アイテム間の相関を考慮したナイーブベイズ法による協調フィルタリングに関する研究

1X08C023-7　大井貴裕
指導教員　後藤正幸

1　研究背景・目的

近年，情報技術の進展により，インターネット上には多くの EC サイトが存在し，扱われるアイテム数も膨大となっている．EC サイトでは，ユーザの購買履歴データがデータベース上に蓄積されるため，この膨大なデータを活用したプロモーション手法が可能であり，特に各ユーザの嗜好，特性に合わせて自動で推薦を行う推薦システムの重要性が増している．その代表的な手法として，ユーザ間の購買履歴の類似性から被推薦ユーザの好むアイテムを予測し提示する協調フィルタリング（以下 CF）がある．

CF には購買履歴を基に確率モデルを構築するアプローチがあり，その中でも文書分類で用いられているナイーブベイズ法（以下 NB 法）を CF に適用した手法が提案されている．一般に，アイテム A を購入したユーザは，アイテム B も購入する傾向がある, というように何らかの相関があるが, NB 法は全アイテム同士が独立であると仮定しているため, 購買活動を現実的に表現したモデルとは言い難い．この仮定を緩和するため，Wang ら [1] は独立性を仮定して展開した NB 法の尤度項に重みを乗じ，独立性を仮定しない場合の尤度をシンプルな形で精度良く推定する方法を提案している．しかしアイテムは，いくつかの相関のあるアイテム集合（クラスタ）に分けられ，そのサイズは多種多様である．この点を考慮せずどの尤度項にも同一の重みが乗じられているため，サイズの大きいクラスタに推薦結果が依存してしまうという問題がある．そこで，本研究では，アイテムをユーザの購買履歴データでクラスタリングしたもとで，NB 法による CF において各クラスタのサイズを考慮した重みを各尤度項に乗じることで，より精度の高い推薦を可能とするモデルを提案する．提案手法を推薦システムのベンチマークデータに適用し，提案手法の有効性を示す．

2　NB 法による CF

NB 法は，未購買アイテムの購買確率の予測を目的とした確率モデルである．いま，ユーザ集合を $\mathcal{U} = \{U_1, U_2, \cdots, U_J\}$，アイテム集合を $\mathcal{I} = \{I_1, I_2, \cdots, I_M\}$ とする．被推薦ユーザ U_j の既購買のアイテム数が L_j 個のとき，被推薦ユーザ U_j の既購買アイテム集合を $\mathcal{Y}_j = \left\{I_1^j, \cdots, I_{L_j}^j\right\}, (1 \leq L_j \leq M)$ とし，被推薦ユーザ U_j の未購買アイテムを $I_x^j \in \mathcal{I} \setminus \mathcal{Y}_j$ とする．また，アイテム I_m が購買されることを $R_m = 1$，アイテム I_m が購買される確率を $\Pr\{R_m = 1\}$ とする．NB 法では，被推薦ユーザの未購買アイテム I_x^j を購買している他ユーザが，被推薦ユーザの既購買アイテム $I_l^j \in \mathcal{Y}_j$ を購買する条件付き確率が独立であると仮定する．このとき，未購買アイテムが購買される確率は，ベイズの定理により，以下の式のように求められる．

$$\begin{aligned}&\Pr\{R_x^j = 1 | R_1^j = 1, \cdots, R_{L_j}^j = 1\},\\ &\propto \Pr\{R_x^j = 1\} \times \Pr\{R_1^j = 1, \cdots, R_{L_j}^j = 1 | R_x^j = 1\},\end{aligned} \tag{1}$$

$$= \Pr\{R_x^j = 1\} \times \prod_{l=1}^{L_j} \Pr\{R_l^j = 1 | R_x^j = 1\}. \tag{2}$$

NB 法では (1) 式の第二項の独立性を仮定して，(2) 式へと展開されている．推薦にあたり，(2) 式の各項 $\Pr\{R_x^j = 1\}$，および $\Pr\{R_l^j = 1 | R_x^j = 1\}$ は学習データを用いて推定し，この確率の高いアイテムを候補とする．しかし，例えば「好きなアーティストの CD は全て購入する」等，ユーザの購買するアイテムには相関があると考えるのが一般的であるが, NB 法はこれらの独立性を仮定しているため，購買活動を現実的に表現したモデルとは言い難い．

