

社会状況の変化に適合したイベント予測手法

川崎 仁嗣[†] 豊田 正史^{††} 是津 耕司^{†††}

[†] 東京大学情報理工学系研究科 〒113-86563 東京都文京区本郷 7-3-1

^{††} 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

^{†††} 情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

E-mail: [†]{kawasaki,toyoda}@tkl.iis.u-tokyo.ac.jp, ^{††}zettso@nict.go.jp

あらまし 新型コロナウイルス (COVID-19) の感染状況により、スポーツやコンサート等のイベントの開催が自粛されたり、観客数制限の下で開催されたりする状況が生じたためイベント参加者の規模は大きな影響を受けた、今後も新たな感染症等により同じような事態が起きうる可能性がある。イベント発生地での人口予測は、混雑を避けるナビゲーションや開催地周辺における混雑対応などに重要であるが、こうした社会状況の変化が激しい中では、過去のデータに基づく予測では十分な精度が得られない。本研究では、イベント発生地の数日先までのイベント有無および種類を、未来のイベントに関する言及を含む Twitter 投稿と過去の人口時系列データから予測するタスクについて、社会状況の変化に適応して予測器を更新する手法を提案する。

キーワード イベント予測, Twitter, COVID-19

1 はじめに

2020年2月頃からの新型コロナウイルスの感染拡大により、多数の人々が集まるイベントは主要な感染源とみなされスポーツやコンサート等の大規模なイベントは長期間中止され、街における人の流れは大きく変化した。COVID-19感染症の感染経路特定や感染地域の予測などにおいて感染者の発生傾向と人の移動傾向との相関が高いことが報告[1][2]されており、移動制限やリモートワークなど様々な対策が取られたことも人流の変化に寄与している。その後様々な感染対策やワクチン接種等により感染状況が改善されると、人数制限等の元でイベントも再開されるようになった。しかし、コロナ以前の状況に戻ったわけではなく、現在も第8波の感染が広がっているなど終息したとは言い切れない状況であり、今後も感染拡大の状況によって人流が大きく変化する可能性がある。

新型コロナウイルスの感染拡大後においては、イベントにおける人数制限の適用やオンライン配信とのハイブリッド開催など、感染症流行前のイベント開催時とは人口変動が大きく異なっており、イベントの有無や規模の予測は以前と同様にはできなくなっている。プロ野球については2020年7月末から観客を入れた試合開催が行われるようになったが、イベント開催が完全になくなったり、ナイトゲームのみが実施されたりというように、感染状況の変化によって開催状況が大きく変化しており、イベント開催日の数、イベントの種類、およびイベントの規模の分布が時期によって大きく異なる。このため機械学習に基づくイベント予測の精度の劣化は避けられず、モデルの頻繁な再学習が必要となる。

図1は、東京ドームにおける翌日のイベントの有無を、Twitter データを用いて予測するという単純なタスク設定において、コロナ禍以前(2014~2018)に学習したモデルを使用し続け

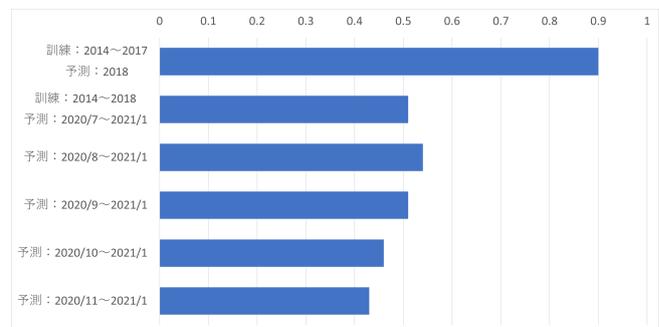


図1 東京ドームにおける翌日のイベント有無の予測精度の劣化

た場合、有観客試合が再開された2020年7月以降の予測精度(イベント有を正解とするF1値)がどのように変化するかを示している。分類器としては入力としてイベント会場である東京ドームとイベントの日付の両方に言及する複数のツイートから抽出したBoW (Bag of Words) を取り、翌日のイベントの有無を出力するGradient Boosting Classifier (GBC) を用いた。コロナ禍以前には0.9だったF1値が、イベント再開以降については0.5以下に落ち込んでいる。実際にはイベントがない多くの日をイベント有りと予測することにより適合率が大幅に低下している。この予測結果をナビゲーションに用いた場合、利用者が過剰にイベント会場を避ける行動を取る要因となりうる。

