で可視化した例であり、励起波長、蛍光波長の 2 軸に対応する光の強度を色で可視化している。図 1 において、ある励起波長における蛍光波長方向の信号を蛍光スペクトル、ある蛍光波長における励起波長方向の信号を励起スペクトルと称する。蛍光指紋は単一の蛍光スペクトルと比較して広範囲の励起波長帯に渡る蛍光特性の情報を有しており、化学・医薬品、食品、工業製品等、様々な分野で試料の解析に用いられている。

蛍光波長は励起・蛍光波長方向に広がる様々な信号が重畳されており、この中から解析に有効な特徴を抽出することが重要となる。近年では、深層学習モデルを用いることで、蛍光指紋から重要な特徴を自動抽出し、効率的かつ高精度な推論を実現している [1, 2]。

画像や自然言語処理において、目的の回帰や分類タスクに対する学習を行う際、事前に大規模なデータセットを学習した深層学習モデルのエンコーダ(特徴抽出器)を利用することがある[3, 4]。大量のデータを学習したモデルの特徴抽出器はドメインに対する強力な埋め込み表現を獲得しているため、目的のタスクの学習を容易にし、ランダムな初期値から学習する場合と比較して高精度化も期待できる。

我々の知る限り、蛍光指紋に対して、大規模データによる事前学習済みモデルを活用した回帰や分類手法は提案されていない。これは、蛍光指紋の大量収集が困難であるからと考える。画像と異なり、Web での収集が容易ではなく、また、実際に収集することを考えても、多様な試料を準備して測定することは非常に大きな時間とコストを要する。

本稿では、蛍光指紋を実際に収集することなく、蛍光指紋のための特徴抽出器を学習する手法および特徴抽出器を用いた回帰方式を提案する。図 2 は提案手法の概要図である。本提案手法ではまず、励起・蛍光スペクトルの波形情報と蛍光指紋の物理的特性に基づいて、実際の蛍光指紋をシミュレートした疑似蛍光指紋を大量に生成する。生成した大量の疑似蛍光指紋を用いて表現学習を行うことで蛍光指紋解析のための特徴抽出器を構築する。その後、実際の蛍光指紋に対して特徴抽出を行い、任意の回帰モデルで回帰を行う。複数の化合物濃度推論タスクに対して提案手法による回帰を行った結果、蛍光指紋を入力とする従来の回帰と比較してノイズに頑健かつ高精度な回帰を実現した。

† 日立製作所 研究開発グループ Hitachi、 Ltd. Research & Development Group

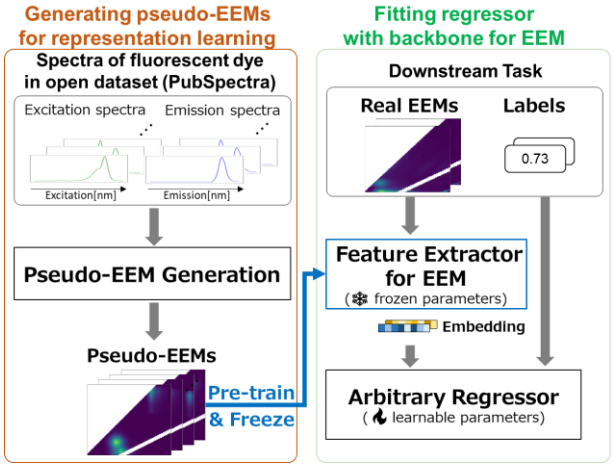

図 2 提案手法の概要図

2. 提案手法

提案手法は①疑似蛍光指紋の生成、②疑似蛍光指紋を用いた表現学習、③埋め込み表現を用いた任意の回帰の 3 つのステップで構成される。①、②によって蛍光指紋の埋め込み表現を獲得し、③で埋め込み表現による回帰を行う。蛍光指紋自体を入力とする(説明変数とする)従来の回帰に対し、図 2 右に示すように、埋め込み表現を入力としている。以降、①、②に関して説明を行う。

2.1 疑似蛍光指紋の生成

実際の蛍光指紋をシミュレートするために、tri-linear モデル[5]を用いる。tri-linear モデルは、蛍光指紋の物理的特性[6]に基づいた、蛍光指紋を複数の信号の重み付き和で表すモデル式であり、蛍光指紋を複数の信号に分解し、構造を解析する用途に用いられる。本提案手法では tri-linear モデルを蛍光指紋のシミュレートに応用することで、疑似蛍光指紋を大量に生成する。

