The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

半教師学習と特異値分解によるCold-Start問題へのアプローチ

Combining semi-supervised learning and singular value decomposition to Cold-Start problem

内田 匠 *1*3 Takumi Uchida　中川 慧 *2*3 Kei Nakagawa　吉田 健一 *3 Kenichi Yoshida

*1*3筑波大学大学院 ビジネス科学研究科 University of Tsukuba Graduate School of Business Sciences

*2野村アセットマネジメント株式会社 Nomura Asset Management Co., Ltd

機械学習を用いてWebマーケティングには多くの課題があり、その一つにCold-Start問題がある。例えば通販サイトなどでユーザーに商品を推薦する場合、ユーザーと商品の購買ログやレビューログを取り扱うことになる。しかし、Webのマーケティングデータはロングテールになる傾向があるため、多くのユーザーや商品のログデータは少なく、数件程度しかないことも多い。商品の推薦システムではこの過去ログに基づいてユーザーに対して商品を提示するため、ロングテールの商品は提示されにくく一部の人気商品ばかりが提案される事になる。本研究では、このCold-Start問題に対して半教師学習と特異値分解を組み合わせた手法を提案する。また、提案した手法をMovieLensが提供している映画の評価スコアデータで検証した結果を報告する。

1. はじめに

Cold-Start問題は特にWebマーケティングの推薦システムにおける問題として、重要な研究課題となっている [Shi 14]。Cold-Start問題とは、推薦システムにおいて、推薦するためのデータが少ない新規ユーザーや新規アイテムに対する評価が難しいことをいう。Cold-Start問題を解決する意義は新商品やマイナーな商品の適切な提案を可能にする点にある。未だユーザーの目に触れた事のない商品を高い精度で提示できるのであれば、企業はユーザーに対してより多様な価値提供が可能になる。

Cold-Start問題を解決するためのアプローチとして、コンテンツベースの協調フィルタリング [Schein 02] やユーザークラスタを応用するもの [Lika 14]、半教師学習を用いたもの [Zhang 14] など多くの研究がある。その中でも本研究では半教師学習を用いたアプローチの改善手法を提案する。半教師学習とは、観測されたデータをLabeled Data、まだ観測されていないデータをUnlabeled Dataとし、Labeled DataとUnlabeled Dataの両方を利用した学習法全般をいう。具体的には、[Zhang 14] が提案した半教師学習を用いた推薦システムに特異値分解を組み合わせることを提案する。[Zhang 14] の手法は、例えばwebページの分類をする際の画像、文章や、推薦システムにおけるユーザーとアイテムのようなデータが二つの素性に分解できることを仮定する必要がある [Blum 98]。一般に、データどうしに相関がある場合、それらのモデル内での役割は同じであり、そのままでは利用できない。そこで、データをいったん特異値分解により次元削減することでデータ間の相関を取り除く。これが特異値分解を組み合わせる動機である。したがって本手法は、推薦システム以外のデータの素性が容易に分解できないような領域にも適応が可能である。

本論文の構成は次の通りである。2章で、先行研究 [Zhang 14] の手法を概観し、3章で提案手法の紹介を行う。そして、4章で、MovieLensが提供している映画の評価スコアデータで検証した結果を報告する。

連絡先: 内田 匠, 筑波大学大学院 ビジネス科学研究科, 〒112-0012 東京都文京区大塚 3-29-1

2. 先行研究

[Zhang 14] は半教師学習の [Blum 98] の Co-training を Cold-Start 問題の解決のために応用した手法 (CSEL) を提案した。Co-training は、データが2つの素性に分解できると仮定し、それぞれの素性について分類器を個別に用意し、観測されたLabeled Dataのみでそれぞれ学習する。1つ目の分類器で高い確信度で分類できたUnlabeled Dataを、2つ目の分類器の次回の学習に用い、逆に2つ目の分類器で高い確信度で分類できたUnlabeled Dataを1つ目の次回の学習に用いる。これを繰り返して学習データを拡大する手法である。Cold-Start問題の原因はロングテールには十分なログが存在しない事にある。CSELはその不足したログをCo-trainingによって推定追加することで、ロングテールの推薦精度を改善した。

CSELの概要をまとめる。CSELは推薦に利用する変数をA) ユーザーの属性変数、B) 商品の属性変数、C) ユーザーと商品の相性変数に分解した上で、2つの推定モデルを別々の変数を使って学習する。具体的には、ある推定モデルではAとCの変数を使って学習し、もう片方の推定モデルではBとCの変数を使って学習する。次に、それぞれの推定モデルがUnlabeled Dataを予測しその予測値の信頼度を評価する。上位の信頼度をもつデータを学習データに追加した上で再び学習を行う。このように学習と推定追加を繰り返し行い、ロングテールのデータを増加させた上で最終的な学習を行う。CSELを映画の評価スコアデータであるMovieLensのデータで検証した結果、ロングテールにおいて特に予測精度の改善が見られた。

