時系列データマイニングによる動的ヘテロなシステムからの知識発見 -宇宙天気研究における大規模帰納処理システム構築へ向けて

knowledge discovery from heterogeneous dynamic systems by time series data mining - inductive computing systems for space weather

徳永 旭将 [1]; 中村 和幸 [2]; 樋口 知之 [3]; 池田 大輔 [4]; 大久保 翔 [5]; 藤本 晶子 [6]; 吉川 顕正 [7]; 湯元 清文 [8]; MAGDAS/CPMN グループ 湯元　清文 [9]
Terumasa Tokunaga[1]; Kazuyuki Nakamura[2]; Tomoyuki Higuchi[3]; Daisuke Ikeda[4]; Kakeru Ookubo[5]; Akiko Fujimoto[6]; Akimasa Yoshikawa[7]; Kiyohumi Yumoto[8]; Yumoto Kiyohumi MAGDAS/CPMN Group[9]

[1] 九大・理・地球惑星; [2] 統数研; [3] 統数研; [4] 九大・システム情報; [5] 九大・理・物理; [6] 九大・理・地球惑星; [7] 九大・理・地球惑星; [8] 九大・宙空環境研究センター; [9] -
[1] none; [2] ISM; [3] Inst. Stat. Math.; [4] ISEE, Kyushu Univ.; [5] none; [6] Earth and Planetary Sci.,Kyushu Univ.; [7] Earth and Planetary Sci., Kyushu Univ.; [8] Space Environ. Res. Center, Kyushu Univ.; [9] -

近年、太陽地球系科学の分野においては、太陽-惑星間空間-磁気圏-電離圏-地上での同時観測が可能となり、大規模なデータ群が蓄積されつつある。これらのデータセットから変数間の依存関係を抽出することは、太陽風 - 磁気圏 - 電離圏間の相互作用を理解する上で、とりわけシミュレーションによる再現が困難な領域間のダイナミクスに関する知見を獲得する上で、非常に重要である。ところが、現在までの宇宙天気研究においては、これら大量データセットを十分に活用しきれていないのが現状である。 そのため我々は、動的ヘテロな時系列データ群からデータ同士の依存関係のルールを高速にマイニングすることを目的とした、次世代の宇宙天気研究のためのシステム開発を行っている。

我々が対象といているシステムの特徴として、(1a) システム全体のスケールが巨大、(2a) システムが非定常な太陽風により駆動（データが非定常）、(3a) 磁気圏・電離圏など異なる領域が連動する（システムが動的）、(4a) ヘテロなデータセット、(5a) 変数の数が多い、の 5 点が挙げられる。これら (1a)〜(5a) の特徴が、宇宙天気研究におけるデータ解析を困難なものにしていると言える。「困難な要素」とはより具体的には：(1b) データ間の依存科関係は線形な相関関係のみではない（非線形・時間遅れを含む）、(2b) 解析に使用するデータ区間の選び方により解釈が大きく異なる可能性がある（解析区間の選出において解析者の主観が入り込みやすい。また、データ全体の傾向を抽出する従来の統計解析はほとんど無意味）、(3b) それぞれの変数は無限の変動パターンを持つ、(4b) データ間の依存関係を定量的に評価することが困難、(5b) 解析できるデータ数（event 数）に限界がある。過去行われて来た宇宙天気研究におけるデータ解析は、上記 (1b)〜(5b) の「困難な要素」に対し、どれ一つとして十分な対応がなされていない。本研究の目的は、これらの問題点に対して柔軟に対応できるような、次世代の宇宙天気研究のための帰納的データ解析システムを構築することである。

我々は宇宙天気研究におけるデータ解析について、時系列データマイニングという文脈から新たな枠組みを与える。データマイニングとは、膨大なデータの中から部分的な傾向を抽出する方法である（全体的な傾向を抽出する従来の統計解析とは異なることに注意）。本研究における時系列データマイニングの枠組みは以下の通り：Step1: データの変化点検出、Step2: 局所時系列モデリング、Step3: 時系列モデルを有限数に分類、Step4: 3 で分類したモデルの発生ルール抽出、Step5: 変数間の依存関係抽出

問題点 (2b) および (3b) への対応策として、データを比較的均一なセグメントに分割することが重要である（データの分節化）。近年、時系列の変化点検出は、工学の分野において重要な課題の一つとなっている。その代表的な方法として、局所定常 AR モデルに基づく変化点検出がある [e.g., Kitagawa, 2005]。ところが、我々が対象としているデータは前述の (2a) および (3a) の特徴を持つため、AR モデルは必ずしもよい出発点とは言えない。そこで我々は、パラメトリックな時系列モデルによる方法と比べ、より柔軟に変化点を検出出来得る方法を第 1 ステップとして用いることにした。具体的には、 テキストデータマイニングおよび [Moskvina and Zhigljavsky, 03; Ide and Inoue, 05] で提案されている SSA(特異スペクトル解析) を応用した変化点検出を採用する。第 2 ステップとして、第 1 ステップにより事前に抽出した変化点の出現パターンを拠り所として、局所定常時系列モデリングを行う。時系列モデルのパタメター値は最尤法によって、またパタメター数は情報量規準によって統一的に決定される。第 3 ステップとして、局所モデリングにより再分節化された変動パターンを有限な変動パターンにクラスタリングする作業を行う。その後、第 4 ステップとして有限数に分類された変動パターンの発生ルールを抽出、第 5 ステップとして変数間の依存関係のルール抽出、という流れになっている。Step2~5 のベースとなる枠組みは、Kinjo et al., [2007] で提案されたものである。これらの作業の目的は、有限数に分類された時系列モデルの出現パターンから変数間の依存関係を抽出することで、(1b) および (3b) の問題に対応することである。とりわけ、ヘテロで時間遅れを伴う非線形な相関関係を抽出することが重要な目的である。講演では、問題全体の枠組み、現時点での進捗状況および将来的な構想について紹介する。