2005－MPS－56（1）
2005／9／21

光量子科学研究におけるLVQを用いたデータ分類・可視化処理の自動化

○上島　豊，斎藤　寛二，[1)]松山　仁美，[2)]城　和貴
日本原子力研究所 関西研究所 光量子科学研究センター
1)　日本総合研究所
2)　奈良女子大学大学院　人間文化研究科

光量子科学研究において，研究過程で生じる画像データを分類したり，多次元データを適切に可視化したりすることは研究者にとって大きな負担である．にもかかわらず，研究者が管理しなければならないデータは増え続けるばかりである．一方，データの分類や可視化のノウハウは，研究者個人のみに蓄積され続けている．それゆえ，データの管理，分類，可視化，保全性の品質は，全く均一ではなくなっている．その結果，過去のデータ処理を再現させようとしても，その保証はどこにもないのが現状である．我々は，このような問題を解消するために，ニューラルネットワークの一種である Learning Vector Quantization（LVQ）を利用して研究者のノウハウを知識化し，画像の分類や適切な解析を研究者が行い易くするための画像データ解析支援システムの開発と評価を行った．

Automatic data classification and visualization by LVQ in advanced photon research

○Yutaka　UESHIMA, Kanji SAITOH, [1)]Hitomi MATSUYAMA and [2)]Kazuki JOE
Advanced Photon Research Center, Kansai Research Establishment,
Japan Atomic Energy Research Institute
1)　The Japan Research Institute, Limited.
2)　Graduate School of Humanitites and Sciences, Nara Women's University.

It makes the burden too heavy for a researcher that a researcher must classify image data and visualize multi-dimensional data opportunely in the advanced photon research. The amount of data which a researcher must manage continues to increase. On the other hand, the know-how of classification and visualization is kept only in an individual researcher.　Therefore, quality of data management, classification, visualization and conservation, is not only homogeneous. Of course, there is no guarantee that treatment of data is presented in exactly the same form as previous treatment. To dissolve these problems, we have developed and tested a support system of data analysis using the Learning Vector Quantization（LVQ）which is a kind of neural network model.

1. 緒論

医療・バイオ・光・環境など多くの研究分野で，データを可視化することによる解析が行われている．情報抽出や実験の評価といった必要に応じ，様々な角度から目的にあった可視化処理を行うことが重要である．解析手順のなかでデータを効果的に可視化できるかどうかは，研究進展に大きく影響を及ぼす．可視化処理を含めた解析は，通常，実験やシミュレーションの終了後，インタラクティブに行っていることが多い．この解析手順は，各分野において細かな違いがあるが，概ね以下のような手順である．

- 実験によって生成されたデータの分類
- データ種類・実験種類ごとに可視化手法の選択
- 可視化のためのパラメータ設定
- 可視化
- 可視化処理によって生成したデータの分類
- 各種データの目視・比較による解析

解析を効率よく行うためには，研究過程で生成されたデータの起源や実験情報などを関連付けて記録しておく必要がある．しかし，実験計測器や計算機の高性能化やネットワーク環境の普及，可視化手法の多様化によって，生成されるデータは多様化し，加速度的に増加している．また，多くの大学・研究機関ではコストの問題などから，原子炉・人工衛星分野で利用されているような研究活動を支援する包括システムは構築されていない．そのため，データの管理は，研究者個人が記憶やメモなどによって行われている．そのため，稀にデータ自身の紛失という最悪の事態が起こるだけでなく，データの付随情報等は，日常的に欠損してしまっている．結果として，データ分類・探査作業に研究の多くの時間を費やしてしまっている．

一方，データの可視化処理における手法選択やパラメータ設定は，研究者の個人的経験的な判断によって行われている．しかし，全ての研究者が可視化処理について専門的な知識をもっているわけではないので，可視化処理が系統的に実施されず，重要な現象を見逃してしまうことも多くなってきている．また，大量なデータから有用な情報を探査しなければならないことも多くなり，インタラクティブな可視化処理を行う時間が増大し，実験そのもの時間よりそれら作業的な時間が大きくなってきている．

