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

広島大学 M2 古賀柚希

どのような問題？

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する

Nova(新星):298
DN(矮新星):1041
WZ(WZ Sge型矮新星):165
Mira(ミラ型変光星):779
UV(フレア星):421

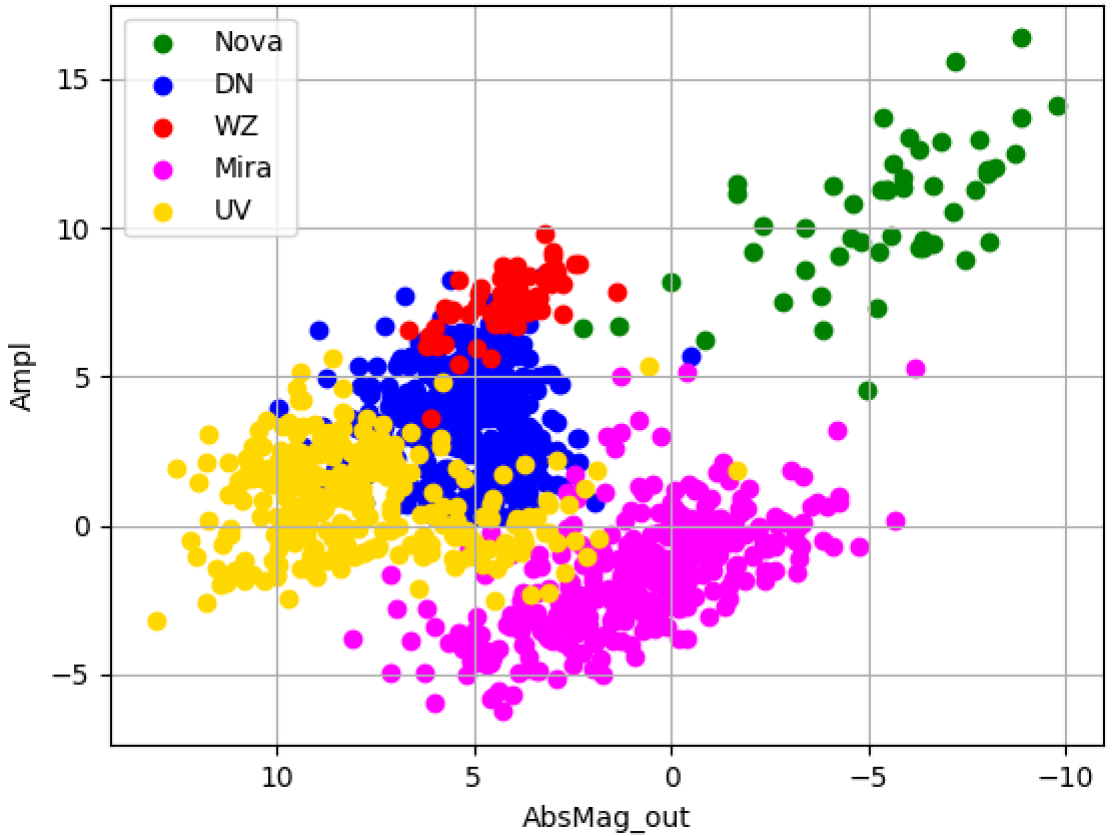
①座標：l,b(銀経・銀緯)　←常に利用可能
②距離：d_kpc(距離),gal_abs_z(銀河面からの距離),AbsMag_out(増光時絶対等級)
③静穏時[可視光]：Ampl(振幅),g-r(色),r-i(色),i-z(色)
④静穏時[近赤外線]：j-h(色),h-k(色)
②③利用可能：AbsMag_qui(静穏時絶対等級[可視光])
②④利用可能：AbsMag_J(静穏時絶対等級[近赤外線])
③④利用可能：i-k(色)

