東京工業大学工学部

学士論文

離散値特徴量や特異な嗜好ケースを考慮したコピュラを用いた推薦システム

指導教員 宮崎純 教授

平成30年2月

提出者

学科　　情報工学科
学籍番号 14_00368
氏名　　朝日 諒

指導教員認定印		
学科長認定印		

平成29年度　学士論文内容梗概

離散値特徴量や特異な嗜好ケースを考慮したコピュラを用いた推薦システム

指導教員　　宮崎純　教授

情報工学科
14_00368　　朝日 諒

本研究では,コピュラを用いた既存の推薦システムの問題点を改善することで,ユーザのもつより複雑な嗜好を汲み取れるような推薦システムの構築を目指す.

情報推薦システムは,膨大な情報の中からユーザの嗜好にあった情報を提示するシステムである. 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から嗜好モデルを構築し,それに基づいて推薦を行う.

コンテンツベースフィルタリングで既存の推薦システムの一つに,コピュラという確率モデルを用いて嗜好モデルを構築するものがある.これは,ユーザが関心を示したアイテムを教師データとして嗜好モデルを構築するものであり,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をもつ. しかし,このシステムの問題点として特徴量の分布が正規分布であると仮定している点,離散値の特徴量が扱えない点,特異な特徴量分布に対応できない点,などがある.

本論文では,これらの問題点に対し,カーネル密度推定や,ノイズを考慮した関心度式や,許容範囲フィルターを用いた手法などを組み合わせた手法を提案する. 評価実験の結果,既存の手法で用いられたデータセットでは,提案手法が既存手法と同等の結果を示し,離散値特徴量を追加した新たなデータセットでは,通常のケースと,アイテムが2値の離散値をとる特徴量をもつ場合や特徴量の分布が前述の特異な分布になるような場合を含めて,提案手法が有意水準1%以下で既存手法よりも結果が良いことが確認できた.

目　次

第1章

序論

近年,インターネットが普及しweb技術が進化することで,日々大量の情報が発信されている.一方で,大量の情報の中から人々が自らにとって有益な情報を選択することは困難になっている.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[2][3]とコンテンツベースフィルタリング[4]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.学習ベースの手法は,ユーザ自身が嗜好情報を回答することで嗜好モデルを構築するような手法と比べて,ユーザ自身が把握しきれないような嗜好を汲み取れたり,ユーザへの負担が小さいというメリットがあるため,本研究では学習ベースのコンテンツベースフィルタリングを扱う.

学習手法には,機械学習手法[5][6]や確率モデル[10]を用いるものがある.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推薦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,累積分布を定義できないため,離散値の特等量が扱えない. 第三に特異な分布,例えば特徴量値の範囲にこだわりをもつような場合,その特徴量の分布を正しく扱えないという問題がある. 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. しかし,特徴量は,特徴を数値化したものなので,その数値が高ければ良いとは限らない.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.

こういった問題に対処するこで,ユーザのより複雑な嗜好を汲み取れるような推薦システムの構築が期待できる. よって,本研究では特徴量の分布モデルにカーネル密度推定を採用し, 特異な分布をもつ特徴量に対しては,特徴量のスコア値から,ユーザが強い嗜好を示す範囲を抽出するフィルターを定義し, これを用いてより複雑な嗜好を汲み取れるような推薦手法を提案する. さらに評価実験により提案手法が有効であることを示す.

第2章

基本的事項

2.1 情報推薦システム

情報推薦システムとは,ユーザの過去の行動履歴からユーザの嗜好を汲み取り,ユーザにとって有用と思われる情報や商品などを提示するシステムである.推薦対象の商品や情報などをアイテムと呼ぶ.動画配信サービスなどで目にすることが多い「あなたにおすすめの動画」と提示されるリストは,情報推薦システムの出力の1例である.このようなシステムに用いられるアルゴリズムは,協調フィルタリング[2][3]とコンテンツベースフィルタリング[4]に大別される.

