# かなた望遠鏡による情報理論・機械学習を用いた突発天体現象観測の自動意思決定システム構築

古賀柚希, 植村誠(広島大学), 他、 Smart Kanataチーム

## システム

### 研究背景

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が行えることが期待できる。

➡ **キーワード：情報エントロピー**　**不確実性**を表す量で、定義を右に示す。$p(k)$は$k$である確率、$k$は今回の場合は「天体の型」を表す。

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

### システムの全体像

1．突発現象が報告されるASAS-SN,ZTF, TNS,TOCPの4つのデータベースを見張る。
- ASAS-SN:超新星の発見を目的とした、オハイオ州立大学を中心とする研究グループのプロジェクト
- ZTF:カリフォルニア工科大学などが参加するプロジェクト
- TNS:ZTF,Gaia,Master等、様々なプロジェクトから報告される
- TOCP:アマチュア発見の明るい天体が報告され、速報性が良い

2．カタログ(表1)から既知天体と照合する。既知天体の場合は観測対象にしない。
3．新天体の情報(表2)をオンラインデータベース(Vizier)から収集する。
4．事前に用意してある教師データから判別モデル(SMLR)を構築、$p(k)$を得る。
5．$p(k)$と、事前に用意した$p(x_i|k)$から**相互情報量**を計算し、追跡観測のモード$i$を決定する。
→ $i$ = B-V(色)、 logEW(輝線の等価幅)、 dm/dt(連続観測)
6．時刻、天体位置、天気などのステータスを確認する。
7．観測可能であれば望遠鏡を駆動させる。

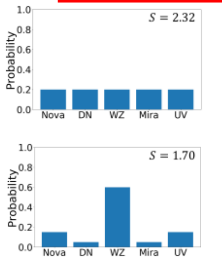


図1:情報エントロピーでの評価例

| | |
|---|---|
| 変光星カタログ | AAVSO VSX、ATLASから2秒以内 |
| 銀河カタログ | NGC、GLADE、SDSS DR8 galaxies から5分以内 |
| AGNカタログ | WISE AGN、FIRST-NVSS-SDSS AGN sample catalog、Blazar Radio and Optical Surveyから2秒以内 |

表1:照合カタログ

| | |
|---|---|
| 全てのサンプルで利用可能 | 銀経(l)、銀緯(b) |
| 静穏時対応天体があれば利用可能 | 振幅(Ampl)、色(g-r, r-I, i-z) |
| 近赤外線対応天体があれば利用可能 | 色(J-H, H-L, j-K) |
| X線対応天体があれば利用可能 | HR(HR1, HR2)、X-Opt |
| 距離が分かれば利用可能 | 距離(d_kpc)、銀河面距離(gal_abs_z)、極大時絶対等級(AbsMag_out)、静穏時絶対等級(AbsMag_qui)、Jバンド絶対等級(AbsMag_J) |

表2:特徴量

**相互情報量について**

天体発見直後の情報エントロピー$S_0 = -\sum_k p(k) log_2 p(k)$、追跡観測で観測量$x_i$が得られた場合の情報エントロピー$S_{x_i} = -\sum_k p(k|x_i) log_2 p(k|x_i)$、ベイズの定理などを用いて計算された$S_{x_i}$の平均的な情報エントロピー

$$S_i = \int p(x_i) S_{x_i} dx_i = -\int \sum_k p(k|x_i)\, p(x_i) log_2 p(k|x_i) dx_i = -\int \sum_k p(x_i|k)\, p(k) log_2 \frac{p(x_i|k)p(k)}{p(x_i)} dx_i$$

を用いて計算される$M_i = S_0 - S_i$を**相互情報量**といい、これを最大化する追跡観測を行うことになる。

**本研究の目的**

① 情報理論・機械学習を用いた型判別・観測提案 ➡ ② 型同定のための追跡観測 ➡ ③ 型同定後の追跡観測 ➡ ④ 新星・矮新星の物理解明！！

## 機械判別

### 判別モデル

研究としては①→②→③→④と進むが、①の判別精度が上がれば②はスキップ可能

➡ 判別モデルの見直し

判別モデルで求めるもの
→ $p(C_k|\mathbb{x})$ : 特徴量ベクトル$\mathbb{x}$が得られた時、型$C_k$である確率
$(C_1 = Nova,\ C_2 = DN,\ C_3 = WZ,\ C_4 = Mira,\ C_5 = UV)$
注意：ここでの$p(C_k|\mathbb{x})$はポスター左半分における$p(k)$に対応する

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

$$p(C_k|\mathbb{x}) = \frac{p(\mathbb{x}|C_k)p(C_k)}{\sum_i p(\mathbb{x}|C_i)p(C_i)} = \frac{\exp(a_k)}{\sum_i \exp(a_i)} \quad \cdots (★)$$

ベイズの定理

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

$p(\mathbb{x}|C_k),\ p(C_k)$の形が分からないので、$a_k = \mathbb{w}_k^T \mathbb{x}$の形で求められるとして(厳密には$p(\mathbb{x}|C_k)$が正規分布なら成立)、モデルパラメータ$\mathbb{w}$をデータから(最尤)推定し、$p(C_k|\mathbb{x})$を求める。

**◆生成モデル(Generative Model)**

式(★)において$p(\mathbb{x}|C_k),\ p(C_k)$をモデル化し、ベイズの定理から直接$p(C_k|\mathbb{x})$を求める。現段階では$p(\mathbb{x}|C_k)$は正規分布、 $p(C_k) = 1/5$としてモデル化している。

➡ 特徴量によって正規分布によるモデル化の向き・不向きがあるか？

向いていると思われる例

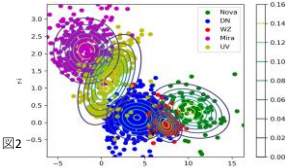


図2

向いていないと思われる例

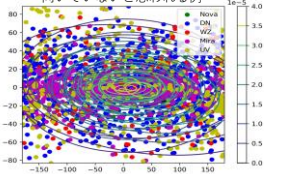


図3

判別例

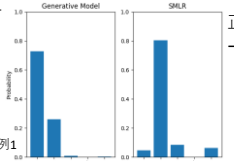


図4:判別例1

正解:DN
→SMLR成功

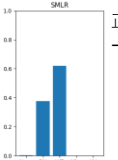


図5:判別例2

正解:DN
→GM成功

それぞれのモデルの得手不得手を調べ、全体としての判別性能の向上を目指す