それゆえ，いくら実験がスムーズに行われ，大量のデータ計測が可能であっても，データ分類や可視化処理といった作業が研究の大きなボトルネックとなってしまっている．このような背景から，我々はデータ分類や可視化処理を支援するデータ解析支援システムの開発に着手した．これらのデータ分類や可視化処理を支援するシステムを開発するためには，研究者個人のノウハウの抽出が必要である．しかし，これは安易に抽出可能なものではなく，単純なルールベースとして記述できるものではない．そこで，演繹的なアルゴリズムを組み立てるのではなく，パターン認識手法を用いて研究者の分類・可視化ノウハウを吸収することを試みた．

パターン認識とは，認識対象がいくつかのカテゴリに分類できるとき，観測されたパターンをそれらのカテゴリのひとつに対応させる処理である．これまで，最尤法，ニューラルネットなど様々なパターン認識手法が提案され,利用されている.最尤法は，各カテゴリの確率分布が多次元正規分布に従い，かつその出現確率が等しいという仮定の下で，ベイズの定理を利用して分類を行う方法である．しかし，研究過程で生成されたデータを分類するカテゴリが前述した仮定を満たすとは限らない．そこで，前述した最尤法の問題点を回避するため，パターン認識手法としてニューラルネットワークを用いることにした.

ニューラルネットワークの代表とも言える誤差逆伝播法(BP 法 : Error Back-Propagation Method)にはローカルミニマム問題があり，これらが未解決問題として残されている．一方，学習ベクトル量子化法(LVQ : Learning Vector Quantization)は，教師あり学習によって入力データを類似度に応じて分類する能力を自律的に獲得していくという特徴をもっている[1]-[3]．この手法は，ローカルミニマム問題を回避できる，単純なアルゴリズムで分類能力を獲得することができる，という利点がある．よって LVQ を本システムのコアエンジンとして採用した.

本研究では，日本原子力研究所 関西研究所 光量子科学研究センターで実施されている X 線レーザ研究で生成された計測データを用いて，本提案手法がデータ分類・可視化処理において有用であるかの検証を行う．特に，解析支援の種類に応じた良い分類精度をもつニューラルネットワークを構築するために，LVQ の学習ベクトルの次元や参照ベクトル，学習回数の選定について検討も行った．これらの結果に基づいて，提案手法の有用性を検証する.

2. 光量子科学研究データに関する現状と課題

日本原子力研究所 関西研究所 光量子科学研究センターでは，極短パルス超高ピーク出力小型レーザなどの先進的レーザの開発やこれらを用いた利用研究が行われている[4],[5].これらの研究における実験や計算機シミュレーションは大規模，高精度であるので，大量の計測データが発生する．発生した計測データは，様々な可視化処理および統計解析が行なわれ，その可視化データや数値データを目視，グラフ化を行い，研究が進められる.

現在，この計測データや可視化データおよび統計データが大量となってきているため，それぞれのデータの生成過程および派生関係を把握することが難しくなってきている．そこで，それらの問題を解消するために，様々な研究支援システムが開発されている[5]．光量子科学研究センターでは，実験・シミュレーションの過程およびデータ解析履歴をデータベース化する「光量子科学理論・実験データベースシステム(実験 DB システム)」が開発され，利用されている．この実験 DB システムの導入によって，より効率的に解析が行われ，研究の品質も向上した.

しかしながら，解析手順の効率化による実験数の増加とともに，データの分類や可視化処理が研究者の負担となってきており，研究者が創造的研究に専念することができなくなってきている．現状，研究者の経験によって行われているデータの分類や可視化は，演繹的なルールベースとして記述することができず，自動化が行われていない．そのため，データ分類や可視化の技術を吸収し，それらの作業を支援することで，研究者の負担を軽減する仕組みが切望されている.

