PLAIN News
宇宙航空研究開発機構 宇宙科学研究本部
科学衛星運用・データ利用センター
http://www.isas.jaxa.jp/docs/PLAINnews/
Center for Science-satellite Operation and Data Archive
SODA ISAS/JAXA
2010.3.26/ No. 197

衛星データとデータマイニング
第３回 データマイニングの活用例

本田 理恵（高知大学自然科学研究系理学部門 /JAXA 宇宙科学情報解析研究系 客員）

1. はじめに

2009 年９月から開始した本連載も最終回となりました。今回は衛星データへのデータマイニングへの活用例を紹介します。衛星によって取得されるデータには，複数の属性を持つ数値，カテゴリ値データ，時系列データ（多次元），スペクトル（多次元），画像（マルチチャンネル），時系列画像（動画）等が有り，データの形式と目的（タスク）によって適用する手法を選択することになります。今回は衛星データに特有の（時系列）画像，スペクトルについて，筆者と共同研究者が行った研究を中心に扱っていきます。なお，紹介する事例には地球観測衛星の例が多く含まれますが，宇宙科学データにおいても共通する問題を多く含むため，一般性を損なう事はないと考えています。また，実例では，前回までに紹介した代表的な手法だけでなく多様な手法が複合的に使用されていますので，初出の手法についてはその都度簡単な紹介を加えていく事にします。

2. 時系列気象画像に対するデータマイニング

筆者の所属する高知大学では 1996 年以来，東京大学生産技術研究所，気象業務支援センター等から配信された気象衛星画像ひまわり５号（GMS5），GOES，ひまわり６号（MTSAT）の日本周辺の画像をアーカイブしています [1]。本節ではこのデータをテストベッドとして実施された３つの検討事例について紹介します。

2.1 クラスタリングと時間相関ルールの抽出

我々が気象画像（雲に感応する IR 画像）を観察するとき，“典型的な夏の画像”，“梅雨の時期の画像”，“冬の画像”など，経験によって特徴に応じた画像のタイプを判断する事ができます。このようなラベル付けを自動的に計算機に実施させて大量の時系列画像を記号系列に変換することができれば，そこから時間変動のパターンなどの知識を抽出する事ができると期待されます。

図１は，このタスクのために開発したシステムの概要です [2][3]。まず特徴に応じた画像のラベル付けを行うために画像集合に対してクラスタリングを実施します。この際，雲塊の位置にはこだわらず，“台風と前線を含む画像”といった画像の意味で大まかにグループ化するために，２段階のクラスタリングを実施しています。まず，画像をブロック化し，ブロック毎の特徴ベクトル（輝度ベクトル，または FFT パワースペクトル）に対してクラスタリングを実施します。得られた結果から１枚の画像中に含まれるブロックのクラスタについての頻度分布を求め，この頻度分布を画像の特徴ベクトルとして，再度クラスタリングします。最終的に得られたクラスタは，画像中に含まれる台風や前線などの領域がしめる面積によってグループ分けされ，それぞれの台風の位置などには影響を受けない事になります。

クラスタリングの結果，時系列画像は（A, A, A, A, B, B, A, C …）クラスタラベルの時系列に変換されます。ここから“A の後には B が発生する”，“B の後には C が発生する”，といったルールを抽出するには，第２回に説明した相関ルールに類似した手法を用いる事ができます。時系列の時間窓をバスケット，時間窓中に連続して存在するクラスタラベルをアイテム（イベント）と見なして，相関性の高いルールを抽出することができます。また，こうして抽出されたルールやクラスタラベル，画像をデータベースに格納することによって，ユーザーが対話的に高次の知識発見を支援することも可能となると考えられます。

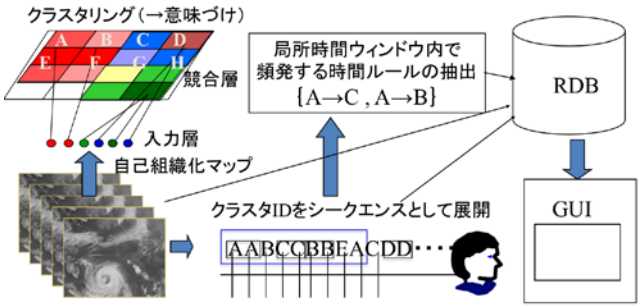

図１ 時系列気象画像のクラスタリングと時間相関ルール抽出 [2][3]