予測モデルの更新は必須であるが、時間とともに各訓練データが精度に寄与するかどうかが大きく変化することから通常の更新方法を用いることは難しい。単純にはオンライン機械学習の手法を用いる方法や、直近の一定の期間のデータを移動窓により選択して再学習を行う等の方法が考えられるが、コロナ禍においては数か月にわたってイベントがない期間が続くなど過去のデータを忘却するとイベント開催時のデータがなくなるような不具合が生じ、実際に予測精度が悪化することを確認して

いる。

人口変動の状況を取得するために GPS を搭載したスマートフォンなどの携帯端末から収集された位置情報を活用し、都市全体の規模で人口予測を行う手法はこれまでに研究されてきているが [3], [4], 現在の状況や数時間先の予測をおこなうものでコロナ禍のような非常に大きい人流の変化に適応した予測に関してはほとんど研究がなされていない。我々は、Twitter における未来のイベントへの言及を利用して、不定期に発生するイベントの人口を数日前までに予測する手法を提案してきた [5], [6]。コロナ禍におけるイベント開催状況へ適応するため再学習の際に使用する訓練データをサンプリングする手法も提案したが [7], 精度向上は限定的であった。本論文では、再学習の際に使用する訓練データの期間を選択する比較的単純な方法を提案し、再学習時に精度向上が可能である事を確認した。

2 問題設定

本論文では再学習の基礎的な検討を行うため、イベント発生地については東京都内最大規模のイベント会場である東京ドームを選択し、翌日のイベントの有無を Twitter データを用いて予測するシンプルなタスク設定を想定する。イベントの有無を予測する対象となる日を d とし、予測を行う日をその前日 $d-1$ とする。予測器の入力は $d-1$ の終わりまでに得られるツイートとし、 d におけるイベントの有無を 2 値分類して出力する。予測器への入力となるツイートデータは [5], [6] と同様に、予測対象日に対する未来の言及を含むツイートを使用する。具体的には d の日付を表す文字列および会場の名称（今回は「東京ドーム」「ビッグエッグ」など）を両方含むツイートを抽出し、その集合を入力とする。

東京ドームでは 2020 年 3 月から 7 月中旬まで一切のイベントが中止されていたため、イベントが再開された 2020 年 7 月 28 日の 8 日前である 7 月 20 日から翌日のイベント有無の予測を開始するものとする。再学習は毎日行うものとし、 $d-1$ の終わりまでに得られた情報 ($d-1$ のイベント有無およびツイートデータ) を用いて再学習を行ったモデルを用いて d におけるイベントの有無の予測を行う。また、コロナ禍以前の Twitter データ (2014 年 12 月~2018 年 11 月) および各日のイベントの有無のデータが存在するものとし、訓練データとして使用できるものとする。

3 再学習手法

コロナ禍においては、イベント開催の状況がその時々での感染対策の方針に従って変化していくため、再学習を行う際に追加することで精度向上に寄与する訓練データの期間が状況に応じて変化する可能性がある。再学習に使用する訓練データの期間を変化させることで、予測精度が大きく変わることを予備実験において確認している。

そこで本論文では、再学習の際に使用する訓練データの期間のパターンをいくつか用意し、各パターンのデータで再学習した複数の弱分類器によるアンサンブル学習を行う手法を提案す

