Positive-Unlabeled Learningを用いた位置情報とチェックインログに基づく滞在店舗推定

白井　僚† 今井　遼†† リュウセンペイ†† 高橋　翼†† 天方　大地††† 原　隆浩†††

† 大阪大学工学部　〒 565–0871　大阪府吹田市山田丘 2-1
††† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
†† LINE 株式会社　〒 160–0004 東京都新宿区四谷一丁目 6 番 1 号 四谷タワー 23 階
E-mail: †{shirai.r,amagata.daichi,hara}@ist.osaka-u.ac.jp,
††{imai.ryo,sengpei.liew,tsubasa.takahashi}@linecorp.com

あらまし　本論文は，GPS からユーザが実際に滞在している店舗を予測する滞在店舗推定の問題を扱う．GPS は測定誤差を持ち，誤差範囲内に通常多数の店舗が存在するため，ユーザが実際に滞在した店舗を正確に特定することは困難である．本研究ではまず，GPS をチェックインログと結びつけ，滞在時の特徴を学習することによって予測を行うモデルを作成する．このように，ユーザの位置情報を予測する問題では，チェックインログのような正例データのみが用いられる場合が多い．しかし，正例データのみでの予測では Precision の評価を行えないため，偽陽性率を評価できない．そのため，作成した予測モデルに加え，GPS の周辺店舗を Unlabeled データとして利用することにより，Precision と Recall のバランスが取れた滞在店舗推定モデルを構築する．また，本問題において Precision を予測するための適切な指標を導入する．実世界のデータを用いた実験により，導入した指標と Recall において提案モデルが高い値となることを示す．

キーワード　滞在店舗推定，PU Learning

1 はじめに

1.1 動　機

近年，GPS 機能を備えたスマートデバイスの普及により，多くのユーザの位置情報を取得可能となった．これに伴い，取得した位置情報を解析して新たな知見を得る，位置情報解析に関する研究が盛んに行われている [1, 16–18]．特に，リアルタイムに取得したデータからユーザの行動を推測することは，地図やナビゲーション，ローカル検索や広告配信等の多くの位置情報サービスにとって有用である [1]．また，ユーザの行動を理解するために，ユーザが現在滞在している店舗を推定する研究 [16, 18] や，将来滞在するであろう店舗を予測する研究 [3, 10, 17] が行われている．これらの手法はいずれも，店舗へのチェックインログといった，ユーザが店舗に滞在したデータのみに基づいて行われる．すなわち，これらの予測では正例データのみを用いており，負例を用いていない．通常，負例データを考慮しない予測では Precision の評価を行えない．例えば，ユーザの滞在履歴を基に広告配信を行うアプリケーションでは，ユーザに対して誤った広告配信を行う可能性（偽陽性率）を評価できない．

このような背景から，位置情報解析の分野において滞在した情報のみで評価可能な Recall に加えて，正例データのみでは評価できない Precision を考慮する手法が必要である．本研究では，ユーザの滞在店舗を推定する問題において，Precision を適切に考慮・評価可能なフレームワークを考える．

1.2 課　題

滞在店舗推定の問題では，GPS データからユーザの滞在店舗を特定することを目指す．GPS は図 1 に示すように測定誤差を持ち，ユーザが滞在する店舗は，誤差範囲内のいずれかである．これらの候補から滞在店舗を推定する素朴な方法は，GPS が示す緯度経度から距離の近い店舗を予測結果とする手法である．しかし，GPS には誤差があるため，ユーザが訪問した真の店舗が GPS の位置座標に近いという保証はない．

本研究ではベースラインとして，GPS とチェックインログが紐づいたものを学習データとして使用する推定モデルを用いる．チェックインログによりユーザが店舗に滞在した情報（正例）は取得できるが，滞在していない情報（負例）は取得が困難である．そのため, ベースラインの Precision を評価する手段がない．また，偽陽性率を減少させるために，このように負例がない状況で Precision を考慮した予測モデルを構築する必要がある．

1.3 貢　献

提案手法では，ベースラインに加えて GPS の誤差範囲内の店舗を，ユーザが滞在しているかどうか不明な Unlabeled データとして追加し，Positive-Unlabeled (PU) Learning [2] を導

