機械学習を用いた光電子収量分光 (PYS) の閾値の判定

柳生進二郎*・吉武道子・知京豊裕

国立研究開発法人 物質・材料研究機構 305-0003 茨城県つくば市桜 3-13

(2017 年 11 月 13 日受付；2018 年 2 月 5 日掲載決定)

Determination of Threshold of Photoelectron Yield Spectroscopy (PYS) Using Machine Learning

Shinjiro YAGYU*, Michiko YOSHITAKE and Toyohiro CHKYOW

National Institute for Materials Science, Tsukuba, Ibaraki 305-0003

(Received November 13, 2017 ; Accepted February 5, 2018)

In order to analyze the threshold of photoelectron yield spectroscopy (PYS) automatically, we compared the accuracy of the analysis between a fitting (least squares method) method and a machine learning method.

Both methods are comparable in terms of the performance of analysis, but machine learning shows better performance in data with high noise components.

In the analysis of measured Au data, machine learning shows close to the analyzed value by an analyst, but the fitting method shows the different value due to the influence of a high signal intensity.

KEYWORDS : machine learning, least-square fitting method, photoelectron yield spectroscopy

1. はじめに

有機・無機半導体や電極材料開発において、電子が移動する価電子帯上端や HOMO レベルの調整が重要である。このエネルギーレベルの測定方法の一つとして、光電子分光収量分光 (PYS) がある。PYS では、測定試料に単色化された光を照射し、そのエネルギーを徐々に増加させ、光電効果によって放出される電子を計測する。そして、その光電子放出閾値が価電子帯上端や HOMO のエネルギーレベルとなる。この閾値の求め方にはいくつかの方法が提案されているが^{1)~4)}、一般的には、Fowler の光電効果の理論をもとに¹⁾、PYS 強度の平方根をとり、その傾きとバックグラウンドとの交点から閾値を求める⁵⁾。この解析方法については、自動化されたものではなく、解析者が一つ一つのスペクトルを見て、閾値を求めている。そのために、解析者の経験の差などから値に違いが生じることがある。Fig. 1 に多結晶の金を PYS 測定した結果を解析用に再プロット (強度の平方根) し、解析者が引いた補助線を示す。見ての通り、補

助線についてはいくつか引き方があることがわかる。(図では、閾値は 4.47 eV と 4.92 eV の 2 つが考えられる。) このように、引き方にいくつかの可能性があり、解析者によりどちらを取るかで閾値が大きく変わる。一方で、今後、高速測定による計測のビックデータ化に伴

Fig. 1. (color online). Measured Au spectra. Y-axis is converted by the square root of PYS intensity. The dotted lines are extended line by an analyst. We can recognize two threshold (cross point of background) values.

*E-mail : YAGYU.Shinjiro@nims.go.jp

い、大量のデータを人が解析するには限界が生じると考えられる。そこで、高速、かつ人に依存しない自動解析方法が必要である。自動解析が進まない理由として、解析の理論体系がまだまだ発展段階にあるためでもあるが、自動解析の実装方法が限られるためでもある。

自動解析の実装方法の一つとして、パラメータを含むモデル式による最小二乗 Fitting が考えられる。Fitting では、Fitting パラメータに初期値を与え、測定値とモデル式の差の二乗の値が最小になるようにパラメータの探索を行う。初期値が最適値からかけ離れていると探索に時間がかかり、探索が収束しないことが起こる。また、最小二乗法の性質として、ノイズや外れ値に対して過度に反応してしまうことなど考えられる。そこで、もう一つの実装方法として最近急速に発展してきた、機械学習の手法を検討した。

機械学習は、大量のデータを機械に学習させ、データから規則性や知識を見つけることである。機械学習の種類として、教師あり学習（特徴量とラベルのデータセットを持つもの）と教師なし学習（特徴量のみのデータセット）とがある。教師あり学習の中には、連続値を扱う回帰と、分類などの離散値を扱うクラス分類がある。教師なし学習では、特徴量を分類するクラスタリングや、特徴量の相関関係や主因子を見つける主成分分析などがある。