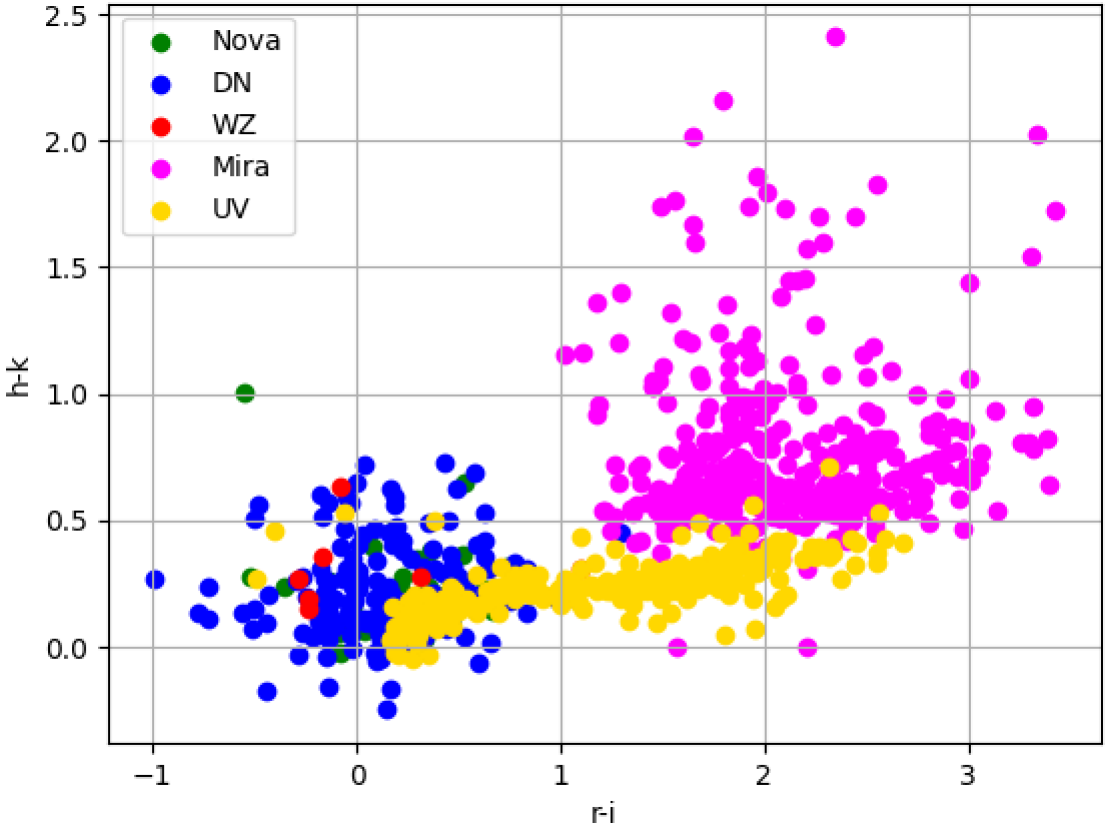
	type	l	b	d_kpc	gal_abs_z	AbsMag_out	AbsMag_qui	AbsMag_J	Ampl	g-r	r-i	i-z	j-h	h-k	i-k
0	Nova	121.119811	-22.099995	NaN	NaN	NaN	NaN	NaN	4.7638	NaN	NaN	0.1446	NaN	NaN	NaN
1	Nova	198.449204	9.249622	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Nova	318.535741	8.628934	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Nova	313.280192	-8.394845	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Nova	0.123455	7.215529	NaN	NaN	NaN	NaN	NaN	NaN	1.2938	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2699	UV	84.534443	-67.443400	0.093274	0.086139	10.851197	12.505997	9.694197	1.6548	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

[2704 rows x 15 columns]

どのような問題？

特徴量の分布





教師データについて

	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	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	NaN
2	Nova	318.535741	8.628934	NaN	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	NaN
4	Nova	0.123455	7.215529	NaN	NaN	NaN	NaN	NaN	NaN	1.2938	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2699	UV	84.534443	-67.443400	0.093274	0.086139	10.851197	12.505997	9.694197	1.6548	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

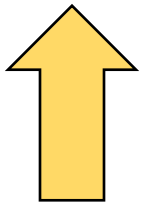
[2704 rows x 15 columns]

欠損値を多く含む

→天文学の他のケースでも生じる共通の問題

機械判別で求めるもの

$p(C_k|\mathbb{x})$：特徴量$\mathbb{x}$をもつサンプルが属するクラスがC_kである確率
$(k = 1,2,\dots,K$(クラス数)$)$



ベイズの定理　$p(C_k|\mathbb{x}) = \dfrac{p(\mathbb{x}|C_k)p(C_k)}{p(\mathbb{x})}$

$p(C_k|\mathbb{x})$を確率として扱う条件$(\sum_j p(C_j|\mathbb{x}) = 1)$より、

$$p(\mathbb{x}) = \sum_j p(\mathbb{x}|C_j)p(C_j), \qquad \therefore p(C_k|\mathbb{x}) = \frac{p(\mathbb{x}|C_k)p(C_k)}{\sum_j p(\mathbb{x}|C_j)p(C_j)}$$

