離散値特徴量と特異な分布を考慮したコピュラを用いた推薦システム

朝日　諒† 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 2-12-1
†† 東京工業大学情報理工学院　〒151–8550 東京都目黒区大岡山 2-12-1
††† 国立研究開発法人科学技術振興機構 ACT-I　〒332–0012 埼玉県川口市本町 4-1-8
E-mail: †{asahi,keyaki}@lsc.cs.titech.ac.jp, ††miyazaki@cs.titech.ac.jp

あらまし　本研究では，コピュラを用いた既存の推薦システムがもつ単調な嗜好しか扱えないという問題点の改善を試みる．情報推薦システムの代表的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いて推薦を行う．既存の推薦システムでは嗜好モデルの構築にコピュラという確率モデルを用いているため，機械学習手法よりも学習結果の分析が容易であるというメリットがある．しかし既存のシステムの問題点として，特徴量の分布が正規分布で特徴量値が高いほど良いといった単調な嗜好ケースを想定しており，離散値の特徴量や特徴量値にこだわりをもつといった複雑な嗜好ケースに対応できない点がある．この問題点に対し，カーネル密度推定や，ノイズを考慮した関心度，許容範囲フィルターなどを組み合わせた手法を提案する．評価実験の結果，本研究の提案手法がアイテムが 2 値の離散値特徴量や特徴量値にこだわりをもつようなケースを含むデータセットにおいて，既存手法よりも統計的に有意に ($p < 0.01$) 性能がよいことが確認できた．

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

1. はじめに

近年，インターネットが普及し web 技術が進化することで，日々大量の情報が配信されている一方で，人々が自らにとって有益な情報を選択することが困難になっている．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ユーザ自身に適したアイテムを推薦するための情報推薦技術が研究され，注目を集めている．

情報推薦の代表的なアルゴリズムに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] がある．前者はユーザの評価履歴から求めたユーザ同士の類似度を用いて推薦を行うもので，後者はアイテムが持つ特徴量とユーザの嗜好情報から推薦を行うものである．後者は特徴量ベースで推薦を行うため，前者と比較して新しいアイテムを推薦する場合や推薦システムの利用者が少ない場合でも機能するというメリットがある．そこで，本研究はコンテンツベースフィルタリングを用いる．

コンテンツベースフィルタリングには，ユーザの評価履歴を教師データとした学習ベースで嗜好モデルを構築する手法がある．学習手法には機械学習手法 [4] [5] や確率モデル [10] などが用いられる．機械学習手法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のに対して，情報推薦分野においては学習結果を解釈しそこから新たな知見を得ることには大きな意味がある．

そこで鈴木ら [10] は結果の解釈が容易である，確率モデルのコピュラを用いた学習手法を提案した．鈴木らのシステムは，コピュラに特徴量の累積分布を入力することで特徴量を統合するものである．また，鈴木らは特徴量毎にユーザが持つ関心度が異なることに着目している．鈴木らのシステムはこの関心度とコピュラに基づいて推薦を行うため，高い精度で嗜好モデルを構築できる．

しかし鈴木らのシステムにはいくつかの問題点がある．第一に特徴量の分布モデルに正規分布を仮定しているため，実際の特徴量の分布が多峰の分布の場合に学習精度が落ちる可能性がある．第二に離散値の特徴量を扱えない点がある．鈴木らのシステムは特徴量の累積分布を利用するが，特徴量が離散値の場合は累積分布を定義できないため，これを扱うことができない．第三に特異な分布，例えば特徴量値の数値に関心をもつようなケースに対応できない問題点がある．鈴木らのシステムでは特徴量の累積分布をコピュラに入力したスコア値が累積分布の単調増加になることを利用して，スコア値の高いものを優先的に推薦する．よって，特徴量の数値にこだわりを持つケース，例えば高低の両端に関心をもつが，それ以外の区間には関心をもたないケースや，逆に両端ではなくある特定の区間に関心をもつようなケースの場合，その特徴量分布を適切に扱うことできない問題が生じる．本研究はこれらのケースでも安定的に高精度な推薦手法の提案を目指す．

2. 基本的事項

2.1 コピュラの性質

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 $cdf_k(x) = prb[X_k \leq x]$ 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．

$$U = (u_1, u_2, ..., u_k) = (cdf_1(x_1), cdf_2(x_2), ..., cdf_k(x_k))$$

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

$$cdf(x_1, x_2, ..., x_n) = C(cdf_1(x_1), cdf_2(x_2), ..., cdf_k(x_k)) = C(U)$$

と表せることがスクラーの定理 [14] 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．

2.1.1 代表的なコピュラ

コピュラのモデルには様々なものがあり，分布の特性に応じたモデルを選択するのが望ましい．代表的なモデルとして式 (1)〜 式 (3) のようなものがある．θ は依存関係の度合いを表すパラメータであり，θ が大きいほど変数間の依存関係が強いことを意味する．

- *Gumbel* **コピュラ**

C_{Gubmel} は u_i が 1 付近で相関関係が高くなる．

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

- *Frank* **コピュラ**

C_{Frank} は u_i が 0.5 付近で相関関係が高くなる．

$$C_{Frank}(U) = \frac{1}{\theta}\log(1 + \frac{\prod_{i=1}^{k}(\exp(-\theta\ u_i) - 1)}{\exp((-\theta) - 1)^{k-1}}) \quad (2)$$

