

# 欠損の多い教師データを用いた銀河系内突発現象の機械判別

古賀柚希, 植村誠, 佐崎凌佑(広島大学), 池田思朗(統計数理研究所)

## 研究背景

・・・詳細は2枚目

激変星の突発現象の研究には、迅速な型判別と適切な追跡観測が求められる  
 →自動化するシステムを開発中 ↑ ↑ 情報理論を利用  
 →機械判別(迅速な型判別)について調査する

## 機械判別

・・・詳細は3枚目

問題設定：  
 5つの型、14の特徴量を教師データとして、突発現象の天体の型を判別

- ◆新星(Nova), 矮新星(DN), WZ型矮新星(WZ), ミラ型変光星(Mira), フレア星(UV)
- ◆座標、距離、静穏時等級などに関連する14の特徴量

機械判別で求めるもの： $p(C_k|\mathbf{x})$   
 →特徴量ベクトル $\mathbf{x}$ をもつサンプルがクラス $C_k$ に属する確率( $k = 1, 2, \dots, K$ (クラス数))

キーワード：ベイズの定理

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}$$

☆教師データには欠損値が多数 →天文学の他のケースでも生じる共通の問題

type	l	b	d_kpc	gal_abs_z	AbsMag_out	AbsMag_qui	AbsMag_J	Ampl	g-r	r-i	i-z	j-h	h-k	i-k	
0	Nova	121.119811	-22.099995	NaN	NaN	NaN	NaN	4.7638	NaN	NaN	0.1446	NaN	NaN	NaN	
1	Nova	198.449204	9.249622	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	Nova	318.535741	8.628934	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	Nova	313.280192	-8.394845	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	UV	0.123455	7.215529	NaN	NaN	NaN	NaN	1.2938	NaN	NaN	NaN	NaN	NaN	NaN	
2699	UV	84.534443	-67.443400	0.093274	0.086139	10.851197	12.505997	9.694197	1.8548	1.3175	2.0239	0.9544	0.570001	0.356000	3.7378
2700	UV	107.228646	-36.355429	0.111575	0.066141	8.962170	11.364770	8.895170	2.4026	1.3086	1.7187	0.7933	0.616000	0.254001	3.3396
2701	UV	111.583013	-23.057892	NaN	NaN	NaN	NaN	NaN	-0.5807	0.7255	1.6170	0.5775	0.638000	0.191000	3.0573
2702	UV	64.950517	-75.031686	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.540000	0.283999	NaN
2703	UV	96.327929	-60.173109	0.583090	0.505850	5.471321	6.152521	4.921321	0.6812	0.6796	0.2980	0.1381	0.556000	0.054000	1.8412

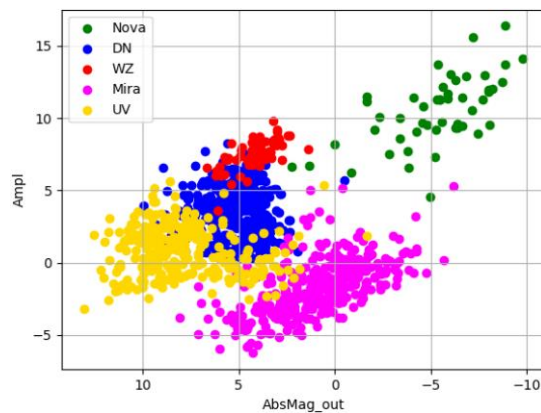


図1:特徴量の散布図例

## 調査内容

・・・詳細は3枚目

どのモデルがシステムに合うか調査  
 →性質の異なる3つの判別モデルを用いて比較

表1:モデル毎の特徴

モデル	教師データ	カーネル化
LR(ロジスティック回帰)	特徴量が揃っているサンプル	なし
SMLR(スパース多クラスロジスティック回帰)	特徴量が揃っているサンプル	あり
GM(生成モデル)	全サンプル	なし

- ◆特徴量が揃っている、について判別したいデータがもっている特徴量と同じ特徴量が揃っているサンプルのみを、教師データとして扱える
- ◆カーネル化について教師データに非線形な変換を施し、複雑な決定境界を引けるようになる

