コピュラを用いたユーザプロファイリング手法の提案

鈴木 崇弘† 欅 惇志†† 宮崎 純††

† 東京工業大学工学部情報工学科 〒 151–8550 東京都目黒区大岡山 2 丁目 12-1
†† 東京工業大学情報理工学院 〒 151–8550 東京都目黒区大岡山 2 丁目 12-1
E-mail: †{suzuki,keyaki}@lsc.cs.titech.ac.jp, ††miyazaki@cs.titech.ac.jp

あらまし 本研究では, Copula を用いたユーザプロファイリング手法の提案する. 情報推薦システムの代表的なアルゴリズムの一つであるコンテンツベースフィルタリングでは, ユーザが好むアイテムを教師データとして嗜好モデルを構築し, それに基づき推薦を行う. 嗜好モデルは, SVM やニューラルネットワークといった機械学習手法により構築することが可能だが, これらの手法は学習結果に対する理由付けが困難であるという問題がある. このような問題に対し, 変数間の依存関係を捉えることが可能かつ, 学習結果の理解が容易な確率モデルである Copula を用いて構築した嗜好モデルに基いてアイテムのランキング付けを行う推薦手法を提案する.

キーワード コピュラ, 情報推薦, コンテンツベースフィルタリング, 機械学習

1. はじめに

近年, インターネットの普及, 及び, Web 技術の進歩により大量の情報が発信され, 人々はそれらを容易に取得できるようになった. その一方で, 大量の情報の中から自分にとって有益な情報を選ぶことの困難さは増してきている. このような問題を解決するために, ユーザの嗜好を汲み取り個人に適したアイテムを提供する情報推薦技術が考案され, 近年注目を集めている.

情報推薦のためのアルゴリズムはこれまでに様々なものが提案されているが, 代表的なものとして, 協調フィルタリング [1] [2] とコンテンツベースフィルタリング [3] がある. 前者には大規模なユーザ数とそのアイテムの評価履歴が必要となるため, 新規アイテムの推薦を行いたい場合や推薦システムの利用者が少ない場合は有効な推薦が行うことができない. 後者はアイテムが持つ特徴とユーザの嗜好情報（ユーザプロファイル）を照合して推薦を行うため, そのような問題は起きない. 本研究はこのコンテンツベースフィルタリングに焦点をあてる.

ユーザプロファイルの作成は, ユーザが過去に関心を示したアイテムの特徴パラメータ（商品の価格, 属性など）を機械学習手法を用いて学習することで行えるが, 多くの機械学習には学習結果の解釈が容易でないという問題が存在する. 情報推薦を行うサービスにとって, 学習結果から新たな知見を得ることは大きな意味があるといえる. このような従来の機械学習手法が抱える問題点に対し, 近年, コピュラを用いて複数パラメータの統合を行う試みが情報検索の分野で行われている [7] [8] [9]. コピュラとは金融工学で用いられてきた確率モデルである. 変数間の依存関係を考慮することができ, また各変数とそれらの依存関係を別々に分析することができるため学習結果の解釈も容易となる.

このコピュラを情報推薦におけるユーザプロファイルの作成に用いることで, 前述の機械学習の問題点を解消することが期待できる. しかし, 推薦アイテムの特徴全てをユーザが考慮しているとは限らないため, コピュラを用いて全ての特徴パラメータを統合すると, ユーザが全く考慮していないパラメータの値が全体のスコアに影響し推薦精度が落ちる可能性が考えられる.

そこで本研究では, ユーザの各特徴パラメータへの関心度を KL ダイバージェンス [10] を用いて定義し, それらの値を元に特徴パラメータの次元削減や重みを大きくする特徴パラメータを決定した後にコピュラを用いてアイテムのランキング付けを行う手法を提案する.

評価実験を行った結果, 提案手法度は, 全特徴パラメータを用いて構築したコピュラによるランキング手法と比べ, 少なくとも有意水準 5%で統計的有意に精度が向上したことを示した. また, 従来の機械学習を用いた手法との比較も行い, 同程度かそれ以上の精度で推薦が行えることを示した.

2. コピュラの概要

2.1 基本的な性質

k 次元の確率変数ベクトル $X = (x_1, x_2, ..., x_k)$ を考える. それぞれの累積分布関数を $F_k(x) = P[X_k \leqq x]$ とすると, 確率変数ベクトル X を以下のように k 次元単位立方空間 $[0,1]^k$ に写像できる.

$$U = (u_1, u_2, ..., u_k) = (F_1(x_1), F_2(x_2), ..., F_k(x_k))$$

このとき k 次元同時累積分布 $F(x_1, x_2, ..., x_n)$ はある関数 C を用いて,

$$\begin{aligned} F(x_1, x_2, ..., x_n) &= C(F_1(x_1), F_2(x_2), ..., F_k(x_k)) \\ &= C(U) \end{aligned}$$

と表せることがスクラーの定理 [11] で知られている. この関数 C がコピュラであり, 周辺分布間の依存関係を表す. 同時分布を構築する際, 各周辺分布のパラメータとコピュラのパラメータは個別に推定することができるため, 柔軟なモデリングが行うことができ, モデルの解釈も容易となる. 以下にコピュラが持つ性質を示す.

- $C(u_1, u_2, ..., u_k)$ は単調増加関数である.

- U の要素のうち, ある一つの要素 $u_i(i = 1, ..., k)$ 以外の要素が全て 1 ならば, C の値は u_i と一致する. すなわち,

$$C(1, ..., 1, u_i, 1, ..., 1) = u_i$$

- U の要素のうち, 少なくとも一つの要素が 0 ならば, C の値は 0 となる. すなわち,

$$C(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_k) = 0$$

2.2 代表的なコピュラ

コピュラが満たすべき性質を持つ関数には様々な種類があり, それぞれ性質や利点が異なる. 同時分布を推定する際には, 分布の特性に応じて適切にコピュラを選択する必要がある.

