The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

「土地勘」は学習できるのか –機械学習によるエリアラベリング

Learning a sense of locality?: labeling local areas by machine learning

宗政一舟 *1　深見俊和 *2　森脇大輔 *2
Isshu Munemasa　Toshikazu Fukami　Daisuke Moriwaki

*1明治大学　*2株式会社サイバーエージェント
Meiji University　Cyberagent, Inc.

To improve the performance of location-based advertising, a model of "sense of locality" is estimated, where the output variable is the "label" of each location and inputs are geographical and demographic information associated with the location. As the input variables are all taken from the Internet, the output is unique dataset that we collects from people who know well the location. The model is estimated with three methods and XGBoost outperforms over logistic regression and SVM. The results show fairly predictive power with f1 score of 0.66.

1. はじめに

近年インターネット広告においては, 人々の行動履歴から適切な広告をプログラマティックに判断して配信する「行動ターゲティング広告」が主流となっている *1. その中で, オンライン上での行動に止まらずオフライン上での行動履歴からターゲティングを行う広告配信手法も登場している.

オフライン上での行動履歴は, スマートフォンなどの端末から得られる位置情報を人間が理解できる文脈に変換することで, ターゲティングに活用できる. 例えば, 土日に海辺に行ってる人は, 何らかのマリンスポーツを楽しんでいる可能性があり, 釣り道具やサーフボードへの関心が高いかもしれない. また, テーマパークによくいく人は, 他の面白そうなテーマパークに関心が高いかもしれないといった具合である.

ただし, 海辺やテーマパークなどのように行動が即座に読み取れる場所はそれほど多くない. 建物が密集する地域では, 住宅や商店, オフィスなど様々な用途の場所が混在しており, 何のためにそこにいるのかを位置情報だけから読み取るのは難しい. 他方で, 古書店街や道具屋街, 電気屋街など, 業種が偏ったまちが存在することはよく知られている. こうしたまちに頻繁に来る消費者は, ある特定の物事に関心が高いことが容易に予想され, 関心に応じた広告を配信することが可能である.

そのまちがどのような属性を持つかは, そのまちに長く住んでいる「土地勘」を持った人間には判別可能だが, そうでない場合は, 客観的なデータに基づいて分類することが必要である. 実際, 広告配信の現場では, 店舗の存在などから土地を分類し, その土地に来訪したユーザーをある属性を持つとみなすということ行われている. しかしながら, そうしたルールによる分類は, 実際の土地の雰囲気を見誤っている可能性や本来収集したい場所を見落としている可能性があり, 効率的とは言えない. データに基づいて長年そこに住んだ人がもつ「土地勘」を予測モデルに学習させることができれば, オフラインの行動履歴からより正確なターゲティングが可能である. こうした問題意識から, 本研究では, 土地勘のあるものによる土地のラベリングを元にオープンデータを用いた機械学習を行うことで, 日本全国の土地を分類するモデルを推定した.

本稿の構成は以下のようになっている.2. では, 関連研究について言及する.3. では, 分析手法について説明し,4. では研究に用いたデータについて解説する.5. では, 結果を示し考察を述べる.6. でまとめを行う.

連絡先: 宗政一舟，明治大学理工学研究科基礎理工学専攻，
〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1,
E-mail: munemasa@cs.meiji.ac.jp

*1 [Book 2015] は主要なアプリ広告プラットフォームの一つである Google 社の AdMob (https://www.google.co.jp/admob/) を解析することで, ユーザーの属性や行動をもとにしたターゲティングを用いた広告配信は 4 割と推定している

2. 関連研究

人々の行動と土地の特性を有機的に結びつけて土地の分類を行った先行研究としては,[Noulas 2011] があげられる.[Noulas 2011] はロンドンとニューヨークのある区画を格子状に区切った上で, その小区画を SNS におけるユーザーの「チェックイン情報」に基づいてクラスタリングしている. このチェックイン情報は明示的なユーザー行動を示しており, その小区画がどのような土地なのかという情報を抽出することができる.[Noulas 2011] はさらに得られた土地のカテゴリーに基づいてユーザーのクラスタリングを行っている. また,[Kurashima 2013] は, レビューサイトと SNS から, ユーザーの行動履歴を抽出し, 潜在トピックモデルを当てはめることで, 地点の特徴を推定できることを示している.[Kurashima 2013] の手法は, 明示的に地点の特徴の情報がなくても, ユーザーの行動履歴から推定することができることを示している. 本研究は土地を格子状に区切っていることや, 明示的な土地の属性情報を用いている点で [Noulas 2011] に近いと言えるが, 分析者が直接, 土地の分類ラベルを設定し, それを土地勘のあるものにつけさせている点で異なる. 分析者が直接分類ラベルをつけることで, 実際にラベルを活用する際に使いやすい分類を行うことができることが利点である一方, 分析者が予期しないカテゴリはその他のカテゴリに埋没するため, 新たな知見が得られにくいという欠点があると考えられる.

