

自己教師あり学習を用いた消費者の購買予測方式の提案 Proposal of a Consumer Purchase Prediction Method Based on Self-Supervised Learning

姜 友顕[†] 山崎 憲一[†]
Woohyun Kang Kenichi Yamazaki

1. はじめに

近年、消費者の購買・行動データを活用した予測モデルが提案されているが、その多くは大量かつ高品質のデータを前提として構築されている。十分な量のラベル付きデータを収集することが困難な場合には適用が難しい。特に、企業が保有する購買データの大多数は非公開であり、汎用的なデータセットの不足が課題となっている。

一方、従来の時系列データにおけるデータ拡張 (Augmentation) は、データ全体の整合性を維持する必要があるため、全てのデータに対して一律な変化を加えることが一般的であり、柔軟性に乏しい課題があった。本研究では、購買データを用いた部分的かつ意味的なデータ拡張方式を提案する。これらの拡張データを活用し、自己教師あり学習による高精度な購買予測を目指す。自己教師あり学習は、大量のラベル付きデータを必要とせず、少量のデータでも学習可能なモデルとして注目されている。本研究では、提案するデータ拡張方式により消費者の多様な変動を事前に反映し学習データを増やす。これにより、限られた購買データでの消費者行動の多様性を反映した高精度な予測モデルの構築が期待される。

2. 関連研究

従来の購買予測においては、予測精度がデータセットの質に大きく依存する傾向がある。[1]は購買履歴における次の購入カテゴリを予測する Sequential Event Prediction (SEP) モデルを提案した。購買履歴を系列データとして扱い、Association Rule に基づく信頼度ルールを活用した教師ありランキング学習を構築している。個人別の信頼度行列を導入することで、消費者ごとの購買傾向の違いを考慮し、予測精度の向上を図っている。ただし、十分な購買履歴データを必要することが課題として残る。

少量かつ一般の品質のデータから高精度な予測を実現する手法として、データ拡張の活用が注目されている。特に、時系列データにおけるデータ拡張では時間領域の拡張 (Cropping, Wrapping, Jittering など) や周波数領域での拡張など、多様な手法が紹介されている[2]。これらの多くは、系列全体の連続性を保つ必要があるため、全データに一律の変換を加える手法が主流となっている。

しかし、購買データにおいては、時系列の特性を持ちながらも、ある購買イベントが他の購買に直接的な影響を及ぼさない場合も多い。このため、購買データに対しては部分的かつ意味的な拡張を柔軟に施す手法が必要となる。

さらに、近年では少量のラベル付きデータで高精度な表現学習を可能にする自己教師あり学習 (Self-Supervised Learning) が注目されている。[3]は、時系列データに対して

自己教師あり学習を適用する新たなモデルである TS-TCC (Temporal and Contextual Contrasting) を提案し、その有効性を示している。TS-TCC では、時系列固有の特性に適合するよう弱拡張 (Weak Augmentation) と強拡張 (Strong Augmentation) を分離して適用する手法を導入している。弱拡張ではスケーリングやタイムシフトなどの軽微な変動を、強拡張では Permutation などによる系列順序の大幅な変動を加えている。これにより、多様な擬似ビューを生成しつつ、頑健な特徴抽出を促進している。学習モデルとしては二つの学習課題を統合している。Temporal Contrasting では弱拡張と強拡張のクロスビュー間で未来予測を行わせ、時系列の時間的依存構造を効果的に捉えている。また、Contextual Contrasting では時系列サンプル内の局所セグメント間の文脈的な一貫性を学習する。ここでの文脈は、系列の先頭からある時点までの履歴を集約して得られる要約ベクトルを指す。同一系列から異なる時点までの履歴を取り出し、それぞれの文脈ベクトルを正例として対照学習を行う。これにより、局所的ばらつきや一時的な変動に頑健な特徴表現が得られ、少量のラベルでも高精度な分類が可能になる。

3. データ拡張と購買予測方式の提案

本章では、本研究の提案方式について述べる。3.1 節に本研究におけるデータ拡張の考え方および拡張方式、3.2 節に自己教師あり学習を用いた拡張済みデータによる購買予測方式について示す。

3.1 データ拡張の考え方

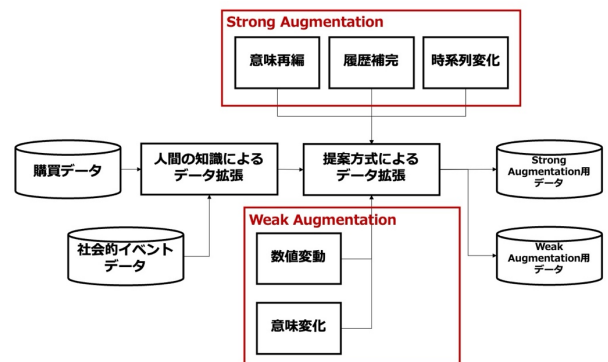


図1 データ拡張の構成

本研究では、購買データに対して部分的かつ意味的なデータ拡張を導入する。図1はデータ拡張の構造図である。大きく人間の知識に基づく拡張とアルゴリズムによる拡張に分類される。前者としては、社会的イベントの影響を考慮する。例えば、ブラックフライデーでは割引によって価格が下がり、通常より高額商品の購買が増加する。このような現象を模倣し購買数量・回数の増加、価格の割引模倣といった拡張を行う。