今回、閾値（連続値）をラベルデータとし、特徴量としてスペクトルデータ（エネルギーに対する取量の値）と考え、教師あり学習の回帰問題として機械学習を行った。実際に得られている Au の PYS データを用いて Fitting による解析と機械学習による解析との比較検討を行った。

2. Fowler の理論式と計算データ

PYS の解析モデルとして用いられる Fowler の光電効果の理論式は以下の通りである。

$$PYS^{1/2}(x, T, BG, Nor) = (Nor \cdot f(\mu(x, T)) + BG)^{1/2} \quad (1)$$

$$(hv - x)/kT = \mu \leq 0$$

$$f(\mu) = \left[\frac{\pi^2}{6} + \frac{1}{2}\mu^2 - \left\{ e^\mu - \frac{e^{2\mu}}{2^2} + \frac{e^{3\mu}}{3^2} - \dots \right\} \right] \cong 0 \quad (2)$$

$$(hv - x)/kT = \mu \geq 0$$

$$f(\mu) = \left\{ e^{-\mu} - \frac{e^{-2\mu}}{2^2} + \frac{e^{-3\mu}}{3^2} - \dots \right\} \quad (3)$$

(1) 式に示すようにパラメータは 4 つあり、閾値 (x)、温度 (T)、バックグラウンド (BG)、規格化 (Nor) である。Fitting では、測定データとモデル式との 2 乗差が最小となるようにこの 4 つのパラメータを動か

して最適な値を求める。

機械学習では訓練データが必要になるために、この 4 つのパラメータを変化させて計算データを作成した。機械学習においてどの程度データセット（特徴量とラベルを 1 セットとする）が必要かの目安がある⁶⁾。50 以下では機械学習はほぼできない。100 程度ではできないこともないがデータを増やす必要がある。1000 オーダー以上で機械学習による予測が機能する。このために、1000 オーダー以上のデータセットを準備する必要がある。そこで、閾値 (x) 4.5 eV から 5.9 eV まで 0.1 eV 刻み、温度 (T) 300 K から 900 K まで 100 K 刻み、バックグラウンド (BG) 5 に固定、規格化 (Nor) 0.1 から 2 まで 0.1 刻み、でパラメータを変えて計算データを作成した。また、ノイズとして、ガウシアンノイズ、ポアソンノイズを重ねたスペクトルも作成した。計算データとして、合計 6300 個（ノイズあり（ポアソンノイズ、ガウシアンノイズ）：各 2100 個、ノイズなし：2100 個）用意した。計算結果で得られたスペクトルの一部（6300 個のうちランダムに 7 個を選んでいる。）を Fig. 2 に示す。Fig. 1 と同様に閾値を境にエネルギーが高くなるにつれて強度が増加する。また、ノイズを付与したものではノイズ成分が強く閾値を判別するのは難しいと思われるものもある。（実際の PYS 測定では、ノイズが大きいスペクトルも多い。）

3. 機械学習

計算で作成したデータのうち特徴量として、4.0 eV から 6.2 eV まで 0.05 eV 刻みの強度の値、ラベルとして閾値 (x) とした。(Fig. 2 で示した各エネルギーに対する $PYS^{1/2}$ の強度が特徴量で、グラフの凡例にあるエネルギー

Fig. 2. (color online). Calculated the square root of PYS spectrum based on the Fowler formula. The parameters are followed ; threshold (x), temperature (T), background (BG) and normalized factor (Nor).

