

情報推薦のための機械学習のバイアスの可視化

栃木 彩実* 伊藤 貴之† Xiting Wang‡

概要. 映画などのコンテンツ推薦システムでは近年、その推薦エンジンに機械学習を適用することがある。一方で近年、機械学習における公平性やバイアスについての議論が活発化しており、機械学習にとって重大な課題の1つとなっている。このような問題の一要因として学習データのバイアスが挙げられる。学習データにバイアスが存在することで、意図せずとも不公平な学習結果を生じることがある。そこで本研究では学習データと学習結果を比較可視化することで、そのバイアスの発見につなげるシステムを提案する。具体例として本報告では、ユーザ群の映画鑑賞履歴を学習データとし、機械学習による映画推薦結果と比較可視化することで、ユーザ間の推薦のバイアスがどのように分布するかを観察した事例を報告する。

1 はじめに

スマートフォンなどの普及に伴い、ユーザ好みのコンテンツを提案する推薦システムが活躍するようになった。一方で、過度にパーソナライズ化された推薦システムは、偏った推薦結果を引き起こす。そのため、推薦システムを構築するには公平性や多様性を考慮しなければならない。また近年、コールドスタートや高負荷の問題に対処するために推薦エンジンに機械学習技術 [5] を取り入れる事例が増えている。しかし、バイアスを含む学習データや学習モデルの利用によって、不当な学習結果を引き起こす恐れがある。そこで本報告では学習データのバイアスに着目し、学習データと推薦結果を比較可視化することで学習データのバイアス発見を支援する可視化システムを提案する。

2 関連研究

説明可能な推薦システムにとって可視化は有力なツールである。推薦システムの可視化に関する研究は、機械学習の発展と共に活発になっている。一方で、機械学習の公平性の可視化はまだ新しい分野であり、先行研究もまだ少ない [1][2]。これらの先行研究では、ユーザ評価を予測する機械学習に焦点を当てており、ユーザの公平性にのみ着目した可視化システムを提案している。一方で、本研究では推薦システムに特化して機械学習の公平性に着目し、ユーザと推薦されるコンテンツの双方を可視化する。

3 提案手法

ここでは、各コンテンツの鑑賞履歴の学習データと、推薦結果すなわち各ユーザに推薦される一連のコンテンツ情報の比較を支援する可視化システムについて紹介する。本研究では、可視化における課題を以下のように定義する。

- T1:** ユーザ/コンテンツによって構成される各クラスターパターンの可視化。
- T2:** 鑑賞結果と推薦結果に大きな差があるクラスター、または偏った推薦を引き起こす恐れのある外れ値を含むクラスター等の発見。
- T3:** ユーザ/コンテンツの特定のクラスターにおける属性の詳細な分布の観察による、偏った推薦をもたらす要因の探索。
- T4:** 特定のユーザ/コンテンツにおける属性の詳細な分布の観察による、バイアスの要因に関する議論の展開。

図1に示すように、我々のシステムは4つの可視化要素から構成される。まずシステム利用者は **View 1** を観察をすることで **T1** と **T2** を実現する。散布図上では、特定範囲の拡大、クラスターやノードの選択、さらに以下に示す2種類の配色機能が利用可能である。

- (1) **Similarity:** 散布図上のノードを選択すると、選択したノードとのコサイン類似度に基づいて他のノードを10段階に配色する。ここで彩度をコサイン類似度に比例させる。
- (2) **Difference:** 各ノードに対する学習データと推薦結果の特徴量ベクトルのコサイン類似度に基づいて10段階に配色する。学習データと推薦結果との差が大きいノードほど彩度が高くなる。

View 2 は特定のクラスターの属性分布を表示する棒グラフである。**View 2** を用いて各クラスターを観察することによって、**T3** の実現が可能になる。さ

Copyright is held by the author(s). This paper is non-refereed and non-archival. Hence it may later appear in any journals, conferences, symposia, etc.