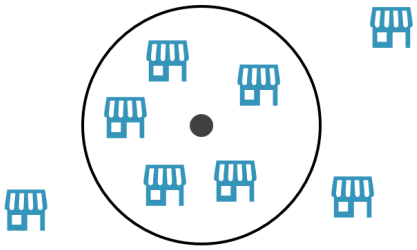

図 1: ある地点で取得された GPS の例．図中の円は GPS の誤差範囲，中心点は取得された GPS の位置座標を表す．この範囲内に存在する店舗が，ユーザが滞在した店舗の候補となる．

入することで，Precision を考慮した予測を行う．また，本手法が Precision において優位であることを確かめるために，滞在店舗推定において適切な指標を提案する．我々の手法は Recall と提案した指標において，ベースライン手法よりも高精度であることを実験的に示す．本研究の主な貢献は以下である．

- 滞在店舗推定において Precision を考慮した予測手法を提案する．
- 予測結果の Precision を評価するための適切な評価指標を導入する．
- 実世界のデータセットを用いて提案手法が Recall と Precision において高精度であることを示す．

本稿の構成は以下の通りである．まず，2 章で本研究で取り扱う問題を定義する．3 章では，滞在店舗推定において Precision を考慮するために必要な関連研究について述べる．4 章では，実際に Precision を考慮した予測・評価方法について述べ，5 章で実験結果を示す．6 章で，これらをまとめ，考察を行う．

2 予 備 知 識

2.1 問 題 定 義

ある時点で取得された GPS は，位置座標 $\mathbf{l}$，時刻 t，および誤差範囲 h から構成されるとする．これを $g_t = (\mathbf{l}, t, h)$ と表す．本研究では，各店舗についてのコンテキスト情報を事前に取得可能であるとする．すなわち，店舗の集合 $V = \{v_1, v_2, \cdots, v_m\}$ の内，任意の店舗 v_j に対して，その位置座標 $\mathbf{l}$ および店舗カテゴリ d は事前に取得可能 $(v_j = (\mathbf{l}_j, d_j))$ である．このことから，GPS が得られたとき，その GPS の誤差範囲内に存在する店舗集合 Ω_t を計算可能であり，これを $\Omega_t = \{v_a, v_b, \cdots, v_p\}$ とする．以上を用いて，滞在店舗推定に関する定義を行う．

定義 1 (滞在店舗推定)．Ω_t の内，ユーザが実際に滞在した店舗の集合を T とする．滞在店舗推定の問題は g_t が与えられたとき，Ω_t 内でユーザが t の前後 τ 分以内に滞在した店舗の集合 T^* を推定することである．このとき，T^* は次のように表される．

$$T^* = \{v \mid p(y = 1|v) > \theta, v \in \Omega_t\}$$

ここで，y は滞在しているかを示す 2 値変数であり，この値が 1 の時に滞在していることを表す．また，θ は閾値であり，滞在確率がこの値を超える店舗を T^* に含める．

本研究では店舗のコンテキスト情報に加えて，各店舗のチェックインログの集合 R を使用可能であるとする．R の各要素 r_j は，店舗の位置座標 $\mathbf{l}$，チェックイン時間 t を持ち，これを $r_j = (\mathbf{l}, t)$ と表記する．

2.2 ベースライン

GPS は位置座標，時刻，および誤差範囲で構成されるデータであるため，滞在点抽出 [14] 等の意味的な情報の抽出を行わない場合，ユーザの店舗滞在に関する特徴を学習できない．本研究では，滞在情報としてチェックインログ R が使用可能であるため，R の要素と結びついた GPS の特徴を学習し，滞在店舗を予測するモデルを作成する．この予測モデルは，入力として GPS とチェックインログの特徴を選択し，各店舗の滞在確率を出力する．

このベースライン手法は，正例データのみの学習であるため，Precision を考慮していない．我々の目的は，滞在店舗推定の問題において Precision と Recall のバランスが取れた推定モデルを構築し，適切に評価することである．

2.3 Positive-Unlabeled (PU) Learning

本節では，我々の研究で必要となる PU Learning を紹介する．正例データのみが明示的に取得可能な状況において，正例か負例か不明なラベル無し (Unlabeled) データを分類する PU Learning は，2 値分類の変種として Liu らによって導入されて以降多くの研究がなされてきた [2]．しかし，位置情報における解析手法については，あまり研究が進んでいない．チェックインデータと結びついていない GPS データは Unlabeled データと見なせるため，本研究では PU Learning を用いて本問題を解く．