## 結果・今後

・・・詳細は4枚目

14個全ての特徴量を使ったモデル

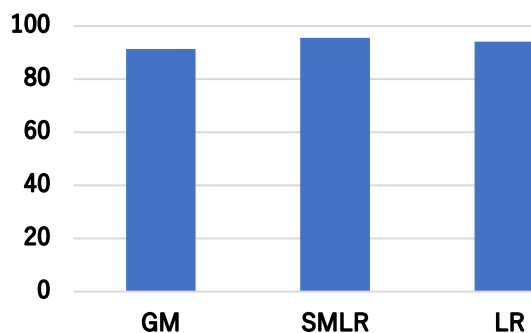


図3.1:モデル毎の正解率

座標と静穏時(近赤外)の4特徴量のみ

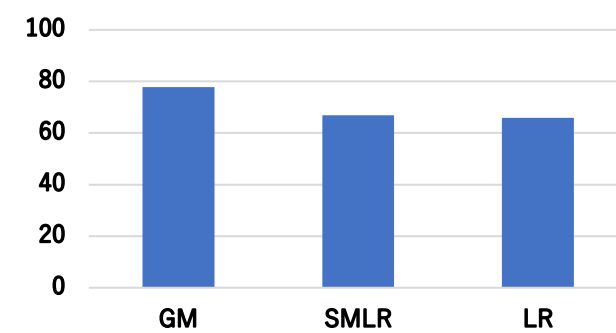


図3.2:モデル毎の正解率

◆3つのモデル、いくつかの特徴量の組み合わせについて、正解率に大きな差は見られない  
 →複雑な決定境界は不要

◆全体の性能はGMが若干高い  
 →少なくとも、静穏時の近赤外データのみが利用可能な場合は、GMを用いるべき

➡ 特徴量の組み合わせを細分化し、判別に効く特徴量を調査する

図2:教師データ

[2704 rows x 15 columns]