* お茶の水女子大学

† お茶の水女子大学

‡ Microsoft Research Asia

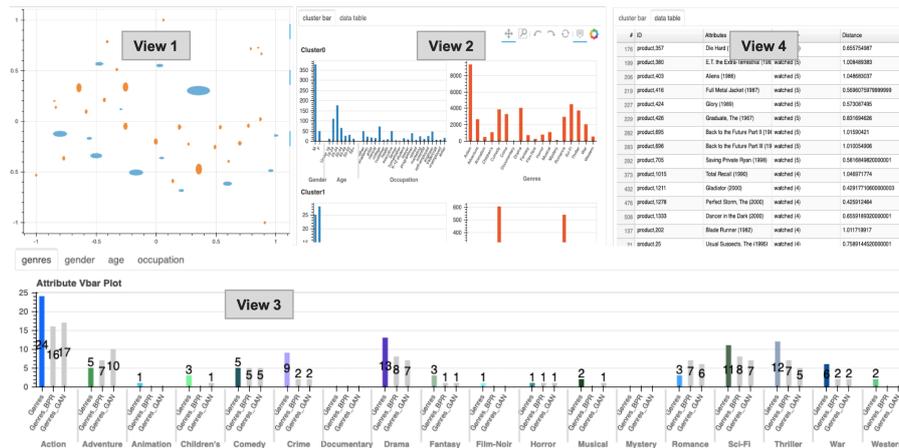


図 1. View 1: ユーザ群とアイテム群を同一座標平面上に描画した散布図, View 2: 特定のクラスタの属性分布を示す棒グラフ, View 3: 特定のノードの属性分布を示す棒グラフ, View 4: 特定のノードに関する詳細な情報を示すデータテーブル.

らに, View 3 は特定のユーザ/コンテンツの属性分布を示し, View 4 は特定のユーザ/コンテンツに関するより詳細な情報を提供する. これらを用いて散布図で選択したノードに対する可視化結果を観察することで, T4 の達成につながる.

4 実行結果

4.1 データの前処理

本研究では映画鑑賞データセット [4] を可視化に用いた. データセットは 3883 本の映画と 6040 人の顧客を含む. 映画はジャンル, 顧客は性別, 年齢, 職業などの属性を持つ. また学習モデルとして BPR[5], および BPR に GAN[3] を適用したモデルを使用した. データセットを訓練データとテストデータに分割した上で, 訓練データに対して機械学習を適用した. 得られた学習結果に基づき, テストデータに含まれる各顧客に対して 20 本の映画を推薦した. 最終的に, テストデータからサンプリングした 1000 人の顧客と 512 本の映画を可視化した.

4.2 可視化画面の観察

上記のデータセットを可視化し, 分析した結果の一部について紹介する. ここで, View 2 の青い棒グラフは顧客, オレンジの棒グラフは映画の属性分布を表している. また, View 3 はグレーの棒グラフが推薦結果, 寒色の棒グラフが視聴結果を表す.

まず Difference モードで配色された View 1 を観察し顧客クラスタに着目すると, 推薦結果と視聴結果の差が大きいクラスタ 1 と 5 (図 2 の散布図) が確認できる. さらにこれらのクラスタ内の顧客属性を View 2 にて分析すると, クラスタ 1 は男性より女性が多く, クラスタ 5 では子供の割合が多いことがわかる. 使用したデータセットでは女性や子供と

いった属性は少数であり, 視聴結果と推薦結果の差が大きい要因として顧客の属性の違いが考えられる.

次に上記クラスタから 1 人の顧客を選択し, View 3 で視聴結果と推薦結果を比較した. 図 3 を見ると, Children's や Animation を多く視聴しているのに対し, Comedy や Sci-Fi が多く推薦されている. これは視聴傾向が大衆と異なることが要因と考えられる. このように本システムを用いることで, まずユーザやコンテンツの分布やクラスタの特徴をとらえ, 特定のクラスタやノードの属性分布に焦点を当てたり, データテーブル内の詳細情報を読み取ることで, 特定のユーザ/コンテンツのバイアスを分析することが可能である.

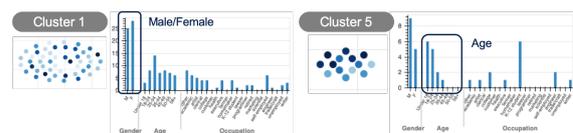


図 2. View 2 における顧客クラスタ 1 と 5 の属性分布.

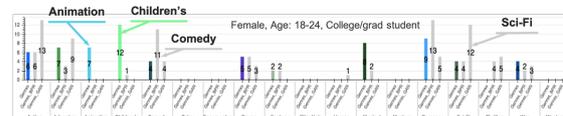


図 3. View 3 における視聴結果と推薦結果の差が大きい顧客の可視化結果.

5 むすび

本報告では推薦システムの公平性に着目し, 複数の可視化画面を組み合わせることで学習データと推薦結果を比較することで, 学習データに存在するバイアスの発見を支援するシステムを提案した.

参考文献

- [1] Y. Ahn and Y.-R. Lin. FairSight: Visual Analytics for Fairness in Decision Making. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 26, pp. 1086–1095, 2020.
- [2] A. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. P. Chau. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *IEEE Conference on Visual Analytics Science and Technology*, 2019.
- [3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *In Proceedings of Neural Information Processing Systems*, pp. 2672–2680. NIPS, 2014.
- [4] F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. In *ACM Transactions on Interactive Intelligent Systems*, Vol. 5. TiiS, 2015.
- [5] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.