- **特殊なコピュラ**

各確率変数が独立である場合, コピュラは積コピュラ C_{indep} で表される.

$$C_{indep}(U) = \prod_{i=1}^{k} u_i \tag{1}$$

各確率変数間に完全な正の相関がある場合, コピュラは以下の式で表される.

$$M(U) = min\{u_1, u_2, ..., u_k\} \tag{2}$$

各確率変数間に完全な負の相関がある場合, コピュラは以下の式で表される (ただし, $k = 2$ のときのみ).

$$W(U) = max\{\sum_{i=1}^{k} u_i + 1 - k, 0\} \tag{3}$$

また任意のコピュラ C に対し, 以下が成り立つ.

$$W(U) \leqq C(U) \leqq M(U) \tag{4}$$

これはフレシェ-ヘフディング境界と呼ばれ, M と W はコピュラの上限と下限であることを表している.

- **パラメトリックなコピュラ**

上記のコピュラはエッジケースであり, 一般にコピュラはパラメータを持つ関数として表現される. 応用上よく用いられているものとして, アルキメデス型コピュラと総称されるコピュラがある. その中で代表的なものを以下に示す.

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k} \frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \tag{5}$$

$$C_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k}(-log(u_i))^{\theta})^{\frac{1}{\theta}}) \tag{6}$$

$$C_{Frank}(U) = \frac{1}{\theta} log(1 + \frac{\prod_{i=1}^{k}(\exp(-\theta\ u_i) - 1)}{\exp((-\theta) - 1)^{k-1}}) \tag{7}$$

これらは順にクレイトンコピュラ, グンベルコピュラ, フランクコピュラと呼ばれる. パラメータ θ は依存関係の度合いを表しており, 例えばグンベルコピュラでは $\theta = 1$ のとき C_{indep} に一致し, $\theta \to \infty$ のとき M に一致する. つまり変数間の依存関係が強いほど, θ の値は大きくなる.

またこれらのコピュラはそれぞれ異なる性質を持っている. U の任意の成分 u_i について,

– クレイトンコピュラ, u_i が 1 付近よりも 0 付近で相関関係が高くなる.

– グンベルコピュラは, u_i が 0 付近よりも 1 付近で相関関係が高くなる.

– フランクコピュラは u_i が 0 付近と 1 付近での相関は等しく, 0.5 付近で最も相関が高まる.

クレイトン, グンベルのように分布の裾部分で相関関係が高まるような性質を裾依存性という.

アルキメデス型以外によく用いられるコピュラとして, 正規コピュラがある.

$$C_{Gaussian}(U) = \Phi_{\Sigma}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_k)) \tag{8}$$

この式は多変量正規分布を変形することにより得られる. ここで Φ は標準正規分布の累積密度関数を,Φ^{-1} はその逆関数を示す. パラメータとして共分散行列 $\Sigma \in R^{k \times k}$ を持つ.

- **パラメータの推定方法**

コピュラのパラメータを推定は, まず各周辺分布のパラメータ推定を行い, その後コピュラのパラメータ推定を行うという順序で行えばよい. コピュラのパラメータ推定は最尤推定で行うことができる.

3. 関 連 研 究

本節では, 情報推薦に関する先行研究と, 情報検索分野で用いられているコピュラによる適合度統合式について述べる.

3.1 情報推薦に関する研究

情報推薦のアルゴリズムは協調フィルタリング [1] [2] とコンテンツベースフィルタリング [3] に大別される.

協調フィルタリングは推薦対象ユーザがアイテムにつけた評価値を利用して, 推薦対象ユーザと似た嗜好を持つ近傍ユーザを特定する. その後, 近傍ユーザが好むアイテムを推薦対象ユーザを推薦する. ユーザがつけた評価値のみを用いるため推薦対象アイテムについての知識がなくとも推薦が行える, ユーザにとって意外性のあるアイテムの推薦を行うことが可能など様々な利点がある. しかし, 大規模なユーザ数とアイテムへの評価履歴がない場合, 有効な推薦を行うことができない.

コンテンツベースフィルタリングは, 推薦対象ユーザがどのような特徴を持つアイテムを好むのかという情報（ユーザプロファイル）を獲得し, 嗜好情報とアイテム特徴の類似度を算出して推薦を行う. 協調フィルタリングとは異なり, 推薦システムの利用者が少数の場合でも推薦を行うことができる. ユーザプロファイルの獲得方法は明示的手法と暗黙的手法に分けることができる.

明示的手法は, いずれの特徴・属性を持ったアイテムを好むのかをユーザに質問し, 回答データをユーザプロファイルとする. このような手法は, ユーザ自身が自分の嗜好を正確に把握しきれていない可能性があることや, これらの情報を元にして行う各特徴パラメータの重み付け線形和スコアリングではユーザの嗜好を表現しきれないことが岸田らの研究 [4] で示唆されている.

一方, 暗黙的手法はユーザが過去に関心を示したアイテムや, ユーザの嗜好情報取得のための一部アイテムへの評価結果を教師データとして SVM [5] やニューラルネットワーク [6] といった機械学習手法を用いて, ユーザプロファイルを作成する. 代表的な手法として, ランキング SVM [15] がある. このアプローチでは, ユーザの潜在的な嗜好を汲み取ることが可能であり, また特徴パラメータ間の非線形な関係を捉えることできるため前述の問題点を解決することができる. しかしながら, 学習結果からユーザの傾向の分析を行うことはである.

3.2 コピュラを用いた適合度統合式

情報検索の分野では, 多様化された検索要求に応えるために, 複数の検索モデルで算出された適合度を統合することによって検索精度を向上させる研究がなされてきた. Eickhoff ら [7] はコピュラを適合度統合に応用し, 以下の式を統合式として提案した.

$$C_{prod}(U_{rel}) = C(U_{rel}) \prod_{i=1}^{n} u_{rel,i} \tag{9}$$

ここで, U_{rel} は正解文書の適合度の周辺確率分布の n 次元ベクトルである. 尤度を表す周辺分布の積にコピュラを掛け合わせることで, 変数間の依存関係を考慮した適合度の統合が可能となる. 評価実験を行った結果, いくつかのデータセットでコピュラ統合式が線形結合よりも有効であることを示した. また, 様々なコピュラを用いて比較を行った結果, 情報検索のタスクにおいてはグンベルコピュラを用いることが適切であることを示した [8].