る。訓練データの期間のパターンとしては以下を用いる。

- All: コロナ禍以前のデータおよび $d-1$ までに得られたすべてのコロナ禍後のデータを用いる。
- MW: 一定期間 (t) の移動窓を用い、 $d-1-t$ から $d-1$ までに得られたデータを用いる。
- Mix: コロナ禍以前のデータおよび MW のデータを併せて用いる。

All は使用できるすべてのデータを用いたもので、平均的な性能を示すことが期待される。MW は一定期間の移動窓に含まれるデータのみで学習を行うことで、一定期間コロナ禍以前とは異なる状況が続く場合に良い性能が得られることを期待する。Mix は、コロナ禍以前のデータと MW を組み合わせることでコロナ禍中の特殊な状況だった期間を忘却していくことを意図している。

アンサンブルの方法としては、以下の 2 通りを用いる。

- 多数決: 3 つの弱分類器の予測結果の多数決で予測を行う。
- 失敗時切替: 弱分類器を 1 つ選択して使用し、予測を失敗した場合には次の日から他の予測に成功した弱分類器に切り替える (成功した弱分類器が複数ある場合は上記の順番で選ぶ)。予測に成功した弱分類器がない場合は現在のものをそのまま使用する。

多数決は最も良く用いられるアンサンブルの方法として用い、失敗時切替については 1 つの弱分類器だけが暫くの間正解を続けるような状況があることを想定している。

4 実験結果

提案した再学習手法を東京ドームを対象としたイベント有無予測に適用した。2020 年 7 月 20 日から翌日のイベント有無の予測を開始し、1 日毎に予測器の再学習を行いながら 2020 年最後のイベントがあった 12 月 15 日まで予測を行った。

分類器への入力は $d-1$ の終わりまでに得られるツイートであり、 d の日付および東京ドームを表す文字列を両方含むものの集合とする。今回は基礎的な検討であるため、ツイートの集合から形態素解析によって得られる形態素の BoW (Bag of Words) を入力としている。形態素の頻度については無視し、各形態素の有無のみを 0, 1 のベクトルとして表現したものをを用いる。分類器としてはロジスティック回帰 (L2 正則化を使用) を用いている。

まず、訓練データの 3 つのパターンの性能の違いを評価する。イベント有を正解とした直近 30 日の F1 値を日毎にプロットしたものを図 2 に示す。All, MW, Mix の各パターンの F1 値の他、比較のためコロナ禍前のデータのみを訓練データとして用いた分類器を使用し続けた場合の性能 (PreCOVID) も示している。MW パターンにおける移動窓の期間 (t) としては 60 日を定数として用いており、チューニングは行っていない。

図 2 を見ると一番良い性能を示すパターンがある程度の期間毎に入れ替わっており、アンサンブル学習が有効であることを示唆している。イベントが開始された直後は Mix パターンが