- *Clayton* **コピュラ**

$C_{Clayton}$ は u_i が 0 付近で相関関係が高くなる．

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k}\frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \quad (3)$$

2.2 カーネル密度推定

カーネル密度推定 [12] は，標本からその密度関数 pdf を推定するノンパラメトリックな手法である．$x_1, x_2, .. x_n$ を確率密度関数 pdf をもつ独立同時分布からの標本とする．カーネル関数 K，バンド幅 h のカーネル密度推定量 $\hat{pdf}$ は式 (4) である．

$$\hat{pdf}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}K(\frac{x - x_i}{h}) \quad (4)$$

バンド幅 h の選択は，カーネル密度推定の結果に影響する．バンド幅に採用される値の中で，代表的なものには以下の h_1，h_2 がある [11] [12].

$$h_1 = \left(\frac{1}{n}\sigma^5\right)^{\frac{1}{5}} \quad h_2 = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}}$$

3. 関 連 研 究

3.1 コピュラによる適合度統合

情報検索の分野では，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 [20] [21] [22]．Eickhoff ら [6] はコピュラを適合度統合に応用し，式 (5) を適合度統合式として提案した．

$$C_{prod}(U_{rel}) = C(U_{rel})\prod_{i=1}^{n}u_{rel,i} \quad (5)$$

U_{rel} は正解文書の適合度の累積分布の n 次元ベクトルである．各適合度の尤度の積とコピュラを掛け合わせた式 (5) は評価実験を行った結果， いくつかのデータセットで統合式 (5) 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 C_{Gumbel} を用いることが適切であることを示した [7].

3.2 混合コピュラ

Komatsuda ら [8] 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き線形和で同時分布を表現する混合コピュラを用いた統合式 (6) 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．

$$C_{mix}(U_{rel}) = \sum_{c=1}^{k}p_cC_c(U_{rel,c}) \quad (6)$$

ここで， k は文書集合のクラスタ数， 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 (6) に加え， Eickhoff らの式 (5) を混合コピュラ用に拡張した式

$$C_{mix-prod}(U) = C_{mix}(U_{rel})\prod_{i=1}^{n}\sum_{c=1}^{k}p_cu_{rel,c,i} \quad (7)$$

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 番目の適合度の累積分布を表す．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 [9].

3.3 関心度を考慮した特徴量統合式

情報検索分野での適合度統合手法を情報推薦分野に適用する場合，適合文書を適合アイテム，各適合度をアイテムの各特徴量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嗜好モデル構築に適用することができる．

文書の適合度はユーザに依存せず客観的に表現され，高精度検索に貢献することが検証されている統計量により定義される．これに対して，ユーザの嗜好はユーザ依存であり，全ての特徴量を対称に扱うのは不適切である．よって，適合度統合式を嗜好モデル構築に適用する場合，ユーザがもつ各特徴量への関心度を考慮する必要がある．

鈴木ら [10] は KL 距離 [13] を用いてユーザの i 番目の特徴量に対する関心度 Att_i を式 (8) のように定義した．

$$\begin{aligned} Att_i &= \log_{1p}(D_{\mathrm{KL}}(ALL_i \| User_i)) \\ &= \log_{1p}(\int_{-\infty}^{\infty} pdf_{all}(x_i)\log\frac{pdf_{all}(x_i)}{pdf_{user}(x_i)}\,dx) \end{aligned} \quad (8)$$

KL 距離は分布差を示す指標であり，$D_{KL}(ALL_i \| User_i)$ はユーザが関心を示したアイテムと全アイテムの分布差を表している．

鈴木らは関心度 Att を用いて，関心度集合 S_{Att} から関心度が高い特徴量のみを抽出した S_{emp} と，S_{Att} から関心度が低い特

微量を除いた S_{rdc} を式 (11) と式 (12) のように定義した．ここでは S_{Att} に関して平均と分散を推定し，それらを用いて検出した外れ値を特徴量抽出に利用している．
外れ値の検出に必要な平均と分散は外れ値の影響が小さいロバストな方法 [17] でメジアン Med と $MADN$(式 (10)) として推定する．鈴木らの評価実験では $cns_a = 2.5$ で最高の結果を示した．

$$MAD(S_{Att}) = Med(\{|Att_i - Med(S_{Att})|\}) \tag{9}$$

$$MADN(S_{Att}) = \frac{MAD(S_{Att})}{0.675} \tag{10}$$

$$S_{emp} = \{i | Med(S_{Att}) + cns_a \cdot MADN(S_{Att}) \leqq Att_i\} \tag{11}$$

$$S_{rdc} = \{i | Att_i \leqq Med(S_{Att}) - cns_a \cdot MADN(S_{Att})\} \tag{12}$$