2.3.1 SCAR 仮定

データから実際に取得される正例データをラベル付き (Labeled) データとする．このとき，Labeled データは全正例データからランダムに選択されたと仮定する．これを Selected Completely At Random (SCAR) 仮定と呼び，次のように表す．

$$p(s = 1|x, y = 1) = p(s = 1|y = 1) = c$$

SCAR 仮定の下では，正例データがラベル付きデータとして選択される確率は定数 c となる．ここで，x は特徴量，$y \in \{0, 1\}$ は目的変数を表す．また，$s = 1$ のときデータはラベル付きであることを表し，$s = 0$ のときはラベルがついていない，すなわち Unlabeled データであることを示す．

2.3.2 PUF Score

正例と Unlabeled データしか得られない状況下では，予測結果を Precision といった評価指標で評価できない．そこで，Lee らは F 値が Recall と Precision が共に高いときに，値が大き

くなる指標であることに基づいて，次に示す PUF Score を提案した [9].

$$\begin{aligned}\frac{Precision \cdot Recall}{p(y=1)} &= \frac{Precision \cdot Recall^2}{Recall \cdot p(y=1)} \\ &= \frac{p(y=1|\hat{y}=1) \cdot Recall^2}{p(\hat{y}=1|y=1)} \\ &= \frac{Recall^2}{p(\hat{y}=1)}\end{aligned}$$

このとき，SCAR 仮定の下で $Recall = p(\hat{y}=1|s=1)$ と表せるので，正例と Unlabeled データのみで PUF Score は評価可能である．ここで，$p(\hat{y}=1)$ はデータを正例と予測した割合を示す．また，PUF Score が高いほど Precision も高いことが期待できる．

3 関 連 研 究

本研究は我々の知る限り，位置情報解析において Precision を考慮した初めての研究である．以下では，特に関連深いテーマである，滞在点抽出，GPS の誤差を考慮した位置情報解析，および PU Learning について述べる．

3.1 滞在点抽出

GPS の軌跡など連続的に得られる位置座標の内，ユーザが特定時間滞在した地点は重要であると考えられる．そのため，これらの地点を抽出するための滞在点抽出手法が数多く提案されている [20]．滞在点を抽出するための素朴な手法は，ユーザから得られる連続的な GPS に対して，時間的および距離的な閾値を設け，この値を超えない GPS の集合を滞在点とする方法である．Yuan らはこれに加え，地点の密度を考慮して滞在点の抽出を行った [19]．Cao らは，地点-地点間，ユーザ-ユーザ間，および地点-ユーザ間の関係性を考えることにより，各地点の重要度を考慮したグラフベースの滞在点抽出のアプローチを提案した [4].

このような滞在点抽出手法の発展に伴い，抽出した滞在点に意味づけを行う手法に関しても，多くの研究が行われてきた．滞在点はある店舗周辺など，特定の意味を持つ地点に複数存在する場合が多い．そのため，まずこの滞在点を K-means 等のクラスタリング手法によって滞在クラスタとし，このクラスタ単位で意味的情報を付与する方法がよく用いられる [22]．Kang らは取得した滞在点に対して、時間ベースのクラスタリングを行うことで，ユーザにとって重要な場所を推定するアルゴリズムを提案した [7]．また，Ye らはより高い精度で滞在クラスタを推定するために密度ベースのクラスタリング手法を提案した [21].

これらの手法は GPS からユーザの行動を理解するために有用な方法であるが，いずれも GPS の誤差は考慮していない．そのため，GPS の誤差範囲内の店舗滞在について考えるためには別のアプローチが必要となる．

3.2 GPS の誤差を考慮した位置情報解析

GPS から取得可能な位置座標は誤差を持つことを前提とした位置情報解析に関する研究はいくつか行われてきた．例えば，Wu らはユーザの滞在店舗を推定するために，マルコフ確率場を用いたモデルを導入することで、GPS の誤差を考慮した店舗推定を行った [16]．また，Yi らは GPS の誤差範囲からユーザが滞在した店舗のカテゴリを推定するために，事前に取得した GPS からユーザ，時間，および店舗カテゴリを軸とする 3 次元テンソルを作成した．そして，このテンソルを分解することでユーザ間の類似性を捉えた潜在因子を抽出し，特定時間におけるユーザの滞在店舗カテゴリを推定した [18]．しかし，この手法は店舗カテゴリ毎の滞在情報を基に推定を行うため，店舗単位の予測は想定していない．また，これらの手法は正例データのみで評価可能な Recall において高精度を目指すものであり，Precision は考慮していない.

