F-015

分散重み付き相関係数を用いた協調フィルタリング手法

A Method of Collaborative Filtering Using Correlation Coefficient Weighted by Variance

岡 遼太郎† 長井 隆行‡
Ryotaro Oka Takayuki Nagai

１．はじめに

近年,膨大なデータから利用者や消費者のユーザの要求を自動的に推定し,その要求を満たす情報を推薦するシステムが研究されている.そのシステムの手法として協調フィルタリングがある.これは,多くのユーザの嗜好情報を蓄積し,利用者が嗜好の類似した他の利用者の情報を用いて自動的にその利用者が望むであろう情報を決定する方法である.この協調フィルタリングの代表的アルゴリズムに相関係数法[1]がある.この手法は,あるアイテムが好きか嫌いかという多数のユーザの多段階評価値について,評価値の相関係数を用いて特定のユーザの評価値を予測する.しかし評価値が多いアイテム,得点が高いアイテムなどアイテムに投票された情報に差異があるにもかかわらず,全てのアイテムが同様とみなしている問題がある.そこで多様な評価値であるアイテムにより重みを置く entropy 手法、評価値にかかわらず評価した人数が少ないほどよりアイテムに重みを置く inverse 手法がある[2].本研究では各アイテムの評価値の標準偏差に基づく重みを導入する手法を提案し,よく使われるベンチマークデータでの有効性を示し,その理由を示した。

2. 相関法に基づく協調フィルタリング

ここでは相関法に基づく協調フィルタリングで最も基本である手法（これを本論文で **basic** または**ベースライン手法**とよぶ）を示す[1].

行がアイテムを表し,列がユーザを表す行列を $M=(M_{ik})$ と書く.その (i,k) 要素を i 番目のユーザ u の k 番目のアイテム v_k への評価値とする.また評価対象ユーザ i,データベース中のユーザ j とする.以上の値を用いて相関係数 C_{ij},未知の評価値 $\hat{M}_{ik}$ を以下に算出する.

$$C_{ij}=\frac{\sum_k (M_{ik}-\overline{M}_i)(M_{jk}-\overline{M}_j)}{\sqrt{\sum_k (M_{ik}-\overline{M}_i)^2}\sqrt{\sum_k (M_{jk}-\overline{M}_j)^2}} \quad (1)$$

ただし,ここでの k に関する和は評価対象ユーザ i とデータベース中のユーザ j の両方で評価値が既知のアイテムについてのみとる.以上の値を用いて未知の評価値の推定値 $\hat{M}_{ix}$ を以下のように算出する.

$$\hat{M}_{ik}=M_i+\sum_j \frac{C_{ij}(M_{jk}-M_j)}{|C_{ij}|} \quad (2)$$

ただし,ここでの k に関する和は評価対象ユーザ i とデータベース中のユーザ j の両方で評価値が既知のアイテムについてのみとる.

3. 重み付け相関係数法

相関係数法はすべてのアイテムを同等にみなしているが,実際には推測されるアイテムに大きく影響を与えるアイテムとそうでないアイテムがあると考えられる.このようにアイテムによる差異を考慮する手法として,エントロピー,inverse user frequency[2]を用いた手法が提案されている[2].それぞれアイテム k に対応する重みを w_k として与えている.その式を式(3)に示す.標準的な相関係数法は全てのアイテムを同等とみなし,式(3)においての重み w を 1 としている.

$$C_{ij}=\frac{\sum_k w(M_{ik}-\overline{M}_i)(M_{jk}-\overline{M}_j)}{\sqrt{\sum_k w(M_{ik}-\overline{M}_i)^2}\sqrt{\sum_k w(M_{jk}-\overline{M}_j)^2}} \quad (3)$$

3.1 Entropy 手法

評価のばらつきが大きいアイテムほど予測にとって有益なアイテムであると仮定,ばらつきを測る尺度として entropy[2]を用いる.まず,アイテム k に対する entropy を以下の式のように定義する.

$$H_k=-\sum_k \frac{N_v(k)}{N(k)}\log\frac{N_v(k)}{N(k)} \quad (4)$$

ただし, $N_v(k)$ はアイテム k を評価値 v と評価したユーザ数, $N(k)$ はアイテム k を評価したユーザの総数.

さらに,entropy を用いて,アイテム k に対する重みを以下のように定義する.ただし, $H_{\max}$ は全アイテムの entropy の最大値.

$$w_k=\frac{H_k}{H_{\max}} \quad (5)$$

3.2 Inverse user frequency 手法

多数のユーザに見られたアイテムとほとんど見たユーザがいないアイテムがある.その評価は,映画タイトル集合の中で偏って現れるアイテムの方がその映画タイトルを特徴づけると仮定する.ここで重みを以下に定義する.

$$w_k=\log\frac{N}{N_k} \quad (6)$$

ただし N は全ユーザの数, N_k はアイテム k を評価した全ユーザの数とする.

