大規模データ解析に向けた光度曲線および AGB 星の SED の分類

安部 太晴（広島大）
植村 誠（広島大） 板 由房（東北大） 松永典之（東京大）

概要

近年、サーベイ観測などにより、取得できる天体データの量が多くなってきた。これら大量のデータを効果的に扱うよう整理することは重要であるが、この作業を人間の手で行うことは難しい。そのため、自動分類の重要性が高まっている。本研究は、線形判別を用いて分類軸を選択することにより、AGB 星の赤外線データと OGLE の光度曲線データの自動分類を試みた。

1 2MASS と AKARI データによる AGB 星の C-rich と O-rich の分類

本研究では、2MASS と AKARI の 2 プロジェクトのブロードバンド測光データをマージしたものを使用して、AGB 星の炭素過剰星 (C-rich 星) と酸素過剰星 (O-rich 星) の分類を行った。2MASS のデータは、J-band、H-band、Ks-band の 3 バンドの等級データ。AKARI のデータは $9\mu m$、$18\mu m$ の 2 バンドであり、マージして計 5 バンドの等級データになる。本研究では、この 5 バンドのデータ内でそれぞれ Ks-band に対して差をとり、最終的に 4 つの色指数データとして扱った。

解析の手順は以下の通りである。まず分類が判明しているサンプル (教師データ) を基に線形判別によって判別器を作成し、その判別器の性能を交差検定によって評価する。今回は SIMBAD での分類を教師データとして使用した。

サンプル数はマージできたバンドによって異なるが、総計 6960 天体である。うち C-rich は 5250 天体で、O-rich は 1710 天体である。

分類を行うにあたって、分類に用いる色データを全ての組み合わせを使用して分類を行い、その中で最も分類成績のよい組み合わせを選ぶという手法を用いた。例えば、今回用いる色指数は [K]-[J]、[K]-[H]、[K]-[9]、[K]-[18] の 4 色である。これを、ある試行では [K]-[J] のみ色指数を用いて、他の指数は用いない。また、別の試行では [K]-[J] と [K]-[H] の 2 色を用いて他は用いない。手持ちの色指数が 4 色なので、これを合計 $2^4 - 1$ 回行って、最もうまく分類できる色指数の組み合わせを調べるというものである。

1.1 線形判別理論

今回行う分類では、簡単のため、分類の境界は多次元平面で定義する線形判別と呼ばれるものを用いる。その中でも、フィッシャーの線形判別と呼ばれるモデルを使用する。

このモデルは、各クラス (ここでは O-rich と C-rich) 間の分散とクラス内の分散のみを参考に判別器を作成する。ここで言うクラス間分散とは、各クラスの平均値の分散のことで次の式 1 で定義される。

$$\boldsymbol{S}_{\mathrm{B}} = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\mathrm{T}} \tag{1}$$

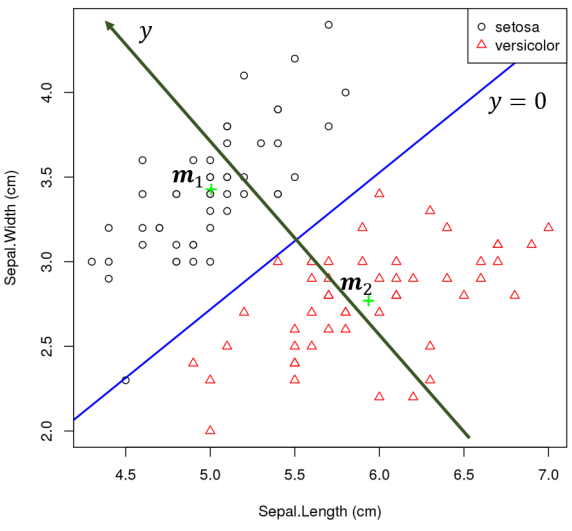

図 1: フィッシャーの線形判別によりアヤメの、がくデータを分類した図。丸と三角の 2 種類の品種を分類している。y 軸は参考のために付け足したもので、正確なものではない。

ここで、$\boldsymbol{m}_1$ と $\boldsymbol{m}_2$ は各クラスの入力の平均値で、今回の場合だと C-rich と O-rich の [K]-[J]、[K]-[H]... の平均値である。また、クラス内分散は次の式 2 ように定義される。

$$\boldsymbol{S}_{\mathrm{w}} = \sum_{i=1}^{2} \sum_{n \in C_i} (\boldsymbol{x}_n - \boldsymbol{m}_i)(\boldsymbol{x}_n - \boldsymbol{m}_i)^{\mathrm{T}} \tag{2}$$