半教師学習での推定追加における課題は、その推定値の信頼度である。推定値を学習データに追加してそのままの同じモデルで再学習する場合、推定誤差を強化する方向に作用する懸念がある。CSELはユーザーと商品の組合せ情報という推薦システムに特有のデータ構造を利用し、それぞれ違う推定モデルがデータを評価するCo-trainingを活用することでこの問題を解決した事例といえる。

3. 提案手法 - CSEL with SDV

提案手法のアルゴリズムはCSELを基本としつつ、それぞれのCo-training推定モデルが学習に用いるデータが特異値分

解を応用して生成されている点が大きく異なる。本研究で提案する手法は、Co-training における推定モデルの多様性を、特異値分解をデータに適用しておくことで、データ間の相関を取り除き、モデル間の多様性を担保することを着眼点としている。また、どの推定値を学習に追加するかの判断基準となる信頼度については、一つのサンプルに対する異なる推定モデルからの推定値の標準偏差の逆数を用いた。つまり、複数のモデルからの推定値の分散が小さいほど信頼度が高いとしている。これは異なるモデルであるのに同じような推定値を出力してきた場合はその推定値の信頼度は高い、という仮定を根拠としている。この提案手法の概略を図 1 にて示す。

まず始めに、すでに目的変数が観測出来ている Labeled Data と観測出来ていない Unlabeled Data を同じ説明変数行列 X として扱う。次に、この説明変数行列を特異値分解を応用して、以下の数式のように複数の行列に分解する。

$$X = U\Sigma V^T \quad (1)$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_q u_q v_q^T \quad (2)$$

σ_i は固有値、u_i, v_i は X の左特異値ベクトルと右特異値ベクトルである。q は X が $n \times m$ の行列とすると $min(n, m)$ である。$u_i v_i^T$ は行列 X と同じ次元をもつ行列となりそれぞれが行列 X を構成する要素となっている。また、$u_i v_i^T$ は互いに基底である別々の特異値ベクトルから生成されるため、他の $u_i v_i^T$ との間には Co-training に必要な十分な多様性を有していることが期待できる。

Co-training で採用できる推定モデルの数は 2 ～ q 個まで任意に選択することができ、その予測アルゴリズムについても説明変数に行列データをとれるものであれば何を用いても良い。

このような処理を繰り返すことで、観測できなかったデータの推定値の信頼性を確認しながら学習データに追加する。追加されたデータにロングテールがあれば、ロングテールの推定精度を改善する可能性がある。

4. 実験

提案手法の有効性を確認するために、MovieLens[Harper 16] の映画のレビュースコアデータに適応して検証を行った。このデータにはユーザー ID と映画 ID に 1〜5 の評価スコアが観測されている。ユーザーの属性情報として性別、年齢、職業、映画の属性情報としてジャンルと制作年度がある。これらの属性情報に加えて、推薦システムでよく用いられる潜在因子分析によるユーザーと映画の相性因子も説明変数とした。さらにこれらの説明変数間の二次の公差項も追加している。検証では、過去の 7 日間のレビューデータを学習データとし、次の 7 日間を教師データとしてユーザーの評価スコアを予測した。指標値は予測誤差の絶対値の平均とした。

また本データでは、7 日間の間に 100 本以上の映画を視聴しているユーザー ID が存在していた。現実的に個人による視聴とは考えにくく本検証では不適切なデータだと判断したため、学習期間である 7 日間中に 20 本以上の映画を視聴したユーザー ID は検証から除外した。このデータ除外をすると 7 日間で約 14,700 件のレビューログが存在している。

評価の際には学習と教師データを 7 日間スライドして 10 週分の平均値を確認した。比較対象のモデルとして、観測できた Labeled Data のみで学習したモデルとの比較を行った。Co-training に用いた推定モデルの数は 2 と 5 を検証してそれぞれの予測精度に違いがあるかを比較した。Unlabeled Data の推定追加は、信頼度が最も良かった 1,000 件の追加を 10 回繰り返した。これにより学習データは平均して約 68% 増加することになる。予測の手法としては二次の正則化項つきの線形回帰モデルを採用した。Co-training のモデルも最終予測のモデルもこの手法を採用した。