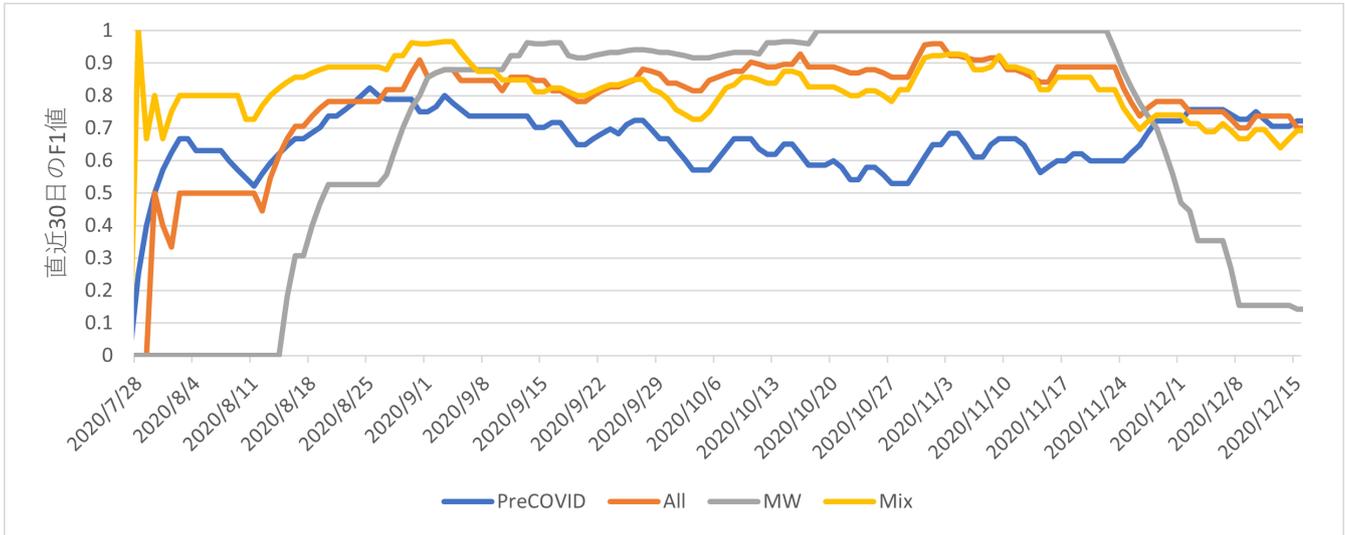


図2 各訓練データパターンによる東京ドームにおける翌日のイベント有無の予測精度 (F1 値)

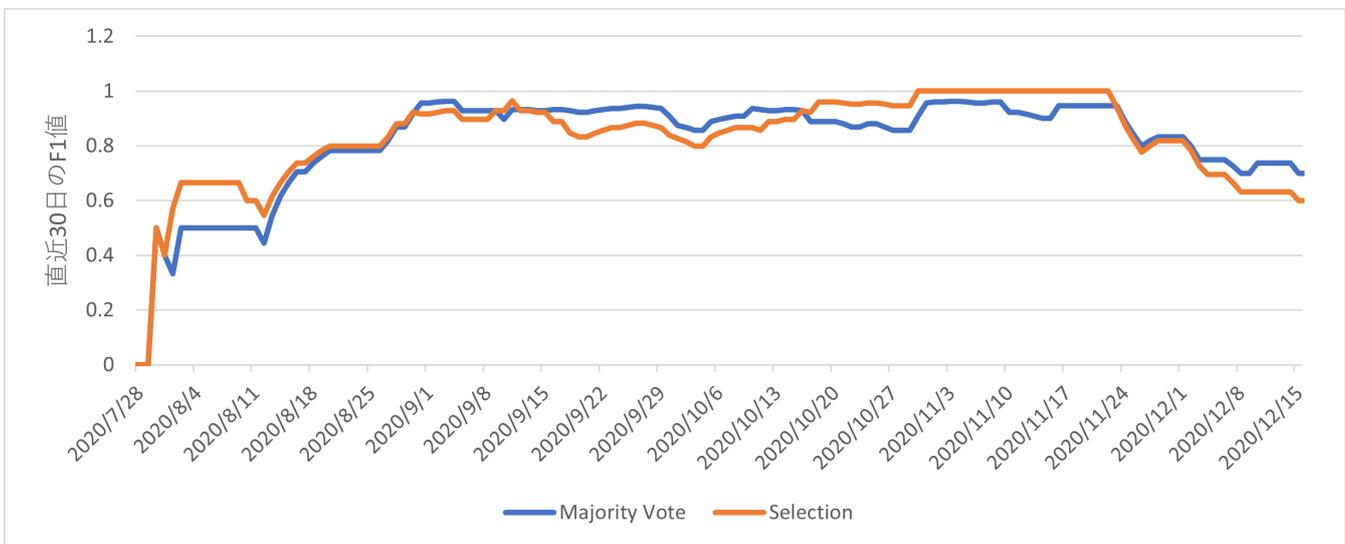


図3 アンサンブルの方法による東京ドームにおける翌日のイベント有無の予測精度 (F1 値)