# 研究背景

## ◆ 新星 ～突発現象を起こす天体～

- ・ 恒星と白色矮星からなる連星系
- ・ 恒星からのガスが白色矮星表面に降着し、水素の核燃焼が始まり**新星爆発**を起こす。

## ◆ 矮新星

- ・ 恒星と白色矮星からなる連星系
- ・ 恒星からのガスが白色矮星表面に降着する際、すぐには降着せず円盤外縁部に溜めこまれる

→溜めこまれた質量がある**臨界値**を越えると円盤が不安定になり、それまでに溜めこまれた物質が一気に降着し爆発

## →矮新星アウトバースト



新星・矮新星などの突発現象を起こす天体は、その発見時には天体の型の不確実性が高く、専門家の判断に伴う適切な追跡観測の判断・実行が要される

→情報理論や機械学習の枠組みを用いてこれらを自動化するシステムが構築できれば、激変星のより効率的な研究を行えることが期待できる

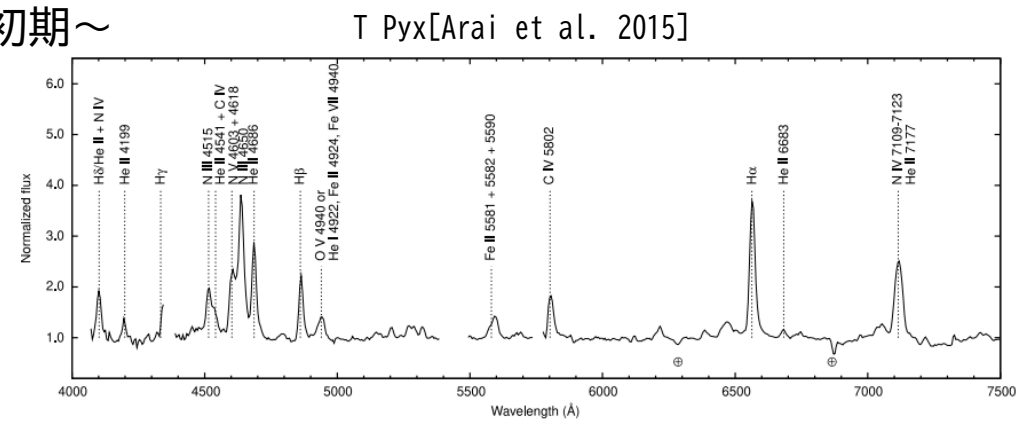
キーワード：**情報エントロピー**

$$S = - \sum_k p(k) \log_2 p(k)$$

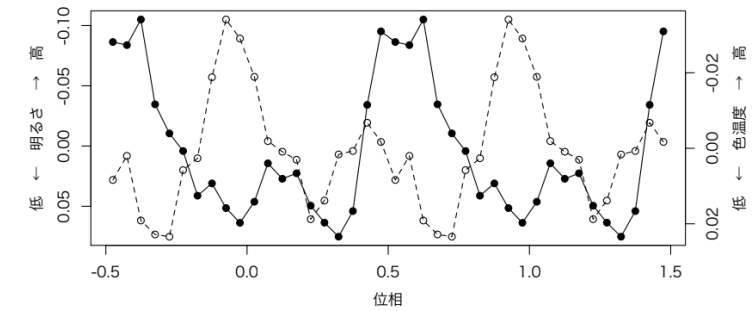
不確実性を表す量  
 $p(k)$ :  $k$ である確率  $k$ : 天体の型



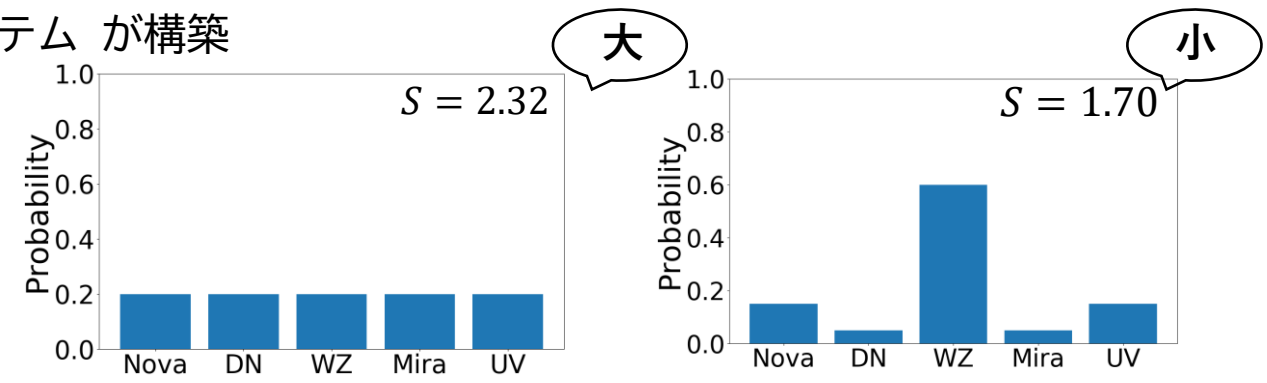
## ～爆発初期～



初期特有のスペクトル→白色矮星表面の物理  
 (岩波データサイエンスVol.6 p89 図3)「アンドロメダ座V455」



明るさと温度の振動→降着円盤の構造



# 機械判別・判別モデル

## ◆ ロジスティック回帰(Logistic Regression)

ベイズの定理を以下のように変形する。

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_j p(\mathbf{x}|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \log(p(\mathbf{x}|C_k)p(C_k))$$

$C_k (k = 1, 2, \dots, K)$ : クラス,  $\mathbf{x}$ : 特徴量ベクトル ( $M$ 次元)

さらに  $a_k$  を  $\mathbf{x}$  の線形結合で以下のように表せるとする。

$$a_k = \mathbf{w}_k^T \mathbf{x}$$

このモデルパラメータ  $\mathbf{w}_k (k = 1, 2, \dots, K)$  をデータから最尤推定して構築する判別モデルを **ロジスティック回帰** という。

尤度関数、目的関数(負の対数尤度)はそれぞれ以下

$$p(\mathbf{Y}|\mathbf{W}) = \prod_i \prod_k \left\{ \frac{\exp(\mathbf{w}_k^T \mathbf{x}_i)}{\sum_j \exp(\mathbf{w}_j^T \mathbf{x}_i)} \right\}^{y_{ik}}$$

目的変数を、 $K$ 次元ベクトル  $\mathbf{y}_i$  とする。

(例)  $\mathbf{y}_i = (1, 0, 0, \dots, 0)$  →  $i$  番目のサンプルのクラスが  $C_1$

$$E(\mathbf{W}) = -\log p(\mathbf{Y}|\mathbf{W}) = -\sum_i \sum_k y_{ik} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}_i)}{\sum_j \exp(\mathbf{w}_j^T \mathbf{x}_i)}$$

