機械学習による瀬田キャンパス青志館の売上予測

知能情報メディア課程 和氣匠汰

1. 概要

本調査は，瀬田キャンパス・青志館にある食堂の売上を予測する時系列モデルの作成を目的としたものである．モデルは未完成であるため，作成にあたり調査を行った食堂の現状，データの状態，気候といった説明変数として扱えるデータセットの入手方法，商品名の質的データへの変換方法，モデルの学習に用いる事のできる類似のデータセットをここに掲載する．

2. 瀬田キャンパス・青志館食堂の現状

本章では対象の食堂の現状及び保管されるデータセットの形式について述べる．

2.1 食堂売り上げデータの形式，状態

食堂の売上データはシステムの変更により 2023 年 4 月以降もののみ保管されている．月毎，週毎，日毎の売上をエクセルの形式で出力することが可能であり，「日付，決済時刻，商品名，単価，カテゴリー」といった項目がある．そのため Python の Pandas ライブラリを用いて「日付，商品名，日ごとの売上個数，カテゴリー」といった具合に時系列解析が可能な形式への整形が容易である．

2.2 仕入れの現状

食堂が採用する会計システムに搭載された分析ツールはあるものの，機械学習を含む予測ツールは用いておらず，職員が曜日や過去の売り上げ個数から仕入れ数を決定しているのが現状である． 既存の予測ツールがない以上，ユーザーにとって扱いやすいソフトウェアである必要があるだろう．

3. 気象情報の取得方法

売上や来場者数の予測には気象情報が広く用いられる．本調査において気象情報を取得する手法は以下の条件を満たす事が理想である．

1. 商用利用可能である．
2. 有料でない．
3. 滋賀県という地域を指定できる．
4. データセットを手動で更新する必要がない．
5. 訓練データの期間内の気象データを提供する．

以上の 4 つ全てを満たす API やサービスは 2023 年 10 月時点では確認できない．よって訓練データには気象庁で公開されるデータを使用し，逐次的なデータの取得を要求される実用時には Weather API を用いるという手法を採用する．Weather API は商用利用可能な API であり，過去 14 日分の気象データにアクセスできる[1]．また気象庁のホームページに公開される気象データは商用利用可能であり[2]，CSV による出力も可能である．

4. 商品のカテゴリ区分

今回は提供されるデータ数が少ないため，商品をカテゴライズして質的データに変換する手法が適切である．食堂で取り扱う商品の名称は期間によって変わるための商品名をそのまま質的データとして扱うのは適切ではない，また商品は会計データの時点で「主菜，副菜，丼・カレー，デザート」といった形で大まかに区分されているが，カテゴリごとに売上予測を行うにあたり，より詳細な区分，例えば主菜ならば「唐揚げ，ハンバーグ，魚」のような区分が行われている事が理想である． 学習データセットの説明変数にこのより詳細なカテゴリーを追加には，以下のようなアプローチが考えられる．

1. 商品名に用いられる語の抽出による区分
2. 分散表現のクラスタリングによる区分
3. 分散表現の次元圧縮による区分

以上三つのうち，2 の「分散表現のクラスタリングによる区分」のみ実験を行った．以下ではそれぞれの手法の詳細と特徴を説明する．結論から述べると，要求するデータセット数やその正確性から， 1 の「商品名に用いられる語の抽出による区分」の問題点を解消した手法が最も本タスクに適していると考察する.問題点とその解消方法については 4.1 節で述べる．

4.1 商品名に用いられる語の抽出による区分

これは商品名から語を探索し，ルールベースでカテゴリを区分する手法である．例えば「唐揚げブラックペッパー」「チキンカツカレー大」からは「唐揚げ」「カレー」といった語を抽出し，カテゴライズする．このような分類を行い，その分類に漏れたものを手動により区分すれば，現在入手可能な売上データ過去 6 ヶ月分のカテゴライズは十分可能である．しかしこの手法には以下の問題がある．