一番良い性能を示しているが、プロ野球の試合が継続的に行われた9月、10月には移動窓を用いた MW パターンが最も良い性能となっている。11月下旬からは都市対抗野球が12日間に亘って行われており、移動窓内のデータのみでは予測ができなくなったため MW パターンは大きく性能を落とし、All パターンが一番良い性能を示すようになってきている。

次に、アンサンブルの2通りの方法の違いを評価する。同様にイベント有を正解とした直近30日のF1値を日毎にプロットしたものを図3に示す。多数決 (Majority Vote) と失敗時切替 (Selection) を比較すると全体としては大きな差はないが、イベント開始直後においては正解を出力できる弱分類器が1つしかない期間があったため失敗時切替の方が良い性能を示している。その後は若干の入れ替わりがあるもののほぼ同程度の性能を示している。全期間の累積のF1値は、多数決が0.85、失敗時切替が0.84となっておりほぼ同等の性能となっている。

5 まとめと今後の課題

コロナ禍による社会状況の変化に適応してイベント予測モデルを再学習する手法を提案した。コロナ禍における大きな人流の変化にあたっては、時期ごとに訓練データの質が変化するためオンライン学習等の直近のデータを重視する再学習の方法は適しておらず、使用する訓練データの期間を適切に変更することが有効であることを示した。

今回示した結果では、移動窓の期間や分類器のパイパーパラメタのチューニングを実施していない。イベント再開直後など正解データの少ない時期にどのようにハイパーパラメタのチューニングを行うかは今後の課題として残っている。また、訓練データの期間のパターンは移動窓の期間のバリエーションなど複数考えられ、パターンを増やした場合にはアンサンブルの方法による性能の違いも大きくなることが考えられる。

今回はイベントの有無というシンプルな問題設定で実験を

行っているが、野球場でのイベントにはデイゲーム、ナイトゲーム、展示会、コンサートなど複数の種類があり、イベント種別の分類やイベント人口の時系列の予測などが次のステップとして必要となる。さらに長期間のイベント中止期間からイベント再開となった時、また再度イベントが中止となった場合等にどの程度早く状況に適応できるかも今後の課題として重要である。再学習時点でのイベントの開催方針、非常事態宣言の有無など事前に知り得る情報に基づいて半自動でチューニングを可能とする方法を検討していく予定である。

謝 辞

本研究成果は、国立研究開発法人情報通信研究機構の委託研究「ウイルス等感染症対策に資する情報通信技術の研究開発」により得られたものです。

文 献

- [1] R. Jiang, Z. Wang, Z. Cai, C. Yang, Z. Fan, T. Xia, G. Matsubara, H. Mizuseki, X. Song, and R. Shibasaki, “Countrywide Origin-Destination Matrix Prediction and Its Application for COVID-19,” *ECML/PKDD* (4), pp. 319–334, 2021.
- [2] C. Alket, L. Marco, M. Marco, and Z. Franco, “Sensing and Forecasting Crowd Distribution in Smart Cities: Potentials and Approaches,” *IoT*, vol. 2, no. 1, pp. 33–49, 2021.
- [3] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, “CityMomentum: An online approach for crowd behavior prediction at a citywide level,” in *UbiComp*, pp. 559–569, 2015.
- [4] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu, “Deep sequence learning with auxiliary information for traffic prediction,” in *KDD*, pp. 537–546, 2018.
- [5] R. Tsukada, H. Zhan, S. Ishiwatari, M. Toyoda, K. Umamoto, H. Shang, and K. Zettsu, “Crowd forecasting at venues with microblog posts referring to future events,” in *BSD*, 2020.
- [6] 塚田涼太郎, 詹浩森, 石渡祥之佑, 豊田正史, 梅本和俊, 商海川, and 是津耕司, “未来のイベントに言及するマイクロブログ投稿を用いた人口変化の予測,” in *DEIM*, 2021.
- [7] 川崎仁嗣, 塚田涼太郎, 豊田正史, 是津耕司, “COVID-19 によるイベント開催状況の変化に適合した人口変化の予測,” in *DEIM*, 2022.