F-022

テンソル学習を用いたトランザクションデータの機械学習手法

Machine Learning Method for Transactional Data Using Tensor Learning

石崎 諒† 新宮 理史† 新田 泉† 等々力 賢‡ 丸橋 弘治† 中島 哲†
Ryo Ishizaki Masafumi Shingu Izumi Nitta Masaru Todoriki Koji Maruhashi Satoshi Nakashima

1. はじめに

業務上で蓄積されたいわゆるトランザクションデータを機械学習で活用し、新たな知識を発見することで業務課題の解決に役立てたいという要求が増えている。例えば、オンライン教育システムにおいて受講者の学習行動履歴を記録・管理するだけでなく、過去事例と照合して成績予測を行うことにより、リモート学習で困難になりがちな教育を支援したいといったものである。

トランザクションデータは図 1 の例に示すように、過去の行動履歴が複数行に渡って記録されたものであるが、これを機械学習で利用するためには何らかの方法で情報を集約し、1 次元配列のベクトルデータ表現に変換する必要がある。しかし、その集約作業には、多くの場合、前以て特徴量設計が必要で、そのためには多くの業務経験や様々な知識を要するため、業務システムでの機械学習利用促進を阻害する一因となっている。

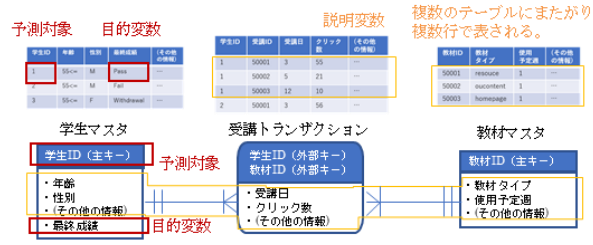

図 1　トランザクションデータの例

本発表では、この前処理問題に対する解決策として、筆者らが開発したグラフ AI 技術 NNDT(Neural Network Deep Tensor)[1]利用する。

2. NNDT

NNDT は、先行するグラフ AI 技術である Deep Tensor[2]を発展させた独自技術である。Deep Tensor はテンソル表現を利用してグラフデータの学習を行い、分類・回帰を実行する教師あり学習器である。グラフデータはノードとノード間のエッジ、及び、ノードやエッジの属性で構成されているが、グラフ情報のノード間の接続関係をテンソルで表現するアプローチは、データマイニングで有効とされるテンソル分解などの技術を適用できることに利点がある。さらに、テンソルのモードを追加していくことでグラフの接続関係だけでなくノードやエッジの属性情報の関係も自動的に抽出できることが期待される。テンソル表現を用いることにより、多くの経験や知識を要せずとも情報集約が容易になり、業務システムでの機械学習利用促進が図れることが期待される。NNDT は、Deep Tensor で使用するテンソル分解に着想を得たノード情報の埋め込み処理を用い、さらに、自然言語処理分野で使用されるトランスフォーマー手法を組み合わせたグラフデータの機械学習手法である。Deep Tensor と比べ、適用可能なグラフ形式とタスクの増加、精度や学習速度の向上、動作安定性の改善などが得られている。

グラフデータの機械学習問題のうち、グラフデータ間の相違を学習するグラフ分類・回帰問題では、予測対象のグラフデータのサイズ（ノード数やエッジ数）は様々である場合が多い。テンソルを対象とする教師あり機械学習手法として知られる Tensor regression[3]では、適用可能なテンソルサイズが均一であることが想定されているため、グラフ分類・回帰問題への適用には工夫が必要である。その点 Deep Tensor や NNDT は、サイズの異なるグラフデータをそのままテンソルで表現し、そのまま学習が可能であるため、グラフ分類・回帰問題への適用が容易である。本発表でも、グラフ分類問題を取り扱い、NNDT を使用する。