ーがラベルデータとなる。) 実際に行った機械学習の手順は以下の通りであり, 実行プログラムは Python の Scikit-learn を使用した⁷⁾。

- (1) データの入手: 上記の通り計算したデータ (6300 セット) を訓練用データ (全データの 90%) とテストデータ (検証用) (10%) に分割した。
- (2) 手法の選択: 連続値の回帰問題として取り扱う。
- (3) モデルの学習・評価・調整: 試したモデルは線形モデル (Linear Regression), 非線形モデル (カーネルサポートベクトルマシン SVM (kernel = rbf)), 決定木 (Decision Tree), アンサンブル (決定木を使ったもの Gradient Boosting) である。なお, グリッドサーチを行い, モデルパラメータの最適化を行った。
- (4) テストデータ (検証) を用いて正答率の評価を行う。なお, ここでの正答率は, Appendix に記載しているように, 残差変動をラベルデータと予測値との差の 2 乗として, 計算している。
- (5) 実際の測定データを用いて予測との違いを評価する。

それぞれの学習モデルのテストデータを用いた正答率, 予測結果及び, 予測と実験値との差を Table 1 に示す。表中の Au(exp.) 値とは, Fig. 1 の Au データの低エネルギー側の交点の 4.47 eV である。Fowler の理論式では, 価電子帯状態密度の立ち上がり付近を近似していることから, 閾値に近い低エネルギー側の立ち上がりを取ることが望ましいと考えられる。また, 実験的にも閾値近傍では Fowler の式と一致するがそれよりもエネルギーが高くなるとずれることが報告されている^{8,9)}。線形モデルの正答率が悪いが, 他のモデルでは 94% 以上の正答率が得られた。線形モデルでは, モデルが単純で非線形の効果を取り入れることができないために正答率が悪いと考えられる。一方で, Fitting によって得られた正答率は, 一部の機械学習より低い値を示した。実際の測定データの予測でも線形モデル, Fitting 以外では Au(exp.) の値に近い予測を示している。

Table 1. The accuracy, the predicted value of Au and the difference between the measured value and the predicted value, in each regression model and fitting.

Model	Accuracy (Test Data)	Prediction [eV] (Au(exp.) Data)	Residual Au(exp.) : 4.47 eV
Linear Regression	0.713	3.97	0.49
SVM (kernel = rbf)	0.981	4.50	-0.03
Decision Tree	0.944	4.50	-0.03
Gradient Boosting	0.989	4.50	-0.03
Fitting	0.947	4.96	-0.49

Fig. 3. (color online). Residual plots of Gradient boosting regression and fitting. The X-axis is predicted energy (threshold) and Y-axis is residual (difference of predicted energy and label (teacher) data).

4. 機械学習と Fitting との比較

テストデータを用いて, 正答率の高いモデル (Gradient Boosting) と Fitting それぞれ得られた残差プロットを Fig. 3 に示す。残差プロットとは, ラベル (教師データ) とモデルが予測した値の差を残差と呼び (残差 = 教師データ - 予測値), 横軸に予測値, 縦軸に残差をプロットしたものである。残差プロットでは, 多くのデータが残差 = 0 に近いところに集まれば, よいモデルということになる。Fitting では, 前述の通り初期値の与え方によっては Fitting が収束しないことなどあるために, 初期値としてラベル (x) を与えて Fitting を行っている。機械学習と Fitting とともに, テストデータの 98% 以上が, ± 0.2 eV 以下で値の予測ができています。5.5 eV よりもエネルギーが高い領域では, 残差が大きくなるものが増え, 機械学習に比べ Fitting の方がその割合が多い。これは, ノイズ成分が信号成分に比べて相対的に大きくなるために Fitting がうまくいかなかったためである。Table 1 の Fitting における正答率の低さは, Fig. 3 にあるように, 極端に予測がずれるものがあり, この影響である。

実データの Fitting 結果を Fig. 4 に示す。Fitting 結果は 4.96 eV であった。決定係数 (R^2) は 0.99 であり, Fitting 曲線はよくデータを再現していると考えられる (決定係数の計算式は Appendix に記載)。しかし, 4.3 eV 付近があまり一致していないのがわかる。これは, 4.8 eV より大きいエネルギーの PYS 強度が大きいためそちらの値にひきずられたためであると考えられる。一方で, 機械学習では, 4.5 eV と Fig. 1 のエネルギーの低い閾値に近かった。機械学習では, Fowler の式より計

Fig. 4. (color online). Fitting results of Au data. The predicted value is 4.97 eV.