鈴木らは，S_{rdc} と S_{emp} から関心度を反映させた統合式 (14) を提案した．統合式 $C_{kl-emp-rod}$ は特定の評価指標について既存の統合式と比較して最高の結果を示した．

$$C_{kl-emp}(U_{rdc}) = C_{mix}(U_{rdc}) \prod_{i \in S_{emp}} \sum_{c=1}^{k} p_c u_{rel,c,i} \tag{13}$$

$$C_{kl-emp-prod}(U_{rdc}) = \begin{cases} C_{mix-prod}(U_{rdc}) & \text{if } S_{emp} = \emptyset \\ C_{kl-emp}(U_{rdc}) & \text{otherwise} \end{cases} \tag{14}$$

4. 提 案 手 法

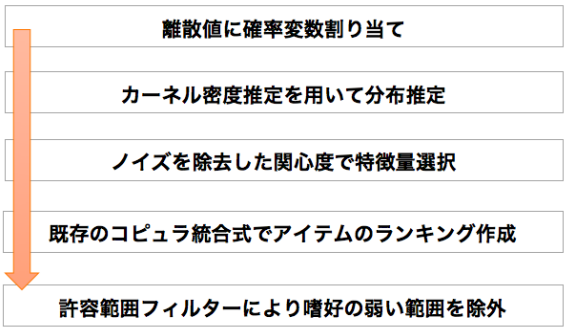

図 1　提案手法の処理チャート

本提案手法は 4 つの構成から成っており，鈴木らのシステムでは扱えない離散値特徴量や特異な分布に対しても高精度な推薦を目指す．

提案手法の処理チャートは図 1 であり各プロセスの概要は以下である．

- **離散値へのマッピング**　(4.1 節)
  ユーザの嗜好を反映した離散値特徴量への数値マッピング
- **カーネル密度推定を用いた分布推定**　(4.2 節)
  カーネル密度推定を利用して矩形型や多峰の分布を推定
- **ノイズを除いた関心度**　(4.3 節)
  既存の関心度では連続値と離散値でノイズ差が大きいため両者を比較できるようにノイズを除去
- **許容範囲フィルター**　(4.4 節)
  許容範囲フィルターにより，特徴量値からユーザが関心を示さない区間を推薦対象から除去

4.1 離散値のマッピング

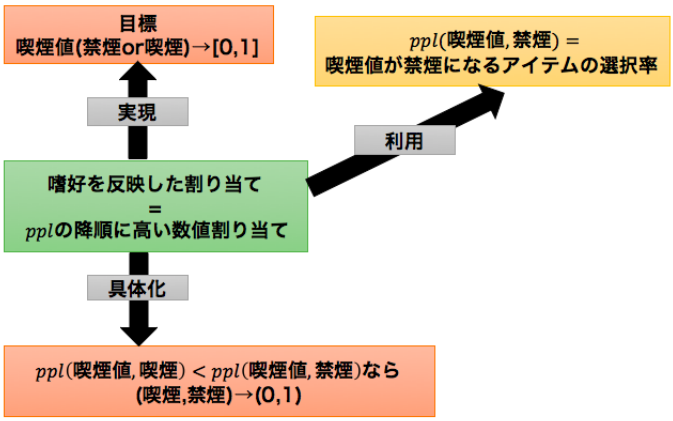

図 2　離散値のマッピング

特徴量には，喫煙可能かどうかを示す喫煙値のように離散値をとるものがある．このような離散値特徴量を扱う場合，喫煙可能のアイテムのみを推薦するような方式は適切ではない．喫煙値ではユーザの嗜好と反するものの，他の特徴量との兼ね合いで嗜好アイテムとして選択されるようなケースを想定する必要がある．よってこのようなケースを想定して，提案手法では離散値特徴量も連続値特徴量と同様にその特徴量の累積分布を推定し，コピュラにより他の特徴量と統合するといった方式を用いる．

特徴量の累積分布を求めるためには，特徴量を数直線上の実数値にマッピングする必要がある．特徴量の累積分布は特徴量に対して単調増加するため，特徴量が離散値の場合数値を離散値に割り当てる順番が累積分布の値に影響する．この際にユーザがもつ離散値特徴量への嗜好順を考慮すべきである．

例えば喫煙可能か禁煙かを表す離散値特徴量として喫煙値というものを定義し，これに 0 と 1 の数値を割り当てる場合を考える．ユーザが禁煙家の場合は禁煙に 1，喫煙可能に 0 を割り当てるべきなのに対し，ユーザが喫煙家の場合は喫煙可能に 1，禁煙に 0 を割り当てるべきである．

そこで，式 (15) のように定義した ppl を利用するマッピング手法を提案する (図 2)．i 番目の特徴量の離散値 v に対して，ユーザがもつ人気度 $ppl_i(v)$ は式 (15) で定義される．

$$S_{ScoreUser}(i, v) = \{item | score_i(item) = v\} \cap S_{User}$$

$$S_{ScoreAll}(i, v) = \{item | score_i(item) = v\} \cap S_{All}$$

$$ppl_i(v) = \frac{|S_{ScoreUser}(i, v)|}{|S_{ScoreAll}(i, v)|} \tag{15}$$

$score_i(item)$ は $item$ の特徴量 i のスコア値を返す関数，S_{User} はユーザが選択した $item$ 集合，S_{All} は全 $item$ 集合であることに注意する．$ppl_i(v)$ は i 番目の特徴量が v の $item$ 集合から，ユーザがどれだけの $item$ に関心を示したかを表している．