表 1: Co-traing with 2 and 5 models

映画の観測数	モデル数:2	モデル数:5	差
000〜	1.058	1.059	-0.001
001〜	1.028	1.065	-0.037
002〜	1.023	1.05	-0.027
003〜	0.94	0.956	-0.016
004〜	0.929	0.917	0.012
005〜	0.932	0.916	0.016
006〜	1.023	1.021	0.002
007〜	1.043	1.031	0.012
008〜	0.881	0.874	0.007
009〜	0.91	0.87	0.04
010〜	0.902	0.933	-0.031
020〜	0.792	0.79	0.002
030〜	0.785	0.814	-0.029
040〜	0.999	0.983	0.016
050〜	0.85	0.831	0.019
060〜	0.825	0.824	0.001
070〜	0.795	0.785	0.01

Co-training に 5 つの推定モデルを使った結果を図 2 に示す。図 2 の結果から、提案手法が観測されたデータだけの学習と比較して全体的な予測精度が良いことが分かった。学習データ中には存在しなかった商品 ID の予測精度が大きく改善していることが確認できた。本手法は新商品のレビュー予測については有効であると言える。しかし、よく見られていた映画の推定精度が改善している一方で、1 から 6 件程度のレビューがついた映画については予測精度は悪化する傾向が見られた。これは当初の仮説に反する結果となった。

次に Co-training に用いる推定モデルの数を 2 と 5 にした場合の比較を表 1 に示す。表 1 からは、Co-training で用いるモデルの数を比較しても優位な改善は見られなかった。

5. まとめ

本研究では半教師学習の Co-Training について特異値解析を組み合わせた Unlabeled データの推定追加の手法について提案し検証を行った。

その結果、MovieLens の映画の評価スコアの予測精度を改善することに成功した。特に学習データでは観測されなかった映画の予測精度を大きく改善できた。一方で、当初の仮説に反しロングテール映画での予測精度は改善しなかった。また、推定追加するデータの信頼度計算を Co-traning の異なる推定モデルの推定値の標準偏差の逆数としたことから、モデルの数を増やせばより予測精度が改善すると仮説したが、それを実証する結果は得られなかった。

今後の研究としては、Unlabeled Data の選出方法についてより合理的な手法について検討したい。また本手法は Web マーケティングの推薦システム以外にも適応が可能であることに着目し、他の領域での類似課題が解決可能かについても検証していきたい。