算したデータを学習させているので、Fowlerの式の形に近いものから閾値を予測していると考えられる。Fittingでは、初期値の与え方による収束性（計算時間）、ノイズ、強度の強い値の影響を受けるなどの問題がある。特にFittingでは、初期値の与え方自体がノウハウである。機械学習では、あらかじめ学習させおくコストがかかるものの、予測性能、ノイズ耐性などの点でFittingよりも優れていると考えられる。今後は、より多くの実験データを用いて機械学習やFittingの検証を行いモデルの改良などを行うことが必要であると考えられる。

5. ま と め

最小二乗によるFittingと教師あり機械学習を用いてPYSスペクトルの自動解析の検討を行った。Fittingのモデル式に用いるFowlerの理論式よりパラメータを変えてFittingを評価するとともに、パラメータを変えて作成した計算データ（6300スペクトル）を用いて機械学習を行った。機械学習の識別モデル（Gradient boosting：GB）では、98%以上の識別性能を示した。Fitting及び機械学習による残差プロットから、機械学習では、初期値の推定の不要な点、ノイズ耐性の点でFittingより優れていると考えられる。実際Auのデータにおいては、機械学習では解析結果に近い値を導き出したのに対し、Fittingでは大きな信号成分の影響を受け解析結果より大きな値を示した。今後、より多くの実験データを用いて機械学習やFittingの検証比較を行い機械学習モデルの

改良などを行うことが必要であると考えられる。

Appendix

正答率（Accuracy）と決定係数（ R^2 スコア）

本論文の解析で用いたPythonのScikit-learnでは、正答率は決定係数と同じであるが、Fittingにおける決定係数との混乱を避けるため本文中では、ラベルデータと予測の差を評価するものを正答率とし、FittingにおけるデータとFitting関数との差を評価するものとして決定係数とした。

なお、正答率（決定係数： R^2 ）は、 $R^2 = 1 - (\text{残差変動} / \text{全変動})$ と表される。分母の全変動（総変動）は、データのそもそものばらつき具合を表し、分子の残差変動は、モデルと実測値とのずれを表す。

正答率（Accuracy）の計算では、機械学習の場合の残差変動 $= (\text{ラベル} - \text{予測値})^2$ 、Fittingの場合の残差変動 $= (\text{ラベル} - \text{Fittingパラメータの} x)^2$ 、全変動 $= (\text{ラベル} - \text{ラベルの平均})^2$ としている。

Fittingにおいては、個々のスペクトルについてデータとどれくらいずれが生じているかを評価するため

$$\text{残差変動} = (y_i - f(x_i))^2, \text{全変動} = (y_i - \mu Y)^2,$$

としている。なお、 (x_i, y_i) はデータ、 μY は y_i の平均、 $f(x_i)$ はFittingモデルを表す。

文 献

- 1) H. Fowler : *Phys. Rev.* **38**, 45 (1931).
- 2) E.O. Kane : *Phys. Rev.* **127**, 131 (1962).
- 3) G.W. Gobeli and F.G. Allen : *Phys. Rev.* **127**, 141 (1962).
- 4) M. Kochi, Y. Harada, T. Hirooka and H. Inokuchi : *Bull. Chem. Soc. Jpn.* **43**, 2690 (1970).
- 5) C. Adach, H. Oyamada and Y. Nakajima : "Data Book on Work Function of Organic Thin Films" (CMC Books, 2004).
- 6) http://scikit-learn.org/stable/tutorial/machine_learning_map/
- 7) A.C. Müller and S. Guido : "Introduction to Machine Learning with Python: A Guide for Data Scientists" (O'Reilly Media, 2016).
- 8) C.N. Berglund and W.E. Spicer : *Phys. Rev.* **136**, A1044 (1964).
- 9) F. Wooten and R.N. Stuart : *Phys. Rev.* **186**, 592 (1969).