本研究は，実験・シミュレーション研究活動において研究者がデータの分類や適切な解析を容易に行えるためのデータ解析支援システムを開発し，研究者の負担を軽減するとともに,研究者のデータ解析,分類ノウハウのデジタル化を試み，多くの研究者の共有知識とすることを目的とする.

3. データ解析支援システムの評価対象

ここでは，前章で述べたような研究者支援と研究品質の保証を目指したデータ解析支援システムについて述べる．本システムは，パターン認識技術を用いて，研究者の専門的な判断を必要としてきたデータの分類や解析処理といった作業から研究者を解放し，創造的研究に専念できるようにする．そのパターン認識技術として教師あり学習アルゴリズムである LVQ を採用する.

LVQ の利用方法は，データ分類と可視化処理とで異なる．データの分類作業には，カテゴリは各ケースによって違うものの，LVQ のカテゴリ分類を行う性質を直接適用することができる.可視化処理には,その作業に直接 LVQ を適用することはできない．しかし，可視化処理では，可視化パラメータが対象のデータに,適当なものであるか,などの判断を伴う.これらの判断を LVQ に代行させる形で，可視化処理に LVQ の学習アルゴリズムを利用する.

LVQ の学習アルゴリズムとして,Kohonen によって提案された基本アルゴリズムである LVQ1 の改良アルゴリズムである最適化学習率 LVQ1(OLVQ1 : Optimized-learning-rate LVQ1)を用いる[1]-[3].OLVQ1 は，学習の収束速度が速いが，そのアルゴリズムが複雑ではないのが特徴である[3].本研究では,他のアルゴリズムに比較して定数パラメータが少ない点も考慮し，OLVQ1 を採用した.

具体的な評価対象は，データペアリングの真偽判定，データの分類，可視化処理パラメータの最適解の探査の３つとした．以下，それぞれについてもう少し詳しく説明する.

試行１：

光量子科学実験では，実験計測データの他に実験前に環境ノイズ（バックグラウンド）を計測しておき，実験の評価時には，それらの差分データを作成し，その差分データを元に詳細な分析を行う．しかしながら，このデータのペアリング（実験計測データ：バックグラウンドデータ）は，熟達した研究者でも間違ってしまうことがある．この試行では，実験計測データとバックグラウンドデータの適切でない組み合わせを検出し，研究者にそれを高い確率で提示できるかを評価する.

試行２：

実験で計測されたデータは，様々な形で分類できるように属性値が割り当てられる．例えば，この計測データは，高次高調波実験のものであるとか，焦点をずらした撮像結果であるとかである．そして，一般的にその属性値も複数あり，ひとつの計測データは複数の意味情報をもっている．本試行では，同じデータ群に対して，異なる 2 種類（実験種類，視野種類）の分類をさせ，研究者の分類作業の支援ができるかどうかを評価した．

試行 3：

計測データが 2 次元の場合，より定量的な判断を容易にするため断面を抽出する可視化処理を行うことが多い．しかしながら，断面生成のためには，いくつかの可視化パラメータを決定する必要がある．X 軸と Y 軸方向の 2 種類の断面生成であると限定しても，断面の位置およびその前後の平均化範囲などの自由度がある．本試行では，X 軸と Y 軸方向の 2 種類の断面生成に対して，研究者が適切と判断する可視化パラメータを選び出すことができるかどうかを評価した．

最後に LVQ 学習の入力データと LVQ による学習終了規則について説明する．LVQ 学習では，入力データをベクトルとして特徴ベクトル空間に表現し，これにカテゴリ情報を加えてトレーニングデータとする．本試行では，2 次元画像データの強度値から統計的情報（平均値，標準偏差，3，4，5 次モーメント，最大値，最小値，最大値の座標，最小値の座標）および断面生成情報（断面生成座標，平均化幅）を抽出し，これを LVQ 学習の入力データとした．

