4M-07

動物園の来場者予測におけるSNSデータの貢献

鈴木耀司† 櫻井義尚‡

†明治大学大学院先端数理科学研究科

‡明治大学総合数理学部

はじめに

屋外イベントを行っている企業やイベント主催者にとって，その日の天気は来場者数に影響を及ぼし，さらに売上に直結する非常に重要な要素である。また，天気情報を含め将来の来場者を予測することは収益を最大化するために重要である。

Yap 等の研究[1]では，大規模な美術館から収集したデータを使用して，天候が訪問者の出席に大きな影響を与えるかどうかを調査した。その結果では天候は，より極端な気象環境や屋外会場でより顕著な影響を与える可能性があると結論付けられていた。また，布施 等の研究[2]では熱中症救急搬送者数を予測するモデルを作成する際に，Twitter データを用いたことで予測精度が向上したという結果が発表された。

そこで本研究では，札幌市の円山動物園の来場者予測に気象データだけでなく関連するツイートデータを加える事による予測精度への貢献と予測モデルによる違いについて明らかにしたので報告する。

対象・方法

2016 年 1 月 1 日～2018 年 12 月 31 日における日別の札幌市円山動物園の来場者数データと気象データ，屋外イベントや行事の有無を含むイベントデータとツイートデータを対象とした。

札幌市円山動物園の来場者数データは一般財団法人さっぽろ産業振興財団が運営している札幌市 ICT 活用プラットフォーム DATA-SMART CITY SAPPORO [3]，気象データは気象庁のホームページから収集した。

イベントデータは札幌市円山動物園の公式ホームページに記載されているイベント情報を元に，イベントや行事などがある日に「1」を，何も無い日に「0」のフラグを立てデータを作成した。

ツイートデータの取得に関しては，「円山動物園」の文字列を含むツイート 62,078 件から「円山動物園に来場すると考えられる」ツイート（以下、来場者数ツイート）の 11,799 件を抽出し，分析対象とした。円山動物園に来場すると考えられるツイートは表 1 に示す文字列を含むものとした。本論文の分析で用いたツイートデータは Twitter API を使用して取得した「ツイート本文」のテキスト情報と「アカウント情報」である。リツイートされたツイートデータは取得段階で削除済である。

気象データから取得された天気概況データについては，佐々木等によって提案された天気数値化簡易手法[4]を参考に天候の数値化を行い、特徴量を作成した。気象庁が提供している天候データは，昼（6:00～18:00）と夜（18:00～翌日 6:00）の 2 回あるが、このうち昼のデータを使用した。天候データは，札幌市円山動物園に最も近い札幌市地点のデータとした。

また，質的データである週データに関してダミー変数化処理を行い，周期性を考慮できるようにした。ダミー化処理を行う方法として Python の pandas.DataFrame のメソッドである get_dummies 関数を用いた。

以上より，本研究では気象データ，来場者数ツイートデータ，イベントデータ，週データを含む 27 の特徴量を用いて予測モデルを作成した。本研究では，予測モデルの当てはまりの良さを検証する為に重回帰，XGBoost，LightGBM，CatBoost の 4 つのモデルで実験を行った。モデルの学習方法では，トレーニングデータとテストデータを 8:2 に分割し，random_state は 2 に設定した。XGBoost のハイパーパラメータは，目的関数に回帰，学習率は 0.3 を採用した。LightGBM のハイパーパラメータは，目的関数に回帰，学習率は 0.1 を採用した。CatBoost のハイパーパラメータは，学習率は 0.05 を採用した。評価方法ではモデルの当てはまり良さを示す決定係数（R2）を用いた。

表 1　来場すると考えられるツイート

「Contribution of SNS data for predicting zoo visitors」

†「Yoji Suzuki・Meiji University Graduate School of Advanced Mathematical Sciences」

文字列	
円山動物園（に）行った	円山動物園（に）行きたい
円山動物園（に）行く	円山動物園（に）行って
円山動物園（に）来た	円山動物園（に）行きたい
円山動物園（に）来て	円山動物園（に）行きました
円山動物園（に）居る	円山動物園（に）居て
円山動物園（に）寄る	円山動物園（に）寄って
円山動物園（に）向かう	円山動物園（に）向かった