4. 提案手法

4.1 分散重み付き相関係数

従来の相関係数法とは異なり,多数の人が同程度の評価をしているアイテムが予測にとって有益だと仮定し,その尺度として, アイテム評価値の標準偏差の逆数と定め,式(1)に重みとして取り入れた式(7)を新たなアルゴリズムとして提案する.これはマハラノビス距離の概念に類似する考えを取り入れたものである.従来法はベクトル間類似度

†電気通信大学大学院情報システム科情報システム基盤学専攻
‡電気通信大学大学院電子工学専攻

を計算していたものであるが,標準偏差の逆数を取り入れることによりベクトルとベクトル集団との類似度の定義にする.

$$C_{ij}=\frac{\sum_k \frac{1}{\sigma_k(j)}(M_{ik}-M_i)(M_{jk}-M_j)}{\sqrt{\sum_k (M_{ik}-M_i)^2}\sqrt{\sum_k \left(\frac{1}{\sigma_k(j)}\right)^2 (M_{jk}-M_j)^2}} \quad (7)$$

4.2 クラスタ化

ユーザ間の差異をより反映したアイテムの評価値を得るために類似するユーザをクラスタ化し,それぞれのクラスタの標準偏差を用いて,相関係数を計算した.クラスタ化の手法としては,ユーザの属性（性別,職業,etc）と評価値に基づいた k-means 法を用いた.

5．性能評価実験

5.1 利用データベース

評価データには benchmark databse ある Movie Lens データ[3]を用いた.Movie Lens データは 943 人の 1682 本の映画に対する 5 段階評価データである.評価されている項目は約 100,000 個あり欠損率は 93.7%である．欠損率の高い理由として実環境を想定しているためである.

5.2 実験方法

本実験では,テストデータのうち一つをマスクし,その箇所を予測し,評価する方式で,既知のユーザの評価情報が十分にある状態での推薦システムの振る舞いをみた.

5.3 評価方法

提案手法を絶対平均誤差(MAE)を用いて評価した.

$$MAE=\frac{\sum_{l=1}^{N}\left|\hat{M}_{ik}(l)-M_{ik}(l)\right|}{N}$$

ここで N は実験回数，$\hat{M}_{ik}(l)$ は l 回目の実験の予測値，$M_{ik}(l)$ は l 回目の実験の正解値である．i はユーザ番号，k はアイテム番号，l 回目の評価用テストアイテムと決めると，i と k が定まる.

5.4 実験結果と考察

[実験１]

データに付くラベルによるグループ分けを行い,提案手法を用いて推定値をもとめた.図１に従来手法（basic, entropy, inverse）からの精度向上比率の結果を示す.図１から,従来法(basic, entropy, inverse)に比べ,全ての場合において提案手法による精度の向上が見られた.

[実験２]

k-means 法でクラスタリングを行い,[実験１]同様推定値を求めた.実験結果は図２に示す.[実験１]同様に提案手法の有効性が確認された.またクラスタ数が小さいほど精度の向上が見られた.

[考察]

[実験１][実験２]から明らかのように,従来手法より性能の向上がみられた.最も良い結果がクラスタ１のときで11.2%の精度向上が見られた.その理由として標準偏差による重み付けの有効性を示す.またクラスタ数が増加により推測値が下がる理由として,対象データベースの欠損率の多さから,クラスタ数増加により分散の信頼度が低下しているためと考えられる.

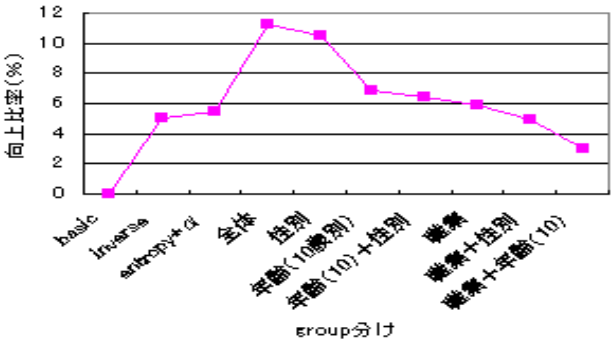

図1　提案手法の従来法からの向上比率

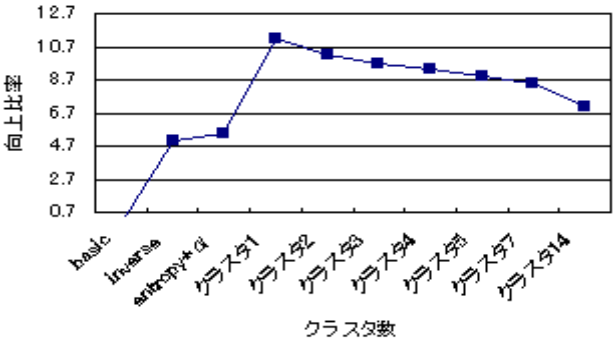

図2　提案手法の従来法からの向上比率

おわりに

本研究では,新しい相関係数法を提案した.また実験で従来手法に比べ予測精度の向上を示した.提案方式が有効である理由はマハラノビス距離がユークリッド距離より識別で有効である理由と同様であると考えられる．マハラノビス距離がより大きなベクトルの距離要素の割合を全体の中でより小さくすることにより,分離性を高くしているの同様、通常の相関係数計算(ユークリッド距離に対応)に比べ,提案方式は,より分散の大きな要素の貢献度がより少なくなる．また提案手法の重みと他の重みとの違いは,ばらつき抽出の違いにある．今後はさらに他のデータへの応用を考えていきたい.

参考文献

[1]John S. Breese, David Heckerman, Carl Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, 1998 , in Proc. of the 14th Conference on Uncertainly in Artificial Intelligence, pp.43-52, 1998.

[2] Kai Yu, Xhong Wen, Xiowei Xu, Martin Ester, “Feture Weighting and instance Selection for Collaborative Filtering”, Proc. of the 2ndInternational Workshop on Management of Information on the Web, 2001.

[3] http://www.cw.umn.edu/Research/GroupLens

[4] 高島秀佳,山岸英貴,平澤茂一“欠損値推定による協調フィルタリング手法”,FIT2005（第 4 回情報科学技術フォーラム），A-008，2005