蛍光指紋は、試料に含まれる各蛍光分子が放出する光が重畳された光であり、試料に含まれる蛍光分子の種類、濃度等の要因で変化する。tri-linear モデルは、この光の重畳をモデル化しており、励起波長数Ex、蛍光波長数EmのK個の蛍光指紋群$\boldsymbol{X} \in \mathbb{R}^{\mathrm{Ex}\times\mathrm{Em}\times\mathrm{K}}$を式(1)のように表現する。

$$\hat{\boldsymbol{X}} = \sum_{r=1}^{\mathrm{R}} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \qquad (1)$$

式(1)は蛍光分子の種類の数Rと、R組のベクトル$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$で構成される。$\boldsymbol{a}_r$は$r$番目の蛍光分子由来の励起スペクトル(Ex次元ベクトル)、$\boldsymbol{b}_r$はr番目の蛍光分子由来の蛍光スペクトル(Em次元ベクトル)、$\boldsymbol{c}_r$はr番目の蛍光分子に対するK個の濃度(K次元ベクトル)を表す。

提案手法では、tri-linear モデルのパラメタに様々なパターンを当てはめることで、疑似蛍光指紋を生成する。蛍光分子の種類の数Rと、対応する各濃度$\boldsymbol{c}_r$はランダムな値を当てはめる。しかし、励起・蛍光スペクトル$\boldsymbol{a}_r$, $\boldsymbol{b}_r$に関しては、滑らかな波形であり形状も様々であるため、ランダムな定義が難しい。ガウス過程のような生成モデルで波形を生成することも考えられるが、実際の蛍光指紋と形状が乖離する可能性がある。そこで、実際の励起・蛍光スペクトル情報を使用することを検討した。具体的には、図 2 左上に示すオープンデータセット PubSpectra[7]に含まれる蛍光色素の励起・蛍光スペクトルを用いた。蛍光色素は蛍光塗料として使用される素材であり、発光特性を調べるために、励起・蛍光スペクトルが測定・公開されている場合が多く、PubSpectra には励起・蛍光スペクトル 267 組が含まれている。$\boldsymbol{a}_r \circ \boldsymbol{b}_r$を励起・蛍光方向にランダムにシフトさせつつ、濃度$\boldsymbol{c}_r$をランダムに定義することで大量のパターンの疑似蛍光指紋を生成しつつ、ノイズへの頑健性を獲得するために、ガウシアンノイズを付与した。

2.2 疑似蛍光指紋を用いた表現学習

生成した疑似蛍光指紋を用いた表現学習により、埋め込み表現を獲得する。式(1)から分かるように、蛍光指紋は、どのような励起・蛍光スペクトル$\boldsymbol{a}, \boldsymbol{b}$がどのような割合(濃度 $\boldsymbol{c}$)で重畳された信号かによって特徴づけられる。実際の蛍光指紋は蛍光分子の種類の数Rを含め,$\boldsymbol{b}, \boldsymbol{c}$は未知であるためこれらを直接学習することは難しい。一方で、疑似蛍光指紋はこれらの要素をランダムに定義していることから自明である。そこで、濃度パターンを活用した蛍光指紋解析のための表現学習手法を提案する。

表現学習手法を図 3 に示す。ある疑似蛍光指紋$\boldsymbol{X}_a$と、濃度パターンが類似する$\boldsymbol{X}_p$、異なる$\boldsymbol{X}_n$の 3 つの疑似蛍光指紋に対して CNN (Convolutional Neural Network) σ_θを用いて特徴抽出し、$\sigma_\theta(\boldsymbol{X}_a)$と$\sigma_\theta(\boldsymbol{X}_p)$の距離を近付け、$\sigma_\theta(\boldsymbol{X}_a)$と$\sigma_\theta(\boldsymbol{X}_n)$を遠ざけるように、triplet loss[8]を用いて CNN のパラメタθを学習する。

3. 実験

まず、triplet loss 計算用の 3 件 1 組の疑似蛍光指紋を合計 2,670,000 組生成し、CNN を用いて表現学習を行った。学習後、CNN の重みを固定し、図 2 右のように、実際の蛍光指紋を用いた回帰の実験を行った。実験には化合物を含んだ溶液から測定された蛍光指紋のオープンデータセット[9]を用いた。蛍光指紋には 5 種の化合物濃度がラベル付けされている。各化合物濃度を推論するタスクに対して、蛍光指紋を入力とする従来の回帰モデルと、埋め込み表現を入力とする回帰モデルを用いて各化合物濃度の推論を行った。