2.1.1 代表的な手法

協調フィルタリング

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,推薦対象ユーザまたはアイテムと似た要素からなる近傍を特定する.その後,近傍を利用してアイテムを推薦する.[1]

表2.1のように各ユーザが各アイテムに評価値を与えていたとき,ユーザ1のアイテム5の評価値を協調フィルタリングを用いて予測することを考える.協調フィルタリングは,近傍形成と評価値予測の方法によりユーザベースの手法[2]とアイテムベースの手法[3]に大別される.近傍とはある要素に似たものを指す.

ユーザベースの場合,ユーザ間の類似度を算出し,推薦対象のユーザとの類似度が高いユーザから形成した近傍の各ユーザについて,式(2.1.2)のように推薦対象アイテムの評価値と推薦対象ユーザとの類似度の重みつけ和により,評価値を予測する.近傍形成にはユーザaとユーザbの類似度にピアソンの相関係数(式(2.1.1))を用いて,推薦対象ユーザと正の相関をもつユーザから形成する方法がある.

$$sim_u(a,b) = \frac{\sum_{p\in P}(r_{a,p}-\overline{r_a})(r_{b,p}-\overline{r_b})}{\sqrt{\sum_{p\in P}(r_{a,p}-\overline{r_a})^2}\sqrt{\sum_{p\in P}(r_{b,p}-\overline{r_b})^2}} \tag{2.1.1}$$

$\overline{r_a}$ はユーザ a の評価値の平均を表し,P はアイテム集合である. ユーザ 1 との相関係数はユーザ 2 が 0.85, ユーザ 3 が 070, ユーザ 4 が 0.00, ユーザ 5 が-0.79 であるため, ユーザ 1 の近傍ユーザはユーザ 2 とユーザ 3 である.
次に, この近傍ユーザ 2 と 3 の評価履歴からユーザ 1 のアイテム 5 の予測評価値 $pred_u(1,5)$ を式 (2.1.2) のようにして求める.

$$pred_u(a,p) = \overline{r_a} + \frac{\sum_{b\in N_u} sim_u(a,b)(r_{b,p}-\overline{r_b})}{\sum_{b\in N_u} sim_u(a,b)} \tag{2.1.2}$$

N_u は推薦対象ユーザの近傍ユーザの集合である. よってユーザ 1 のアイテム 5 の予測評価値は $4 + \frac{1}{(0.85+0.7)}(0.85(3-2.4)+0.7(5-3.8)) = 4.87$ となる

Table 2.1: ユーザの評価値行列

	アイテム 1	アイテム 2	アイテム 3	アイテム 4	アイテム 5
ユーザ 1	5	3	4	4	?
ユーザ 2	3	1	2	3	3
ユーザ 3	4	3	4	3	5
ユーザ 4	3	3	1	5	4
ユーザ 5	1	5	5	2	1

アイテムベースの場合, アイテム間で類似度を算出し, 推薦対象アイテムの類似度が高いものから近傍を形成する. アイテム間の類似度には式 (2.1.3) のような, アイテムベクトル間のコサイン類似度などが用いられる.

$$sim_i(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\times|\vec{b}|} \tag{2.1.3}$$

近傍形成後は, ユーザベースと同様に近傍 N_i を対象に式 (2.1.4) のようにして, 推薦対象アイテムの近傍 N_i の各アイテムについて推薦対象ユーザによる評価値と推薦対象アイテムとの類似度の重みつけ和により未評価アイテムの予測評価値 $pred_i(1,5)$ を算出する.

$$pred_i(a,p) = \frac{\sum_{b\in N_i} sim(b,p)r_{a,b}}{\sum_{b\in N_i} sim(b,p)} \tag{2.1.4}$$

実際には, 全ての未評価アイテムの評価値を計算し, 予測された評価値の高い順番にアイテムを提示することで推薦を行う. このように協調フィルタリングではユーザがつけた評価値のみを用いて推薦を行うため, 推薦対象アイテムについてのドメイン知識は必要ない. また, 他のユーザの評価値を利用するため, 推薦対象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となる. 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は有効な推薦を行うことができず, このような問題はコールドスタート問題として知られている.

コンテンツベースフィルタリング

コンテンツベースフィルタリングは, アイテムが持つ性質やユーザ個人の好みを利用して推薦を行う手法である. ユーザの好みを表す情報はユーザプロファイルとよばれる.