この人気度 ppl の高い離散値から降順で高い数値を割り当てることで，ユーザの嗜好を考慮して，離散値特徴量に数値を割り当てられることが期待できる．

4.2 カーネル密度推定を用いた分布推定

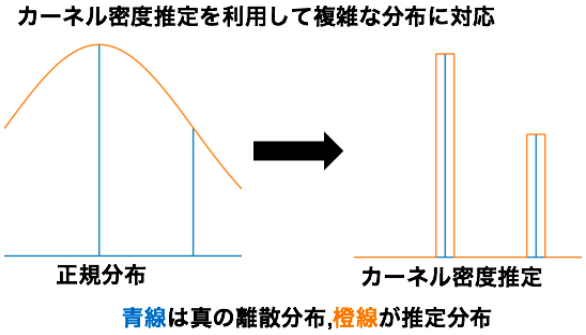

図 3　カーネル密度推定を用いた分布推定

鈴木らのシステム [10] が特徴量分布の推定に正規分布を仮定していたのに対し，本研究では多峰性の分布など複雑な分布を考慮してカーネル密度推定を用いる．

これにより，正規分布以外の複雑な分布も推定することができると期待できる (図 3)．式 (4) のパラメータであるカーネル関数式には式 (16) を用いた．

$$K(x) = \begin{cases} gaussian(x) & (\text{特徴量 } x \text{ が連続値}) \\ tophat(x) & (\text{特徴量 } x \text{ が離散値}) \end{cases} \tag{16}$$

$$gaussian(x) = \frac{1}{2\sqrt{\pi}\sigma} \exp\left(-\frac{(x-u)^2}{2\sigma^2}\right) \tag{17}$$

$$tophat(x) = \begin{cases} \frac{1}{2} & (-1 \leqq x \leqq 1) \\ 0 & (otherwise) \end{cases} \tag{18}$$

続いてバンド幅の決定方法について述べる．

$$h_{opt} = \begin{cases} h_{G\_opt} & (K = gaussian) \\ 0.1 & (K = tophat) \end{cases} \tag{19}$$

$$S_{lct} = \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, ..., 5 \cdot 10^{-1}\} \tag{20}$$

$$S_{opt} = \{silverman, scott\} \tag{21}$$

$$S_{Grd} = S_{lct} \cup S_{opt} \tag{22}$$

カーネル密度推定式 (式 (4)) のパラメータであるバンド幅 h の選択には式 (19) のようにカーネル関数 $K(x)$ の特性に合ったものを用いた．連続値の場合，最適値の位置を探るための S_{lct} と 2.2 の h_2，h_1 からなる S_{opt} から構築した探索集合 S_{Grd} から $GridSearch$ で最適値 h_{G_opt} を選択する．S_{lct} は $K(x)$ が $gaussian(x)$ で累積分布が求まるという条件と，常用対数が負であるという条件を同時に満たす要素の集合である．

以下に $Gridserach$ の概要を載せる．

（1） 訓練用データ集合 S_{trn} からこれを k 分割した各集合 S_{Nscr_i} を除いた評価用集合 S_{scr_i} を得る．

（2） S_{Grd} の要素 h を考える．

（3） 式 (24) のようにして S_{scored_i} で h の評価値を求める．

（4） 式 (25) のようにして全評価用集合での評価値平均を h の評価値とする．

（5） (2) から (4) を全 h に対して行い，h の評価値が最大となるものを h_{G_opt} とする．

$$S_{scr\_i} = S_{trn} \backslash S_{Nsrc\_i} \tag{23}$$

$$score(h, i) = \sum_{x \in S_{scored\_i}} \log f(x) \tag{24}$$

$$score(h) = \frac{1}{k} \sum_{i=1}^{k} score(h, i) \tag{25}$$

離散値の場合，カーネル関数が $tophat$ なので各離散値点で $tophat$ が干渉しないようなバンド幅を選択すれば推薦に必要な累積分布を得ることができる．本研究で用いる離散値は 0 と 1 の値をとるため，バンド幅は 0.5 未満になればよい．よって，離散値のバンド幅には，0.5 未満で，常用対数が整数であるという条件をみたす実数の最大値である 0.1 を採用した．

4.3 ノイズを除いた関心度

既存の関心度 Att_i(式 (8)) では積分区間が無限区間であり，スコア値の範囲である区間 [0,1] 以外の計算結果が含まれていた．しかし区間 [0,1] 以外の計算結果をノイズとすると，離散値のカーネルは $tophat$ なのでノイズが 0 になる．よって既存の関心度で連続値と離散値を比較する場合，ノイズ差を考慮していない問題が生じる．よって新たな関心度 (式 (26)) を Att_{i_Shr} として提案する．

$$\begin{aligned} Att_{i\_Shr} &= D_{\mathrm{KL}}(User_i \| ALL_i) \\ &= \left(\int_0^1 pdf_{user}(x_i) \log \frac{pdf_{user}(x_i)}{pdf_{all}(x_i)}\, dx\right) \end{aligned} \tag{26}$$