Komatsuda ら [9] は単一のコピュラでは多峰的な同時分布を表現できないことを指摘し, 複数のコピュラの重み付け和で同時分布を表現する混合コピュラを用いた統合式を提案した. 混合コピュラを構築する手順を以下に示す.

（1） 適合文書のクラスタリングを行う.

（2） クラスタごとに周辺分布, コピュラのパラメータ推定を行う.

（3） 各クラスタのコピュラを足し合わせ混合コピュラを算出する. 以下に混合コピュラの式を示す.

$$C_{mix}(U_{rel}) = \sum_{c=1}^{k} p_c C_c(U_{rel,c}) \tag{10}$$

ここで, k はクラスタ数, p_c はクラスタ毎の重みで, c 番目のクラスタに属する適合文書の割合である.

Komatsuda らは式 (10) に加え, Eickoff らの式 (9) を混合コピュラ用に拡張した式

$$C_{mix-prod}(U) = C_{mix}(U_{rel}) \prod_{i=1}^{n} \sum_{c=1}^{k} p_c u_{rel,c,i} \tag{11}$$

を適合度統合式として提案した. $u_{rel,c,i}$ は c 番目のクラスタに属する適合文書の i 番目の適合度の周辺分布を表す. 評価実験の結果, これらの統合式は Eickhoff らの式よりも精度が高く, C_{mix} よりも $C_{mix-prod}$ の方が精度が高いことが示された.

4. 提 案 手 法

推薦対象のアイテムが持つ各特徴パラメータは, ある属性を満たすかどうかの二値 {0, 1} の特徴パラメータ（「期間限定」や「特典あり」など）や, 商品の価格やユーザレビュー値などの連続値で表現されることが多い [13]. 二値属性の特徴パラメータは, 推薦アイテムのスコアリングよりもむしろ, 推薦対象アイテムのフィルタリングに利用されることが想定されるため, 本研究で扱う推薦対象アイテムの持つ特徴パラメータは, 連続値で表現されているものとする.

ここで, コピュラを用いた情報検索の情報推薦への適用を検討した場合, 適合文書を適合アイテム, 各適合度をアイテムの各特徴パラメータと読み替えることで, コピュラによる適合度統合式を情報推薦のユーザープロファイリングに適用することができる. その際, 情報推薦のコンテキストにおいては, 全ての特徴パラメータを使用することが必ずしも適切とは限らない. 文書の適合度はユーザに依存せず客観的に表現することが可能であり, 高精度検索に貢献することが検証されているスコア関数やスコア関数算出に用いられる統計量により定義される. それに対して, ユーザの嗜好はユーザごとに異なり, 更に必ずしも全ての特徴を考慮しているわけではない. 従って, 特徴パラメータ全てを適合度として読み替えることは適切ではない. また, 一般に機械学習では各特徴パラメータに重み付けを行うが, コピュラによる統合式は対称式になっており, 特定の特徴について重み付けされていない.

そこで本研究では, ユーザの各特徴パラメータへの関心度を KL ダイバージェンスを用いて定義し, 関心度の値を元に特徴パラメータの次元削減及び重要な特徴パラメータの検出を行った後に, それらを元にしたコピュラ統合式によるスコアリングを行う手法を提案する.

提案手法の概要は以下の通りである.

（1） 各特徴パラメータへのユーザの関心度を算出する.

（2） 関心度の値を元に特徴パラメータの次元削減を行う. また, ユーザが特に重要視している特徴パラメータを特定する.

（3） 混合コピュラの推定を行う.

（4） 混合コピュラと手順 (2) で特定したユーザが重要視している特徴パラメータを用いて提案スコアリング式を構築する.

以降では, 各手順の詳細を述べる.

4.1 関心度の算出

ユーザが好むアイテム (以下嗜好アイテムとする) を用いて, ユーザがどの特徴パラメータにどの程度関心があるかを算出する. 関心を示していないパラメータの場合, ユーザがそのパラメータを考慮せずに嗜好アイテムを選定しているため, そのパラメータの値は無作為に選ばれてるとみなせる. そのため, 嗜好アイテム中の関心を示していない特徴パラメータの値の分布は, 全アイテムのその特徴パラメータの値の分布に比較的近くなると考えられる. 逆に, ユーザがあるパラメータに注目してアイテムの良し悪しを判断している場合, 嗜好アイテム中のその特徴パラメータの値の分布は, 全アイテムのその特徴パラメータの値の分布と異なる分布になると考えられる. 以上のことか

ら, 分布間の距離を測る KL ダイバージェンスを用いて, i 番目の特徴パラメータに対するユーザの関心度 att_i を以下のように定義する.

$$\begin{aligned} att_i &= \log_{1p}(D_{\mathrm{KL}}(ALL_i \| User_i)) \\ &= \log_{1p}(\int_{-\infty}^{\infty} f_{all}(x_i) \log \frac{f_{all}(x_i)}{f_{user}(x_i)}\, dx) \end{aligned} \tag{12}$$

ここで, $f_{all}(x_i)$ は全アイテムの i 番目の特徴パラメータの確率密度関数, $f_{user}(x_i)$ はユーザ嗜好アイテムの i 番目の特徴パラメータの確率密度関数を表す. $log_{1p}(x)$ は $log(1+x)$ を表し, 関心度の取りうる値が大きくなりすぎないようにするための補正として用いている.

