ノート

機械学習によるラマン分光分析のスペクトルデータ解析

天城　裕子*

Raman Spectrum Analysis Using Machine Learning

AMAKI Yuko*

1. 緒　　言

ラマン分光分析は近年低価格化や小型化により普及が進み，異物分析や材料解析の分野における強力なツールとなっている。鉄さびや DLC，セラミック材料などの無機物から樹脂系などの有機物，一部の液体など適用範囲が広いこと，複雑な前処理が不要で簡易にスペクトルデータが得られることなどが長所である。しかしながら当所の装置には標準データベースが付属していないため，未知試料の測定では解析が困難な場合がある。現状は想定される物質からの過去データや文献データの検索を行い，解析を行っているが，これらは煩雑な作業であり，また測定者の経験に依存する部分が多い。本研究では過去の標準データとの照合プログラムを作成，簡易的な検索を行えるようにすることを目的とする。

2. 検討及び結果

2.1　標準データの準備および前処理

測定データは解析プログラム用の独自形式のため，ソフトウェアのエクスポート機能で csv または txt 形式に変換し，各種作業を行った。比較用の標準データは 1000 程度を準備した。同等サンプルの重複があるため，有効数としては半数程度である。今回は重複分を削除せずそのまま使用した。

ラマン分光分析では蛍光の影響などで不定形かつ高いバックグラウンドノイズが現れる場合が多い。生データの状態ではこの高いバックグラウンドが解析結果に影響する。これを除去する手法について検討した。温度計の赤色色素のデータについて，パラメータを適当に設定した Asymmetric Least Squares Smoothing 法 [1]により処理を行った結果を図 1 に示す。データにより補正効果がやや異なるが，おおむね解析可能な程度に補正することができた。

あわせて，データポイント数の調整や，測定データの中心波数による分類などを行い，次項以降の解析用データとした。

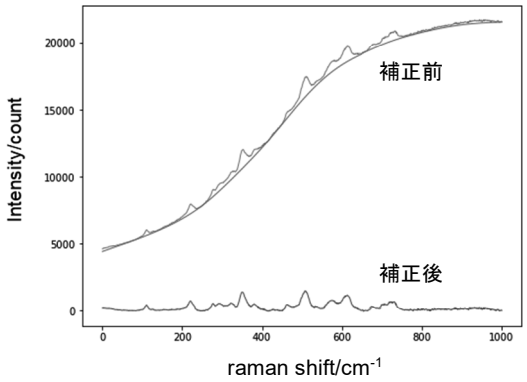

図 1　ベースライン補正

2.2　比較法の検討

前項で変換したサンプルデータと，比較用の標準物質データとの類似度を，Python の scipy ライブラリによる以下の 2 手法で検討した。

(1) データ全体の相関係数

(2) メインピークをいくつか検出し，検出ピークの位置と相対強度での比較

(1) 法については，scipy.stats.pearsonr，scipy.stats.spearmanr でそれぞれピアソン，スピアマンの相関係数を算出したところ，両者の傾向に大

* 下越技術支援センター

きく差はみられないため，暫定的にピアソンの相関係数の降順で比較を行った。(2)法についてはscipy.signal.argrelmaxを用いてピークを検出した。図1の補正後データでの，(2)法の検出結果を図 2 に示す。矢印に示すように判定しないピークがみられる。これらはパラメータの調整で検出可能であるが，ノイズをピークと誤判定するようになる。ラマンはサンプルによりFWHM や S/N が大きく変わるため，(2)法における汎用的なパラメータの設定は難しいと思われる。以降の(1)法との比較については，ノイズを拾わない程度のパラメータを用いて行った。

同じデータについて，図 3 に(1)法，図 4 に(2)法でのそれぞれ一致度上位3データとの比較結果を示す。なお，図3，4とも一番上がサンプルデータである。図3の(1)法では上位3データとも赤色色素系の物質であったのに対し，図 4 の(2)法ではやや精度が落ちる。メインのピークがいくつか未検出なのが原因と思われる。(2)法ではピーク検出用のパラメータの選定が困難な課題があり，(1)法ではより安定した結果が得られると考えられる。

2.3 測定用PCでの実施方法の検討

実際の測定中に使用することを想定し，測定用 PC で使用可能にするための方法を検討した。非 python 環境での動作方法として，pyinstallerなどを用いた windows 用の exe 化による方法があるが，作成したファイルはエラーにより動作を確認できなかった。python のバージョン，モジュールの対応/非対応などの要因が考えられるが詳細は不明である。プログラムを持ち歩いて他の PC で使用することは基本的になく，exe 化のメリットはあまりないと考えられるため，python を導入することとした。移行元 PC にminicondaと最小限のモジュールをインストールした仮想環境を設定し，この環境をコピーした測定用 PC で動作を確認した。

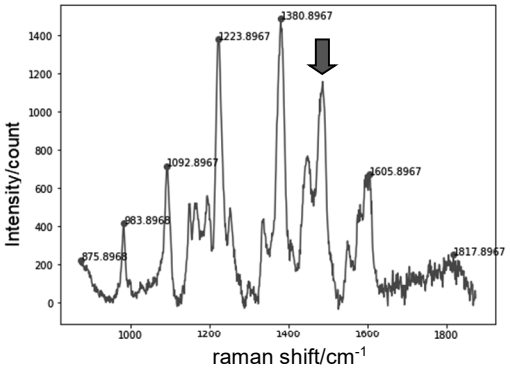

図 2　ピーク検出結果

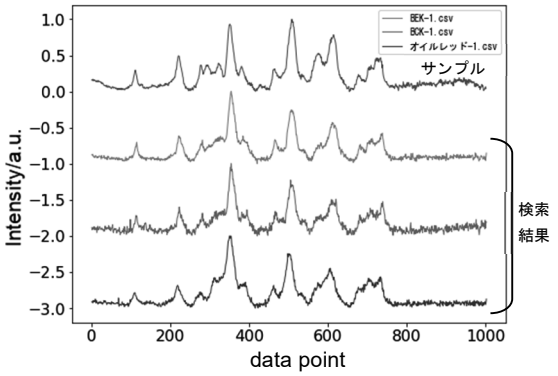

図 3　(1)法による比較結果

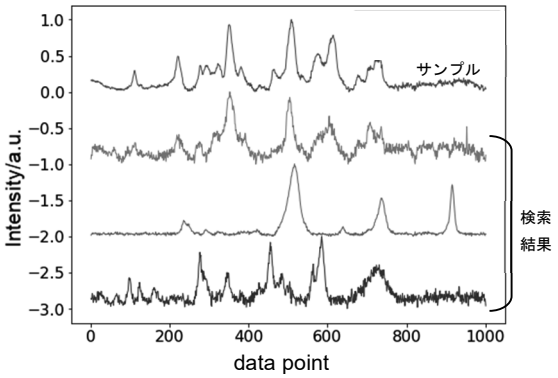

図 4　(2)法による比較結果

3. 結　　言

(1) ラマンデータの簡易的な検索システムを構築することができた。

(2) 比較用データの種類が精度に影響するため今後比較用データの拡充を行っていく。

参考文献

1) PHC Eilers, HFM Boelens - Leiden University Medical Centre Report, 2005