3　従来研究

前述の問題を解決するためには，ユーザの購買するアイテムの独立性の仮定を緩和する必要がある．しかし，(1) 式の第二項の独立性が仮定できない場合，これを直接推定することは計算量の問題で困難である．そこで Wang らは，一般に，

$$\begin{aligned}&\Pr\{R_1^j = 1, \cdots, R_{L_j}^j = 1 | R_x^j = 1\},\\ &\geq \prod_{l=1}^{L_j} \Pr\{R_l^j = 1 | R_x^j = 1\},\end{aligned} \tag{3}$$

が成り立つことに着目し，(3) 式の右辺に対し，重みを乗じることで左辺と近い値に補正する手法（improved naive bayes 法）を提案した．improved naive bayes 法による未購買アイテムの購買確率は，重み w_0 を用いて，(2) 式の第二項に重みを乗じた形となり，(4) 式のように近似される．

$$\begin{aligned}&\Pr\{R_x^j = 1 | R_1^j = 1, \cdots, R_{L_j}^j = 1\},\\ &= \Pr\{R_x^j = 1\} \times \prod_{l=1}^{L_j} \Pr\{R_l^j = 1 | R_x^j = 1\}^{w_0}.\end{aligned} \tag{4}$$

(4) 式における重み w_0 は全ユーザで同一の値 ($w_0 \leq 1$) を取り，実験的に良い値を求めている．

4　提案手法

Wang らの手法では，既購買アイテム同士を区別せず等しい重みを乗じている．しかし，相関のあるアイテムをクラスタリングすると，多種多様なクラスタができると考えられるため，これらのサイズを正規化するような重みづけをせず同一の重みを乗じた場合，推薦アイテムがサイズの大きいクラスタに依存してしまう可能性がある．そこで，そのサイズに応じた重みを (4) 式の第二項に乗じることで，推薦精度が向上すると考えられる．

アイテム集合 $\mathcal{I}$ を購買履歴を基に重複のない C 個のクラスタに分割し，クラスタ c に属しているアイテムには条件付き確率項に重み w_c を乗じることで，サイズの大きいクラスタの影響を緩和する．クラスタの集合を $\mathcal{C} = \{g_c : 1 \leq c \leq C\}$，クラスタ g_c に含まれるアイテム数を $|g_c|$ とし，そのクラスタに属するアイテム毎への重み w_c を $w_0/|g_c|$ とすると，求める確率は以下の式のように表わせる．

$$\Pr\{R_x^j = 1 | R_1^j = 1, \cdots, R_{L_j}^j = 1\},$$

$$= \Pr\{R_x^j = 1\} \times \prod_{c=1}^{C} \prod_{I_l^j \in g_c} \Big\{ \Pr\{R_l^j = 1 | R_x^j = 1\} \Big\}^{w_c}, \tag{5}$$

$$w_c \quad := \quad w_0 / |g_c|. \tag{6}$$

一方，相関のあるアイテムをクラスタリングする手法を考える．本研究では，階層的クラスタリング手法であるウォード法と，非階層的クラスタリング手法である k-means 法の 2 手法 [2] を適用し，推薦精度の差を確認する．

5 実験

5.1 実験条件

本実験では，推薦システムのベンチマークであるデータセット MovieLens を用いた．このデータは 1997 年 9 月から 1998 年 4 月までに集められた映画の評価データである．ユーザ数 $J = 943$，映画数 $M = 1682$，総データ数 10 万件であり，学習データ 8 万件とテストデータ 2 万件に分けられている．評価値は 5 段階評価であり，評価済みのデータを 1，未評価のデータを 0 とする購買・非購買のデータを用い購買確率の推定を行う．当該データにおいて，ユーザは全てのアイテムのうち最低 20 件以上に評価を与えている．従来手法において重み w_0 は経験値が与えられていたが，本研究ではデータセットに最も当てはまる重みを用いるために，クロスバリデーションにより重み w_0 を推定し，$w_0 = 0.1$ を得た．

実験は，提案手法におけるクラスタ数の変化による推薦精度の違いを確認するため，クラスタ数を $10 \leq C \leq 120$ とし，各クラスタリング手法による Top10 精度を確認した．提案手法の有効性を示すため，最も推薦精度の高くなるクラスタ構成を用いて Top1,10,20 精度を求めた．なお，比較手法として NB 法，重み $w_0 = 0.1$ としたときの従来手法 (improved naive bayes 法) を用いた．