3. 分析手法

本節では, 分析手法を解説する.3.1 で学習に用いた教師データの作成方法について説明し,3.2 において用いたモデルについて説明する.

3.1 教師データの作成

ほとんど人の訪れない山中などを除けば土地は特定の目的にラベルづけすることができる. 例えば, オフィスが密集している

場所は, オフィス街としてラベルづけすることができ, 日中そこにいる人は一定の確信を持って「サラリーマン」とか「OL」としてラベルづけすることができる. あるいは, 土日に海や山に滞在することが多い人は, アウトドア好きという推測ができる. 広告配信の観点からは，趣味嗜好が類似したユーザーが集まる場所を特定することが出来れば，該当地にいるユーザに対し，興味を持つ可能性が高いと推定される商品を効率的に訴求することができる.

筆者らはまずその土地にいる人の特性が解釈しやすいラベルを用意した. ラベルの選定にあたっては, 普段広告主とコミュニケーションを取っている営業担当からヒアリングを行い, 広告配信ターゲットを想定しやすいものとするとともに, あらゆる土地に一つ以上ラベルがつけられるように網羅性を確保するようにした. 例えば,「アウトドア」や「観光地」,「高級ショッピング街」といったその人の趣味嗜好や経済力が現れるようなラベルを用意する一方,「住宅街」や,「ビジネス街」,「工業地帯」といったそれだけでは人の属性が抽出しづらいものの, 多くの面積を占めている場所のラベルも用意するようにした. ラベル数は結果的に 25 となった.

土地のラベルづけの単位は概ね $100m \times 100m$ と, 全国 27 万地点を対象とした. 先行研究と比較すると [Noulas 2011] は $625m \times 625m$ のサイズのやや広いエリアを単位とする一方,[Kurashima 2013] はレストランなどの地点を分析と単位としており, 本研究はその中間にあるといえる. 分析単位を小さくするとラベルづけする地点の数が膨大となりデータ収集が難しくなる一方, 分析単位を広くとると, 本来の目的である広告配信の最適化に使えない精度の低いものとなってしまうため, そのバランスをとった.

上記の通り生成された 27 万地点の区画は, 協力者 15 名によって任意に選ばれた上でラベルづけされた. それぞれの協力者にはこれまで住んだことがある土地や勤務地など「土地勘」のある場所に限定してラベルづけをお願いすることで, ラベルに「土地勘」が反映されるようにした.「土地勘」には, 土地の用途や地価といった静的, 客観的な情報だけではなく, その場所に対する主観的な見方, その場所に集う人の属性など動的かつ主観的な情報をも加味されて生成されていると考えられる.

ラベルづけされた土地は,15,292 地点となった. その分布は, 東京都, 神奈川県, 大阪府, 愛知県, 奈良県, 沖縄県となっている.

3.2 モデル

こうして得られた「土地勘」を機械的に生成するため, 人がつけたラベルを公開されている土地情報で学習をさせ予測を行う. 人が土地を評価する際に明示的, 暗黙的に用いる情報としては, 主に住居, 商業地, 歓楽街といったその土地の用途, どのような人がその土地を利用するかに分けられると考えられる. 用途を学習させるためには, 存在する店舗や施設など建造物の情報, さらに駅からの距離などのアクセシビリティが必要となる. また, どのような人が利用しているかについては, その地域に住んでいる人の属性や地価などが有用であると考えられる.

これらを表す属性として, 本稿ではウェブ上で取得できる店舗のデータや, 公的機関が公表する統計情報を用いることとした. 単純な分類問題となるため, モデルは, ロジスティック回帰, サポートベクターマシン, 勾配木ブースティングを用いた. 勾配木ブースティングについては, 広く使われている Xgboost[*2] の Python ライブラリを用いた. モデルの推定にあたっては, マルチクラス分類ではなく, 土地がある特定のクラスに属しているかどうかを判定する 2 値分類問題として推定した. これは, 実際に広告配信を行う場合には, 全ての土地を分類するというより, 特定の属性をもつエリアを特定することがタスクとなるためである.

4. データ

学習に用いるデータセットは,3. に述べた協力者によるラベルデータ, ウェブ上で取得できる有力チェーンの店舗一覧などの商業施設のデータ, 国土交通省が公表している国土数値情報 [*3],2017 年国勢調査とした.