今回は全アイテム, 嗜好アイテムの各特徴パラメータの分布は共に正規分布を用いて推定を行った. そのため, att_i は全アイテムの平均 $\mu_{all,i}$, 分散 $\sigma_{all,i}$, 嗜好アイテムの平均 $\mu_{user,i}$ 分散 $\sigma_{user,i}$ を用いて以下のように表せる.

$$\log_{1p}(\log \frac{\sigma_{user,i}}{\sigma_{all,i}} + \frac{\sigma_{all,i}^2 + (\mu_{all,i} - \mu_{user,i})^2}{2\sigma_{user,i}^2} - \frac{1}{2})$$

4.2 次元削減と重要な特徴パラメータの決定

算出した関心度を元に, パラメータの次元削減と重要な特徴パラメータの決定を行う. 具体的には, 算出した関心度のうち値の低いものはユーザが無関心なパラメータとみなし, コピュラによる統合対象から外すこととする. また関心度が他のパラメータに比べ高い場合は, ユーザがそのパラメータをとても重要視しているとみなし, 他のパラメータと比べより大きな重みを付与する. そこで各特徴パラメータへの関心度を降順に並べ, 下側の外れ値を特徴パラメータから取り除き, 上側の外れ値を重要なパラメータに分類することを目標とする.

外れ値の検出には平均と標準偏差を用いて行うことが可能だが, いずれの値も外れ値の影響を受けやすいため, 代わりにロバストな推定量を用いる. 平均 μ の推定量として中央値 $Med(X)$ を, 標準偏差 σ の推定量として中央値からのばらつきを表す中央絶対偏差 $MAD(X)$ を正規化した値 $MADN(X)$ を用いた.

$$MAD(X) = Med(\{|x_i - Med(X)|\}) \tag{13}$$

$$MADN(X) = \frac{MAD(X)}{0.675} \tag{14}$$

$MADN(X)$ は, データの分布を正規分布と仮定した際, 標準偏差の不偏推定量になることが知られている [14]. これらと正定数 a を用いて, $\mu - a\sigma$ 以下の値を下側の外れ値として特徴パラメータから除外し, $\mu + a\sigma$ 以上の値を上側の外れ値として重要な特徴パラメータに分類する. 今回の実験では $a = 2.5$ とした際に最適な結果となったため, これらの値を用いることとした.

外れ値検出の例を示す. 表 1 では 8 つの特徴パラメータ {A, B, C, D, E, F, G, H} に対し, それぞれの関心度が算出されている. 平均値の推定量は 0.71, 標準偏差の推定量は 0.19 となる. このとき下側の閾値は 0.235 であるので, A は特徴パラメータから除外される. また上側の閾値は 1.19 であるので, H をユーザが特に重要視しているパラメータとみなす. 以上より, このユーザのプロファイリングに用いる特徴パラメータは {B, C, D, E, F, G, H} となり, 特に重要視しているパラメータは {H} となる.

表 1 あるユーザの関心度の値

	A	B	C	D	E	F	G	H
関心度	0.033	0.54	0.57	0.64	0.79	0.81	0.83	1.94

4.3 混合コピュラの推定

パラメータの次元削減が完了すれば, コピュラを用いて同時分布の推定を行う. 単峰的なコピュラを用いた場合, 周辺分布の推定は既には関心度の計算時点で行っているので計算の効率化が図ることが可能であるが, 推薦精度の観点から本研究では混合コピュラを用いることとする. 混合コピュラの推定方法は, Komatsuda ら [9] の方法に従った. なお, クラスタリングの結果, クラスタに属する適合アイテムが一つとなった場合, そのクラスタは外れ値として除外する.

4.4 提案スコアリング式

最後に, 推定した混合コピュラ C_{mix}, 手順 2 で算出した, ユーザが特に重要視しているパラメータを用いて提案スコアリング式を構築する. ここでは四種類の式を提案する. そのうち二つは, 従来のスコア式 C_{mix} と $C_{mix-prod}$ を次元削減して構築した式 C_{kl-mix} と $C_{kl-mix-prod}$ である. U_{rdc} は手順 2 において次元削減が行われた特徴パラメータの確率ベクトルを表す.

$$C_{kl-mix}(U) = C_{mix}(U_{rdc}) \tag{15}$$

$$C_{kl-mix-prod}(U) = C_{mix-prod}(U_{rdc}) \tag{16}$$

三つ目の提案式は, 混合コピュラに重要なパラメータの周辺確率をかけ合わせた $C_{kl-emp}U_{rdc}$ である. S は手順 2 で算出した重要な特徴パラメータに対応する添字の集合である.

$$C_{kl-emp}(U) = C_{mix}(U_{rdc}) \prod_{i \in S} \sum_{c=1}^{k} p_c u_{rel,c,i} \tag{17}$$

重要な特徴パラメータの周辺分布のみを掛け合わせることで, その特徴パラメータの重みを大きくする. またこの式は重要な特徴パラメータの集合 S が空集合のとき, C_{mix} に一致する.

四つ目の提案式は, C_{kl-emp} を拡張し, S が空集合の時は $C_{mix-prod}$ に一致するように設計した $C_{kl-emp-prod}$ である.

$$C_{kl-emp-prod}(U) = \begin{cases} C_{mix-prod}(U_{rdc}) & \text{if } S = \emptyset \\ C_{kl-emp}(U) & \text{otherwise} \end{cases} \tag{18}$$

Komatsuda らの研究では C_{mix} よりも $C_{mix-prod}$ の方がより高い精度を示したため, S が空集合の場合に限り全ての特徴パラメータを重要なパラメータとみなすことで, 重要な特徴パラメータが検出されない場合での精度向上が見込める.

5. 評 価 実 験

本節では, 提案手法の有効性検証のために行った評価実験の内容について述べる.

5.1 実験準備

5.1.1 データセット

データセットとして楽天トラベルのホテルデータのうち, 東京 23 区内のホテル 245 件を対象とした. 各ホテルは, 価格, サービスレビュー, 施設レビュー, 部屋レビュー, 立地レビュー, 風呂レビュー, 食事レビュー, 最寄り駅からの直線距離の計 8 種類の情報を持つ. 各レビュー値は楽天トラベル利用者が評価した 1-5 の五段階評価の平均値で, 未評価の場合は 0 となる. 価格はそのホテルの全宿泊プランの価格の中央値である. これらの値の取りうる範囲が [0, 1] になるように正規化を行い, これを特徴パラメータとして用いた.