1. カテゴリ間の関連性の表現
   例：「ハンバーグ」「唐揚げ」などのカテゴリには共通の「肉」という区分が存在する．しかし単にカテゴリを説明変数に加えるだけではこの「ハンバーグ」と「唐揚げ」の間にある「肉」という関連性を表現できない．
2. 既存のカテゴリ外の商品への対応
   例：新しく「麻婆丼」という既存のカテゴリ外の商品が観測された時，対応できない．

これらの問題を解決するために，カテゴリを複数用意する手法を採用する．例えば「ハンバーグ」と「唐揚げ」には「肉」という共通のカテゴリが存在する．しかし単に単一のカテゴリにより区分を行えば，これらの関係性を表現する事ができない．よって具体カテゴリ，抽象カテゴリのように複数のカテゴリを設けることで「ハンバーグ，肉」,「唐揚げ，肉」のように商品同士の関連性を表現する事ができる．その上，例えば 4.1.1 にて前述した「麻婆丼」という既存のカテゴリに存在しない商品が観測されたとしても，より抽象的な他のカテゴリに「丼もの」「中華」というカテゴリーを用いることで，ユーザーが新しく商品を追加する際，「麻婆」というカテゴリがなくともその商品の特徴を表す事ができる．

4.2 分散表現のクラスタリングによる区分

4.1 の手法は新規商品のカテゴライズをユーザーが手動で行う一方で，本節の手法はユーザーが新しい商品を追加する際にカテゴライズを自動的に行う事ができる．BERT エンコーダ[3]により分散表現に変換し，それをクラスタリングすることでカテゴリの自動生成を行う．表 1 は実験の結果である．表 1 からわかるように，この手法である程度カテゴリを分類できる．しかしクラスタごとの要素数の偏りやカテゴライズの不完全性が見られる．ハイパーパラメータの調整などにより改善の見込みはあるが，正確性の面で前述した 4.1 の手法に劣る．

4.3 分散表現の次元圧縮

この手法は BERT により出力される 768 次元の埋め込み表現をパラメータ更新可能なレイヤもしくは主成分解析などにより圧縮し，説明変数として組み込むという手法である．埋め込み表現を圧縮して用いる理由は，説明変数の数が膨大になり，商品名以外の説明変数がモデルに与える影響を微弱にしてしまう危険性を回避するためである．以下では次元圧縮の手法として，パラメータ更新可能なレイヤと主成分解析の両方の側面から考察を行う．

4.3.1 パラメータ更新可能なレイヤ

パラメータ更新可能なレイヤとして挙げられるのは，Attention 層及び全結合層である．これらの手法ではタスクに適応した形での特徴量抽出が期待できる．また出力層に Softmax 層を採用することでカテゴリ形式にも対応させられる．しかしパラメータ数の増加により膨大なデータセットを要求するため，5 章にて後述する類似のデータセットによる事前学習が必須である．膨大なデータセットが用意されていれば汎用的かつパフォーマンスの高いアーキテクチャの作成が期待できるが，商品名という部分に大きくリソースを割く必要性については検討の余地がある．

4.3.2 主成分解析

パラメータ更新可能なレイヤとは異なり教師なし学習に分類される主成分解析はデータセットの量に関わらず適応可能であるがこのタスクに順応する形で次元圧縮を行う事ができるか疑問が残るため，今後実験により実証していく必要性がある．

5. 公開されている類似のデータセット

本章では事前学習やデータセットの拡張に用いることのできる，売上データに類似した二つの公開データセットについて述べる．

5.1 Kaggle の Predict Future Sales

Kaggle のコンペディションである Predict Future Sales はロシアの 1C Company により提供された商品の売上データセットである．[4]商品名はロシア語で提供されており，2013 年の一年分のデータが公開されている．「日付，商品名，売上個数，カテゴリー」といったパラメータがあり，商品は食料品ではないが，事前学習データセットとして用いることができる．しかしこのデータセットを商用のモデルの学習に用いることが可能かどうか規約に明示されていないため，問い合わせる必要がある．また，Discussion ページに各コンペディション参加者のアプローチが掲載されているため，参考にすることができる[5]．

