

協調フィルタリングによる大域解の推定

大槻有紗[†] 三浦 孝夫[†]

[†] 法政大学理工学部創生科学科

1. 前書き

本研究では評価値が一つに定まるような方法の発見・問題の定式化を行い、さらに大域解の生成手順を示し、実用的な働きを示す手法を提案する。

2. 協調フィルタリング

$N(j;a)$ を、項目 j を評価した利用者 k で $s_{[ak]}$ が定義されている(類似していると思える)人の集合とする。このとき、 $N(j;a)$ のメンバ k の評価値 $r_{[kj]}$ を類似度 $s_{[xk]}$ で加重平均した値を、推定値 $r_{[aj]}$ する。 b_{xi} は R の規準を仮定しており、 $b_{xi} = \mu$ (平均判定値) + b_x (商品 x の平均評価値 - μ) + b_i (利用者 i の評価値分散) で求められるとすると次の式 1 のようになる。

式 1

$$r_{aj} = b_{xi} + \frac{\sum_{j \in N(j;a)} S_{xk} \cdot (r_{kj} - b_{xi})}{\sum_{j \in N(i;a)} S_{xk}}$$

3. 大域解の計算式の提案

利用者類似度 S を用いた(利用者 x の項目 k に対する)加重平均値 rxk は、 x と類似した利用者 i と、 i による項目 k の値の類似度を用いた補正值 s_{xi} 、 r_{ik} で得られ、これを全ての類似者の和 $\sum_{i \in N(k;x)} s_{xi}$ で調整する。各行をこの合計値で割った値を成分とする(修正)類似行列 S^* を用いると、式 2 は次のように表せる。

式 2

$$\vec{r}_i = S^* \vec{r}_i$$

すなわち、大域解とは S^* の(固有値 1 の)固有ベクトルであり、高々 N 個存在する。

4 実験

4-1 実験手順・評価

本研究ではまず協調フィルタリングが正しく行われるかについて実験を行う。次にクラスタごとの正答率推定を行い、回答者特性に推定値が依存するのかを検証する。最後に大域解の推定をデータ全体・各クラスタ全体データを用いて行い、大域解の存在を確認する。また同時にクラスタの影響を確認する。

評価については、協調フィルタリングによる正答率の推定時は 10 件を無作為に抽出し解を隠して推定を行い推定値と ± 1 の誤差ならば正答であるとする。大域解の推定時は、固有値 1・固有ベクトルが存在した場合大域解の存在を確認する。

4-2 実験データ

本実験では NII 楽天トラベルデータ 341(件)で実験を行う。今回は目的変数評価 1(立地) 評価 2(部屋) 評価 3(食事) 評価 4(風呂) 評価 5(サービス) 評価 6(施設・ア

メニティ) 評価 7(総合評価) 価格を用いる。さらに、 k -Means 方式によるクラスタ作成し予備実験より $k=4$ としてクラスタリングする。

4-3. 実験結果

協調フィルタリングによる全体・各クラスタの正答率は全体:30%、クラスタ 1:40%、クラスタ 2:70%、クラスタ 3:80%、クラスタ 4:100%となる。

表 1 各クラスタの正答率と改善率

	正答率(%)	全体からの改善率(%)
全体	30	
クラスタ1	40	10
クラスタ2	70	40
クラスタ3	80	50
クラスタ4	100	70

大域解の推定を行ったところ全体・各クラスタで S^* に固有値 1 の固有ベクトルが存在した。全体と各クラスタの固有ベクトルをカイニ乗検定において検定をすると、数値の誤差は 10^{-9} よりも小さくなる。

表 2 カイ 2 乗検定結果

全体	クラスタ1	クラスタ2	クラスタ3	クラスタ4
V1	V1	V1	V1	V1
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143
0.142857143	0.142857143	0.142857143	0.142857143	0.142857143

4-4 実験考察

今回の実験により、協調フィルタリングによる推定を行う際にはクラスタリングを行い属性分けをしてから行うことが正答率を上げる一つの方法になると考えられる。大域解の推定においては、本実験により大域解が存在していることがわかる。さらに全体と各クラスタの固有ベクトルを比べると数値に差がないことから、クラスタリングと大域解の関係性はないと考えられる。

5 結論

本実験より評価の正答率が 30%から平均 72.5%上昇したことからデータの属性分けを行うことで正答率が上がるということがわかった。大域解においては協調フィルタリングによるアンケートの大域解の存在を発見することに成功した。大域解の推定の場合クラスタリングと大域解に関係性は見られず、クラスタリングを行う必要がないことがわかった。

参考文献

情報推薦システム入門 監訳 田中克己, 角谷和俊