5.1.2 評価実験のデザイン

被験者は研究室内の大学生及び大学院生 12 人 (平均年齢 23 歳) である. ホテル推薦時のコンテキストの違いによる推薦精度への影響を調査するため, 表 2 に示す四種類のシナリオを用意し, 各被験者はいずれかのシナリオにおいてホテル全件の適合・不適合判定を行った. 各シナリオを担当した人数は, シナリオの番号順に 2 人, 4 人, 3 人, 3 人となっている.

表 2 四種類のシナリオのガイダンス文

シナリオ 1	出張用に泊まるホテルを探している. 会社規定があるため, なるべく安いホテルに泊まりたい.
シナリオ 2	友人と旅行に行く際に宿泊するホテルを探している. 観光に重点を置いた旅行を想定している.
シナリオ 3	観光で宿泊するホテルを探している. 金銭面はかなり余裕があるので, よいホテルに泊まってみたい.
シナリオ 4	恋人と宿泊するホテルを探している. 金銭面はそれほど余裕はないが, 喜んでもらえそうなホテルがよい.

また各人の嗜好を把握するために, 実験終了後に, 表 3 のように, 適合判定の際の各特徴パラメータの優先度を合計が 100 になるように回答してもらった.

表 3 あるユーザの嗜好回答データの例

	価格	サービス	施設	部屋	立地	風呂	食事	距離
回答	50	0	0	0	0	0	10	40

5.2 比較手法

提案手法との比較に用いるスコアリング式について述べる. 以下, X は特徴パラメータを表すベクトルであり, U はそれらの周辺確率ベクトルである.

- **嗜好回答情報を利用した重み付け線形和**

$$LIN(X) = \sum_{i=1}^{n} w_i x_i \tag{19}$$

i 番目の特徴パラメータへの重み w_i 値は, 対応する嗜好回答データの値を 100 で割ったものである. 表 3 を例にすると, 価格への重みは 0.5, 食事レビュー値への重みは 0.1, 距離へ重みは 0.4, それ以外の特徴パラメータへの重みは 0 となる. 比較の目的は, コピュラを用いたパラメータ統合式が, 線形和によるスコアリングよりも有効であるかを検証することである.

- **次元削減を行わないコピュラ統合式**

次元削減を行わず全特徴パラメータを対象にコピュラを推定し, スコアリング式を構築する. スコア統合式として, 式 (10) の C_{mix} 及び式 (11) の $C_{mix-prod}$ を用いる. 比較の目的は, パラメータの次元削減を行うことで推薦制度が向上するかどうかを検証することである.

- **嗜好回答情報に基いて次元削減を行ったコピュラ統合式**

各特徴パラメータへの回答データのうち値が 0 になっているものを無関心な値として除外し, 残った特徴パラメータについてコピュラを推定し, スコアリング式を構築する.

$$C_{u-mix}(U_{rdc}) = C_{mix}(U_{rdc}) \tag{20}$$

$$C_{u-mix-prod}(U_{rdc}) = C_{mix-prod}(U_{rdc}) \tag{21}$$

表 3 を例に出すと, この場合, 次元削減を行ったあとの特徴パラメータベクトルは (価格, 食事レビュー値, 距離) となる. 比較の目的は, KL ダイバージェンスを用いた特徴パラメータの選定が適切に行えているどうかを検証することである.

- **ランキング SVM**

SVM^{rank} (注1) [16] を用いて, ランキング SVM モデルを構築する. コストパラメータ C には SVM^{rank} のデフォルト値である 0.01 を用い, カーネルには RBF カーネルを用いた. カーネルがもつパラメータ γ には, $2^{-10}, 2^{-9}, ..., 2^{9}, 2^{10}$ の候補の中から, 最も精度が高くなった 2^3 を用いた. 比較の目的は, 従来の機械学習手法と比べ, コピュラによる手法が有効かどうかを検証することである.

5.3 評価指標

実験の評価指標として, 推薦結果上位 k 件の精度 ($P@k$) と $nDCG@k$, 及び再現率が i の時点の補間適合率 ($iP@i$) を用いた. $P@k$ は推薦結果上位 k 件のうち, 正しく推薦されたアイテムが占める割合を示す指標. k 件中に含まれる適合アイテムの数を h とすると, $P@k$ は以下の式で表される.

$$P_k = \frac{h}{k} \tag{22}$$

$nDCG@k$ は, 上位 k 件の推薦結果のランキング付けの妥当性を示す指標である. 推薦結果の上位に適合度が高いアイテムが多いほど値が大きくなる指標 $DCG@k$ を, $DCG@k$ の理想値 $iDCG@k$ で割って正規化した値が $nDCG@k$ である. $iDCG$ は推薦結果のアイテムを適合度順にソートしたときの DCG を計算することで求めることができる.

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{23}$$

(注1) : https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 4 実験結果

手法	LIN	C_{mix}	$C_{mix-prod}$	C_{u-mix}	$C_{u-mix-prod}$	SVM^{rank}	C_{kl-mix}	$C_{kl-mix-prod}$	C_{kl-emp}	$C_{kl-emp-prod}$
iP@0.0	0.980	0.995	0.993	0.992	0.991	0.991	0.995	0.987	**0.996**	**0.996**
iP@0.1	0.959	0.991	0.979	0.987	0.991	0.986	**0.992**	0.979	**0.992**	**0.992**
iP@0.2	0.940	0.976	0.964	0.982	0.980	0.981	0.987	0.973	**0.989**	**0.989**
iP@0.3	0.918	0.953	0.944	0.956	**0.967**	0.965	0.949	0.942	0.955	0.964
iP@0.4	0.884	0.922	0.916	0.938	0.945	**0.958**	0.922	0.920	0.933	0.941
iP@0.5	0.849	0.883	0.886	0.922	0.922	**0.945**	0.887	0.889	0.907	0.924
P@5	0.888	0.946	0.946	0.946	0.958	0.946	**0.962**	0.954	**0.962**	**0.962**
P@10	0.863	0.885	0.881	0.913	**0.923**	**0.923**	0.885	0.89	0.898	0.917
P@15	0.799	0.808	0.799	**0.857**	**0.857**	**0.857**	0.817	0.806	0.833	0.844
P@20	0.732	0.740	0.748	0.781	0.782	**0.794**	0.743	0.741	0.763	0.775
nDCG@5	0.960	0.990	0.989	0.987	0.986	0.981	0.991	0.984	0.991	**0.992**
nDCG@10	0.957	0.985	0.982	0.982	0.983	0.980	0.987	0.980	**0.988**	**0.988**
nDCG@15	0.952	0.978	0.976	0.978	0.980	0.977	0.980	0.975	0.981	**0.983**
nDCG@20	0.949	0.973	0.969	0.974	0.976	0.974	0.975	0.969	0.977	**0.979**