アルゴリズムによる拡張は、以下の手法を導入する。

[†] 芝浦工業大学大学院 Shibaura Institute of Technology

- **数値変動**：価格や数量に軽微なノイズを付加し、軽度の価格変動や誤記録に対する頑健性を付与する。
- **意味変化**：同一カテゴリ内での価格帯の変更を行い、消費者の価格嗜好の揺らぎを再現する。
- **時系列変化**：購買履歴の順序を入れ替え、嗜好や購買順序の曖昧性を学習させる。
- **意味再編**：履歴内のカテゴリを意味的に近い他のカテゴリへ置換し、代替購買行動の柔軟性を表現する。
- **履歴補充**：購買回数が少ない顧客に対して、意味的に近い購買履歴を追加する。

3.2 購買予測方式

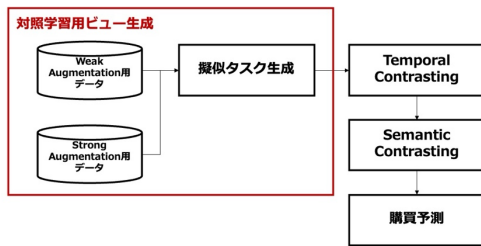


図 2 購買予測の構成

本研究では、前節で構築した拡張済み購買データを用いて、自己教師あり学習に基づく購買予測方式を提案する。図 2 は購買予測の構造図である。本研究では、[3]と同様に拡張済みデータに対して弱拡張と強拡張を適用し、自己教師あり学習用の擬似ビューを生成する。

学習モデルとしては、Semantic Contrastive Module と Temporal Contrastive Module の二つを導入する。Temporal Contrastive Module は、[3]と同様の方法であり、購買履歴に含まれる時間的依存性や周期的変動を学習する。これにより、嗜好の変化が時間に伴って生じる場合にも対応可能な時系列表現の学習を目指す。

Semantic Contrastive Module は本研究の新規提案である。TS-TCC 手法の Contextual Contrastive Module では、時間的変動に着目しているため、購買行動における意味的変動には十分に対応できない。ここでいう意味的変動とは、価格帯の違いに応じて商品を選び直す場合や、用途が近い別のカテゴリ商品に置き換えたりするような、嗜好の柔軟性を指す。そこで本研究では、こうした購買行動の特性を捉えるため、Semantic Contrastive Module を新たに導入する。

本モジュールでは、本提案のデータ拡張方式により生成された履歴系列を用いる。この中から意味的に類似した系列を選別・集約し、文脈ベクトルを構成する。この文脈は、社会的イベントなどの外的要因に影響された購買履歴であっても、その背後にある消費者固有の嗜好パターンを抽出することを目的とする。

意味的類似性では、消費者が示す購買傾向を考慮する。具体的には、「類似した商品を継続して購入する」、「専門的・マニアックな商品を好む」といった傾向に基づき、消費者の行動を分類する。同じ分類の系列を意味的に近い正例とみなす。対照学習では、これらの正例と異なる傾向を示す系列を負例とする。これにより、外的要因に左右されにくい、消費者の本質的な購買傾向を捉える特徴表現の獲得を目指す。

最終的に、学習した特徴を用いて購買カテゴリを予測する。自己教師あり学習済みモデルを初期値とし、少量のラベル付きデータによる微調整（ファインチューニング）を行う。これにより、ラベル情報が限られた環境でも、高精度な購買予測を実現することが期待される。

4. 予備実験

提案方式の実装に先だって、データ拡張が購買予測精度に与える影響を簡易的に評価した。元データと拡張データを用いてそれぞれ学習を行い、予測精度を比較する。拡張には、本研究で提案したアルゴリズムによる拡張に加え、ブラックフライデーの反映を適用した。予測対象は、最終的に購入する商品のカテゴリである。

実験では、一般公開されている架空の購買データセットを使用する。データセットには顧客 ID、商品カテゴリ、注文日、数量、価格などの情報が含まれている。商品カテゴリは 5 種類である。

	元データ	拡張データ
F1-score 平均値	0.1958	0.2508
F1-score 中央値	0.1860	0.2558
F1-score 最高値	0.3023	0.3488

表 1 購買予測の実験結果

表 1 に、実験結果を示す。予測モデルには Random Forest を使用し、F1-score を求めた。データ拡張を適用することで、全体的に予測精度の改善が確認された。一方で、クラスごとのスコアには一部でわずかな変動も見られ、全てのカテゴリにおいて精度向上が得られたわけではない。しかしながら、拡張データの導入により全体的な評価指標のばらつきが減少し、安定した予測性能が確認された。

以上の結果から、本提案のデータ拡張手法は、限られた購買データに対しても予測性能の改善と安定性の向上に有効であることが予想される。

5. おわりに

本稿では、購買データの拡張手法と自己教師あり学習を用いた購買予測方式を提案した。予備実験では、提案したデータ拡張手法を適用した場合に、予測精度がどのように変化するかを検証した。その結果、全体的な予測精度の改善に加え、評価指標のばらつきが抑えられる傾向が見られた。今後は自己教師あり学習との統合による有効性の検証を進める予定である。

参考文献

- [1] Wen, Qingsong, Liang Sun, Xiaomin Song, Jing Gao, Xue Wang and Huan Xu. "Time Series Data Augmentation for Deep Learning: A Survey." International Joint Conference on Artificial Intelligence (2020).
- [2] Rekasiate, L., Palenzuela, A. J. J., Salnaite, N., Cuenca, R. C., & Frasinca, F. "Predicting the category of customers' next product to buy in web shops." Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. (2022).
- [3] Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwok, C. K., Li, X., & Guan, C. "Self-supervised contrastive representation learning for semi-supervised time-series classification." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.12 (2023).