式 (26) は既存の算出式 Att の積分区間を変更し，ノイズが計算結果に含まれないようにしたものである．さらに $pdf_{all}(x) \leqq pdf_{user}(x)$ の部分をより反映させるために D_{KL} のベースを All_i から $User_i$ へ変更した．区間変更に伴い，式 (8) に現れる値調整のための $\log_{1p}$ が不要になったためこれを取り除いた．

4.4 許容範囲フィルター

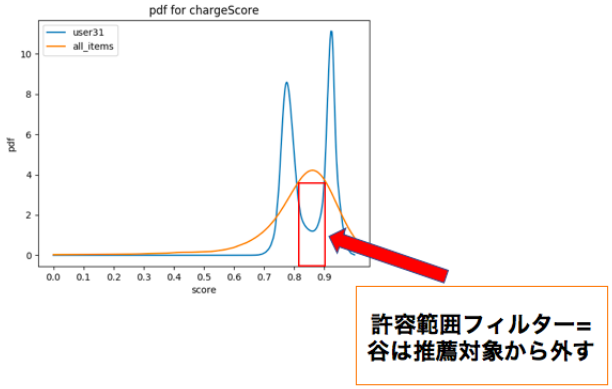

図 4　許容範囲フィルター

鈴木らのシステム [10] はユーザの選択したアイテムを教師

データとして構築したコピュラモデルから，推薦対象のアイテムの累積分布を求め，その累積分布の降順で推薦を行うものである．しかし予備調査の結果，表 1 の ROLE7 や ROLE14 のような特徴量値の特定区間に関心を示すような嗜好ケースに対して問題が生じたため式 (27) のような tlr を定義し，この tlr に含まれるアイテムのみを推薦対象とする許容範囲フィルターを提案する (図 4)．tlr はユーザが選択したアイテムの密度分布 $pdf_{user}(x)$ と全アイテムの密度分布 $pdf_{all}(x)$ を考えるとき次のように定義される．

$$tlr = \{x | pdf_{all}(x) \leqq pdf_{user}(x)\} \tag{27}$$

tlr は，ユーザが関心を示す特徴量値の範囲を抽出するためのものである．

$$S_{filtered} = \begin{cases} S_{sorted}(n) & (S_{emp} = \phi) \\ S_{emp} & (otherwise) \end{cases} \tag{28}$$

$S_{sorted}(n)$ は関心度上位 n 個の特徴量集合であり，n には ROLE7 と ROLE14 で設定した特異な分布になると期待する特徴量を抽出できる整数で最小の 2 を用いた．許容範囲フィルターを $S_{filtered}$ の特徴量のみに適用することで，特異な嗜好ケースに対応できると期待できる．$S_{filtered}$ はフィルターが有効に機能すると期待される特徴量の集合であり，フィルターが過剰に適用されることで有効に機能しなくなる問題を防ぐためのものである．

5. 評 価 実 験

データセットは，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 件を対象としている．各ホテルはサービス，施設， 部屋， 立地， 風呂， 食事のレビュー値と価格， 最寄り駅からの直線距離の計 8 種類の情報をもつ．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したものを連続値特徴量としている．

このホテルデータに 9 個目の新たな特徴量として喫煙値を追加した．喫煙値は喫煙可と禁煙の 2 値で，各ホテルに各離散値が 50% で出現する．

被験者の実人数は大学院生 12 人で，延べ人数が 33 人である．

また, 実験手順は下記の通りである．

（1） 被験者は割り振られた ROLE のシナリオ下でホテルを評価する．全 ROLE は表 1 の通りである．

（2） ホテル評価時に重視した特徴量に合計 100 となるように重みを割り振る．

5.1 比 較 手 法

- **嗜好回答情報を利用した重み付き線形和**

$$LIN(X) = \sum_{i=1}^{n} w_i x_i \tag{29}$$

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．

- **コピュラを用いた統合式**

$$method = C_{X,Y,Z}$$

$$X = \begin{cases} 'Kd' & (\text{カーネル密度推定}) \\ 'Nrm' & (\text{正規分布}) \end{cases}$$

$$Y = \begin{cases} 'Shr' & (att = Att_{Shr}) \\ 'Inf' & (att = Att_{Inf}) \end{cases}$$

$$Z = \begin{cases} 'Tl' & (\text{フィルター有効}) \\ NULL & (\text{フィルター無効}) \end{cases}$$

鈴木らの $C_{kl-emp-prod}$(式 (14)) に，4. で述べた各小手法を部分的に有効にした手法である．用いるコピュラ統合式を $method$ としたとき，$method$ は，$C_{X,Y,Z}$ のように表す．フィルターを用いる場合，Z の文字列は $'Tl'$ であり，フィルターを用いない場合 Z は $NULL$ で空文字である．

例えば，鈴木らの既存手法の場合分布推定に正規分布を用い，関心度には Att_{Inf} を用いるため，その表記は $C_{Nrm,Inf}$ である．提案手法の場合分布推定にカーネル密度推定を用い，関心度には Att_{Shr} を用い，フィルターを有効にするため，その表記は $C_{Kd,Shr,Tl}$ である．