C_i はクラスを表し、$\boldsymbol{x}_n$ は 1 データのベクトル値 (ここでは 1 天体の色指数) である。そして、これらの $\boldsymbol{S}_{\mathrm{B}}$ と $\boldsymbol{S}_{\mathrm{w}}$ をもちいて、次の評価式

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{B}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{W}} \boldsymbol{w}} \tag{3}$$

が最大となる $\boldsymbol{w}$ を求める。その結果、次の式 4 のように $\boldsymbol{w}$ が求まる。

$$\boldsymbol{w} \propto \boldsymbol{S}_{\mathrm{W}}^{-1} (\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{4}$$

この $\boldsymbol{w}$ を用いて、各天体の持つ値 $\boldsymbol{x}$ に対して射影を行い、クラス判別を行う。これを式で表すと次のようになる。

$$y = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + w_0 \tag{5}$$

w_0 はしきい値パラメータと呼ばれる。この式 5 の y の正負でクラスが決定される。図 1 は 2 種類のアヤメをフィッシャーの線形判別で分類したものである。図のように、$y = 0$ が境界となり、分類が可能になる。

1.2 交差検定

本研究では、判別器の性能評価に、10 分割交差検定を用いた。この手法では、サンプルを 10 個のグループに分けて 9 個を解析用とし、残った 1 個を検定用として用いてモデルの予測誤差などの数値で評価する。

作成した判別器により正しく分類できたサンプルの数を T、誤って分類されたサンプルの数を F として、分類の正答率 A を $A = T/(T+F)$ のように定義する。図 2 のように、10 個に分けたグループで、検定用のグループを変えて分類を行う。例えば、1 回目の分類では、グループ 2-9 から判別器を作成し、この作成した判別器でグループ 1 の分類を行い、正答率 A_1 を求める。2 回目の分類では、グループ 1 と 3-9 から判別器を作成し、この作成した判別器でグループ 2 の分類を行い、正答率 A_2 を求める。このようにして 10 回それぞれ出した正答率 A_i を平均したものを最終的な正答率とする。すなわち、分類の最終的な正答率は

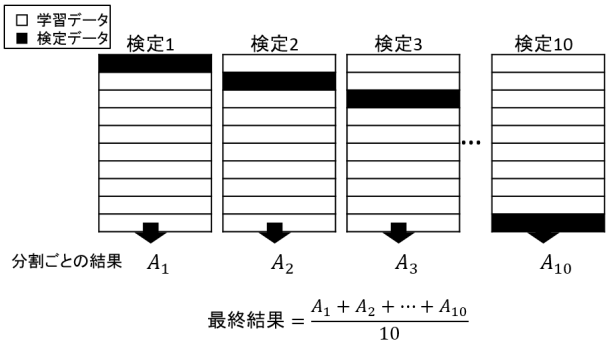

図 2: 10 分割交差検定のイメージ図。図のように多数の学習データから判別器を作成し、一部残しておいた検定データを分類する。これを、検定データと学習データの組み合わせを変えて、計 10 回行う。

$$A = \frac{1}{10} \sum_{i=1}^{10} A_i \tag{6}$$

である。全てのデータから判別器を作成し、その正答率でモデルを評価すると、判別器が過適合なモデルになる可能性があり、モデルの予測誤差が大きくなる。こうした事態を防ぐために、交差検定を用いる。

1.3 結果

使用する変数別の分類正答率を図 3 に示す。図 3 は使用する色指数ごとに分類の正答率を示したものである。図中の分類正答率を比較してみると、4 番目に成績の良いものが正答率 88.6%であるのに対し、5 番目に成績の良い物は 81.9%と、正答率に大きな差がある。また、正答率 1 位から 4 位までは [K]-[9] のカラーを使っているのに対

して、それ以下の順位では [K]-[9] は使っていない。また、正答率 5 位から 8 位の間では [K]-[18] のカラーが、9 位から 12 位の間では [K]-[9] のカラーが使われてる。

$9\mu m$、$18\mu m$ の近くには、Si のダストの吸収線が存在する。正答率上位のモデルにこれら 2 バンドのデータが含まれているのは、O-rich と C-rich の分類に本質的な吸収線の情報を反映しているためと考えられる。また $18\mu m$ のデータを除くと、誤判定が約 1 割増えることが今回の結果からわかる。