3.3 PU Learning

正例と Unlabeled データを基に学習を行うために，Fung らは Unlabeled データの中で正例データと非常に離れたデータを，信頼度の高い負例データとみなし，これらの正例と負例を元に分類器を学習する 2 段階の分類手法を提案した [5]．また，Unlabeled データをノイズを持つ陰性サンプルとみなし，誤分類された正例に高いペナルティを課す手法 [11] や PU データに適した評価指標に基づいて，ハイパーパラメータを調整する手法 [15] も存在する.

このような正例と Unlabeled データのみが取得可能な状況は，実世界において数多く存在する．そのため，それぞれのドメインに対して適切な PU Learning 手法が必要となる．例えば，Yang らは疾患遺伝子を同定する問題において，PU Learning に加えて既知の疾患遺伝子に関する情報を共有し、新しい疾患遺伝子をゲノム規模で検索することを可能とする ProDiGe を提案した [13]．また，Zhou らは推薦システムの問題において，正例と Unlabeled データを分類する分類器とユーザ-アイテム間の関係を同時に学習する PURE を提案した [23]．他にも，文書分類や画像分類，物体検出等の問題において，同様の研究が行われている [6].

4 提 案 手 法

チェックインログは，正例データの取得が容易であるのに対して，負例データは取得が困難である．そのため，ベースライン手法では正例データのみでの予測を行った．提案手法では，Precision を考慮した予測を行うため，GPS の位置座標周辺の店舗を Unlabeled データとして追加した予測モデルを作成する．また，滞在店舗推定の問題における適切な評価指標を導入する．以下では，それぞれについて説明を行う.

4.1 Unlabeled データの追加

Ω_t からユーザが実際に滞在した店舗集合 T^* を推定する際，Precision を考慮するために，本手法では負例データを追加することを考える．負例データとは，ユーザがある店舗に滞在していないことを表すデータである.

負例データを直接作成する方法として，滞在以外を滞在して

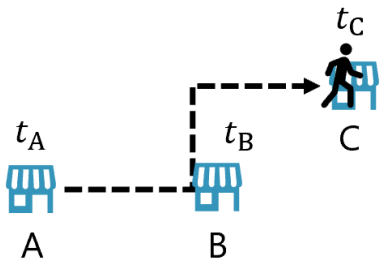

図 2: 一つの GPS が複数の店舗滞在と結びつく例．店舗 A〜C のチェックイン時刻および GPS の時刻をそれぞれ t_A，t_B，t_C，および t_g とする．GPS とチェックインログは時間的な幅 ρ を持って結びつくため，任意の $i \in \{A, B, C\}$ に対して，$|t_g - t_i| < \rho$ のとき，GPS は店舗 A〜C の滞在と結びつく．

いないと考え，負例データとして追加することが考えられる．しかし，GPS とチェックインログは取得タイミングを同期しているわけではないため，厳密には結びつかず，ある程度の時間的な誤差を持って結合している．このとき，ある GPS が複数の店舗滞在と結びつくことがあるため，ある店舗に滞在した場合に他の店舗に滞在していないとは言い切れない（図 2）．このことから，GPS の誤差範囲内に存在する店舗の内，正例以外の店舗は正例か負例か不明な Unlabeled データとし，学習データに追加する．

4.2 特徴ベクトルの作成

Ω_t 内の各店舗に対して，滞在を判定するための特徴ベクトルを作成する．本手法では，以下に示す時間帯情報，位置情報，店舗自身の情報，および周辺店舗の情報を結合した特徴ベクトルを予測モデルの入力とする（図 3）．

4.2.1 時間帯情報

飲食店が昼時に来店数が増加するように，ユーザの店舗滞在は時間帯による影響を受ける．そのため，GPS から取得した時刻の内，時間 h を取り出し，$\cos(2\pi h/24)$ および $\sin(2\pi h/24)$ と変形し，周期性を考慮した特徴量を作成する．

4.2.2 位置情報

GPS は誤差を持つが，距離の近い店舗に滞在している可能性が高いと考えられる．そのため，GPS 位置座標 $\mathbf{l}_i$ と店舗の位置座標 $\mathbf{l}_j$ 間のユークリッド距離 $||\mathbf{l}_i - \mathbf{l}_j||$ を計算し，これを特徴量とする．