5.2 日経 POS 情報

日経 POS 情報とは，日本経済新聞社が独自にスーパーマーケット，コンビニエンスストア，ドラッグストアなどから収集する POS データを公開するサービスである[6]．2019 年~2022 年の食料品の売上が日毎に公開されている．有料であるが，2023 年 10 月現在，龍谷大学のデータベースに公開されており，龍谷大学図書館蔵書検索システムを用いてアクセスが可能である．しかしこちらもデータセットを商用のモデルの学習に用いることが可能かどうか不明のため，別途問い合わせる必要がある．

6. 総括

2 章では，瀬田キャンパス・青志館の現状について，保管されているデータは十分時系列解析を行うことのできる形式として出力可能である事，現在 2023 年 4 月以降のもののみ入手可能である事，また仕入れについても現状機械学習の技術を用いてはいない事を述べた．3 章では，気象データの取得には気象庁のデータセットと Weather API の両方を用いる手法が有効である事を示した．また 4 章では，商品の質的データへの変換については，多くの手法が考えられるものの，語彙抽出によるルールベースの手法，カテゴリの複数作成の手法が有効であるという考察を示した．5 章では，事前学習やデータセットの拡張に有用な Kaggle の Predict Future Sales と日経 POS 情報を紹介した．しかし先述の通り，この双方のデータセットの使用には商用利用の可否を確認する必要がある．

表 1 埋め込み表現のクラスタリングの結果

クラスター	料理リスト
0	豚のスタミナ炒め, 厚揚げ豚生姜炒め, ポーク味噌焼き, 豚肉の生姜焼, 鶏チリソース, …
1	ポークハヤシ, ビーフカレー, キーマカレー, チキンカレー, ポークカレー
2	若鶏ピリ辛焼, やわらかロースのサムジョン, カキフライタルタル, カレイ唐揚げ野菜あんかけ, …
3	肉じゃが, ビーフストロガノフ
4	回鍋肉, 青椒肉絲
5	厚切りイカフライ, ビッグメンチカツ, ハンバーグ, クリームチーズ入りメンチ…
6	白身魚唐揚げ野菜あん, 鶏の唐揚, ポローニャ風カツ, タルタルinソーセージカツ, 鶏のカッシュナッツ炒め, …
7	タンドリーチキン, バーベキューチキン, チキン南蛮, チキンフリカッセ
8	手作りヒレカツ, 手作りロースカツ, 手作りチキンカツ, ランチビュッフェ
9	マーボ豆腐, 酢豚, マーボ茄子, 中華丼, 八宝菜

謝辞

まず，技術的なご指導を賜りました先端理工学部 数理・情報科学課程准教授 樋口教授に対し，深く感謝申し上げます．ご多忙の中，熱心にご指導いただきましたこと，誠にありがとうございました．

次に，調査の前段階において貴重なご意見とアドバイスを提供してくださった先端理工学部教務課の田畑様に感謝申し上げます．また，特に龍谷大学エクステンションセンターの河角様及び関係者の皆様には，心より感謝申し上げます．皆様の協力により，調査の方向性を明確にすることができました．

また，食堂のデータ提供をしていただきました瀬田ショップ SMYLE 店長 若松様並びに青志館食堂店長 佐崎様に感謝の意を表します．皆様のご協力により，実地調査を円滑に進めることができました．

最後に，本調査に携わったすべての方々に対し，心より感謝申し上げます．まだ途中段階ではありますが，本調査は私自身にとって大変貴重な学びとなっております．

[1] WeatherAPI. (2023). Retrieved from https://www.weatherapi.com/
[2] Japan Meteorological Agency. (2023). Commentary. Retrieved from https://www.jma.go.jp/jma/kishou/info/coment.html
[3] Hugging Face. (2023). BERT Base Uncased. Retrieved from https://huggingface.co/bert-base-uncased
[4] Kaggle. (2023). Competitive Data Science: Predict Future Sales - Overview. Retrieved from https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/overview
[5] Kaggle. (2023). Competitive Data Science: Predict Future Sales - Discussion. Retrieved from https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/discussion
[6] Nikkei POS. (2023). Retrieved from https://nkpos.nikkei.co.jp/