→ $p(\mathbb{x}|C_k), p(C_k)$が分かればよい

判別モデル1(LR)

◆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。

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

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

a_kを$\mathbb{x}$の線形結合で以下のように表せるとする。

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

このモデルパラメータ$\mathbb{w}_k(k = 1,2,\ldots,K)$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。

判別モデル1(LR)　～最尤推定について～

目的変数を、K次元ベクトル$\mathbb{y}_i$とする。

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

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

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

尤度関数、目的関数(負の対数尤度)はそれぞれ以下

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

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

→ $\widehat{\mathbb{W}} = \arg min \{E(\mathbb{W})\}$

判別モデル2(SMLR)

◆スパース多クラスロジスティック回帰

(Sparse Multinomial Logistic Regression)

$\rightarrow \widehat{\mathbb{W}} = \arg min \{E(\mathbb{W}) + \lambda \|\mathbb{W}\|_1\}$　　$\|\mathbb{W}\|_1 = \sum_i \sum_j |w_{ij}|$

☆カーネル法を用いた

→ある変換ϕについて、その内積の関数系を与える手法

→複雑な形状の決定境界を作れる

判別モデル2(SMLR) ～カーネル法～

特徴量$\mathbb{x}_i$に非線形な変換$\phi(\mathbb{x}_i)$を施し、$\phi_i = \phi(\mathbb{x}_i)$を新たなデータとして判別する。

$$a_k = \mathbb{w}_k^T \phi_i$$

さらに、 $\mathbb{w}_k$が新たな特徴量ϕ_iの線形結合で表せるとする。

$$\mathbb{w}_k = \sum_j b_j \, \phi_j \qquad \leftarrow \phi_i から \mathbb{w}_k を推定$$

つまり、 a_kは以下のようになる。

$$a_k = \mathbb{w}_k^T \phi_i = (\sum_j b_j \, \phi_j)^T \phi_i$$

$$= \sum_j b_j \phi_j^T \phi_i$$

$$= \mathbb{b}\mathbb{k}(\mathbb{x}_i, \mathbb{x}_j)$$

判別モデル2(SMLR) ～カーネル法～

ここで、$\mathbb{k}(\mathbb{x}_i, \mathbb{x}_j)$は$N$次元ベクトルで、その成分は

$$(k(\mathbb{x}_i, \mathbb{x}_1), k(\mathbb{x}_i, \mathbb{x}_2), \dots, k(\mathbb{x}_i, \mathbb{x}_N))$$

また、

$$k(\mathbb{x}_i, \mathbb{x}_j) = \phi_j^T \phi_i = \phi(\mathbb{x}_j)^T \phi(\mathbb{x}_i)$$

となり、ベクトルϕ_iとϕ_jの内積を表す。

p次ノルム：ベクトル$\mathbb{x} = (x_1, x_2, \dots, x_N)$に対して

$$\|\mathbb{x}\|_p = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_N|^p}$$

(例)ガウスカーネル

$$k(\mathbb{x}_i, \mathbb{x}_j) = \exp\left(-\frac{\|\mathbb{x}_i - \mathbb{x}_j\|_2^2}{2\sigma^2}\right)$$

判別モデル3(GM)

◆生成モデル(Generative Model)

ベイズの定理の表式をそのまま用いる→多変量正規分布

$$p(\mathbb{x}|C_k) = \frac{1}{\sqrt{(2\pi)^M|\Sigma|}}\exp\{-\frac{1}{2}\left(\mathbb{x}-\mathbb{u}_{C_k}\right)^T\Sigma^{-1}(\mathbb{x}-\mathbb{u}_{C_k})\}$$

Σ:分散共分散行列、 $\mathbb{u}_{C_k}$:クラスC_kの平均ベクトル

$$p(C_k) = 1/5$$

～分散共分散行列～

分散・共分散$\sigma_{ij}(i,j = 1,2,\dots,M)$を要素にもつ行列

☆欠損値のため正しく計算できていない場合が発生

→ $\Sigma' = \Sigma + \alpha I(0 < \alpha < 1, I$:単位行列)とした

モデルの特徴

モデル	教師データ	カーネル化
LR	特徴量が揃っているサンプル	なし
SMLR	特徴量が揃っているサンプル	あり
GM	全サンプル	なし

特徴量が揃っている、の意味

	特徴量1	特徴量2	特徴量3	特徴量4
サンプル1	データ11	データ12	データ13	データ14
サンプル2	データ21	NaN	データ23	データ24
サンプル3	データ31	データ32	NaN	データ34
サンプル4	データ41	データ42	データ43	データ44
サンプル5	データ51	データ52	データ53	データ54