3. 実験

オンライン教育システムで蓄積されたトランザクションデータを用いて、NNDT による学習と予測に関する実験検証を行う。学習者の教材受講履歴および最終成績結果の情報を用いて学習を行い、成績不振に陥りそうな学習者を予測して、成績不振に陥る前に教育支援を行う活用シナリオを設定する。

3.1 問題設定

トランザクションデータとして、教育分野の分析利用者向けに公開されているデータセット Open University Learning Analytics Dataset[4]を用いる。オンラインシステムに構築された仮想教育環境で実際に学習者が利用を行った記録に基づいている。データベース上のエンティティである学習者と教材の各事例に ID が割り当てられ、学習者ごとに、どの教材をいつ受講したかといった教材受講履歴および最終成績結果が記録されている。

実際のデータでは、複数のコースを同一の学習者が受講する場合が存在し、取り扱いがやや複雑になるため、本実験では、最終成績結果ごとに別々の学習者が存在するものとして合理化する。最終成績は”Distinction”, ”Pass”, “Fail”, “Withdrawn”の 4 段階であるが、成績不振の学習者を予測する課題として、本実験では”Distinction”と”Pass”を「合格」、”Fail”と”Withdrawn”を「不合格」に割り当てなおし、2 値分類の問題として扱う。最終的に「合格」が 15,278 件、「不合格」が 13,682 件の合計 28,960 件の学習者を訓練・評価対象とした。ここから無作為非復元抽出を行い、80%を訓練事例に、残りの 20% を評価事例に利用する。また、教材受講履歴データは、各学習者ごとにコースの受講開始日を起点とした 29 日以内の履歴データのみを用いる。これは実際の業務支援において、学習者の短期的な成績予測を想定したためである。その結果、受講教材数と受講回数は学習者ごとに一定にはならず、学習者ごとの教材受講履歴データのサイズは不定となる。

3.2 データ構成と提案手法

データ公開元によるデータ記述に関する説明[5]によれば、今回対象とする教育データは、7 つのテーブルを持つ複数のスキーマで定義される。各スキーマは、次の 3 つに大別される（図 1）。まず 1 つ目は学生マスタ(元データでの Student demographics)で、「学習者」の年齢や性別などの情報が存在し、最終成績情報も含まれる。2 つ目は受講トランザクション(元データでの Student activities)で、「学習者」の「教材」の受講日や「教材」選択時のクリック数などの「教材」の「受講」状況情報が含まれる。3 つ目は教材マスタ(元データでの Module Presentation)で、「教材」タイプや「教材」の使用予定週などの「教材」に関する情報が含まれる。これより、「学習者」、「受講」、「教材」が相互に関係する重要項目であることが分かる。これらはグラフデータでのノードとして利用する。また、各項目には属性情報があり、7 つのテーブル情報を総合した結果、合計 24 項目の属性情報が得られることが分かった（ここでは詳細は割愛）。これらは各ノードに関する属性情報となる。尚、本発表で利用する NNDT では、カテゴリ変数を用いるため、連続数値データは適切に離散化し、カテゴリ変数化して処理を行っている。

3.3 比較手法

提案手法の効果を確認するため、以下の各種機械学習手法を比較のため導入する。Python 言語の機械学習用ライブラリ scikit-learn[6]搭載の SVM、ニューラルネットワーク(NN)、決定木、ランダムフォレスト(RF)、および、近年、機械学習のコンペティションプラットフォーム kaggle[7]等で高い予測精度が得られる手法として注目される勾配ブースティング木手法 LightGBM[8]である。これら比較手法は、いずれも特徴量設計を必要とするが、そのためには、結果の違いの根拠となりそうな仮説を立てた上で、その仮説に応じた項目を選択、或いは、新たに特徴量を生成する必要がある。新たに生成する特徴量は、何れかの項目の合計値や平均値などの基礎統計量である場合もあれば、例えば、時間情報を頼りにした行動の頻度や間隔、順序や共起情報を新たに算出した値でも良い。本研究では、仮説として、例えば「期限通りに課題を提出しない傾向のある学習者は、最終試験においても準備が不十分となり、不合格となる可能性が高い」等を考え、これに応じた特徴量を設計した。この作業には一定程度の工数を掛ける必要があった。その具体例を図 2 に示す。