4.2.3 店舗自身の情報

ある店舗について滞在かを判定する際に，店舗自身の情報を予測モデルに入力する必要がある．店舗集合 V の各要素は $v_j = (l_j, d_j)$ と表されるので，店舗カテゴリ d_j を店舗 v_j を表す特徴量とし，これを One-hot ベクトルで表現する．また，各店舗を表す情報として店舗の人気度に着目する．チェックインログの多い店舗は，人気な店舗であるといえる．そのため，各店舗のチェックインログ数を，それぞれの店舗の人気度として

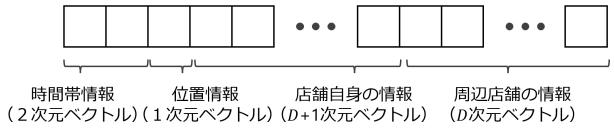

図 3: 予測モデルの入力の例．4.2 節で作成したそれぞれの特徴ベクトルを結合し，$(4 + 2D)$ 次元の特徴量を作成する．また，D は店舗カテゴリの総数を表す．

特徴量に追加する．

4.2.4 周辺店舗の情報

ある店舗 v_j に対して，Ω_t 内の v_j 以外の店舗を周辺店舗とする．このとき，周辺店舗は v_j の滞在確率に影響を及ぼす．例えば，周辺に v_j と同じ店舗カテゴリが存在する場合，v_j の滞在確率は低くなると考えられる．また，周辺店舗の種類や数によっても滞在確率は変化する．このことから，次の手順で特徴量を作成する．

(i) $\mathbf{x} \in \mathbb{N}^D$ の各要素 x_i を，Ω_t 内の店舗カテゴリ d_i の総数とするベクトルを作成する．このとき，v_j は要素数に含めない．(D：店舗カテゴリの総数)

(ii) $\mathbf{x}$ の各要素を GPS の誤差範囲の面積で正規化する．

4.3 予測モデル

滞在店舗の集合 T^* を推定するために，Ω_t 内の各店舗に対して滞在しているかどうかを予測する 2 値分類を行う．追加した Unlabeled データは GPS の誤差範囲内の店舗であるため，作成した特徴量の内，店舗自身が持つ情報および距離情報以外の特徴量は共通である．線形モデルでは，共通の特徴量が多い場合に似た結果を示す．そのため，共通の特徴量が多い場合にも，適切に予測可能な決定木を基にして予測モデルを構築する．

PU Learning では，正例と負例の分類を行う通常の 2 値分類器を用いて，正例と Unlabeled データを分離可能なモデルが提案されている．このうち，本研究では Bagging PU Classifier [12] を，決定木をベースとして使用する．

4.4 Category-aware PUF Score

正例データと Unlabeled データのみが取得可能な状況下において，PU Learning では F 値を予測するために，PUF Score を用いる．PUF Score は SCAR 仮定に基づいて得られる指標である．しかし，チェックインログは通常，店舗カテゴリ毎に取得のされやすさが異なるため，SCAR 仮定を満たさない．例えば，購買履歴から生成されたチェックインログでは，飲食店など，ユーザが購買行動を起こしやすい店舗カテゴリは滞在データとして取得されやすい．そのため，チェックインログを利用した評価に PUF Score は使用できない．

このことから，チェックインログを用いた場合に適切にモデルを評価する指標が必要となる．我々はチェックインログが，ある単一の店舗カテゴリに注目したときに SCAR 仮定を

表 1: データセットの統計

正例データ	Unlabeled データ
329,378	12,286,106

満たす，すなわちユーザのある店舗カテゴリへの滞在情報の内，チェックインログがランダムに取得されることに着目して，店舗カテゴリ毎の PUF Score の重み付き和である以下の指標（Category-aware PUF Score）を導入する．

$$\frac{1}{N}\sum_{d}\frac{Recall_d^2}{p(\hat{y}=1|d)}N_d$$

ここで，d は店舗カテゴリ，N_d は店舗カテゴリ d に属する店舗の総数を表す．また，N は全データ数，$Recall_d$ はカテゴリ d における Recall であり，$Recall_d = p(\hat{y}=1|s=1,d)$ で与えられる．