教師データ(GM)

教師データ(LR,SMLR)

→使える特徴量が多いほど、教師データのサンプルは減る

判別性能比較

→いくつかの特徴量の組み合わせについて、

1個抜き交差検証で正解率を計算＋混同行列も見る

1個抜き交差検証

(全サンプル)−(試験データ1) → 試験データ1 → 判別結果1　○

(全サンプル)−(試験データ2) → 試験データ2 → 判別結果2　×

・
・
・

これを繰り返し、(正解率)=(○の数)/(試験データ数)を求める

判別性能比較

混同行列

サンプルの実際の型と、判別結果の型を行列の形で表したもの

実際＼判別	Nova	DN	WZ	Mira	UV
Nova	45	2	0	1	0
DN	3	312	83	0	4
WZ	0	5	53	0	0
Mira	1	0	0	257	1
UV	3	18	0	0	257

～考えられること～

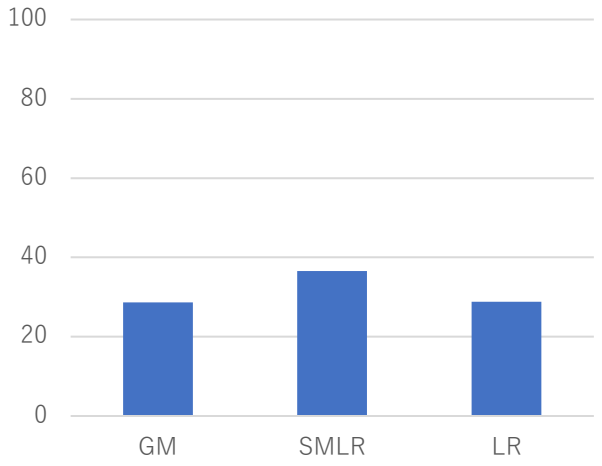
・カーネル化を用いたSMLRは、教師データが多いければ非線形で複雑な決定境界で高い性能を出すと期待される⇔少なければ性能は落ちる？

・GMは多くの教師データを使えるので高い性能を出すと期待される⇔単純な正規分布なので性能は落ちる？

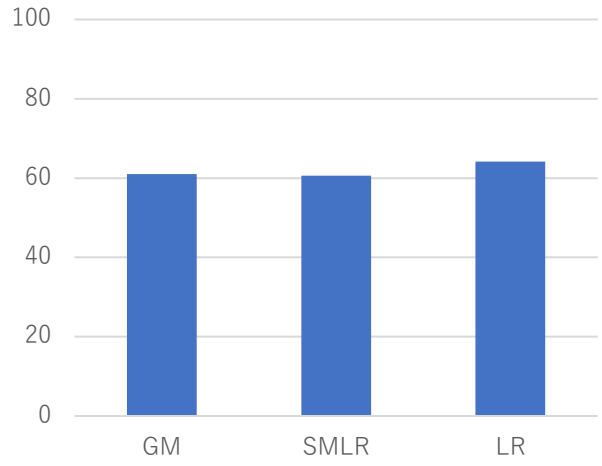
・カーネル化も沢山の教師データもないLRでどこまで性能を出せるか？

結果(正解率)