各試行におけるカテゴリは，試行ごとに違うので後述するとして，ここでは，トレーニングデータとテストデータの選定方法について述べる．本試行の一連の流れとして，対象データから選定したトレーニングデータを用いて LVQ 学習を行い，トレーニングデータ以外の対象データをテストデータとし，これを用いてその識別精度を評価する．本試行では，実際に分類を要求される際のデータ数も，入手可能な対象データの比に近いことを考慮し，対象データの比にあわせた．トレーニングデータは対象データの 9 割，テストデータは 1 割とした．

また，LVQ による学習の終了条件は，学習回数のみで終了を判定する．学習回数は参照ベクトル数の 40 倍程度が良い精度を出すとされている．そこで，学習回数は参照ベクトル数の 40 倍と一意に決定し，参照ベクトル数は，100，550，1000，10000 の 4 つのパターンに関して行った．結果として 550 以上であれば，ほぼ同じ結果が出ることを確認した．

4. 結果

試行1：バックグラウンドノイズ除去が適切ではないデータ検出の評価

本評価では，本来のカテゴリに分類させるべきデータ（間違ったペア＝incorrect）をできるだけ取り落としなく分類させたい．よって，ノイズ（本来は，正しいペア＝correct であるにもかかわらず，間違っていると判定してしまう）を許しても情報落としをしない方が重要なため，この評価では，incorrect 判定の再現率を重視する．

カテゴリ名	データペア数
incorrect	171
correct	122
合計	293

表 1．ペアリングのデータペア数

カテゴリ	精度(%)	再現率(%)	正答率(%)
incorrect	100	96.50327	97.96552
correct	95.49451	100	

表 2．ペアリング判定の精度および再現率

上記結果の通り，96.5%以上の確率で間違ったペアを判定できることになり，また，正しいペアを間違っていると判定する精度も 5%以下であるので，ノイズ，数え落しとも少なく，この問題に関しては，高い性能の判別能力を持っていることが明らかになった．

試行2：研究者の指定するカテゴリによるデータ分類の評価

本評価では，本来のカテゴリに分類させるべきデータをできるだけ取り落としなく分類させたい．よって，ノイズを許しても情報落としをしない方が重要なため，この実験では再現率を重視する．

実験種類の分類カテゴリは，実験の種類によって HHG(高次高調波），X(X 線レーザ)，Seeded(Seed 光)に分けられる．

カテゴリ	データ数
HHG	210
X	107
Seeded	129
合計	446

表 3．実験種類分類のデータ数

カテゴリ	精度(%)	再現率(%)	正答率(%)
HHG	93.5363	95.19737	83.40909
X	77.8181	68.01282	
Seeded	69.2207	79.91622	

表 4．実験種類分類判定の精度および再現率

高次高調波実験の分類は，95%以上の再現率で行えており，人間の分類間違い低減させることを強く支援できるレベルである．X 線レーザおよび Seed 光では，分類能力が上がらなかった．目視確認でもこれら 2 つの分類は，非常に専門性が必要であることがわかっている．高次高調波の分類精度が高いことを考えると，完全のこの 2 つの分類間の未分離であることが原因であり，研究者の分類を支援するためにはより高次の画像統計情報を組み入れ，分類トレーニングを強化しなければならないと思われる．

次に，計測視野の種類による分類結果について説明する．分類は，FarField（焦点のずれている計測）

と Focus（焦点の合った計測）の 2 カテゴリであり，データ数は表 5 の通りである．

カテゴリ	データ数
FarField	119
Focus	327
合計	446

表 5. 計測視野のデータ数

カテゴリ	精度(%)	再現率(%)	正答率(%)
FarField	93.24176	90.00000	95.22727
Focus	96.41041	97.15726	