5 評価実験

5.1 データセット

提案手法の精度を確かめるために，実世界のデータセットを用いて実験を行った．本研究では 2022 年 8 月 1 日〜2022 年 8 月 31 日の期間に日本で取得された店舗情報，GPS，およびチェックインログを使用した．表 1 に本データセットの統計情報を示す．表中の正例データは，GPS とチェックインログが結びついたデータを表す．これらは前後 3 分以内の時間的な幅を持って結合される．そのため，評価実験では $\tau = 3$ として滞在店舗推定の問題を考える．また，各正例データに対して Ω_t を事前に計算し，Unlabeled データとして取得する．店舗の人気度の計算には，2022 年 5 月 1 日〜2022 年 7 月 31 日のチェックインログを使用した．

5.2 比較手法

比較手法として次に示す 4 つの手法を考える．

- ランダム選択．GPS 誤差範囲内の店舗からランダムに滞在店舗を選択する．
- 近傍選択．GPS の位置座標から距離が最も近い店舗を滞在店舗とする．
- ベースライン．2.2 節に示すように，GPS とチェックインログの特徴を選択し，各店舗の滞在確率を出力する．このとき特徴量として，時間および周辺店舗の情報，出力として店舗自身の情報を選択することで，多クラス分類の問題として各クラスの滞在確率を出力する．
- PN Learning．本来 Unlabeled データとして扱う必要のある周辺店舗のデータを負例データとして扱い，この正例と負例を用いて提案手法と同様に 2 値分類を行う．

また，提案手法における通常の 2 値分類器および比較手法の予測モデルは，すべて決定木モデルである LightGBM [8] を使用した．

5.3 評価方法

2 値分類では，予測値 $\hat{y}$ を次のように決定する．

$$\hat{y} = \begin{cases} 1 & (\ p(y=1|v) > \phi\ ) \\ 0 & (\ p(y=1|v) \leqq \phi\ ) \end{cases}$$

このとき，決定境界 ϕ は人為的に決定される．本実験では，決定境界の決め方に依存しない評価を行うために，ϕ によって値を決定するのではなく，GPS の誤差範囲内の店舗に対して，出力値の上位 k 件を 1，その他を 0 とした．また，ある GPS に対してユーザが実際に滞在した店舗は，Ω_t の店舗の内高々5 店舗であると考えられるので，k を 1 から 5 で変化させた．

このように決定した出力値を，正例データのみで評価可能な Recall に加えて，Category-aware PUF Score を用いて評価した．Recall はミクロ平均および店舗カテゴリ毎のマクロ平均を用いた．これらはそれぞれ，正例データを用いて $p(\hat{y}=1|s=1)$，$\frac{1}{D}\sum_d = p(\hat{y}=1|s=1,d)$ で与えられる．

(a) Micro Recall

k	1	2	3	4	5
ランダム選択	0.1772	0.1772	0.1772	0.1772	0.1772
近傍選択	0.3592	0.5135	0.5992	0.6548	0.6954
ベースライン	0.3123	0.4230	0.4994	0.5541	0.5967
PN Learning	0.7385	0.8439	**0.8836**	**0.9051**	0.9192
提案手法	**0.7397**	**0.8443**	0.8834	**0.9051**	**0.9193**

(b) Macro Recall

k	1	2	3	4	5
ランダム選択	0.1661	0.2652	0.3336	0.3916	0.4335
近傍選択	0.3258	0.4381	0.5037	0.5595	0.5965
ベースライン	0.2763	0.3561	0.4184	0.4613	0.4924
PN Learning	0.4350	**0.5780**	0.6408	0.6778	0.7090
提案手法	**0.4368**	0.5777	**0.6411**	**0.6806**	**0.7105**

(c) Category-aware PUF Score

k	1	2	3	4	5
ランダム選択	1.333	1.779	2.006	2.086	2.140
近傍選択	7.375	6.283	5.445	4.9405	4.555
ベースライン	19.827	8.885	6.236	5.327	4.672
PN Learning	**422.125**	26.946	15.785	11.740	9.535
提案手法	401.797	**26.973**	**15.806**	**11.855**	**9.599**

表 2: 滞在店舗数 k を 1 から 5 で変化させたときの（a）Micro Recall,（b）Macro Recall，および（c）Category-aware PUF Score の値．（c）は値が高いほど精度が高いことを表す．