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{iDCG@}k} \tag{24}$$

ここで, rel_i は上位 i 番目アイテムの適合度である. 今回の実験では適合ならば 1, 不適合ならば 0 とした.

$iP@i$ は, 再現率が i の時点での推薦精度を示しており, 再現率が i 以上における精度の最大値で表される.

$$\mathrm{iP@}i = \max_k \{P@k | R@k \geqq i\} \tag{25}$$

再現率は推薦結果の網羅性を示す指標である. 上位 k 件中に含まれる適合アイテムを h, 全適合アイテムの数を a とすると, 上位 k 件を取得した際の再現率 $Recall_k$ は以下の式で表される.

$$Recall_k = \frac{h}{a} \tag{26}$$

今回の実験では, $P@k$, $nDCG@k$ は $k = \{5, 10, 15, 20\}$ の時の値を, $iP@i$ は $i = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ のときの値を四分割交差検定を用いて算出した.

5.4 実験に用いるパラメータ

各スコアリングモデルで用いるコピュラのパラメータとして, 周辺分布に正規分布, コピュラの種類はグンベルコピュラを用いた. C_{mix} を推定する際に行うクラスタリング手法は全て k-means 法を用いて行い, クラスタ数は 1-5 で検証した.

5.5 実験結果

実験結果を表 4 に示す. 表中の値は, 各手法とも $iP@0.1$ の精度が最も優れていたときのクラスタ数における結果を示している. 各評価指標において最も高精度な手法を太字で表す. 表から, 線形結合 LIN に比べ, コピュラを用いた統合手法がいずれの指標においても優れていることが確認できる. また, 次元削減を行わない手法 C_{mix}, $C_{mix-prod}$ は次元削減を行っている手法と比べ, 取得件数を増やしたときの精度の劣化が目立つ. これはユーザが注目していない特徴パラメータを考慮してしまっているため, それらの値がノイズとして働いているためであると考えられる. ランキング SVM とコピュラによる手法を比較すると, 全体の精度ではランキング SVM に劣るものの, $nDCG$ や $P@5$ の値から, コピュラによる手法は推薦結果上位の精度が優れていることが確認できる.

提案手法による次元削減方法は C_{mix} や $C_{mix-prod}$ と比較して高精度であったことからも, 次元削減は有効であるという結果が得られた. また, ユーザの回答データを用いて次元削減を行った手法 C_{u-mix}, $C_{u-mix-prod}$ と比べ, 提案手法は $P@k$ においては劣るものの, その他の以外の評価指標においては遜色のない精度を達成し, $nDCG$ や $iP@\{0.0, 0.1, 0.2\}$ に着目すると $C_{kl-mix-prod}$ 以外の提案手法が回答情報を用いたものよりも高い精度を達成した. この結果から, ランキング上位の結果において, ユーザが自己申告した嗜好情報より, より少ない労力で暗黙的に取得した嗜好情報を用いた提案手法がより有効であることを示唆する.

提案手法の内, 重要な特徴パラメータへの重み付けを行う手法 $C_{kl-emp-prod}$ が iP, $nDCG$ 共に高い値を取っており, $P@k$ も取得件数が上がっても劣化が緩やかで回答データを用いた手法に準ずる精度が出ているため, 最も優れた手法であるといえる. $C_{kl-emp-prod}$ が比較手法に対し, 統計的有意に精度が向上したかを確認するために符号検定を用いて片側検定を行った. 標本には, 12 名の各ユーザへの推薦結果上位 20 件のホテルの適合・不適合のラベルをそれぞれ $\{1, 0\}$ として並べたものを用いた. 今回の実験では四分割交差検定を行ったため, 標本数は $960 (= 12 \times 20 \times 4)$ 件となる. 検定の結果, LIN, C_{mix} に対し有意水準 1%で, $C_{mix-prod}$ に対し有意水準 5%で統計的有意に精度が向上したことが確認できた. なお, この手法は重要と判定された特徴パラメータが存在しない場合に $C_{kl-mix-prod}$ と同値となる式であり, その $C_{kl-mix-prod}$ は提案手法中で最も精度が低い. 従って, 重要と判定された特徴パラメータの有無によって, 採用するコピュラ統合式を選択することがより適切であると考えられる.

5.6 シナリオ別の実験結果

続いて, シナリオ別の実験結果から考察する. 各シナリオ別のユーザの嗜好の傾向を確認するために, 関心度の平均値を表 9 に示す. シナリオ 1 では設定の通り, 価格への関心度が極めて高くなっており, それ以外への関心度は 0 に近い. シナリオ 2 では, 具体的な状況設定が曖昧だったためか他のシナリオと比