2　光度曲線データから行う変光星の分類

先の 2MASS と AKARI データの SED の分類とは別に、OGLE(Optical Gravitational Lensing Experiment) の大マゼラン雲の天体の光度曲線データの分類も行った。さきほどの SED の分類では座標情報を基に同定した SIMBAD の型分類を、教師データとして判別器を作成した。対してこちらは、OGLE が独自に型分類を行っているため、OGLE の分類を教師データとして判別器を作成した。分類する変光星型はミラ、セファイド、食連星、人口ノイズの 4 種類である。

この光度曲線の分類でも、先の SED の分類と同様にフィッシャーの線形判別と 10 分割交差検定を用いた。

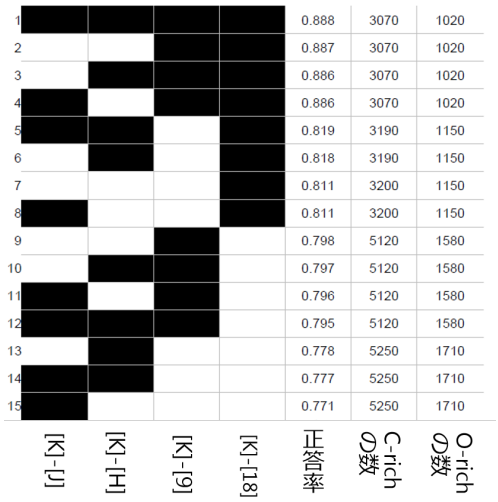

図 3: 使用した色指数別の分類正答率。上から正答率の高い順に並べている。例えば、1 番上の最も正答率の高い全ての色指数を使用したケースでは、正答率が 88.8%であり、正答率の最も低い 15 番目のケースは [K]-[J] の色指数のみを使用したもので、正答率が 77.1%である。右端に使用した C-rich 天体と O-rich のそれぞれのサンプル数を示している。

2.1　使用した特徴量

2.1.1　光度曲線から取得した特徴量

光度曲線から取得した特徴量は、等級の平均値 $L_{\rm mean}$、等級の標準偏差 $L_{\rm sd}$、等級の最大値 − 最小値 $L_{\rm max} - L_{\rm min}$、等級の最大値 − 平均値 $L_{\rm max} - L_{\rm mean}$、そしてそれを $L_{\rm max} - L_{\rm min}$ で正規化した $(L_{\rm max} - L_{\rm mean})/(L_{\rm max} - L_{\rm min})$ の 5 種類である。

2.1.2　パワースペクトルから取得した特徴量

一方、パワースペクトルから取得した特徴量を表 1 に示す。また、ナイキスト周波数は 1.0c/d と 0.5c/d の両方を使用した。これは、ナイキスト周波数が 0.5c/d の場合だとパワースペクトルのピークが 0.5 c/d より大きいところに立っても調べることができず、ナイキスト周波数が 1.0c/d の場合だと天体観測がほぼ 1 日おきに行われていることから折り返しが強く現れてしまうため、特徴量として両方使用することにした。

表 1: パワースペクトルから取得した特徴量

特徴量	意味
$\nu_{\rm max}$	パワースペクトル $P(\nu)$ の最大値 $P_{\rm max} = P(\nu)$ を与える ν。
$\log \nu_{\rm max}$	$\nu_{\rm max}$ の常用対数をとったもの。
$P_{\rm max}/P_{\rm sd}$	パワースペクトルの最大値 $P_{\rm max}$ を標準偏差 $P_{\rm sd}$ で正規化したもの。
$(P_5 - P_1)/P_{\rm sd}$	$\nu = 0.5$ 付近と $\nu = 0.1$ 付近のパワーの差を標準偏差で正規化したもの。
P_5/P_1	$\nu = 0.5$ 付近と $\nu = 0.1$ 付近のパワーの比。