なお，上記の手法のクラスタリングには２層のニューラルネットワークの１種である Kohonen の自己組織化マップ（self organizing map, SOM）[4] を用いています。SOM は本稿で紹介する他の問題でも利用されており，教師データがない状態で，大量のデータ集合の内容を理解するのに非常に有効な手法ですのでここで紹介しておきます。

SOM は図１の左上の略図のように，入力ベクトルを受け取る入力層と２次元のユニットアレイからなる競合層からなり，競合層の各セルは入力データと同じ次元のベクトルを記憶しているものとします。出力層の各ユニットの初期値を乱数で与えた後，各入力ベクトルに対して最も似通ったベクトルを持つ競合層のユニットを“勝者ユニット”として，学習データを割り付け，それと同時に勝者ユニットのデータを学習データに近づけます。この際，勝者ユニット近傍のセルも弱く学習させることによって，競合層は入力ベクトルの分布を徐々に学習し，競合層上で近い位置に似た特徴が配置されるようになります。

SOM は本来クラスタリングを目的とした手法ではありませんが，各ユニットに割り付けられた入力データグループ

［裏へ続く］

をクラスタとしてみなして扱う事ができます。また各グループの類似度が競合層の上での２次元的な距離として視覚化できるため，教師無しデータの発見的な学習にも有効です。

図２に実際に時系列気象画像に対しての学習の結果，得られた競合層の各ユニットに代表的な画像を貼付けたマップを示します。競合層の２次元的なマップ上に特徴の異なるクラスタが特徴に応じて連続的に分布していることがわかります。なお，特にクラスタ間の関係を詳しく評価する必要がないのであれば，第２回で紹介した k-means 法などのクラスタリング手法を使用する事も可能です。

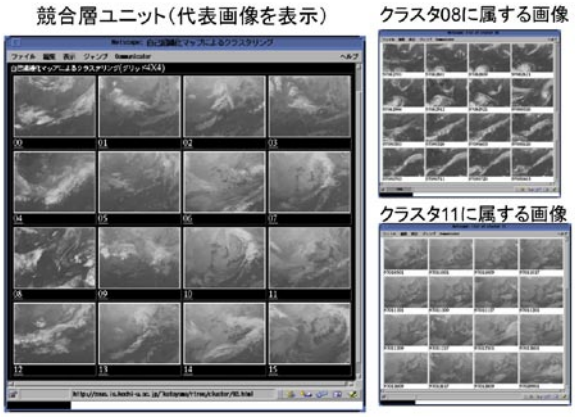

図２　学習後の競合層の各ユニットの代表画像（左右，上から下の順にクラスタ 0,1,2,…15）（左）と 8,11 にクラスタリングされた画像集合（右）[2][3]

2.2 隠れマルコフモデルによる季節変動のモデル化

前節で得られたクラスタラベルの時系列に対して，さらに別の形の時間変動パターンを抽出することを考えてみます。図２において得られた画像クラスタの中身を調べると，“夏の終わりの台風の画像”，“梅雨期の前線の発達した画像”，といったように，画像の特徴と背後にある状態（季節に相当）が密接な結びつきをもっていることが予想されます。

この背後に隠れている状態に着目した隠れマルコフモデル（Hidden Markov Model: HMM）[5]は確率モデルの１種であり，音声認識，自然言語処理などに広く応用されています。HMM は確率的に遷移する状態集合（マルコフモデル）と，各状態に対する確率的な記号出力から構成されます。観測できるのは記号出力系列のみで，背後の状態遷移系列は直接的には見ることができないため，“隠れ“という言葉が手法名に冠されています。HMM では，ある記号系列が与えられたとき，それぞれの記号が生成時の状態や，次の時刻の状態や出力記号を推定することもできます。

今回の問題では，クラスタラベル（画像タイプ）が“記号系列”，その背後にあるなんらかの季節に相当するようなものが“状態“と想定することができます。

モデリングにあたっては，記号系列だけから，状態間の遷移確率，各状態での記号の生成確率などのモデルパラメータを推定します。状態の数は既知である必要がありますが，状態数が未知の場合，複数の状態数に対する試行結果を比較することによって，適切な状態数を発見しなければなりません。

図３に４年分の時系列気象衛星画像に対して得られた隠れマルコフモデルを示します[6]。ここで，隠れた状態の数については最初から常識的な季節の数である“4”とはせず，2-8 の範囲で試行し，情報量基準をもとに最適解として状態数５を選びました。