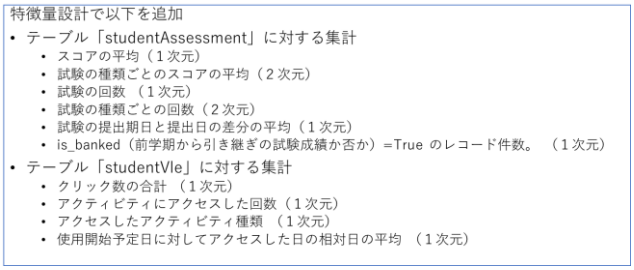

図 2　比較手法での特徴量設計の例

3.4 結果と考察

予測性能指標として代表的な、正解率、再現率、適合率、F 値を表 1 に示す。正解率と適合率では LightGBM が、再現率では Neural Network が一番高い値を示すという結果が得られた。また、総合的な予測指標として信頼性のある F 値では、提案手法である NNDT が他手法を若干上回っており、概して他の手法と同等程度の予測性能を得たものと考えられる。これは、他の手法では、人手を費やし特徴量設計の工夫を行うなどのコストを掛けた上で、ある程度の高い性能を達成している一方、NNDT ではその工数を要さずとも同等の性能が得られたことを意味している。

表 1 ：予測性能指標

比較対象	正解率	再現率	適合率	F 値
NNDT	0.720	0.691	0.705	**0.698**
LightGBM	**0.734**	0.646	**0.752**	0.695
SVM	0.664	0.602	0.654	0.627
NN	0.688	**0.765**	0.640	0.697
決定木	0.646	0.639	0.618	0.629
RF	0.731	0.657	0.740	0.696

4. まとめ

本発表では、トランザクションデータを対象とする機械学習において、ノード情報の埋め込み処理に特徴のある、グラフデータの機械学習手法 NNDT を利用することにより、前以て特徴量設計を必要とせずに、学習、予測を行うことが可能な方法を提案した。実際に、オンライン教育システムにおける学習者の成績予測を実験的に実施し、業務の経験や知識を必要とする作業工程を削減した場合でも、比較的高い精度が維持できることを示した。これは、業務における機械学習利用の有用性を示した結果であると言える。今後は、NNDT のテンソル学習器としての利用方法の検証をさらに続け、業務で得られるトランザクションデータの機械学習への適用可能性を検討していく予定である。

参考文献

[1] Tolmachev A., Sakai A., Todoriki M., Maruhashi K., “Bermuda Triangles: GNNs Fail to Detect Simple Topological Structures” ICLR 2021 Workshop on Geometrical and Topological Representation Learning, (2021).
[2] Maruhashi K., Todoriki M., Ohwa T., Goto K., Hasegawa Y., Inakoshi H., Anai H., “Learning Multi-Way Relations via Tensor Decomposition with Neural Networks,” 32nd AAAI Conference on Artificial Intelligence, (2018).
[3] Zhou H., Li L., Zhu H., “Tensor regression with applications in neuroimaging data analysis”, Journal of the American Statistical Association Vol.108, No.502, (2013).
[4] “Open University Learning Analytics Dataset | Kaggle”, https://www.kaggle.com/rocki37/open-university-learning-analytics-dataset
[5] “Open Learning Analytics | OU Analyse | Knowledge Media Institute | The Open University”, https://analyse.kmi.open.ac.uk/open_dataset
[6] “scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation”, https://scikit-learn.org/stable/
[7] “Kaggle: Your Machine Learning and Data Science Community”, https://www.kaggle.com/
[8] “GitHub - microsoft/LightGBM: A fast, distributed, high performance gradient boosting (GBT, GBDT, GBRT, GBM or MART) framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.”, https://github.com/microsoft/LightGBM