5.4 実験結果

Recall と Category-aware PUF Score を用いて評価した結果を表 2 に示す．

5.4.1 Recall の比較

k の増加に伴って，予測する滞在店舗数が増加するため，予測した店舗の内，実際に滞在した店舗の割合を表す Recall も値が大きくなる．

提案手法はベースラインと比較して，いずれの k においても 40% ほど高精度であった．各店舗に対する滞在確率を予測する際，正例のみでは，2 値分類問題において一方のクラスのみが取得可能な状況となり学習を行えない．そのため，ベースラインでは GPS から取得可能な周辺店舗情報および時間情報を基に滞在時の特徴を学習した．しかし，このモデルでは各店舗が持つ距離情報および店舗の人気度は考慮できないため，Recall において提案手法よりも低い値となる．

また，提案手法は PN Learning と同程度の精度であった．$\tau = 3$ の滞在店舗推定の問題では，ある時間 t の前後 3 分以内のユーザの店舗滞在を考える．この間にユーザが滞在する店舗は多くの場合 1 店舗と考えられ，Ω_t 内の正例以外を負例とする PN Learning の状況に近づく．提案手法は負例を Unlabeled データとして捉えるが，実際にはこれらは多くが負例であるため Recall において PN Learning と同程度になる．一方で，わずかに精度が向上しているのは，ユーザが Ω_t 内の複数店舗に滞在する可能性のある状況を捉えたためであるといえる．

同様に，提案手法はいずれの k においても，ランダム選択および近傍選択よりも高精度であった．近傍選択がランダムに滞在店舗を選択する場合よりも高精度であることから，GPS は誤差を持つが，実際には近傍の店舗に滞在している可能性が高い．

5.4.2 Category-aware PUF Score の比較

k が増加するにつれて Recall は増加するが，多くの手法において Category-aware PUF Score は減少する．これは，指標において分母が増加するためであり，Recall よりも偽陽性率を重視していることを表す．

表 3c は，提案手法がベースラインよりも高精度であり，$k = 1$ の場合を除いて PN Learning よりも高精度となることを示す．このことから，提案手法が正例のみの予測と比較して偽陽性率を考慮しているといえる．また，ランダムおよび近傍選択では偽陽性が多くなるため，この指標において低い値となる．

5.5 特徴量の設計方法による影響

前節で提案手法が，Recall と Category-aware PUF Score において最もバランスの取れたモデルであることを確認した．次に特徴量の設計方法による精度への影響について考察を行う．

4.2 節で設計した特徴ベクトルから時間帯情報，位置情報，店舗自身の情報，および周辺店舗の情報を取り除いて学習する手法をそれぞれ (i)〜(iv) とする．表 3 に，これらを評価した結果を示す．提案手法は (i)〜(iv) のいずれの場合よりも多くの k で高精度であることから，それぞれの特徴量が精度の向上に寄与しているといえる．また，(ii) および (iii) はそれぞれの指標において提案手法との差分が大きいことから，位置情報および店舗自身の情報は精度向上に対する寄与率が高い．

(a) Micro Recall

k	1	2	3	4	5
(i)	0.7386	0.8435	0.8828	0.9046	0.9188
(ii)	0.7042	0.8240	0.8692	0.8934	0.9098
(iii)	0.6297	0.7563	0.8066	0.8344	0.8536
(iv)	0.7290	0.8334	0.8752	0.8988	0.9143
提案手法	**0.7397**	**0.8443**	**0.8834**	**0.9051**	**0.9193**

(b) Macro Recall

k	1	2	3	4	5
(i)	0.4354	**0.5783**	0.6405	**0.6814**	0.7104
(ii)	0.3874	0.5284	0.6066	0.6482	0.6765
(iii)	0.2688	0.3904	0.4489	0.4957	0.5293
(iv)	0.4279	0.5680	0.6249	0.6656	0.6996
提案手法	**0.4368**	0.5777	**0.6411**	0.6806	**0.7105**

(c) Category-aware PUF Score

k	1	2	3	4	5
(i)	400.867	26.955	15.779	11.728	9.525
(ii)	384.0677	25.276	15.5963	11.475	9.361
(iii)	55.644	11.333	8.116	6.591	5.692
(iv)	399.913	**30.601**	**16.235**	11.737	9.265
提案手法	**401.797**	26.973	15.806	**11.855**	**9.599**

表 3: 特徴ベクトルを変化させたときの（a）Micro Recall, (b) Macro Recall，および（c）Category-aware PUF Score の値．