表 6. 計測視野分類判定の精度および再現率

この試行では，分類能力が高く，双方 90%以上の再現率，精度となり，人間の判別能力を大きく上回っていることがわかった．

試行3：可視化処理パラメータの有用性分類

本試行では，2次元データからの特徴的断面生成の自動化を目的とし，可視化パラメータの有用性に関する分類を行う．断面図処理は，X軸とY軸方向の2種類に関して行う．また，断面の処理でのパラメータは，断面の位置および幅であり，適切な位置，幅は，各データによって決めねばならず，一意に決めることができない．その為，LVQによって断面生成に適切なパラメータであるかの判断を学習させる．

この例では，断面図処理におけるパラメータの有用性の度合いを LVQ で分類する．適切であると判定された分類中に適切でない可視化パラメータが混入することがもっとも問題なので，この評価では精度を重視する．

まず，X 軸方向断面処理におけるパラメータ有用性分類カテゴリは， Good(適切)と Bad(不適切)の 2 カテゴリであり，データ数は表 7 の通りである．

カテゴリ	データ数
Bad	542
Good	189
合計	731

表 7. X 軸方向断面候補のデータ数

カテゴリ	精度(%)	再現率(%)	正答率(%)
Bad	90.77259	94.53416	85.49796
Good	89.67677	73.65497	

表 8. X 軸方向断面候補判定の精度及び再現率

Y 軸方向断面処理の場合の分類カテゴリは Good(適切)と Bad(不適切)の 2 カテゴリであり，データ数は表 9 の通りである．

カテゴリ	データ数
Bad	119
Good	327
合計	446

表 9. Y 軸方向断面候補のデータ数

カテゴリ	精度(%)	再現率(%)	正答率(%)
Bad	87.53259	86.86614	83.02817
Good	74.9036	75.1667	

表 10. Y 軸方向断面候補判定の精度及び再現率

精度が 90%近くあり，可視化パラメータ選択の自動化を実現できる可能性があることを示せた．

5. まとめ

本研究では，日本原子力研究所 関西研究所 光量子科学研究センターにて実施されている X 線レーザ研究を対象に，データ分類や可視化処理の一部自動化によってデータ解析支援するシステムの設計を行った．

データ分類と可視化処理機能を実現するために，データ統計情報に LVQ を適用する手法を提案し，提案手法の有効性について検証した．

- バックグラウンドノイズ除去が適切に行われなかったデータの検出と視野種類による分類において 95%以上の再現率を示し，本手法が非常に有用であることが確認できた．
- 可視化パラメータ選択の精度が 90%程度であり，自動化を実現できる可能性を示すことができた．
- データ種類による分類と実験種類による分類においても，学習の設定条件や分類カテゴリをチューニングすることで認識精度を向上させる可能性を示すことができた．

本研究で提案した分類手法は，データの一般的な統計情報を使用するため，対象とする X 線レーザ研究に限定されるものではない．また，本論文で検証した可視化処理もまた，一般的な解析処理である．そのため，他分野の研究で生成されるデータへの応用も期待できる．

謝辞

本研究を進めるにあたり光量子科学研究センター長をはじめとするセンター各位ならびに奈良女子大学大学院　人間文化研究科各位の協力に感謝する．

参考文献

［1］ T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, Germany (1984).

［2］ T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, and K. Torkkola, “LVQ_PAK: The learning vector quantization program package version 3.1”, *Helsinki University of Technology, Laboratory of Computer and Information Science*, Finland (1995)

［3］ H. Senay and E. Ignatius, ”A Knowledge-Based System for Visualization Design”, IEEE CG&A, Vol. 14, No. 6, pp. 36-47 (1994)

［4］ 上島豊，超並列計算機を使った超大規模光量子シミュレーションの現状と課題，Journal of the Japan Society for Simulation Technology, Vol.19, No.4 p.27‐36（2000）

［5］ Y. Ueshima, Y. Kishimoto, Large-scale simulation and advanced photon research, Lecture Notes in Computer Science 1940, pp. 524-534 (2001)