Tomo-e Gozen突発天体探査における半教師あり学習によるReal/Bogus分類

高　橋　一　郎

〈東北大学天文学教室　〒980–8578 宮城県仙台市青葉区荒巻字青葉 6–3〉

e-mail: ichiro.takahashi@astr.tohoku.ac.jp

近年になって機械学習が天文データに応用される中で，学習データが膨大になるとともに，そこに含まれる誤ラベルによる悪影響が問題になってきている．そこで，誤ラベルに対処した半教師あり学習をTomo-e Gozenによる突発天体探査におけるReal/Bogus分類に応用した研究を紹介する．新しい分類器は，自身で学習データに含まれる誤ラベルを検出し，それらを「ラベルなし」にした上で半教師あり学習を行う．人によるラベル付け直しの労力を節約したこの分類器により，実際の観測データに対する分類成績は従来のものと比較して1桁以上改善した．この分類器をTomo-e Gozenの解析パイプラインに実装した結果，本物の突発天体の回収率を保ったまま突発天体候補の数が従来の40分の1にまで減少し，効率的な追観測ターゲットの選定が可能となった．

1.　イントロダクション

近年，天文学によって得られるデータ量が膨大になるにつれて，天文データへの機械学習技術の応用が注目されてきている．天文データには画像データが多く，特に画像認識・分類や画像復元などの機械学習タスクと相性がよい．例えば，銀河の形態分類などは機械学習のよい応用例として挙げられる [1].

突発天体の研究分野においても機械学習の応用が必須になってきている．ここ数十年の間に突発天体サーベイ観測はより広視野，高感度そして高頻度になってきた．その結果，発見される突発天体の数は急増しており，今では年間に数万の発見が報告されるようになってきている．他の望遠鏡で追観測を行い，多波長のスペクトルなど詳細なデータを取得するためには，突発天体をリアルタイムに検出する必要がある．

可視光における突発天体検出は主に画像の差分によって行われる．この方法では新しく取得された画像から同じ場所の過去の画像を差し引くことで突発天体を検出する（図1）．差分画像には明るさが時間で変化しない通常の恒星や銀河などは残らず，突発天体候補のみが残るので，効率的に突発天体を検出することができる．しかし，差分画像からの検出には誤検出が大量に発生しやすいという欠点がある．これらを我々はBogus（日本語で「偽物」の意味）と呼んでいる．

観測規模の増大に伴い，Bogusの数は人間の目やパラメータカットなど従来の選定方法ではチェックすることが不可能なレベルにまで増えている．そこで，これらに代わる手法として機械学習で本物（Real）と偽物（Bogus）を分類する手法が注目されてきている．

これまで機械学習によるReal/Bogus分類に関する様々な手法が提案されており，それらが世界の多くの突発天体サーベイに実装されている．初期には画像から特徴量を抜き出し，それをRan-