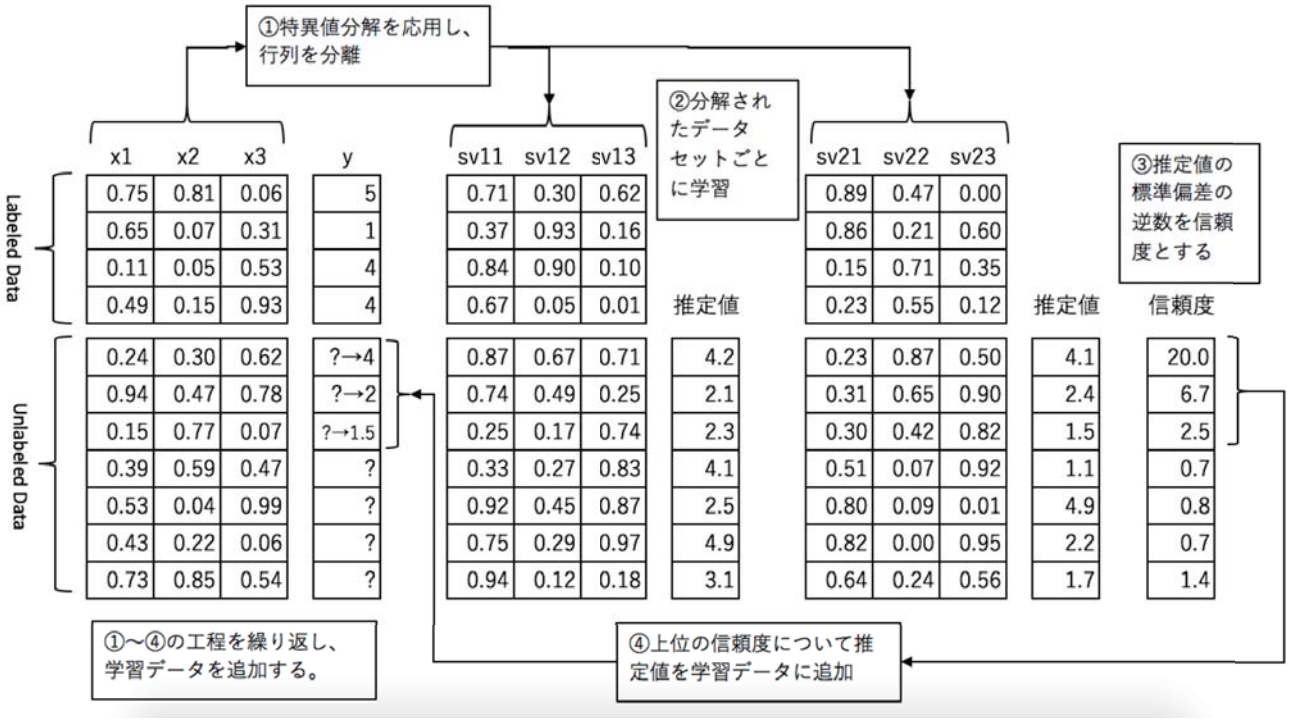

図 1: Proposed Method

参考文献

[Blum 98] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100ACM (1998)

[Harper 16] Harper, F. M. and Konstan, J. A.: The movielens datasets: History and context, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 5, No. 4, p. 19 (2016)

[Lika 14] Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S.: Facing the cold start problem in recommender systems, *Expert Systems with Applications*, Vol. 41, No. 4, pp. 2065–2073 (2014)

[Schein 02] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M.: Methods and metrics for cold-start recommendations, in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260ACM (2002)

[Shi 14] Shi, Y., Larson, M., and Hanjalic, A.: Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, *ACM Computing Surveys (CSUR)*, Vol. 47, No. 1, p. 3 (2014)

[Zhang 14] Zhang, M., Tang, J., Zhang, X., and Xue, X.: Addressing cold start in recommender systems: A semi-supervised co-training algorithm, in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 73–82ACM (2014)

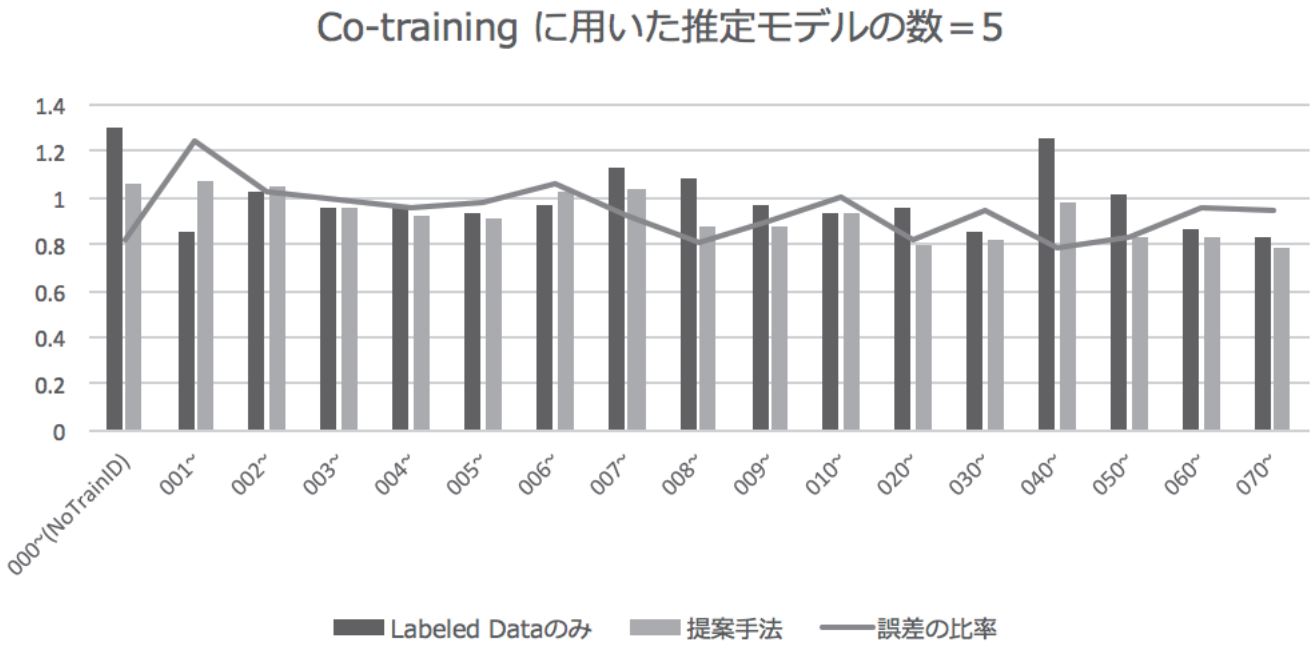

図 2: 棒グラフがそれぞれの予測誤差の絶対値の平均。線グラフが二つの手法の誤差の比率で 1 よりも低い場合は提案手法が良いと言える。