図３には各状態の発生頻度，遷移確率がグラフとともに数値で示しています。例えば状態２（上端）の発生頻度は 0.2244 であり，次の時間には 0.92 の確率で同じ状態へ、0.06，0.03 という小さい割合で状態１，状態３に遷移します。なお､各状態に添えられたグレースケールマップは，その状態におけるクラスタ（画像タイプ）の発生頻度を図３左上のクラスタの代表画像マップに対応させて表しています。このカラーマップと，状態の発生時期の比較から，各状態に意味付けすると，｛冬｝，｛春，秋，梅雨｝，｛春，秋｝，｛夏，秋｝，｛盛夏｝となります。

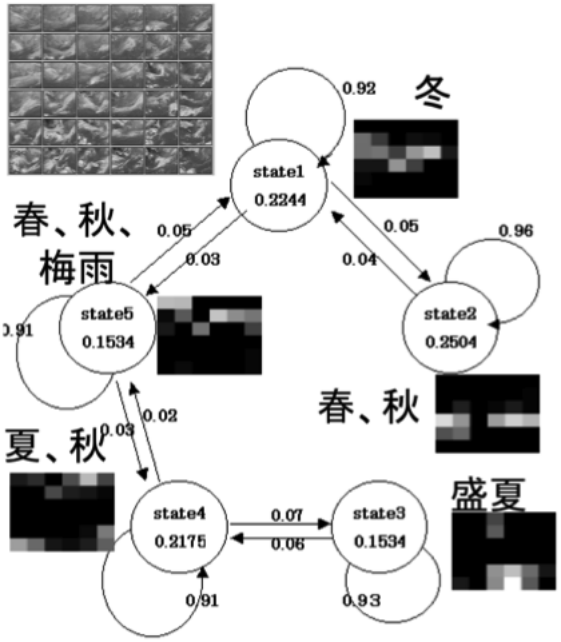

図３ 時系列画像（4 年分）に対して状態数５で得られた隠れマルコフモデル [6]。左上は記号化に使用した自己組織化マップの競合層に代表画像を付与して示したもの。HMM の○は状態，中の数字は初期発生確率，矢印付属の数字は状態遷移確率，画像に添えられたグレースケールマップは，各状態でのクラスタの発生頻度（明るいほど高い）を示す。

また，この結果から各状態は安定でわずかな確率でチェーン状に遷移することや，春や秋の状態が 2-3 種類に分かれること，真冬からの主要な遷移先が２種類あるのに対して，盛夏からの主要な遷移先は一つしかない，といった興味深い知見が得られます。こうした手法は，時間とともに変動するスペクトルの分析などにも適用の可能性があると考えられます。

2.3 雲塊の自動抽出と追跡

さらに視点をかえて，画像の中にふくまれる一つ一つのオブジェクト（この場合は雲塊）に注目してその情報から時空間変動パターンを抽出する問題を考えてみます。対象の形状が決まっていれば，この問題は画像認識分野のテンプレートマッチングやトラッキングの手法で扱う事ができます。しかし，取り出すべき物体の“形状が不定”，そ

の“個数が未知”，また“一部は重なり合う場合もある”といった点が問題となります。こうした問題を解決するには，データマイニングのクラスタリングで使用される確率密度分布の混合分布によるモデリング [7] が有用です。確率密度分布として多変量正規分布を用いることによって，傾きや楕円状の分布も扱うことができます。

図４に具体的な計算過程を示します。まず画像を雲と背景に分かれるように２値化し，雲に相当するピクセルの座標をサンプリングして雲点とします（図４中）。この雲点の分布を混合多変量正規分布でモデル化し，モデルのパラメータを推定します（図４右）。パラメータの推定には EM アルゴリズム [5] を使用することが一般的です。なお，成分数が不定の場合，多数のケースについて試行して最適な結果を選択することが必要になります。

図４　多変量正規分布による雲塊のモデリング。左から原画像，２値化画像，混合分布によるモデリング結果 [8]

また時系列画像での雲塊の追跡には，前の時刻の解に，消滅や，生成を考慮に入れてばらつきを持たせた初期値を導入することによって対処することができます。図５にこの手法で実施された台風の追跡結果を示します [8]。ここでは、一つの雲塊の追跡とともに派生して分離する成分の取得も実現できていることが確認できます。