アイテムが持つ性質は推薦システム内では特徴パラメータとして数値で表される [16]. アイテムの性質に応じて, 以下のような型で表現される.

- **連続値型**

  ［0, 1］のような, 連続値により表現できる.

  例: 価格や駅から店までの距離など

- **二値型**

  二値 {0, 1} で, ある属性を満たすかどうか表現する.

  例: 「期間限定」や「特典あり」など

- **カテゴリ型**

  三次元以上のパラメータを持ち, 該当する子パラメータのみ 1 とし, それ以外を 0 として表現する.

  例: 色 { 青（1, 0, 0）赤（0, 1, 0）, 黄（0, 0, 1） }

コンテンツベースフィルタリングでは各特徴パラメータを入力として, ユーザプロファイルを用いてスコアを算出しランキング付けを行う. スコアの算出方法は, ユーザプロファイルの形態によるが, 各特徴パラメータの重み付け和や, ユーザが好むアイテムとの類似度を計算することで算出できる.

コンテンツベースフィルタリングを用いる場合, アイテムがどのような特徴パラメータを持っているかをシステムが知っている必要があるが, 協調フィルタリングと異なり他のユーザに依存しないため, 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.

2.1.2 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. $P@k$ 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.

$$P@k = \frac{h}{k} \tag{2.1.5}$$

$nDCG@k$ 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.

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.1.6}$$

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{iDCG@}k} \tag{2.1.7}$$

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.

$$Recall_k = \frac{h}{a} \tag{2.1.8}$$

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.

$$\mathrm{iP@}i = \max_k \{P@k | Recall_k \geq i\} \tag{2.1.9}$$

2.2 コピュラ

コピュラとは,多次元の確率変数ベクトルがあるとき1次元の各確率変数の累積分布関数を入力に,多次元の同時分布を出力とする関数のことである.

2.2.1 定義と基本的な性質

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$cdf_k(x) = prb[X_k \leq x]$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,

$$\begin{aligned} cdf(x_1, x_2, ..., x_n) &= C(cdf_1(x_1), cdf_2(x_2), ..., cdf_k(x_k)) \\ &= C(U) \end{aligned}$$

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. 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,

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

2.2.2 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(2.2.1)～式(2.2.3)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.

$$C_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k}(-log(u_i))^{\theta})^{\frac{1}{\theta}}) \tag{2.2.1}$$

- $Frank$ **コピュラ**

  C_{Frank} は u_i が 0.5 付近で相関関係が高くなる.

$$C_{Frank}(U) = \frac{1}{\theta}log(1 + \frac{\prod_{i=1}^{k}(\exp(-\theta\ u_i) - 1)}{\exp((-\theta) - 1)^{k-1}}) \tag{2.2.2}$$

- $Clayton$ **コピュラ**

  $C_{Clayton}$ は u_i が 0 付近で相関関係が高くなる.

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k}\frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \tag{2.2.3}$$

2.3 カーネル密度推定

カーネル密度推定 [12] は, 確率変数の確率密度関数を推定するノンパラメトリック手法の一つである. $x_1, x_2, ..x_n$ 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 (2.3.1) である.

$$\hat{pdf}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}K(\frac{x - x_i}{h}) \tag{2.3.1}$$

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]

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

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

第3章

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 [7] はコピュラを適合度統合に応用し, 式 (3.1.1) を適合度統合式として提案した.

$$C_{prod}(U_{rel}) = C(U_{rel}) \prod_{i=1}^{n} u_{rel,i} \tag{3.1.1}$$

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 (3.1.1) 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 (3.1.1) 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 [8].

3.2 混合コピュラ

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 (3.2.1) を提案した. 以下に混合コピュラの式を示す.

$$C_{mix}(U_{rel}) = \sum_{c=1}^{k} p_c C_c(U_{rel,c}) \tag{3.2.1}$$

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 (3.2.1) に加え, Eickhoff らの式 (3.1.1) を混合コピュラ用に拡張した式

$$C_{mix-prod}(U) = C_{mix}(U_{rel}) \prod_{i=1}^{n} \sum_{c=1}^{k} p_c u_{rel,c,i} \tag{3.2.2}$$

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.