- **ランキング SVM**

SVM^{rank} (注1) [19] 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．

表 1 ROLE 一覧

期待する分布	ROLE 番号	ROLE の説明
特になし	ROLE1	出張で宿泊する．会社規定のため安く済ませたい．
	ROLE2	友達と旅行で宿泊する．
	ROLE3	観光目的で宿泊する．価格は気にせず良いホテルがよい．
	ROLE4	恋人と宿泊する．安くて良いホテルがよい．
離散値で偏った分布	ROLE5	ROLE3+被験者には数日おきに喫煙習慣がある．
	ROLE6	ROLE4+恋人が喫煙者である．
	ROLE9	ROLE1+被験者は極度の嫌煙家である．
	ROLE10	ROLE2+被験者は極度の喫煙家である．
	ROLE11	ROLE2+被験者は極度の嫌煙家である．
	ROLE12	ROLE3+被験者は極度の喫煙家である．
	ROLE13	ROLE4+被験者は極度の嫌煙家である．
特異な分布	ROLE7	駅の騒音と徒歩距離を考慮して，駅から適度な距離のホテルがよい．
	ROLE14	同僚と出張で宿泊する．6,000 円/人まで会社経費．ホテルはルームチャージ制のため，個人利用であれば低価格のホテルしか利用できず，相部屋で利用する場合高価格のホテルまで利用できる．

5.2 評 価 指 標

推薦システムの性能を評価するための尺度について述べる．推薦アイテムのうち， 正しく推薦されたアイテムを適合アイテム， そうでないものを不適合アイテムと呼ぶ．

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

$P@k$ は推薦結果上位 k 件のうち，適合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．

$$P@k = \frac{h}{k} \tag{30}$$

$nDCG@k$ 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．

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{31}$$

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{iDCG@}k} \tag{32}$$

ここで，rel_i は上位 i 番目アイテムの適合度である．適合ならば 1，不適合ならば 0 で表現される．

再現率 $Recall_k$ 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 $Recall_k$ は以下の式で表される．

$$Recall_k = \frac{h}{a} \tag{33}$$

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．

$$\mathrm{iP@}i = \max_k \{P@k | Recall_k \geq i\} \tag{34}$$

5.3 提案手法のパラメータ

提案手法で用いたパラメータは以下の通りである．

- コピュラモデルには，C_{frank}，$C_{Clayton}$，C_{gumbel} の中で，最も良い精度を示した C_{frank} を用いた．
- クラスタ数は 1~5 に変化させた時に最も良い精度を示した 2 を用いた．
- 特徴量選択時の正定数 cns_a には，1.0 ~ 3.5 まで 0.5 刻みで変化させた時に最高の精度を示した 1.5 を採用した．
- $GridSearch$(式 (25)) のパラメータ k には，使用したライブラリのデフォルト値である 3 を採用した．

6. 実験結果

6.1 特異な分布

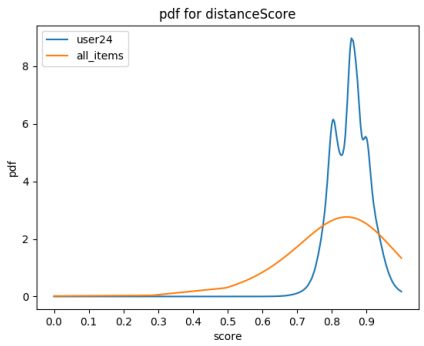

図 5　ROLE7 の距離特徴量分布

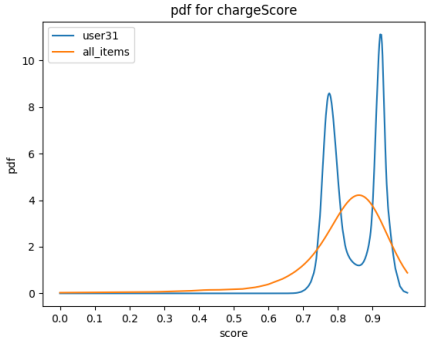

図 6　ROLE14 の価格特徴量分布

評価実験では，特定の特徴量が鈴木らのシステムでは対応できないような特異な分布をとることを想定したケースを表 1 の ROLE7，ROLE14 として用いた．

ROLE7 は駅の喧騒が気にならない程度に駅から離れていて，通勤に不便にならない程度に駅から近いホテルを探すという設定である．よって距離スコアが高すぎるとユーザに選択されないことが想定される．

ROLE7 の分布 (図 5) から，右端付近で許容範囲が途切れていることが確認できる．これはユーザが右端付近は駅の喧騒が気になる範囲で，許容範囲が適度な距離範囲であるとみなしたと考えられる．よって，許容範囲フィルターを有効にすることで右端付近のアイテムが誤って推薦されるというケースを避け，適切なアイテムを推薦できると期待できる．

ROLE14 は，6,000 円付近の低価格帯と 12,000 円付近の高価格帯のホテルを優先的に選択させる設定である．6,000 円以下の低価格帯のホテルには個室利用可能，12,000 円の高価格帯のホテルには高級であるという選択意義があるのに対し，中価格帯のホテルは個室利用もできずホテルのモデルも高価格帯のものより劣るものが多いため選択意義が薄い．よって，ユーザはメリットの薄い中価格帯を避け，その両端の価格帯でホテルの選択をするため価格特徴量の分布が U 字になることが想定される．