	$\nu_{\max}$	$\nu'_{\max}$	$\log \nu_{\max}$	$\log \nu'_{\max}$	$\frac{p_{\min}}{P_{sd}}$	$\frac{p'_{\min}}{P'_{sd}}$	$\frac{p_3-p_1}{P_{sd}}$	$\frac{p_3}{p_1}$	L_{mean}	L_{sd}	$L_{\max}-L_{\min}$	$L_{\max}-L_{\mathrm{mean}}$	$\frac{L_{\max}-L_{\mathrm{mean}}}{L_{\max}-L_{\min}}$	Accuracy
Case 1			●	●	●	●			●	●	●	●	●	0.98953
Case 2			●	●		●			●	●	●	●	●	0.98953
Case 3		●	●	●	●	●			●	●	●	●	●	0.98919
Case 4			●	●	●	●		●	●	●	●	●	●	0.98919
Case 5			●			●			●	●	●	●	●	0.98902
Case 6		●	●		●	●			●	●	●	●	●	0.98902
Case 7			●	●		●		●	●	●	●	●	●	0.98902
Case 8			●		●	●			●	●	●	●	●	0.98885
Case 9		●	●			●			●	●	●	●	●	0.98885
Case 10			●	●		●	●		●	●	●	●	●	0.98885
Case 11			●	●		●	●	●	●	●	●	●	●	0.98885
Case 12		●	●	●	●	●	●		●	●	●	●	●	0.98885
Case 13			●	●	●	●	●	●	●	●	●	●	●	0.98885
Case 14		●	●	●		●	●		●	●	●	●	●	0.98868
Case 15		●	●	●		●		●	●	●	●	●	●	0.98868
Case 16		●	●		●	●	●		●	●	●	●	●	0.98868
Case 17		●	●	●	●	●		●	●	●	●	●	●	0.98868
Case 18		●	●	●	●	●	●	●	●	●	●	●	●	0.98868
Case 19			●			●		●	●	●	●	●	●	0.98851
Case 20		●	●	●		●			●	●	●	●	●	0.98851

図 4: 光度曲線の分類の際に、使用したパラメータと分類の正答率 (Accuracy) を示している。使用したパラメータの欄に黒丸を記入している。また、ナイキスト周波数を 1.0c/d に設定して取得した特徴量は $\nu'_{\max}$ のようにプライム記号をつけている。

2.2　結果・考察

光度曲線の分類を行った結果のうち、成績良かった順に 20 通りを図 4 に示す。図 4 の中の上位 20 位以内のモデルでは、ライトカーブの特徴量と $\log \nu_{\max}$、$P'_{\max}/p'_{\mathrm{sd}}$ が必ず使われている。ν に対して $\log \nu$ が選ばれているのは、図 5 のように、しきい値パラメータを各クラスの平均値の中点に設定する場合に正答率が上がるためである。

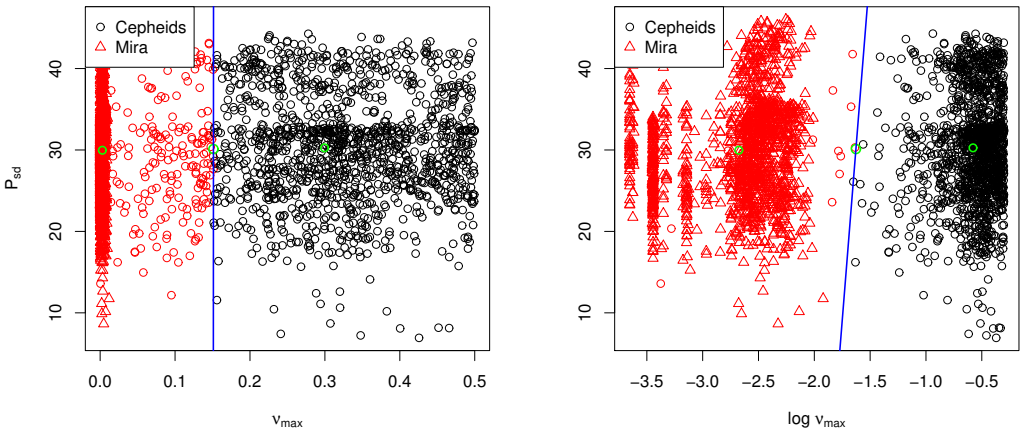

図 5: $\nu_{\max}$ と P_{sd} を使って 2 クラス判別をした結果 (左) と $\log \nu_{\max}$ と P_{sd} を使って 2 クラス判別をした結果 (右)。記号が OGLE 分類による変光星型を示している。太い丸は各変光星型の平均および、それらの平均。

3　まとめと今後

AGB 星の赤外線データと OGLE の光度曲線データの自動分類を行った。今回は分類の境界を線形なものとして定義したが、非線形な境界を設定することにより分類正答率が上がる可能性もある。