演　　題	線形／非線形多変量解析を用いたおうし座分子雲－1における星間分子スペクトル線サーベイデータのシグナル／ノイズ分類―天文データマイニングの一歩として―	
発　表　者 （所　属）	○神部順子, 大石雅寿*, 長嶋雲兵** (大東文化大外国語, *国立天文台, **産総研グリッド研究センター)	
連　絡　先	〒175-8571 板橋区高島平 1-9-1 大東文化大学外国語学部	
キーワード	シグナル/ノイズ分類、天文スペクトル線データ、電波望遠鏡、主成分分析、クラスター分析、判別分析、パーセプトロン、ニューラルネット、再構築学習法	
開発意図 適用分野 期待効果 特徴など	天文データマイニングの一歩として，国立天文台野辺山にある電波望遠鏡で取得した 8－50GHz のスペクトル線サーベイデータを従来の線形多変量解析技法（すなわち主成分分析、クラスター分析、判別分析など）および非線形多変量解析（パーセプトロン型ニューラルネット）を用いて自動分類することを試みた。 おうし座分子雲－1 を対象としたサーベイ観測により、その中に数 10 種類の分子が存在していることが明らかになった。ノイズの混ざった１つのスペクトルから多数の分子スペクトルを分類するためには，微弱な受信信号をノイズから効率的に区別する必要がある。これには熟練を要し、熟練者と言えども分類不能な場合もある。経験の乏しい研究者でも客観性を持った分類を大量のデータに対して行なうことや、人間が判断に困るデータに対してシグナルかノイズかの指針を与えることを多変量解析技術によって可能とすることができれば、「シグナル」と判断されたデータに対する研究を通じて，新物質発見に貢献することが期待される。	
環　　境	適応機種名	
	ＯＳ名	
	ソース言語	
	周辺機器	
流通形態 （右のいずれかに○をつけてください）	・日本コンピュータ化学会の無償利用ソフトとする ・独自に頒布する ・ソフトハウス、出版社等から市販 ・ソフトの頒布は行なわない ・その他　　・未定	具　体　的　方　法

１．はじめに：現代天文学が研究に用いる最新の望遠鏡で観測・取得される世界最先端の観測データはそれぞれの観測所で最も有効な形態でディジタルデータとして格納されている。これらをデータベース化し，かつ，最新のネットワーク技術と組み合わせて相互利用することは世界の大きな潮流となっており，それを活用して複数の観測データや多波長・多領域にまたがるデータを系統的に研究し，統計に基づく処理や，観測の履歴に基づく処理等により精密な天体情報を得るデータベース天文学への期待が高まっている。現代の望遠鏡のデータ生産能力は極めて高く，研究者が個々の観測データを解析するのは不可能な事態となりつつあり、機械学習，パターン認識，データマイニングなどの最新の情報処理技術を活用してデータ解析を行うこ

とが求められている。しかしデータマイニングを天文学に応用した例は少ない。そこで本研究では，天文データマイニングの一歩として，国立天文台野辺山にある電波望遠鏡で取得した 8－50GHz のスペクトル線サーベイデータを従来の線形多変量解析技法（すなわち主成分分析、クラスター分析、判別分析）並びに非線形多変量解析の一種であるパーセプトロン型ニューラルネットを用いて自動分類することを試みた。本研究の最終目的は，発達の著しい計算機・情報処理技術を活用して，研究者を機械的作業から開放し，「思考過程」に専念できるよう補助する機能を実現するものである。

２．データと解析プログラム：本研究で用いたデータは、おうし座分子雲－1 付近のものであり、温度がおよそ 10K，密度がおよそ 104 H2 分子/ cm3 という極低温，極低密度であるが，数千万年にも渡って化学反応が継続している。属性値（周波数、アンテナ温度、半値幅、強度等）によって特徴づけられたピーク値の総数は 678 であり、これらはすでにノイズと信号の分類が行われている。

主成分分析、判別分析には、Exce 統計 2000[1]を用いた。クラスター分析には、Excel 太閤[2]を用い、ニューラルネットワークシミュレータは、我々が開発中の Neco[3]を用いた。

３．主成分分析の結果：全データを用いた主成分分析を行った。第一主成分の寄与率は 46.2%、第二主成分の寄与率は 29.3%である。第三主成分の寄与率は、12.3%、第四主成分の寄与率は、8.3%であった。第一主成分と第二主成分を主軸とする、678 個の主成分得点を用いた K-L プロットを Figure 1 に示した。Aがシグナルとノイズをあわせた全体で、Bがシグナル、Cがノイズである。シグナルとノイズは、K-L プロットでは重なりが大きく明確な分類は不能である。シグナルは第一主成分および第二主成分の 0 付近に集中しているが、第一主成分に沿って-5.0~20.0 の範囲でまんべんなく分布しており、逆にノイズは第一主成分 0 付近に分布している。ノイズは第二主成分の影響を大きく受けており、第一主成分の影響は全く受けていない。これは、少なくとも第一主成分が正で第二主成分が 5.0 以下のものはシグナルである可能性が高く、第一主成分が 0 付近で第二主成分が 5.0 以上のものはノイズである可能性が高いことを示している。

４．ニューラルネットワークを用いたスペクトルピークのシグナル/ノイズ分類：AIC を用いて分類に必要十分な属性値を選択し、さらに再構築学習法を用いてネットワーク構造の最適化を行った。ピークのシグナル/ノイズ分類は、ピークの半値幅を温度で規格化した値と S/N 比の二つの属性値で可能であることがわかった。

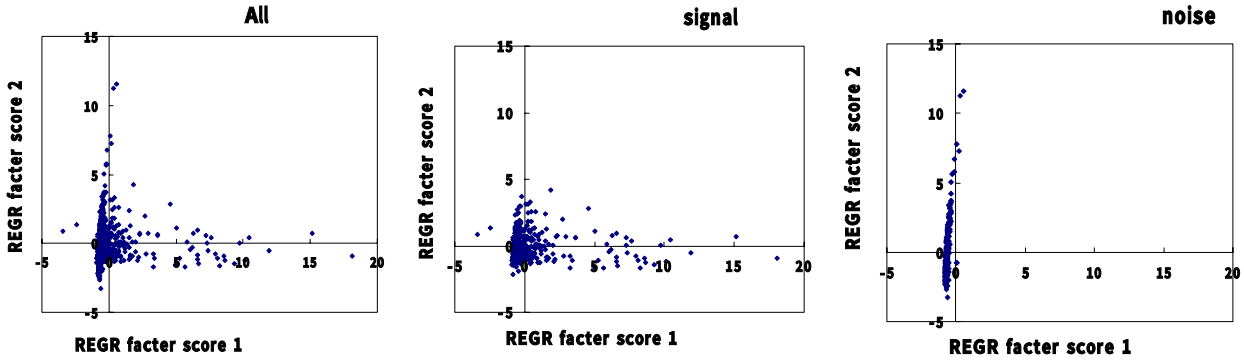

Figure 1 K-L plot of data on the 1st and 2nd principal components. A: all, B: Signal, C: noise.

[1] 多変量解析プログラム Excel 統計、http://www.esumi.co.jp/products_info/toukei/toukei_1.html.
[2] アンケート分析プログラム Excel 太閤、http://www.esumi.co.jp/products_info/taiko/profile_1.html.
[3] ニューラルネットワーシミュレータ Neco、http://www.sccj.net/publications/cssj_jrnl.html.