表 5　シナリオ 1 の実験結果

手法	C_{u-mix}	$C_{u-mix-prod}$	C_{kl-mix}	$C_{kl-mix-prod}$	C_{kl-emp}	$C_{kl-emp-prod}$
iP@0.0	1.00	1.00	1.00	1.00	1.00	1.00
iP@0.1	1.00	1.00	1.00	1.00	1.00	1.00
iP@0.2	0.995	1.00	1.00	1.00	1.00	1.00
iP@0.3	0.995	0.995	0.960	0.912	0.995	0.995
iP@0.4	0.990	0.995	0.960	0.884	0.995	0.995
iP@0.5	0.990	0.987	0.945	0.852	0.995	0.995
P@5	1.00	1.00	1.00	1.00	1.00	1.00
P@10	0.988	0.988	0.95	0.888	0.988	0.988
P@15	0.933	0.933	0.908	0.792	0.975	0.975
P@20	0.894	0.881	0.825	0.763	0.894	0.894
nDCG@5	1.00	1.00	1.00	1.00	1.00	1.00
nDCG@10	0.998	1.00	0.996	0.996	0.999	0.999
nDCG@15	0.996	0.998	0.992	0.983	0.998	0.998
nDCG@20	0.994	0.995	0.990	0.974	0.997	0.997

表 6　シナリオ 2 の実験結果

手法	C_{u-mix}	$C_{u-mix-prod}$	C_{kl-mix}	$C_{kl-mix-prod}$	C_{kl-emp}	$C_{kl-emp-prod}$
iP@0.0	1.00	0.988	0.990	0.969	1.00	1.00
iP@0.1	0.984	0.988	0.990	0.945	0.988	0.988
iP@0.2	0.976	0.960	0.961	0.938	0.972	0.972
iP@0.3	0.962	0.946	0.937	0.913	0.929	0.929
iP@0.4	0.948	0.930	0.896	0.897	0.902	0.902
iP@0.5	0.878	0.881	0.848	0.836	0.858	0.875
P@5	0.950	0.963	0.938	0.925	0.963	0.963
P@10	0.913	0.913	0.900	0.881	0.881	0.888
P@15	0.825	0.817	0.804	0.779	0.813	0.821
P@20	0.769	0.756	0.756	0.725	0.753	0.759
nDCG@5	0.997	0.983	0.985	0.968	0.993	0.993
nDCG@10	0.990	0.979	0.979	0.967	0.988	0.987
nDCG@15	0.986	0.976	0.974	0.963	0.981	0.981
nDCG@20	0.979	0.970	0.967	0.956	0.974	0.974

表 7　シナリオ 3 の実験結果

手法	C_{u-mix}	$C_{u-mix-prod}$	C_{kl-mix}	$C_{kl-mix-prod}$	C_{kl-emp}	$C_{kl-emp-prod}$
iP@0.0	1.00	1.00	1.00	1.00	1.00	1.00
iP@0.1	1.00	1.00	1.00	1.00	1.00	1.00
iP@0.2	0.989	0.992	1.00	1.00	1.00	1.00
iP@0.3	0.962	0.982	0.976	0.976	0.963	0.954
iP@0.4	0.962	0.97	0.96	0.97	0.941	0.943
iP@0.5	0.955	0.955	0.921	0.953	0.935	0.943
P@5	0.950	0.983	0.983	1.00	0.967	0.967
P@10	0.933	0.950	0.933	0.95	0.925	0.933
P@15	0.889	0.889	0.856	0.894	0.883	0.889
P@20	0.817	0.842	0.833	0.854	0.838	0.846
nDCG@5	0.999	1.00	1.00	1.00	1.00	1.00
nDCG@10	0.992	0.997	0.997	0.998	0.994	0.994
nDCG@15	0.988	0.993	0.991	0.993	0.989	0.989
nDCG@20	0.984	0.989	0.984	0.989	0.985	0.985

表 8　シナリオ 4 の実験結果

手法	C_{u-mix}	$C_{u-mix-prod}$	C_{kl-mix}	$C_{kl-mix-prod}$	C_{kl-emp}	$C_{kl-emp-prod}$
iP@0.0	0.981	0.979	0.981	0.975	0.981	0.981
iP@0.1	0.981	0.979	0.981	0.975	0.981	0.981
iP@0.2	0.981	0.979	0.965	0.975	0.965	0.981
iP@0.3	0.936	0.965	0.959	0.971	0.959	0.978
iP@0.4	0.906	0.938	0.938	0.936	0.938	0.943
iP@0.5	0.891	0.910	0.910	0.935	0.906	0.932
P@5	0.900	0.933	0.933	0.933	0.933	0.933
P@10	0.850	0.883	0.833	0.850	0.825	0.867
P@15	0.761	0.789	0.722	0.750	0.733	0.761
P@20	0.629	0.688	0.621	0.671	0.638	0.671
nDCG@5	0.967	0.967	0.974	0.976	0.974	0.978
nDCG@10	0.963	0.965	0.973	0.972	0.972	0.975
nDCG@15	0.960	0.964	0.967	0.968	0.966	0.970
nDCG@20	0.957	0.960	0.959	0.961	0.958	0.965

表 9　シナリオ別のユーザの各特徴パラメータへの関心度の平均値

	価格	サービス	施設	部屋	立地	風呂	食事	距離
シナリオ 1	2.32	0.05	0.06	0.07	0.05	0.07	0.18	0.03
シナリオ 2	0.61	0.13	0.18	0.17	0.05	0.10	0.34	0.24
シナリオ 3	0.10	0.58	0.69	0.68	0.37	0.42	1.29	0.38
シナリオ 4	0.75	0.52	0.57	0.65	0.26	0.38	1.42	0.02

べると関心度が突出した特徴パラメータはなく, 比較的価格に関心があるように見受けられる. シナリオ 3 では金銭的余裕があるという設定のため価格への関心度が最も低い. それ以外の特徴パラメータについては他のシナリオの場合と比べ全体的に関心が高く, 中でも食事への関心が最も高い. シナリオ 4 は関心度の傾向としてシナリオ 3 に近いが, 価格にも高い関心を示していることが分かる. これらの値は各シナリオのガイダンス文と照らし合わせても, 妥当な結果である.