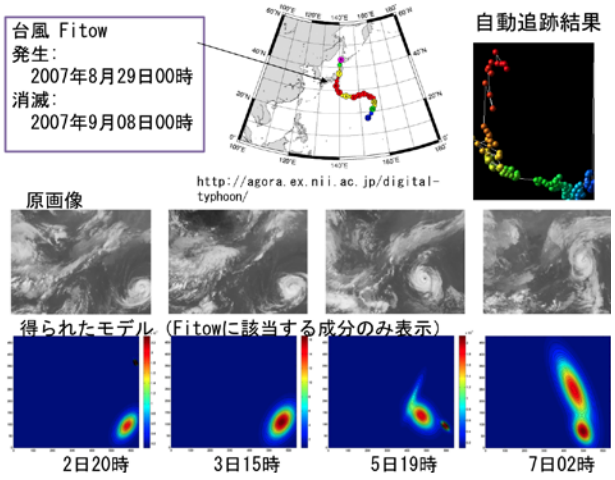

図５　2007 年台風 Fitow の自動追跡事例。上段中央があらかじめ求められている台風の中心経路で，上段右側が混合分布モデルで求められた軌跡 [8]。

このようにして抽出されたオブジェクトの情報に，さらに雲自体のパターンの特徴（テクスチャなど）やその他の観測値を組み合わせることによって，高次の知識発見の支援を行うことができると期待できます。このような手法は，大量画像からの未知天体の自動抽出やその時間変動の追跡，流体シミュレーションの結果の理解にも応用可能と考えられます。

３．惑星画像への適用例

次に月の画像に対する適用例を示します。ここでは惑星画像からのクレーター地形の抽出，およびマルチバンド画像からの地質図の作成について紹介します。

3.1 惑星画像からのクレーター地形の抽出

画像からのクレーターや火山などの特徴地形の抽出には，画像認識の手法の他に，正例と負例（間違った事例）の集合から機械学習の手法によって認識器を作成するというアプローチが検討されてきました。しかし，特に光学的に得られた画像は照明条件の影響を大きく受け，ある条件で作成された認識器が他の条件でも正しく機能するとは限らない，という問題がありました。

図６に示す惑星画像からのクレーター抽出システム [9] では，データマイニング的な視点を取り入れ，まず画像集合を画像の特徴によってクラスタリングしてグループ化し，各グループに対して認識器をチューニングする手法を提案しました。このように従来分析が難しかった手法にも，データマイニング的な観点から，クラスタリングによって自動的なグルーピングを行うというプロセスを導入することによって，問題を系統的に扱いやすくすることができます。

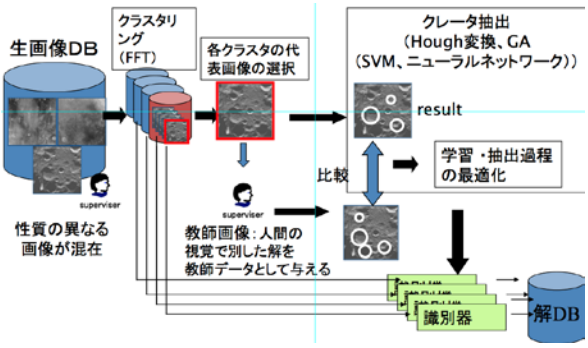

図６　惑星画像からのクレーター抽出の手法 [9]

3.2 マルチバンド画像からの校正の自動化と地質図の作成

近年，かぐや，Chandrayaan などの探査機によって月のマルチスペクトル画像，連続スペクトルが続々と取得されてきています。これらのスペクトル観測の目的は表面の物質分布を知ることですが，そのためには観測条件の影響を取り除くために位相関数などを用いた校正が必要です。しかし，位相関数は理論や観測から物質依存性があることが指摘されていました。

横田 (2003)[10] は，この問題に対して，標準的校正式による校正で仮校正したスペクトルに対するクラスタリング，クラスタごとの校正式の決定という過程を繰り返すことによって，物質グループごとの校正曲線と物質分布図を同時に取得する手法を提案しました。図７にその概念図を示します。この手法のクラスタリングの部分には前節で紹介した SOM が使用されています。この事例

［裏へ続く］

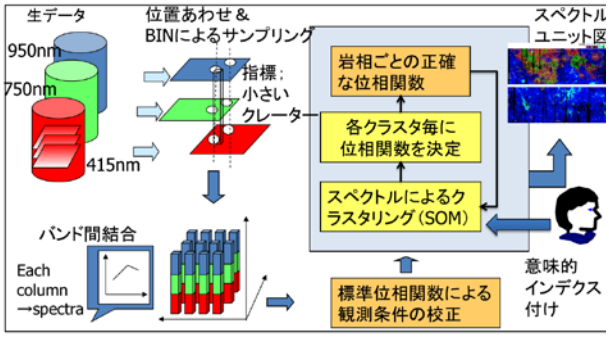

図７　マルチバンド画像からの校正の自動化と地質図の作成 [10]