6 結　　論

本論文では，GPS からユーザが実際に滞在した店舗の集合を予測する滞在店舗推定の問題を扱った．この問題を解くために，まず GPS をチェックインログと結びつけ，滞在時の特徴を学習することによって予測を行うベースモデルを作成した．しかし，チェックインログのみでの予測では Precision の評価を行えないため，偽陽性率を評価できないという問題がある．

我々は，GPS の位置座標の周辺店舗を Unlabeled データとして利用することにより，Precision を考慮した滞在店舗推定モデル構築した．また，この問題において予測結果の Precision を評価するための適切な評価指標を導入した．提案モデルは導入した指標と Recall において高い値となることを示した．

文　　献

[1] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, “Recommendations in location-based social networks: a survey,” GeoInformatica, vol.19, no.3, pp.525–565, 2015.

[2] J. Bekker and J. Davis, “Learning from positive and unlabeled data: A survey,” Machine Learning, vol.109, no.4, pp.719–760, 2020.

[3] P.C. Besse, B. Guillouet, J.M. Loubes, and F. Royer, “Destination prediction by trajectory distribution-based model,” IEEE Transactions on Intelligent Transportation Systems,

vol.19, no.8, pp.2470–2481, 2017.

[4] X. Cao, G. Cong, and C.S. Jensen, "Mining significant semantic locations from gps data," Proceedings of the VLDB Endowment, vol.3, no.1-2, pp.1009–1020, 2010.

[5] G.P.C. Fung, J. Yu, H. Lu, and P. Yu, "Text classification without negative examples revisit," IEEE Transactions on Knowledge and Data Engineering, vol.18, no.1, pp.6-20, 2006.

[6] K. Jaskie and A. Spanias, "Positive and unlabeled learning algorithms and applications: A survey," In IISA, pp.1-8, 2019.

[7] J.H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," ACM SIGMOBILE Mobile Computing and Communications Review, vol.9, no.3, pp.58–68, 2005.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol.30, 2017.

[9] W.S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," In ICML, vol.3, pp.448–455, 2003.

[10] X. Li, M. Li, Y.J. Gong, X.L. Zhang, and J. Yin, "T-desp: Destination prediction based on big trajectory data," IEEE Transactions on Intelligent Transportation Systems, vol.17, no.8, pp.2344–2354, 2016.

[11] B. Liu, Y. Dai, X. Li, W.S. Lee, and P.S. Yu, "Building text classifiers using positive and unlabeled examples," In ICDM, pp.179–186, 2003.

[12] F. Mordelet and J.P. Vert, "A bagging svm to learn from positive and unlabeled examples," Pattern Recognition Letters, vol.37, pp.201–209, 2014.

[13] F. Mordelet and J.P. Vert, "Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples," BMC bioinformatics, vol.12, no.1, pp.1–15, 2011.

[14] H. Peng, Y. du, L. Zhang, J. Yi, Y. Kang, and T. Fei, "Uncovering patterns of ties among regions within metropolitan areas using data from mobile phones and online mass media," GeoJournal, vol.84, 06 2019.

[15] S. Sellamanickam, P. Garg, and S.K. Selvaraj, "A pairwise ranking based approach to learning with positive and unlabeled examples," In CIKM, p.663–672, 2011.

[16] F. Wu and Z. Li, "Where did you go: Personalized annotation of mobility records," In CIKM, pp.589–598, 2016.

[17] A.Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," In ICDE, pp.254–265, 2013.

[18] J. Yi, Q. Lei, W.M. Gifford, J. Liu, J. Yan, and B. Zhou, "Fast unsupervised location category inference from highly inaccurate mobility data," In SDM, pp.55–63, 2019.

[19] N.J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," IEEE Transactions on Knowledge and Data Engineering, vol.25, no.10, pp.2390–2403, 2012.

[20] Y. Zheng, "Trajectory data mining: an overview," ACM Transactions on Intelligent Systems and Technology, vol.6, no.3, pp.1–41, 2015.

[21] Y. Zheng, L. Zhang, X. Xie, and W.Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," In MDM, pp.791–800, 2009.

[22] C. Zhou, N. Bhatnagar, S. Shekhar, and L. Terveen, "Mining personally important places from gps tracks," In ICDE, pp.517–526, 2007.

[23] Y. Zhou, J. Xu, J. Wu, Z. Taghavi, E. Korpeoglu, K. Achan, and J. He, "Pure: Positive-unlabeled recommendation with generative adversarial network," In SIGKDD, pp.2409–2419, 2021.