各シナリオの傾向を踏まえた上で, シナリオ別の実験結果を確認する. シナリオ別の回答データを用いたスコア式及び提案式の実験結果を表 5, 6, 7, 8 に示す. 各モデルのハイパーパラメータは表 4 と同一の値を用いた. 回答データを用いて次元削減を行った C_{u-mix} と $C_{u-mix-prod}$ はどのシナリオでも精度が安定している. 正確に次元削減を行えているため両者の精度の差は小さいが, 価格を最重要視していると思われるシナリオ 1,2 では C_{u-mix} が, それ以外のシナリオでは $C_{u-mix-prod}$ の精度が高い. $C_{mix-prod}$ がグンベルコピュラに変数の独立を表す積コピュラをかけ合わせた式であることを考慮すると, 価格重視の場合は変数間の依存関係が強く, それ以外では変数間の依存関係は弱い傾向にあると解釈することができる. 実際, $C_{kl-mix-prod}$ はシナリオ 1 では他の手法と比べ最も精度が低いが, シナリオ 3 では最も精度が高い手法となっている. このことから, 提案手法 $C_{kl-emp-prod}$ の全体的な精度が高くなった理由は, 価格のような突出した関心度の値を保つ場合には重要な特徴パラメータを中心とした依存関係を考慮した式になり, 関心度が突出した特徴パラメータがなく重要な特徴パラメータが検出されない場合はシナリオ 3 などに適した変数間の依存関係が弱い式になるからだと考えられる.

以上の結果より, $C_{kl-emp-prod}$ がどのシナリオにおいても安定して高精度であることを示した. また, 暗黙的に獲得した嗜好情報を用いて, 明示的に獲得した情報を活用する場合と遜色

ない精度を出せることが確認できた．さらに，ランキング上位においては，暗黙的に獲得した嗜好情報を用いるほうが精度が高いという結果が得られた．推薦システムにおいては，ランキング上位の推薦結果が重要であるため，全体的な精度が高く，ランキング上位での精度も高い $C_{kl-emp-prod}$ が最も優れた手法であるといえる．

6. ま と め

本研究では，情報推薦におけるユーザプロファイリングにコピュラを適用することを試みた．既存のコピュラによる特徴パラメータ統合式を情報推薦のコンテキストで適用する際の問題点を指摘し，その解決策として，特徴パラメータの次元削減と重要な特徴パラメータの決定を行い，その後に統合式を構築する手法を提案した．評価実験の結果，提案手法のうち，次元削減を行った後に重要な特徴パラメータに重み付けを行う $C_{kl-emp-prod}$ は，既存のコピュラによる統合式 C_{mix}，$C_{mix-prod}$ に対し，推薦結果上位 20 件取得時の精度がそれぞれ有意水準 1%，5%で統計的有意に向上することを示した．

今後の課題として，今回の実験は対象の人数が 12 人と少ないため，より多くのユーザを対象にした実験を行う必要がある．また，扱う特徴パラメータが増えた場合，提案手法での特徴パラメータの絞り込みが同様に有効であるか検証を行う必要がある．このような場合について，今回の手法では混合コピュラを構築する際，クラスタリングの前に特徴パラメータの絞り込みを行っていたが，クラスタリングを行った後に各クラスタ毎に特徴パラメータの絞り込みを行うという方法を検討したい．さらに今後の展望として，本研究ではアイテムの特徴パラメータが連続値のみで表されることを仮定したが，スコアリング式を累積密度分布ベースから密度ベースのものに変更することで，2 値などの他の形式で表現される特徴パラメータも含めた，統合的な嗜好モデルの構築が期待される．

謝　　辞

本研究の一部は，JSPS 科研費 JP26280115，JP15H02701，JP16H02908，JP15K20990 の助成を受けたものである．また、本研究のデータセットは，楽天株式会社の提供による．ここに記して謝意を表す．

文　　献

[1] Resnick, Paul and Iacovou, Neophytos and Suchak, Mitesh and Bergstrom, Peter and Riedl, John. GroupLens: an open architecture for collaborative filtering of netnews, Proceedings of the 1994 ACM conference on Computer supported cooperative work,pp.175-184,1994.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl. Item-based collaborative filtering recommendation algorithms, Proceedings of the 10th international conference on World Wide Web, pp.285-295, ACM Press, 2001.

[3] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends, In Recommender Systems Handbook, pages 73-105. 2011.

[4] 岸田 脩平, 欅 惇志, 宮崎 純. スカイライン演算を用いたユーザ嗜好を考慮した情報推薦のランキング手法の精度改善について, 第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2016）論文集, 2016.

[5] C. Cortes and V. Vapnik. Support vector networks. Machine Learning, 20:273-297, 1995.

[6] J. Hertz, A. Krogh and R. G. Palmer. Introduction to the theory of neural computation, Vol. 1, Basic Books, 1991.

[7] Carsten Eickhoff, Arjen P de Vries, and Kevyn Collins Thompson. Copulas for information retrieval. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 663–672. ACM, 2013.

[8] Carsten Eickhoff and Arjen P de Vries. Modelling complex relevance spaces with copulas. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 18311834. ACM, 2014.

[9] Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki. A Score Fusion Method Using a Mixture Copula, 27th International Conference on Database and Expert Systems Applications (DEXA 2016), Volume 9828 of LNCS, pp.216-232, Porto, September 2016.

[10] S. Kullback and R. A. Leibler. On information and sufficiency, The Annals ofMathematical Statistics, 22:79-86, 1951.

[11] Sklar, A. Functions de Repartition an Dimension Set Leursmarges, Publications de L ' In-stitut de Statistique de L ' Universite de Paris, 1959.

[12] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm, Applied statistics, pp. 100-108, 1979.

[13] 奥健太, 中島伸介, 宮崎純, 植村俊亮, 加藤博一. 情報推薦におけるユーザの価値判断基準モデルに基づくコンテキスト依存型ランキング方式, 情報処理学会論文誌, データベース, Vol.2, No.1(TOD 41), pp.57-80（2009）.

[14] Huber, P. J. (1981). Robust statistics. New York: John Wiley.

[15] Herbrich, Ralf and Graepel, Thore and Obermayer, Klaus. Support vector learning for ordinal regression, Proceedings of the 9th international conference on Artificial Neural Networks, IET, 97-102, 1999.

[16] T. Joachims. Training Linear SVMs in Linear Time, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006.