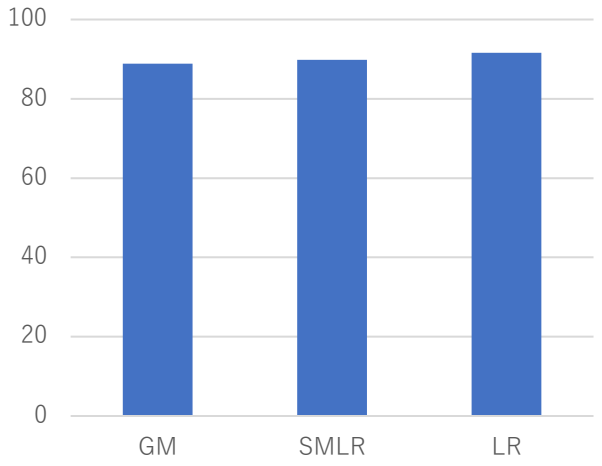
座標のみ



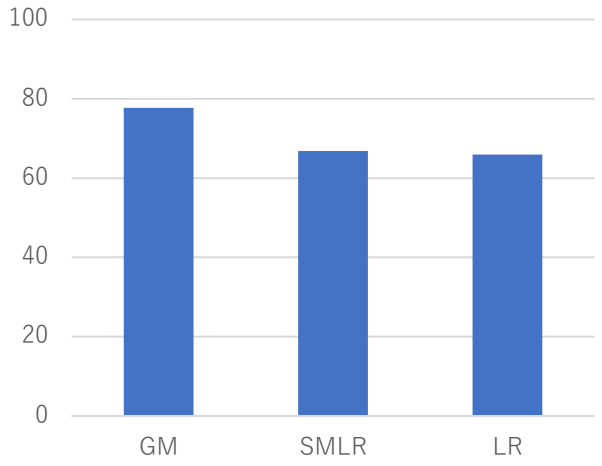
距離



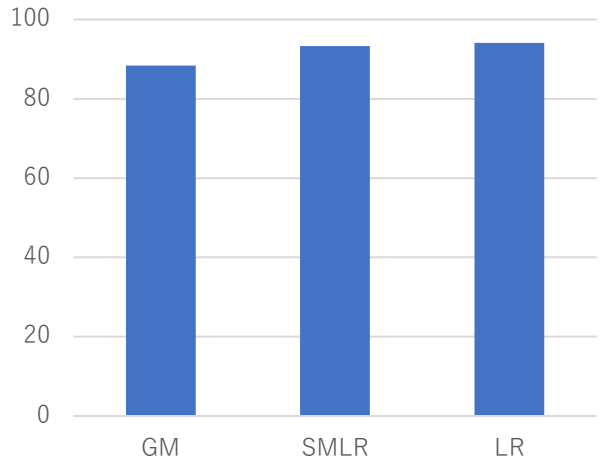
静穏時(可視)



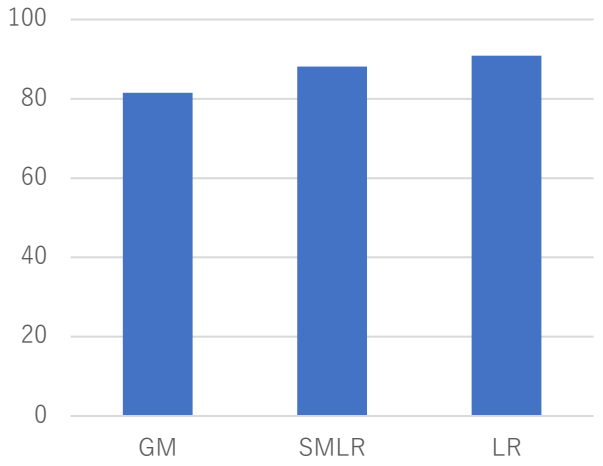
静穏時(近赤外)



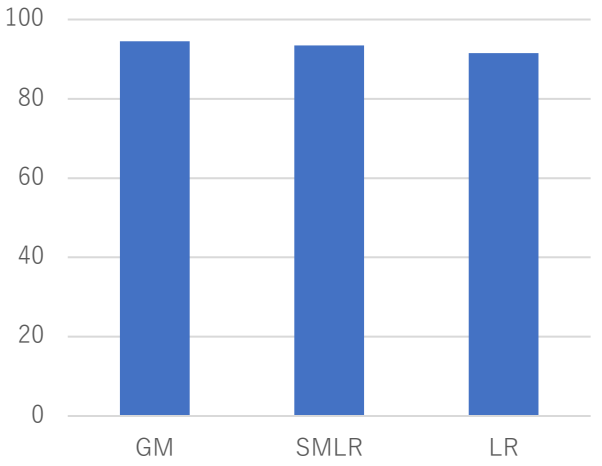
距離、静穏時(可視)



