

自然波動観測データの自動分類アルゴリズムの高精度化

Study on automatic classification algorithm of plasma wave data

田中 祥平[1]; 笠原 禎也[1]; 後藤 由貴[1]

Shouhei Tanaka[1]; Yoshiya Kasahara[1]; Yoshitaka Goto[1]

[1] 金沢大

[1] Kanazawa Univ.

現在多数の人工衛星によって、宇宙環境に関する様々な観測が行われており、日々その観測データが各研究機関に蓄積されている。当機関においても、あけぼの衛星で観測された地球磁気圏の自然波動に関するデータが蓄積されており、総データ容量は 15TByte を越えると考えらる。これら観測データは研究者がの解析により、地球周辺の宇宙環境の解明に用いることが求められるが、データ量が膨大であるため、研究者が必要とするデータを的確に探し出すことが非常に困難な状況にある。そこで、長期間の観測データの中から特徴的な現象を自動的に探し出し、研究者が必要な情報を短時間で検索できるアルゴリズムが開発されている。しかし、現在の手法では後述する複数の要因により、必ずしも適正に分類できていないことが解っている。そこで、本研究ではこの自動分類アルゴリズムについて改良しより高精度な自動分類を行うことを目的とする。

まず、現在の自動分類アルゴリズムの概要を説明する。今回のアルゴリズムでは波動データに含まれる複数の特徴量を用いてクラスタ分析を行い、最終的に 7 つのクラスに分類している。また、データベースと連携させることで、各データのフォーマットに依存せず、また直接検索システムに適用できる設計となっている。

そこで、本研究では実際の波動データではなく、人工的に作成されたダミーデータを用いて分類精度の検証を行った。これまでの検証において明らかになった御判定の要因を示す。まず、今回のアルゴリズムにおいては、波動データの各特徴量はガウス分布をしていることを前提として考えているが、実際の観測データではガウス分布をしているとは限らないことが挙げられる。また、ある 1 つの特徴量の分布が 2 つの特徴的な分布をしている場合、その分布が重なりあっている状態においては、重なりあった部分が原理的に正しく分類できないということも明らかになった。これら以外にも様々な要因が関係しているが、この 2 つが分類結果に大きな影響を与えていると考えられる。

今回、この 2 つの要因による影響を軽減させる方法として、KNN 法を用いる方法を採用している。KNN 法は既に分類されているサンプルに対して、親近性に基づいて未知試料を分類する方法である。未知のクラス予測は K 個の最も近いサンプルの属性に依存し、それぞれの K 個の一番近いサンプルがそのクラスに投票を行う。そして、未知試料は、一番投票数が多いクラスに割り当てられる。今回の手法に KNN 法を適応する場合、クラスタ分析によって分類された結果と、その分類結果を元にした KNN 法による分類結果を比較し、その矛盾を検討することでクラスタ分析が正確であるか否かを判断する。また、KNN 法では親近性の基準としてユークリッド距離を用いているが、今回の手法では各特徴量、及びクラスの分散を考慮したマハラノビス距離を利用する。この方法を用いて従来法に寄る分類結果の妥当性を検証し、分類アルゴリズムの改良に反映させる。

講演においては、KNN 法による現アルゴリズムの検証結果とアルゴリズムの改良の指針について報告する。