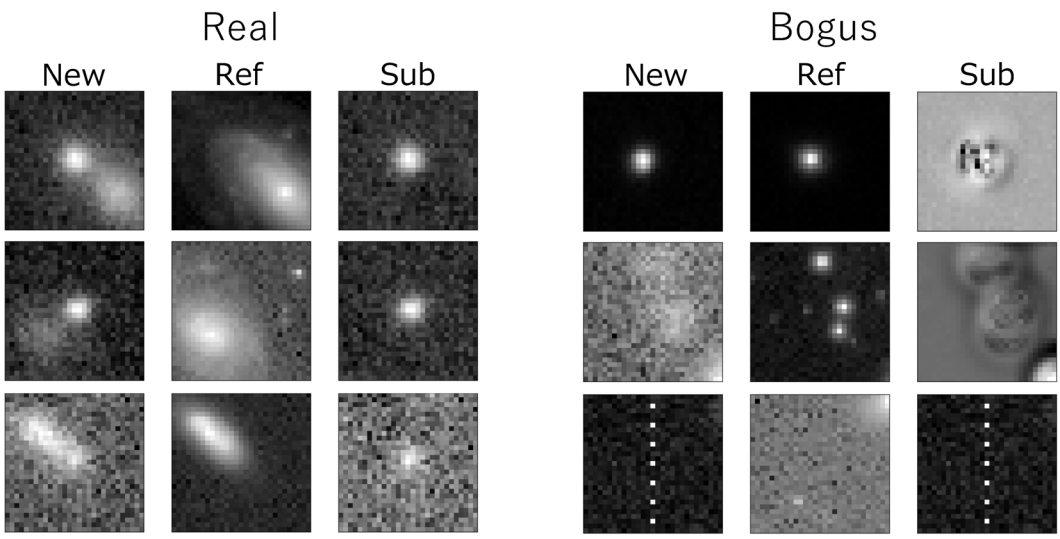

図1　RealおよびBogusの例．それぞれ左から観測画像（New），参照画像（Ref），差分画像（Sub）の順で並んでいる．

dom ForestやNeural Networkに入力して分類が行われていた [2–5]．最近では，Convolutional Neural Network（CNN）を使用して画像データを直接入力し，機械自身に特徴量を学習させて分類させる手法が主流になってきている [6–9]．例えばZwicky Transient Facility（ZTF）の突発天体サーベイでは，CNNベースの分類器であるbraaiが適用されている [10]．

より高性能な分類を行うための手法として，より複雑な機械学習のモデルを使用することが挙げられる．複雑なモデルの学習には大量の学習データが必要となるため，学習には本物のデータではなくシミュレーションデータが用いられることが多い．しかし，何百万もの大量のデータを用意すると，それらを手動でチェックすることは不可能になり，学習データが間違ったラベルによって汚染される可能性がある．学習データに紛れ込んだ誤ラベルは，パフォーマンスに悪影響を与えるため（例えば [11]），これらの誤ラベルに対処することが必要となる．そこで，我々はこれら誤ラベルへの対処に焦点を当て，可視光広視野カメラであるTomo-e Gozenによる突発天体サーベイにおけるReal/Bogus分類器の改善を行った [12]．

2.　半教師あり学習によるラベル誤りへの対処

2.1　使用データ

Tomo-e Gozenプロジェクトでは，東京大学木曽観測所にある1 m望遠鏡と84枚のCMOSセンサによる広視野カメラを用いて可視光での広視野突発天体サーベイ観測を行っている [13]．このサーベイでは1回の露出で20 deg^{2}の視野を約18等の感度で観測し，同じ空を1晩に3〜4回程度の高頻度で探査している．分類器への入力は突発天体の周りを切り抜いた画像で，差分に使われた観測画像と過去の参照画像，そして差分画像の3種類の画像である（図1）．学習データには，Realとしてシミュレートした人工星を埋め込んだもの，Bogusとして実際の観測で検出されたものを使用し，それぞれ数百万のサンプルを用意した．テストデータには，Realとして実際にTomo-e Gozenのデータで検出された突発天体を用い，Bogusとしては学習データと同様に実際のものを使用した．

2.2 分類器

Tomo-e Gozenによる突発天体探査では1日に10^5もの画像が得られており，毎晩10^6もの変動天体候補が検出されている．ほとんどがBogusであるそれらの候補から，CNN分類器によって突発天体候補を選別していたが，分類成績が頭打ちになり，まだ大量のBogusが候補に残ってしまっていた [14]．そこで，分類器の性能改善のため，学習データやモデルの構造を検証したところ，学習データに誤ラベルが存在し，それらは分類器が大きく間違えたサンプル中に多く含まれていることを見出した（図2参照）．見つかった誤ラベルの例を図3に示す．つまり，実際には機械はこれらを正しく分類しており，ラベルが間違っているだけである．これら誤ラベルの混入率はBogusにおいては0.6%，Realで1%程度と見積もられた．

そこで我々は新しい分類手法として，分類を2段階に分けることにした．まず，普通に教師あり学習したモデルで学習データ自体を分類し，そこで誤分類されているものを「ラベル無し」とした上で，次に半教師あり学習を行う．半教師あり学習にはVirtual Adversarial Training（VAT）[15]という学習方法を採用した．