結果

4 つのモデルにおける結果は表 2 に示した。Twitter データなしの場合，4 つのモデルの精度は 55%前後での結果となっている。一方で Twitter データを含めてモデルを作成した場合 80%前後となり，4 つの予測モデル全てにおいて Twitter データを用いてモデルを作成した方が予測式の当てはまりの良さは高くなっていることがわかる。特に XGBoost などの勾配 Boosting のモデルにおいて予測精度の向上が大きいことがわかる。

Twitter データを含めた予測モデルで最も精度の高かった LightGBM の実測値と予測値の結果をプロットしたグラフを図 1 に示した。来場者数が 5000 人付近では予測値と実測値の乖離が少ないが，1 日に 1 万人以上来場するような日では予測値との乖離が大きいことがわかる。

表 2　予測モデルの比較

	Twitterデータなし	Twitterデータあり
	R2	
LightGBM	0.561	0.849
CatBoost	0.552	0.83
XGBoost	0.52	0.833
重回帰	0.577	0.763

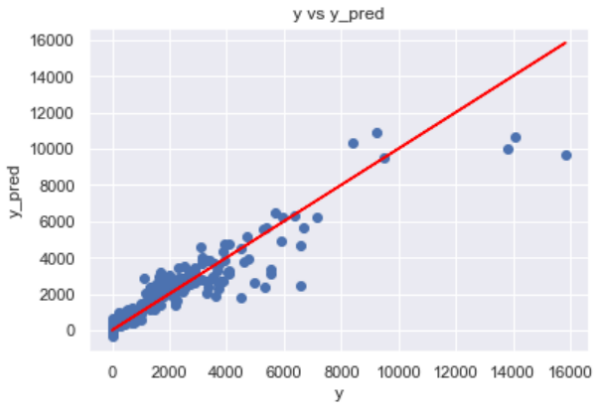

図 1　LightGBM の実測値と予測値

考察

本研究では，来場者予測モデル作成において気象データだけでなく，Twitter データを含めて作成することが来場者予測に有効であるか検討したところ，来場者数が SNS の来場者関連ツイート数と高い相関関係（r=0.848）にあることがわかった。この結果から 4 つのモデル全てにおいて精度が向上したと考えられる。

まとめ

札幌市の円山動物園の来場者予測モデルに気象データやイベントデータだけでなく，ツイートデータを加える事で予測は重回帰、XGBoost、LightGBM、CatBoost の 4 つのモデル全てにおいて精度が向上し，Twitter データが大きく貢献していくことを確認した。特に勾配 Boosting モデルの精度向上への貢献が大きいことを明らかにすることができた。

今後の課題

本研究では予測モデルの当てはまりの良さに着目して検討してきた。今後は予測精度の向上のための施策として、行動履歴データの活用や周辺地域のイベントを考慮することで，より高い精度モデルが実装できると考えている。

文献

[1] N. Yap, M. Gong, R. K. Naha and A. Mahanti, ″Machine Learning-based Modelling for Museum Visitations Prediction,″ 2020 International Symposium on Networks, Computers and Communications (ISNCC), pp. 1-7, (2020)

[2] 布施 明，坂 慎弥，布施 理美，萩原 純，宮内 雅人，横田 裕行，”ーツイッターデータと気象データから熱中症救急搬送者数を予測する”，日本臨床救急医学会雑誌，22 巻 4 号 pp. 573-579, (2019)

[3] 札幌市 ICT 活用プラットフォーム DATA-SMART CITY SAPPORO. URL<https://data.pf-sapporo.jp/>(2022/1/7 閲覧)

[4] 佐々木三郎，福永青空，太田豊，”-天候数値化簡易手法による太陽光発電の 地域別および全国大の発電量の解析”，エネルギー・資源学会論文誌， 38 巻 5 号，pp. 27-35, (2017)