店舗データはチェーン名, 店舗名称と, 緯度経度からなり, 各チェーンに対してコンビニエンスストアや薬局, 服飾店といったカテゴリ名を付与した. こうして得た店舗データは, 総数で 50 万, カテゴリ数は 70 となった.

国土数値情報からは, 地価公示, 施設, 地域資源・観光の項目を用いた. 地価公示については, 駅からの距離, 建ぺい率, 地積, 容積率, 地上地下の階数及び地価を用いた. 施設については, 国・都道府県の機関, 市町村役場等及び公的集会施設, 福祉施設, 学校などの所在地データから土地単位でその数を特徴量とした. 地域資源・観光については, 観光資源の有無を特徴量とした. 地価については必ずしも全ての分類対象地点について地価が存在するわけではないので, 近傍の 5 地点の公示地価を平均することで推定した.

国勢調査については, 地域の就業状況を小地域集計の就業状態等基本集計に関する集計から, 大分類別の産業別男女別就業者数を割合に変換して用いるとともに,5 歳刻み年齢別人口を用いた. さらに, 都道府県単位の産業別年収を賃金構造基本統計調査から, 小地域別に産業別の労働者数を国勢調査から取得し, 加重平均することで小地域ごとに平均年収を推定し特徴量として用いた.

5. 結果

本稿では, 分析結果の一部として,「飲屋街」ラベルを推定した結果を示す.

5.1 節では各推定手法の性能比較を行う. 検証に当たっては学習に用いなかった 10 %のデータを用いた.5.2 節では特徴量を評価する.5.3 節では推定されたモデルを用いてラベルのついていない地点を含めた全国 27 万地点の予測を行った. 最後に 5.4 でモデルの考察をした.

5.1 推定モデルの比較

表 1 では, 各分析手法で「飲屋街」かどうかを推定させた結果を示している.$True$ は正例,$False$ は負例に対する予測を示している. 評価指標は, 精度, 再現率及びそれらを統合した f1 スコアとなっている.

実際の広告配信の際には, 飲屋街に滞在するユーザを特定することが目的になるので, 正例に注目して性能を比較すると, 勾配木ブースティングが他の 2 つのモデルを凌駕している. 特に精度については,87%と十分なものとなった. 他方で, 再現率については,53%とやや低くなっている. これは, 実際に飲屋街だと土地勘のあるものが認識していてもモデルがそのように認識しなかったものが半分程度あることを意味している. 与えた特徴量だけでは, 十分な判定ができていない可能性がある.

5.2 特徴量の重要度

図 1 は特徴量を重要度順に上位 10 まで並べたものである. 重要な特徴量とされたのは駅からの距離, 地積, 女性就業者の

∗2 [Chen 2016]

∗3 国土数値情報 http://nlftp.mlit.go.jp/ksj/

表 1: 分析手法の性能比較

methods	True/False	precision	recall	f1-score
ロジスティック回帰	True	0.50	0.08	0.14
	False	0.86	0.99	0.92
サポートベクターマシン	True	0.73	0.15	0.25
	False	0.87	0.99	0.93
勾配木ブースティング	True	0.87	0.53	0.66
	False	0.92	0.99	0.95

うちの宿泊業・飲食サービス業の割合, 推定所得, 建物の階数, 地価, 男性就業者のうちの金融業・保険業の割合, スポーツ系店舗の数, 女性就業者のうちの製造業の割合及び女性就業者のうちの不動産業・物品賃貸業の割合となった.

繁華街はアクセスのより駅近に密集しており, 駅周辺であることが重要な予測指標となっている. また, 地積は公示地価を調査する際の土地の面積であり, 土地が細分化された地域の方が飲屋街である確率が高いことを表している. 建物の階数については, 高いことが飲屋街率を高めている. 女性の産業比率については, 製造業が少なく不動産業が多いほど飲屋街確率が高まる. 工場などが立地する地域より, 不動産の営業所が並ぶ街中に飲屋街が発生しやすいということを表していると考えられる. ただし, 男性ではなく女性の職業比率がそのまちの属性をよく表していることについては, さらに研究が必要であると考えられる.

5.3 「土地勘」モデルによる予測

学習したモデルを用いて全国 27 万地点の土地を飲屋街かそうでないかについて予測したところ,1342 地点が飲屋街であると判定された. 図 2 は主要都市における推定された飲屋街の分布を示している.