表 1 は訓練データを用いて回帰モデルを学習する際、入力に蛍光指紋(EEM)を用いる場合(従来の解析)と、埋め込み表現(Emb.)を用いる場合(提案手法)の、評価データに対する化合物濃度の推論精度(決定係数)を示している。各列は化合物、各行は実験に用いた回帰モデルを表す。化合物・回帰モデルの組み合わせ 15 条件において 13 条件で精度が向上した。残りの 2 条件に関しては、精度は向上していないものの、蛍光指紋を用いた場合と同程度であった。

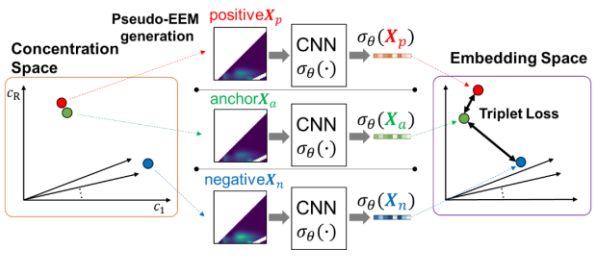

図 3 疑似蛍光指紋を用いた表現学習

表 1 化合物濃度の推論精度(決定係数)
(EEM: 蛍光指紋、Emb.: 埋め込み表現)

	Catechol		Hydroquinone		Indole		Tryptophane		Tyrosine	
	EEM	Emb.	EEM	Emb.	EEM	Emb.	EEM	Emb.	EEM	Emb.
Lasso	0.822	**0.839**	0.932	**0.943**	0.792	**0.806**	**0.872**	0.862	0.818	**0.868**
PLS	0.945	**0.961**	0.942	**0.968**	**0.970**	0.959	0.987	**0.995**	0.896	**0.987**
MLP	0.478	**0.933**	0.771	**0.947**	0.839	**0.951**	0.896	**0.952**	0.701	**0.976**

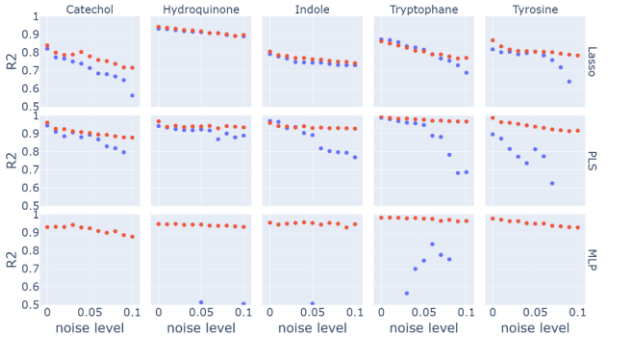

図 4 ノイズを加えた蛍光指紋に対する決定係数
(青: 蛍光指紋、赤: 埋め込み表現)

また、蛍光指紋にガウシアンノイズを加えた場合の推論精度の影響を確認した。図 4 において、各行は回帰モデル、各列は化合物、横軸はガウシアンノイズの分散、縦軸は決定係数を表す。図 4 から分かるように、ノイズが強くなった際の精度低下が、従来の回帰より抑えられている。

4. おわりに

本稿では、蛍光指紋解析に広く使用可能な事前学習の手法として、疑似蛍光指紋を用いた表現学習手法を提案した。従来の解析に対して、学習した埋め込み表現を用いることで、多くの回帰条件で精度が向上しつつ、ノイズへの頑健性を有した推論を実現した。

参考文献

[1] K. Itakura, “Estimation of Citrus Maturity with Fluorescence Spectroscopy Using Deep Learning”, Horticulturae, Vol.5, (2019).
[2] J. Rutherford, et al., “Excitation emission matrix fluorescence spectroscopy for combustion generated particulate matter source identification”, Atmospheric Environment, Vol.220 (2020).
[3] K. He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR2022.
[4] J. Devlin, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in NAACL2019.
[5] R. Harshman, et al., “PARAFAC: Parallel factor analysis”, Computational Statistics & Data Analysis, Vol.18 (1994).
[6] J. Lakowicz, “Principles of Fluorescence Spectroscopy”, Kluwer Academic/Plenum Publishers (1999).
[7] G. McNamara, “PubSpectra - Open Data Access Fluorescence Spectra”, available at http://works.bepress.com/gmcnamara/9/ (2012).
[8] E. Hoffer and N. Ailon, “Deep metric learning using Triplet network”, SIMBAD2015.
[9] B. Rasmus, et al., “Standard error of prediction for multilinear PLS 2. Practical implementation in fluorescence spectroscopy”, Chemometrics and Intelligent Laboratory Systems, Vol.75, (2005).