$\mathbf{Y} (N \times K)$ :  $N$  個の  $K$  次元目的変数  $\mathbf{y}_i$

$\mathbf{W} (M \times K)$ :  $K$  個の  $M$  次元モデルパラメータ  $\mathbf{w}_k$

$$\rightarrow \widehat{\mathbf{W}} = \arg \min \{E(\mathbf{W})\}$$

## ◆ スパース多クラスロジスティック回帰(Sparse Multinomial Logistic Regression)

上の最尤推定に関して、 $\widehat{\mathbf{W}} = \arg \min \{E(\mathbf{W}) + \lambda \|\mathbf{W}\|_1\}$  とした推定方法のこと。  $\lambda$ : パラメータ、 $\|\mathbf{W}\|_1 = \sum_i \sum_j |w_{ij}|$

また、ある変換  $\phi$  について、その内積の関数系を与える **カーネル法** を用いた。

## ◆ 生成モデル(Generative Model)

$$p(\mathbf{x}|C_k) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} \exp\left\{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{C_k})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_{C_k})\right\} \quad p(C_k) = 1/5$$

ベイズの定理の表式をそのまま用いる → 多変量正規分布

$\Sigma$ : 分散共分散行列、 $\boldsymbol{\mu}_{C_k}$ : クラス  $C_k$  の平均ベクトル

# 結果・今後

## 混同行列(座標+静穏時[近赤外])

GM: 77.76%

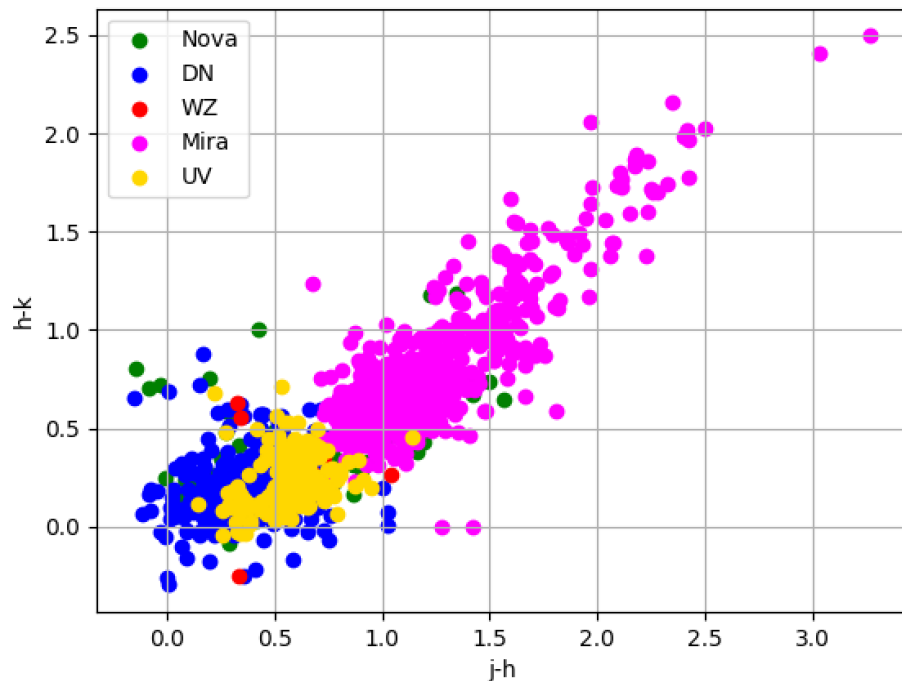
	Nova	DN	WZ	Mira	UV
Nova	49	16	5	8	9
DN	43	145	42	1	71
WZ	2	2	0	1	5
Mira	60	0	18	697	3
UV	10	30	17	3	319

SMLR: 66.83%

	Nova	DN	WZ	Mira	UV
Nova	45	17	11	7	7
DN	63	119	51	4	65
WZ	1	3	2	0	4
Mira	79	3	20	671	5
UV	32	46	93	5	203

LR: 65.93%

	Nova	DN	WZ	Mira	UV
Nova	17	31	18	14	7
DN	48	158	50	4	42
WZ	4	2	0	0	4
Mira	39	0	9	723	7
UV	106	34	110	1	128



特徴量の散布図例

→WZのサンプル数が少ない

特徴量が揃っている少ないWZのサンプルに対して・・・

GM: 全サンプルを教師データとして判別モデルを作れる

LR, SMLR: 特徴量が揃っているサンプルのみで判別モデルを構築

少なくとも、静穏時の近赤外線データが利用可能な場合は、GMを使うべき