ROLE14 の分布 (図 6) は想定通り，U 字型になっている．U 字型の分布の場合，U 字の谷の部分ではユーザが選択を避けているため，全アイテムの分布よりも pdf が下回る傾向がある．よって，許容範囲フィルターをこの特徴量分布に適用し，U 字の谷の部分を推薦対象から外すことで，ROLE14 の中価格帯のアイテムを誤って推薦するというケースは避けられる．

U 字の両峰付近のアイテムについては，鈴木らのコピュラ統合式により価格特徴量とその他のレビュー値を統合できるため，これを用いてアイテムの順位付けができる．よって許容範囲フィルターを用いれば，ROLE14 のような U 字型の分布でも適切な推薦ができることが期待できる．

表 2　データセットでの実験結果

measure	$C_{Kd,Shr,Tl}$(提)	$C_{Nrm,Inf}$(既)	$C_{Kd,Inf,Tl}$	$C_{Kd,Shr}$	$C_{Kd,Inf}$	LIN	SVM
iP@0	0.98	0.706	0.963	0.937	0.908	0.87	0.953
iP@0.1	0.968	0.63	0.939	0.901	0.871	0.862	0.941
iP@0.2	0.955	0.575	0.903	0.867	0.838	0.84	0.927
iP@0.3	0.931	0.539	0.876	0.842	0.808	0.803	0.9
iP@0.4	0.911	0.513	0.852	0.825	0.792	0.787	0.871
iP@0.5	0.875	0.502	0.802	0.804	0.769	0.753	0.844
iP@0.6	0.838	0.482	0.745	0.769	0.736	0.711	0.806
iP@0.7	0.797	0.467	0.698	0.738	0.696	0.675	0.769
iP@0.8	0.731	0.454	0.639	0.691	0.647	0.635	0.731
iP@0.9	0.653	0.43	0.586	0.624	0.599	0.594	0.663
iP@1.0	0.55	0.405	0.515	0.55	0.518	0.527	0.581
MAiP	0.835	0.518	0.774	0.777	0.744	0.733	0.817
nDCG@5	0.963	0.681	0.948	0.892	0.877	0.779	0.941
nDCG@10	0.96	0.702	0.938	0.901	0.875	0.814	0.939
nDCG@15	0.954	0.715	0.932	0.899	0.879	0.832	0.937
nDCG@20	0.951	0.718	0.927	0.899	0.88	0.84	0.934
P@5	0.891	0.455	0.847	0.77	0.755	0.682	0.865
P@10	0.843	0.467	0.768	0.746	0.716	0.671	0.812
P@15	0.781	0.443	0.712	0.703	0.682	0.64	0.762
P@20	0.711	0.445	0.65	0.667	0.634	0.604	0.714

表 3　喫煙値を重視したユーザ集合での実験結果

measure	$C_{Kd,Shr,Tl}$(提)	$C_{Nrm,Inf}$(既)	$C_{Kd,Inf,Tl}$	$C_{Kd,Shr}$	$C_{Kd,Inf}$	LIN	SVM
iP@0	0.997	0.709	0.966	0.986	0.922	0.96	1.0
iP@0.1	0.994	0.624	0.93	0.976	0.899	0.948	0.993
iP@0.2	0.977	0.567	0.898	0.956	0.88	0.933	0.992
iP@0.3	0.962	0.548	0.863	0.95	0.855	0.917	0.982
iP@0.4	0.952	0.517	0.821	0.936	0.841	0.91	0.958
iP@0.5	0.926	0.508	0.718	0.924	0.811	0.892	0.937
iP@0.6	0.892	0.495	0.661	0.902	0.78	0.869	0.899
iP@0.7	0.83	0.48	0.614	0.863	0.738	0.825	0.867
iP@0.8	0.762	0.469	0.576	0.815	0.673	0.773	0.838
iP@0.9	0.641	0.442	0.52	0.704	0.607	0.699	0.781
iP@1.0	0.525	0.386	0.448	0.59	0.478	0.575	0.68
MAiP	0.86	0.522	0.729	0.873	0.771	0.846	0.902
nDCG@5	0.989	0.675	0.949	0.969	0.911	0.899	0.996
nDCG@10	0.985	0.7	0.94	0.967	0.91	0.917	0.989
nDCG@15	0.981	0.711	0.929	0.965	0.908	0.925	0.984
nDCG@20	0.978	0.718	0.925	0.964	0.908	0.93	0.98
P@5	0.95	0.439	0.839	0.922	0.833	0.828	0.961
P@10	0.908	0.456	0.725	0.878	0.761	0.831	0.9
P@15	0.843	0.431	0.669	0.831	0.707	0.791	0.859
P@20	0.749	0.435	0.588	0.774	0.65	0.736	0.799