も従来人による操作では膨大な時間を要して，現実的に不可能であった作業が，データマイニングの視点から，クラスタリングの過程が導入されたことによって解決されうることが示した例ということができます。

4. おわりに

今回は時系列気象画像について３種，惑星画像について２種の例を衛星データへのデータマイニングの活用例として紹介しました。データマイニングの枠組みで扱われる手法には，今回取りあげた手法に限らず多様な手法が含まれ，大量データを扱う研究者にとってはこれらの手法自体が宝の山といえます。

とはいえ，データマイニングや機械学習はコンピュータの非専門家にとってはまだ敷居の高い領域で有るように見えます。一方，データマイニングの成功には専門家の知見が必須であり，データを扱う分野の専門家とデータマイニングなどの専門家の緊密な連携が重要になります。

アメリカのジェット推進研究所には The Machine Learning and Instrument Autonomy (MLIA) Group[11] が存在して，こうした分野の研究を行っています。大量データのアーカイブから利用に向かう流れのなかでは，こうした分野の専門の研究員や常勤のスタッフを迎えたり，外部の研究機関との連携を組織的に形成したりする事も重要かもしれません。最後に今回の連載を通じて少しでもデータマイニングについて興味を持っていただけましたら幸いです。

文献等

[1] 高知大学気象情報 http://weather.is.kochi-u.ac.jp/
[2] 片山幸治，小西修，情報処理学会論文誌:データベース、40-SIG 5 (TOD 2)，69-78. 1999.
[3] Honda R. et al., Principles of Data Mining and Knowledge Discovery: Fourth European Symposium, 204-215, 2001.
[4] Kohonen T.， 自己組織化マップ， シュプリンガーフェアラーク東京，2000.
[5] 北研二，辻井潤一，確率的言語論モデル，東京大学出版会，1999
[6] 勝吉進一，高知大学理学部卒業論文，2004
[7] 元田浩ほか，データマイニングの基礎，オーム社，2006
[8] 石津光洋，高知大学理学部卒業論文，2009
[9] Honda R. et al., Progress of Discovery Science, LNAI2281, 395-407, 2002
[10] 横田康弘，東京大学博士論文，2003
[11] http://ml.jpl.nasa.gov/

平成 21 年度 宇宙科学情報解析シンポジウム
「宇宙科学データの『見せる化』」報告

三浦 昭（宇宙科学情報解析研究系）

平成 21 年度の宇宙科学情報解析シンポジウムを 2 月 23 日に開催致しました。

「宇宙科学データの『見せる化』」をテーマに、宇宙科学にまつわる観測手法から各種データの可視化、映像表現等非常に幅広いご講演を賜りました。またこの度は多くの方々にご参加頂き、宇宙科学情報解析シンポジウムとしては稀に見る盛況となりました。

全天周映像や球体投影等、3 次元・4 次元映像可視化設備の普及や映像化手法の発展により、これらの映像を多くの人が楽しめるようになってきました。3 次元データの平面投影や、ゲーム機・PC・インターネット等、さまざまな場面で宇宙科学データに触れることができるようになってきています。JAXA にまつわる宇宙科学データ、軌道データ等も様々な切り口で、見て、探して、さらには触れる機会も提供されるようになりました。地球上の現象や宇宙の現象など、これまで容易に見えなかったものが、さまざまな映像化技術で目に見えるようになりました。望遠鏡等の観測機器も進化を続け、さらに高精細の宇宙が見えるようになりつつあります。さらには「はやぶさ」の全天周映画等、「見せる化」に携わる方々の情熱が、その先にある感動をもかいま見せてくれたシンポジウムでもありました。

お忙しい中、ご講演をご快諾頂いた皆様方、出席された皆様方に、厚く御礼を申し上げます。

当日の発表資料は、PDF 等によるダウンロードを予定しております。

編集発行：宇宙航空研究開発機構 宇宙科学研究本部 科学衛星運用・データ利用センター
〒 229-8510 相模原市由野台 3-1-1　Tel.042-759-8767　　住所変更等 e-mail：news@plain.isas.jaxa.jp
本ニュースはインターネットでもご覧になれます．http://www.isas.jaxa.jp/docs/PLAINnews