

衛星データとデータマイニング 第1回 データマイニングとは

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

今回から3回にわたり、データマイニングとその衛星データへの応用について紹介していきます。筆者は情報科学分野のデータマイニングや機械学習の手法を衛星データからの知識発見に適用する研究を行っています。きっかけとなったのは1990年代後半の月周回衛星探査計画 SELENE（かぐや）プロジェクトの開始でした。4TBを超える月のデータをおよそ1年の間に日本の科学衛星で取得する、という内容は当時の想像をこえるもので、大量データの効率的解析手法についても新しい枠組みで検討しなければならないと実感させられました。2007年9月に種子島から打ち上げられた“かぐや”は、2009年6月11日に無事その任務を終え、テラバイト級の巨大な月のデータが今や現実のものとなり、11月から一般へのデータ公開も開始される予定です。

また、宇宙科学の分野でもスローン・デジタル・スカイ・サーベイ (<http://www.sdss.org>) や Virtual Observatory (<http://www.ivoa.net>, <http://jvo.nao.ac.jp>) などのプロジェクトで、テラバイト級の観測データベースが提供されてきています。このような状況は我々がこれまでに経験したことのないものであり、従来主流であった仮説検証型の研究手法に加えて、サーベイ型の観測データを前提にその量を質にかえるような新たな解析手法を模索する時期にも来ていると考えられます。

その一つの取り組みが情報分野のデータマイニングや、機械学習などの手法の適用です。これらはデータからの新しい知識取得の支援や、大量のデータの要約、カタログ化、意味付けをして可用性を高めることに適用可能であり、データが大規模化すればするほど重要になると考えられます。

本稿ではこのような状況をふまえて、データマイニングの紹介からはじめて、科学衛星データへの適用例や、実際に小規模な実験を試すことのできるツールの紹介、そしてデータセンタとの関わりについても述べていきたいと思います。

2. データマイニング、機械学習

1990年代、ストレージやCPUなどの低価格化と高性能化、データベース管理システムやインターネットの普及により、様々な分野で大量のデータが蓄積されるようになりました。一方で、データ量の増大に対して人間による処理が追いつかず、データが増えるほど逆に得られる知識の割合が少なくなることが懸念されました。このような背景のもとに1990年代半ばから十数年ほどの間に急速に発展したのがデータマイニングの研究分野です。

データマイニング (data mining) とは、大量のデータからコンピュータを用いて有用な知識を発掘 (mining) しようとする実用志向の技術の総称で、統計学、データベース、情報検索、人工知能分野の機械学習などの分野を源流として発展してきました。特に、機械学習 (machine learning) は、コンピュータが課題解決能力を自ら学習していくために必要なメカニズムを解明することを目的とした学問領域で、決定木学習などのデータマイニングの多くの重要なアルゴリズムがこの分野からもたらされています。

データマイニングはKDD (Knowledge Discovery from Database: データベースからの知識発見) という呼び方で呼ばれることもあります。初期のデータマイニングの発展を牽引したFayyadによれば、知識は、妥当（たまたま見つかったものではない）で、新規性、有用性があり、かつ理解可能であるパターンであり、この考えは一般的に受け入れられています⁽¹⁾。

図1に知識発見のプロセスを示します⁽²⁾。知識発見のプロセスは、大きく分けて、前処理、パターン発見、後処理から構成されます。前処理には、元になるデータベースの統合、不適当なデータの削除や形式の統一などのデータ洗浄、データウェアハウスの構築などが含まれます。さらに、構成されたデータウェアハウスから、パターン発見のタスクに関連のあるデータを選択して学習用のデータセットとする過程が続きます。中心プロセスであるパターン発見では、次節で紹介するような手法を用いて、データから様々な形式のパターンの発見を行います。後処理では得られたパターンを評価し基準を満たしたものを知識として抽出します。なお、この過程は1回限りのプロセスではなく、あらゆる過程において期待する成果が得られなければそれ以前のプロセスにも戻って改善の上試行を繰り返します。厳密には、この知識発見の全過程をKDDと呼び、この中のパターン発見のプロ

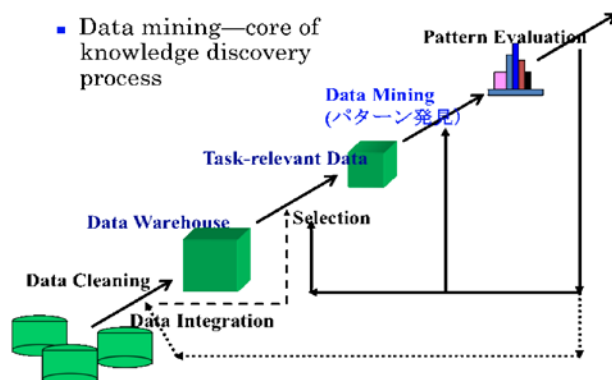


図1 知識発見のプロセス⁽²⁾

セスだけをデータマイニングと呼びます。しかし、実際には前処理や後処理（特に前処理）の過程が結果を大きく左右するため、実用では全過程を視野に入れた検討を行うことが必須となります。