東京では新宿の歌舞伎町や東口周辺, 渋谷のセンター街付近, 新橋, 銀座, 上野の一部など, 主な繁華街は網羅されているほか, 駅周辺が飲屋街として推定されている. 特に, 一般的な繁華街ではない円山町, アメヤ横丁や神楽坂など飲屋街としての色彩が濃いまちも漏れなく判定されている.[*4] これらのまちはチェーン店が密集しているというより, 個人営業の店も多くチェーン店の店舗情報だけでは判定しきれない. 本研究では国土数値情報や国勢調査など, より一般的な土地の属性を捉える工夫を行なっているため, こういったエリアについても特定することができている.

その他の都市を見てみると大阪では難波と梅田を中心に比較的広範囲に飲屋街が広がっていると予測されている. また, 名古屋は名駅周辺及び栄に判定がされている. 横浜については, 横浜駅周辺と, 関内が判定されているものの他の大都市と比較して狭い分布となっている.

5.4 考察

協力者のラベルづけの対象が東京を中心に大都市に偏っていることや, 協力者が「飲屋街」であると判断した地点が 798 にとどまったことなどを考慮すれば, 東京だけでなくその他の都市についても違和感の少ない予測ができていると考えられる. 特にチェーン店が密集している訳ではないむローカルな飲屋街についてもその他の情報から判定されることが示せた.

広告配信の観点からは, ウェブ上の情報だけでは直接知り得ないその土地の性格を推定できることは, 配信対象の拡大に繋がり, 意義が大きい. 従来であれば, 主要な居酒屋チェーンの店舗の位置を収集した上で, ユーザーの位置情報とをマッチングさせることでユーザーの属性を判定する必要があったが, ある程度面的に広がりを持つ土地を飲屋街として判定できれば, 大まかなユーザーの位置情報があれば, ユーザーの属性を捉えることができる.

他方で, 東京以外については飲屋街としての判定がやや辛すぎることが課題であると考えられる. 駅からの距離や地価, 推定所得が強く効いていることから, 鉄道路線が張り巡らされ, 地価や所得水準が群を抜いて高い東京において判定が緩くなっていることが考えられる. こうしたバイアスを除去するためには, 特徴量の正規化が必要になってくると考えられる.

6. おわりに

本研究では, ウェブ上で取得できる店舗の情報, 国土数値情報, 国勢調査のデータを用いることで, そこに住んだことがある人しか得られない「土地勘」を学習させ未知の土地についてその属性を予測を行なった. その結果, 東京に関しては, 居酒屋チェーンなどの店舗情報に頼らずとも飲屋街を判定することができた. 他方で, 学習に使ったデータに地理的な偏りがあることや, 特徴量の正規化がされていないことで東京以外における判定が辛くなっていることが課題であると考えられる.

参考文献

[Book 2015] Book, Theodore and Dan S. Wallach:An Empirical Study of Mobile Ad Targeting, 2015, arXiv:1502.06577 http://arxiv.org/abs/1502.06577

[Chen 2016] Chen, Tianqi and Carlos Guestrin.:XGBoost: A Scalable Tree Boosting System (2016), CoRR abs/1603.02754 (2016). arXiv:1603.02754 http://arxiv.org/abs/1603.02754

[Kurashima 2013] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya and Ko Fujimura.:Geo Topic Model: Joint Modeling of User ' s Activity Area and Interests for Location Recommendation (2013), WSDM 2013.

[Noulas 2011] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo and Massimiliano Pontil.:Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-Based Social Networks (2011), AAAI 2011.

*4 いずれも正解データは存在しない地点.

図 1: 特徴量の重要度

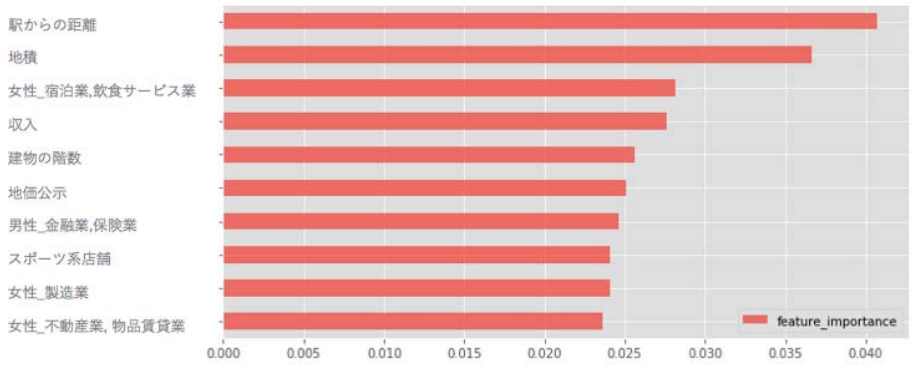

図 2: 各地域において飲み屋街ラベルの付いた地点