VATとは敵対的学習という機械学習方法の一種である．敵対的学習とは分類器が誤りやすい摂動を付加したデータで学習する技術で，これにより摂動に対してよりロバストな分類器になる．VATの大きな特徴の一つが，目的関数の計算にラベル情報を必要としないことで，これにより半教師あり学習が可能となる．新モデルと従来のCNNモデルのテストデータに対する分類性能を比較するのにReceiver Operating Characteristic（ROC）カーブを使用する（図4）．ROCカーブとは分類時のしきい値を変えながら，BogusをRealと誤分類してしまう割合であるFalse positive rate（FPR）と，Realを正しくRealと分類できる割合であるTrue positive rate（TPR）を計算し，それらを繋げたものである．分類性能が良ければこのカーブは左上に近づいていき，誤分類が少なく，正解を正しく拾えているということを表している．

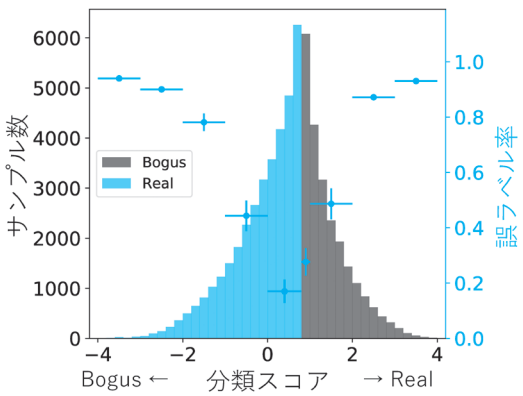

図2　機械に誤分類されたサンプルの分類スコア分布と人間の目による標本調査から推定した誤ラベル率．右にいくほどReal, 左にいくほどBogusと機械は分類している．

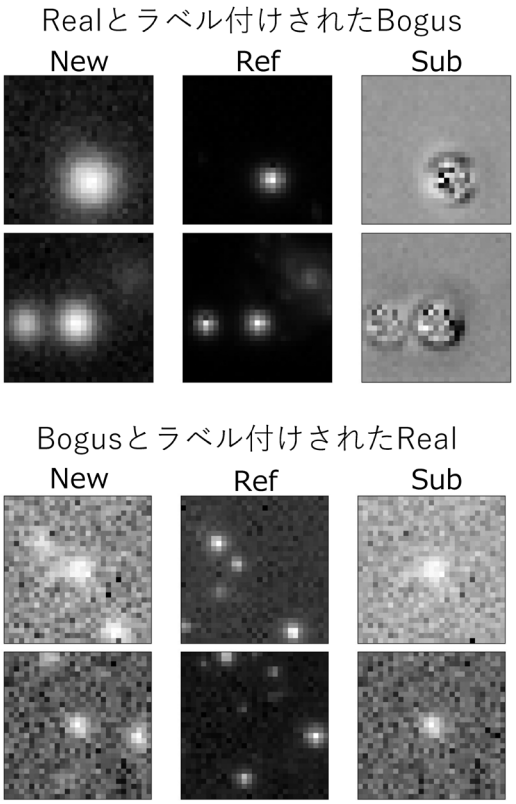

図3　誤ラベルの例．

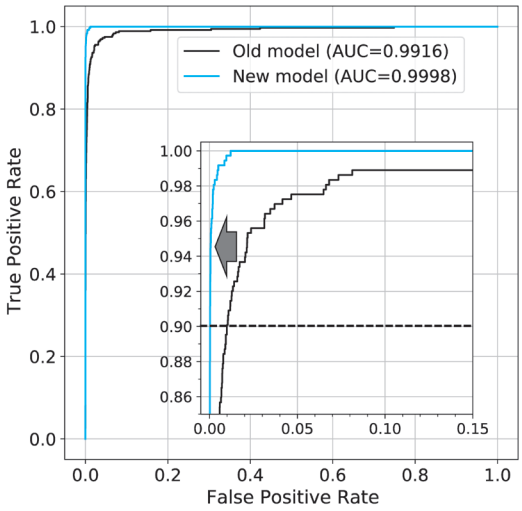

図4　従来モデルと新モデルのROCカーブ．従来モデル（黒）に比べて新モデル（シアン）の成績が高いことが分かる．

図4を見ると，新モデルのROCカーブが従来のものに比べてより左上に位置し，性能が改善していることが確認できる．ROCカーブの下の面積（Area Under the Curve: AUC）は分類器の性能のよさを表し，新モデルではAUCは0.9998まで上昇している．