表 4　特異な分布での実験結果

measure	$C_{Kd,Shr,Tl}$(提)	$C_{Nrm,Inf}$(既)	$C_{Kd,Inf,Tl}$	$C_{Kd,Shr}$	$C_{Kd,Inf}$	LIN	SVM
iP@0	0.958	0.65	0.94	0.869	0.847	0.745	0.898
iP@0.1	0.942	0.588	0.914	0.804	0.783	0.741	0.875
iP@0.2	0.924	0.511	0.857	0.741	0.723	0.708	0.854
iP@0.3	0.884	0.454	0.821	0.694	0.676	0.646	0.807
iP@0.4	0.85	0.423	0.802	0.671	0.654	0.624	0.768
iP@0.5	0.804	0.407	0.76	0.638	0.629	0.572	0.727
iP@0.6	0.753	0.384	0.682	0.587	0.585	0.512	0.679
iP@0.7	0.716	0.368	0.627	0.555	0.538	0.481	0.634
iP@0.8	0.631	0.357	0.548	0.504	0.492	0.445	0.579
iP@0.9	0.561	0.335	0.483	0.454	0.458	0.42	0.495
iP@1.0	0.461	0.324	0.433	0.429	0.431	0.391	0.441
MAiP	0.771	0.437	0.715	0.632	0.62	0.571	0.705
nDCG@5	0.934	0.618	0.922	0.789	0.79	0.589	0.877
nDCG@10	0.93	0.638	0.907	0.81	0.786	0.656	0.878
nDCG@15	0.92	0.659	0.902	0.807	0.795	0.692	0.879
nDCG@20	0.915	0.661	0.894	0.808	0.799	0.708	0.875
P@5	0.813	0.37	0.777	0.56	0.573	0.437	0.747
P@10	0.742	0.363	0.687	0.542	0.547	0.43	0.682
P@15	0.669	0.344	0.619	0.502	0.523	0.413	0.62
P@20	0.598	0.352	0.558	0.481	0.482	0.402	0.573

6.2 実験結果

全てのシナリオに対する実験結果は表 2 である．提案手法が過半数の評価指標で最高の結果を示している．さらに，提案手法と鈴木らの手法を比較するマンホイットニーの U 検定を実施した結果統計的に有意に提案手法が優れていることを確認した ($p < 0.01$)．

次に鈴木らの手法では対応できない各ケースに対する結果について述べる．表 3 では，Att_{Shr} を採用する手法が Att_{Inf} を採用する手法に比べ結果がよい傾向がより顕著である．さらに表 4 では，許容範囲フィルターを採用する手法の結果がフィルターを採用しない結果に比べ結果がよい傾向がより顕著である．よって，提案手法の各プロセスは鈴木らの手法では対応できない各ケースで有効に機能したといえる．

7. まとめ

本論文では，既存の鈴木らの手法の問題点を解決するために主にカーネル密度推定や，ノイズを除去した関心度，許容範囲フィルターを用いた手法を提案し，その評価を行った．

鈴木らのシステムの問題点として，離散値特徴量や特徴量値にこだわりをもつような複雑な嗜好ケースに対応できない点を指摘した．

指摘した問題点の解決手法として，カーネル密度推定を利用する手法や，離散値特徴量と連続値特徴量の関心度を適切に比較できるような新たな関心度 Att_{Shr} を用いる手法，特異な分布に対応するために許容範囲フィルターを用いる手法などを提案した．

評価実験の結果，本提案手法は既存手法と比較して統計的に有意に性能がよい ($p < 0.01$) ことを確認した．

本研究の課題点としては，離散値のマッピング手法に離散値が 2 値という前提のもとで式 (15) で定義される人気度を用いたが，3 値以上の離散値でも有効な手法を考案すべき点と，カーネル密度推定で用いるパラメータであるバンド幅 h の選択手法についてより適切な手法の存在の有無の調査をすべき点，許容範囲フィルターをより適切に用いるために特異な分布をとる特徴量をより高精度に抽出する方法を研究すべき点などがある．

謝辞

本研究の一部は，JSPS 科研費 (JP15H02701，JP16H02908，JP15K20990，JP17K12684)，JST ACT-I の助成を受けたものである．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] C. Cortes and V. Vapnik. "Support vector networks". Machine Learning, 20:pp.273-297, 1995.

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

[6] 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.

[7] 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 1831-1834. ACM, 2014.

[8] 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.

[9] 小松田卓也, 欅 惇志, 宮崎 純. 多峰性のあるコピュラを用いたスコア統合手法の提案及びその検証, 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,C1-1.

[10] 鈴木崇弘, 欅 惇志, 宮崎 純. コピュラを用いたユーザプロファイリング手法の提案, 第 9 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 2017）論文集,A7-1.

[11] D.W.Scott, "Multivariate Density Estimation: Theory,Practice,and,Visualization", John Wiley&Sons, New York, Chicester, 1992.

[12] B.W.Silverman, "Denstyity Estimation for Statistics and Data Analysis.", Vol.26, Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1986.

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

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

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

[16] 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）.

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

[18] 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, pp.97-102, 1999,

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

[20] Ronan Cummins. Measuring the ablity of score distributions to model relevance. In *Information Retrieval Technology,* pp.25-36. Springer,2011.

[21] EdWard A Fox and Joseph A Shaw. Combination of multiple searches. *NIST SPECIAL PUBLICATION SP,* pp.243-243,1994.

[22] Evangelos Kanoulas,Keshi Dai, Virgil Pavlu, and Javed A Aslam. Score distribution models: assumptions, intuition, and robustnesss to score manipulation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval,* pp.242-249. ACM, 2010.