3. データマイニングのタスクと手法

データマイニングの代表的なタスクには、分類、クラスタリング、相関分析、回帰などがあります。これ以外にも統計学的なモデリングなどの高度なタスクが存在します。また、データを記述すること自体を目的の場合と、得られた知識やパターンから予測を行うことを目的とする場合、その両方を目的とする場合があります。

分類はデータの属性からクラスを決定するもので、その手法には決定木学習、ルール学習、ナイーブベイズ学習、最近傍法、ニューラルネットワーク、サポートベクトルマシンなどがあります。分類すべきクラスと属性の両方が既知のデータを用いて学習を行うため、このような手法は機械学習の分野では、教師あり学習と呼ばれるジャンルに属します。

一方、クラスタリングはデータを複数の属性からなる多次元ベクトルとみなし、属性空間の中に分布するデータから特徴の似通ったものを塊（クラス）として検出しようとする手法です。クラスタリングの手法には大きく分けて階層的な手法と非階層的な手法があります。一般的に、クラスタリングはデータだけから手本なしに学習を行う教師なし学習のジャンルに属するものととらえることができます。

関連ルールは“AであればBも高い確率でおこる”というルールで、これをデータ集合から見つけ出すのが関連ルール分析です。A、Bには現象や属性値の組み合わせをとりうるため、属性数の多い問題に対しては組み合わせ爆発により分析困難となりますが、ア・プリオリをはじめとする効率的なアルゴリズムの開発により、大量データに対しても適用が可能になっています。

また、これらの汎用的なアルゴリズムに加え、取り扱うデータの種類によって特化される時系列データのマイニング、画像や動画のマルチメディアデータマイニング、グラフ構造のマイニング、テキストマイニングなどのカテゴリも存在します。

なお、これらの代表的な手法で行うような処理は、科

学的データ分析でも初期段階において一般的に実施されるものであり、大量データに対する科学的発見の支援にも有効と考えられます。

4. データマイニングのソフトウェア

前説で述べた代表的な手法の多くは商用ソフトウェアやフリーソフトウェアなどで利用することができます。このようなソフトウェアの情報はKddnuggets (<http://www.kddnuggets.com>) などのサイトで調べることができます。オープンソースフリーソフトウェアとしては、ニュージーランドのWaikato大学で開発されたWeka (<http://www.cs.waikato.ac.nz/ml/weka>) がよく知られています。Wekaには、前処理、代表的な手法によるパターン発見、評価、そして、データや分析結果の可視化の機能が実装されています。また、使用形態もグラフィカルユーザーインターフェースによる対話的処理、コマンドラインベース処理、Javaコードからの参照などをサポートし、ソースコードも公開されていますので、最初に各種のアルゴリズムによるデータマイニングを体験するには最適なソフトウェアの1つです。

ただし、利用にあたってはデータマイニングの諸過程の様々な手法がどのような原理でどんな効果があるのか理解していることが必要になりますので、現時点で興味を持たれた方は、一度、包括的な教科書（例えば(3)）に目を通した上でソフトウェアを使用ください。

5. まとめ

今回は、データマイニングとは何なのか、代表的な手法とソフトウェアにはどんなものがあるか、といったことについてその概略を紹介しました。次回以降は、個別の手法・具体的な適用例について述べていく予定です。

参考文献等

- (1) U. M. Fayyad et al., Advanced in Knowledge Discovery and Data Mining, 495-515, AAAI Press/MIT Press, 1996
- (2) J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann, 2006
- (3) 元田浩ほか、データマイニングの基礎、オーム社、2006

darts.jaxa.jp をよろしく

海老沢 研 (C-SODA 科学データ利用促進グループ)

にする予定です。

それにともなつて、DARTS を <http://darts.jaxa.jp> でもアクセスできるようにしました。実体はどちらのURLでアクセスしても全く変わりません。日本を代表する科学衛星データベースとして、さらに分野横断的に多くのミッションデータを集約し、より使いやすくしていくつもりです。今後ともDARTSをよろしくお願い致します。

DARTS (Data ARchives and Transmission System) は、C-SODA が管理運営している科学衛星データベースです。そのURLは、<http://darts.isas.jaxa.jp> として広く親しまれてきました。C-SODA は ISAS に所属する組織で、DARTS も ISAS の予算で運営されていますが、今後、ISAS 以外の JAXA ミッションによる科学データ、特に JSPEC (月・惑星探査プログラムグループ) が打ち上げた探査機による月惑星データも DARTS からアクセスできるよう

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>

●編集後記：鵜沼海岸でサーフィンを始めました。(K. E.)