次に新しいモデルのFPRに注目する．実際の運用では，本物の突発天体を逃してはいけないため，TPRが高い場合におけるFPRが重要になる．従来モデルはFPRが高くカーブが右側に寄っているが，新モデルにすることで改善している．TPR＝0.9，つまりRealを9割正しく分類できるようなしきい値の場合のFPRは従来のモデルに比べて，1桁以上小さくなっている．

今回使用した学習データに含まれる誤ラベルの割合は約1%だったが，もしラベル間違いの割合が多かった場合に新手法が通用するのかも検証した．図5はラベル間違いの割合が元のまま，5%に増やした場合，10%にまで増やした場合それぞれのFPRを示している．誤ラベルの割合が高くなるほど誤ラベルを無視した通常の学習の成績が悪くなっている一方，新手法では常にFPRが

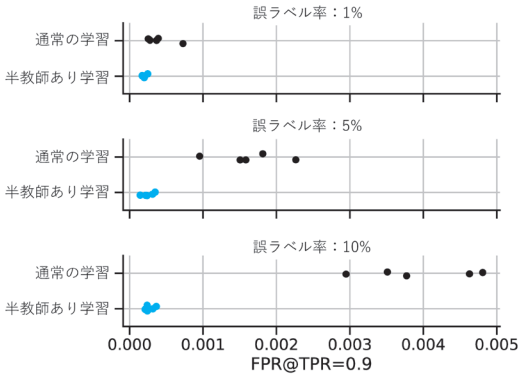

図5　誤ラベルを増やした場合のFPR（TPR＝0.9のとき）．各モデルの5つのデータは，シード値を変えた5回の学習それぞれの成績を表している．

低いままになっている．つまり，新手法がラベル間違いの割合の高いデータセットに対してより効果的であることがわかる．

3.　実際の運用でのパフォーマンス

テストデータの分類性能に問題がなかったので，2021年5月にTomo-e Gozenの解析パイプラインに新しいReal/Bogus分類器を実装した．実装後約1週間の間に検出された分類すべき変動天体候補は約6,000万天体にものぼる．5月初めから約1ヵ月間におけるTomo-e Gozenの突発天体データベースへの登録数の変化を図6に示す．

図の破線が，実装のタイミングを表している．新しい分類器によって，変動天体として登録された天体の登録数は130分の1程度に，それらのうち2回以上同じ座標で検出された確度の高い突発天体の登録数は40分の1程度に減少した．なお，単に数が減っただけではなく，本物の天体の回収率はそのままに，誤分類によって間違って登録されたものが減っていることも確認している．新しい分類器の実装により，最終的な突発天体候補の登録数はこれまで1日あたり6,000天体だったものが150天体にまで減り，人が簡単にチェックできる程度になっている．これにより追観測ターゲットの選定が短時間でできるようになり，Tomo-e