5.2 評価方法

本研究では推薦システムの評価指標として一般的な Top$N$ 精度を用いて，各手法の評価を行う．Top$N$ 精度は，

$$\text{Top}N \quad = \quad \frac{A}{N \cdot J}, \tag{7}$$

によって求めることができる．ただし，N は各ユーザに推薦するアイテム数，A は推薦したアイテムのうち，実際に被推薦ユーザ $U_1, U_2, \cdots, U_J$ が購買しているアイテムの数とする．この評価指標を用いることで推薦の精度を比較することが可能となる．

5.3 実験結果

図 1 にクラスタ数を変化させた場合の Top10 精度を示す．また，図 1 においてクラスタ数が 1 のときは，従来手法である improved naive bayes 法である．ウォード法のクラスタ数は 40，k-means 法のクラスタ数は 30 の場合の Top10 精度が最も高くなっており，以降では徐々に精度が下がっていることが分かる．また，図 2 に最も推薦精度の高いクラスタ構成を用いて Top1,10,20 精度を算出した結果を示す．

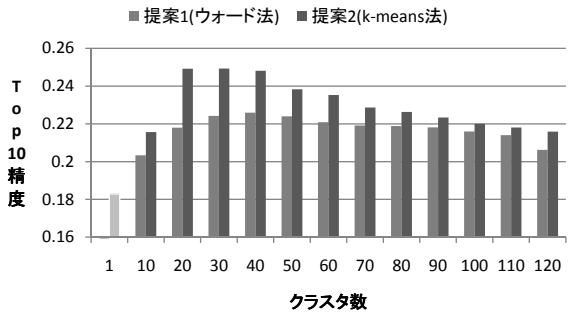

図 1. クラスタ数と Top10 精度

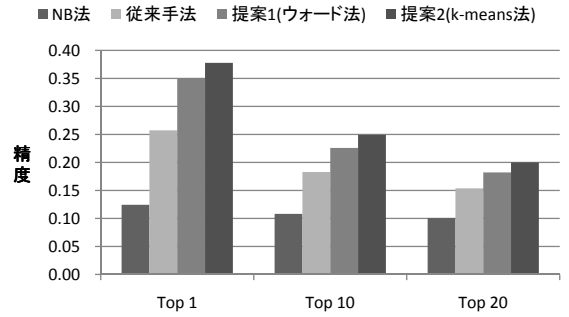

図 2. Top1,10,20 の Top$N$ 精度

図 2 より，アイテムのクラスタリングを行い，クラスタ毎に異なる重みを乗じる提案法の有効性が示せた．また，クラスタリング手法については，実験の結果から，k-means 法の方がより高い推薦精度を得られることが明らかとなった．

5.4 考察

図 1 より，どちらのクラスタリング手法もクラスタ数により推薦精度に差が生じることが分かった．これは，クラスタ数が少ないと相関のあまりないアイテムが同一のクラスタに入り，クラスタ数が多いと相関のあるアイテムが違うクラスタに入るため，いずれの場合もアイテム間の相関を適切に表現できていないためと考えられる．よって，推薦精度を向上させるためには，データセットに応じて最適なクラスタ数を選択する必要があると考えられる．

また，提案手法は Top$N$ 精度の面で従来手法よりも優れた結果を示した．これは，相関のあるクラスタのサイズを反映した重みを乗じることで，より適切に尤度の調節がなされたためと考えられる．

6 まとめと今後の課題

本研究では NB 法を用いた推薦システムに基づくアイテム間の相関を考慮した新たなモデルを提案した．さらにベンチマークを用いた実験により推薦精度が向上することを示した．今後の課題として，ダイス係数等を用いることで，クラスタ内のアイテム間の相関の強弱を考慮するなど，更に適切な重みを決定することが挙げられる．また，最も推薦精度の高くなるクラスタ構成を実験的ではなく自動で求めるようなクラスタリング手法を構築することも今後の課題である．

参考文献

[1] K. Wang and Y. Tan, “A new Collaborative Filtering Recommendation Approach Based on Naive Bayesian Method,” *ICSI 2011, Part II, LNCS 6729, pp. 218-227, 2011.*

[2] 金 明哲，村上 征勝，永田 昌明，大津 起夫，山西 健司，“言語と心理の統計，” 岩波書店*, 2003.*