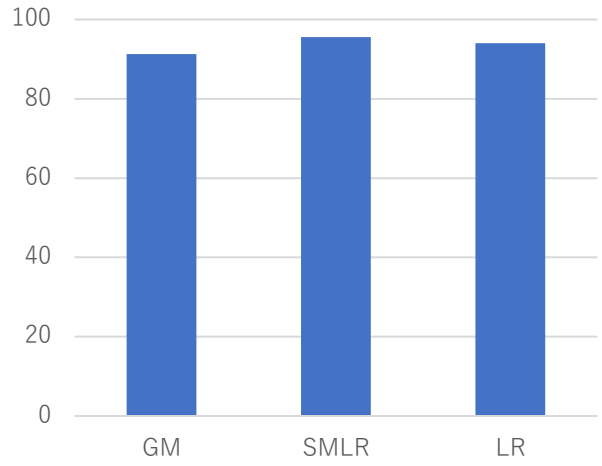
距離、静穏時(近赤外)



静穏時(可視・近赤外)



全部



結果(混同行列)

GM:77.76%

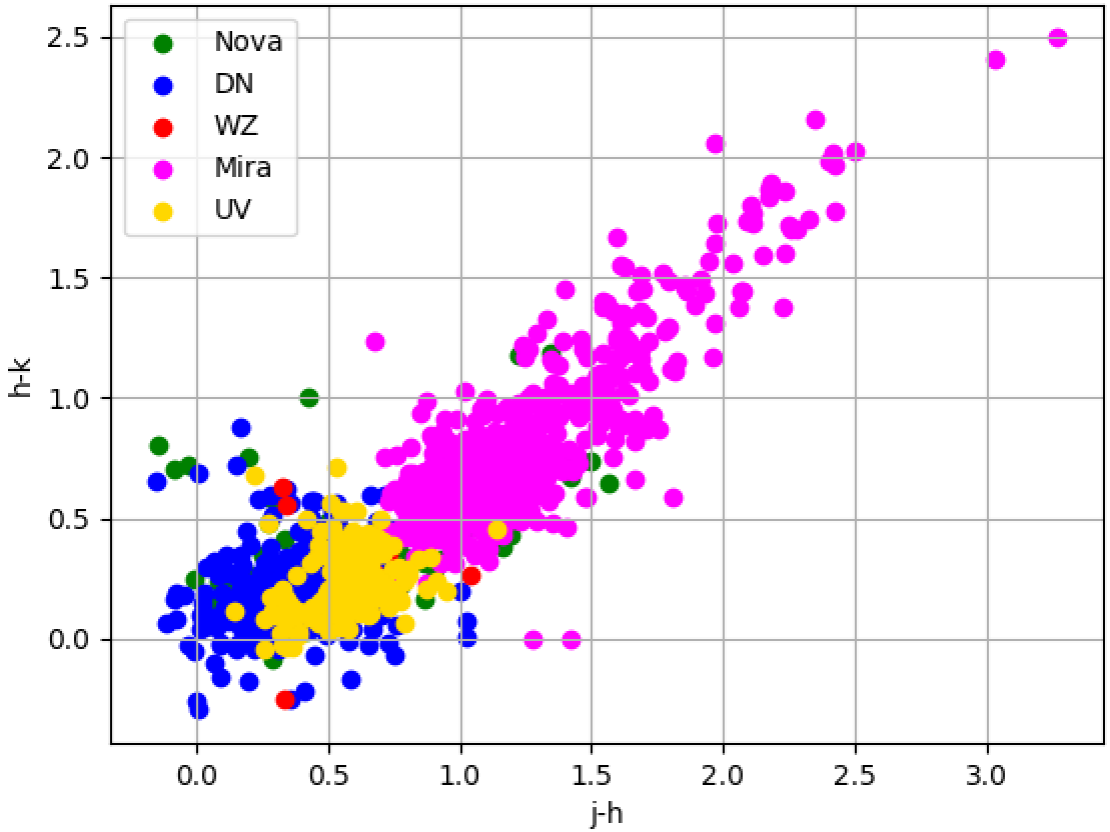
	Nova	DN	WZ	Mira	UV
Nova	49	16	5	8	9
DN	43	145	42	1	71
WZ	2	2	0	1	5
Mira	60	0	18	697	3
UV	10	30	17	3	319

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



→WZのサンプル数が少ない

考察・今後

・カーネル化の有無で正解率に大きな差はない

→このデータにおいて複雑な決定境界は不要

・静穏時の近赤外データをもつWZのサンプルが少ない

・この場合の全体の性能は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より細分化し、どの特徴量がどの天体の型の判別に効くのか調査する