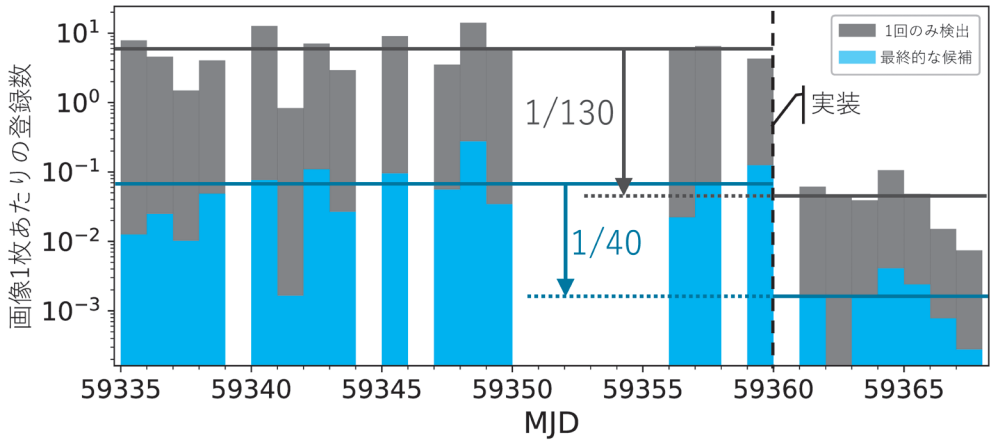

図6　Tomo-e Gozenの突発天体データベースへ登録される候補の数の変化．灰色のデータが主に分類器によって変動天体として登録されたもの．シアンのデータはそれらのうち2回以上同じ座標で検出された最終的な突発天体候補．

Gozenをトリガーとしたせいめい望遠鏡やすばる望遠鏡などによる追観測が以前よりも効率よく行えるようになった．

4. ま と め

我々はTomo-e Gozen突発天体サーベイのデータに機械学習を応用し，従来のCNNでは成績が頭打ちになっていたReal/Bogus分類の成績改善を図った．半教師あり学習を適用することで，人によるラベルの付け直しの労力をかけず，1%の誤ラベルを機械自身が修正することで，最終的に1桁以上の性能向上を実現した．さらに，誤ラベルの混入率が高くてもこの手法で対応できることも確認することができた．

今回紹介した手法は，誤ラベルや，確度の低いラベルを含んだ学習データを使用した様々な分類タスクへの応用が可能である．同様の問題の解決にこの記事が参考になれば幸いである．

謝　辞

この度，本稿の執筆の機会を与えてくださった西澤淳氏に感謝いたします．ここで紹介した研究は筆者らによる投稿論文 [12] に基づいています．共著者の皆様に改めて感謝いたします．また，分類器開発に多大なご協力を頂いた井本康宏氏にここで感謝の意を表します．

参 考 文 献

[1] Barchi, P. H., et al., 2020, Astronomy and Computing, 30, 100334
[2] Bloom, J. S., et al., 2012, PASP, 124, 1175
[3] Brink, H., et al., 2013, MNRAS, 435, 1047
[4] Wright, D. E., et al., 2015, MNRAS, 449, 451
[5] Morii, M., et al., 2016, PASJ, 68, 104
[6] Gieseke, F., et al., 2017, MNRAS, 472, 3101
[7] Turpin, D., et al., 2020, MNRAS, 497, 2641
[8] Killestein, T. L., et al., 2021, MNRAS, 503, 4838
[9] Hosenie, Z., et al., 2021, Experimental Astronomy
[10] Duev, D. A., et al., 2019, MNRAS, 489, 3582
[11] Ayyar, V., et al., 2022, arXiv e-prints, arXiv:2203.09908
[12] Takahashi, I., et al., PASJ, in press (doi:10.1093/pasj/psac047)
[13] Sako, S., et al., 2018, Society of Photo-Optical Instrumentation Engineers（SPIE）Conference Series, 10702, 107020J
[14] 浜崎凌, 2020, 修士論文, 甲南大学
[15] Miyato, T., et al., 2016, arXiv e-prints, arXiv:1507.00677

Real/Bogus Classification in the Tomo-e Gozen Transient Survey With Semi-supervised Learning

Ichiro Takahashi

Astronomical Institute, Tohoku University, Aoba, Sendai, Miyagi 980–8578, Japan

Abstract: This article introduces the application of semi-supervised learning to the Real/Bogus classification of the Tomo-e Gozen transient survey. The new classifier detects mislabeled samples in the training data, and performs semi-supervised learning by setting mislabeled samples to "unlabeled" samples. The classification performance of this new classifier is improved by more than an order of magnitude compared to the previous classifier. In the Tomo-e Gozen transient survey, this classifier has reduced the number of transient candidates to 1/40 of the previous system, while maintaining the recovery rate of real transients.