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 (3.3.1) のように定義した.

$$\begin{aligned} Att_i &= \log_{1p}(D_{\mathrm{KL}}(ALL_i \| User_i)) \\ &= \log_{1p}(\int_{-\infty}^{\infty} pdf_{all}(x_i) \log \frac{pdf_{all}(x_i)}{pdf_{user}(x_i)}\, dx) \end{aligned} \tag{3.3.1}$$

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徴量を除いた S_{rdc} を式 (3.3.4) と式 (3.3.5) 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 *robust* な方法 [17] でメジアン Med と $MADN$(式 (3.3.3)) 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.

$$MAD(S_{Att}) = Med(\{|Att_i - Med(S_{Att})|\}) \tag{3.3.2}$$

$$MADN(S_{Att}) = \frac{MAD(S_{Att})}{0.675} \tag{3.3.3}$$

$$S_{emp} = \{i | Med(S_{Att}) + cns_a \cdot MADN(S_{Att}) \leq Att_i\} \tag{3.3.4}$$

$$S_{rdc} = \{i | Att_i \leq Med(S_{Att}) - cns_a \cdot MADN(S_{Att})\} \tag{3.3.5}$$

鈴木らは,S_{rdc} と S_{emp} と, 小松田らの混合コピュラ式 (3.2.2) から関心度を反映させた統合式 (3.3.7) 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.

$$C_{kl-emp}(U_{rdc}) = C_{mix}(U_{rdc}) \prod_{i \in S_{emp}} \sum_{c=1}^{k} p_c u_{rel,c,i} \tag{3.3.6}$$

$$C_{kl-emp-prod}(U_{rdc}) = \begin{cases} C_{mix-prod}(U_{rdc}) & \text{if } S_{emp} = \emptyset \\ C_{kl-emp}(U_{rdc}) & \text{otherwise} \end{cases} \tag{3.3.7}$$

第4章

提案手法

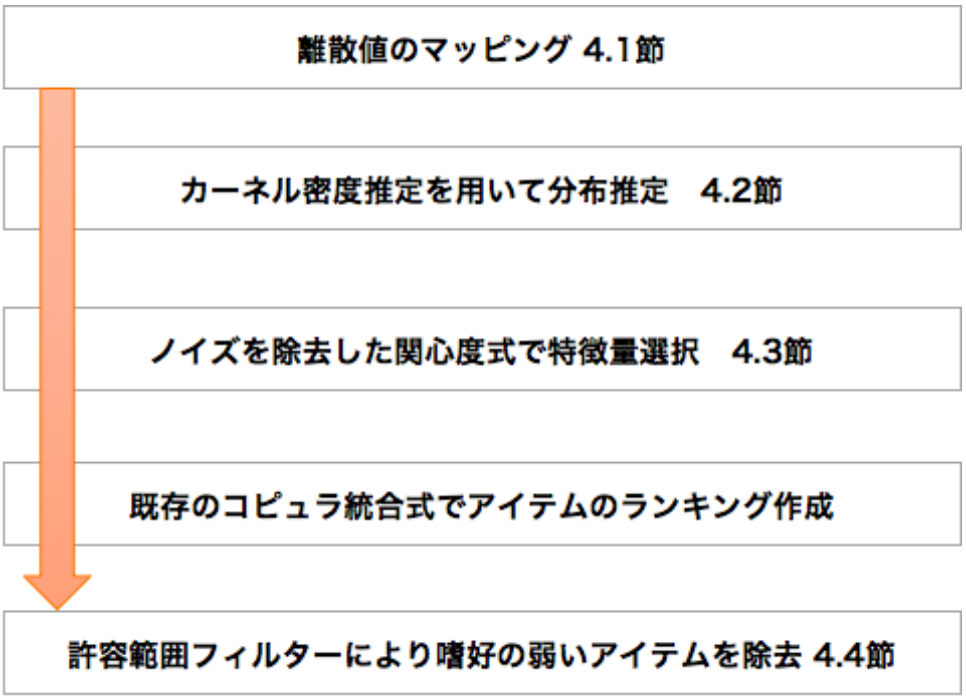

Fig. 4.1: 提案手法の処理チャート

本提案手法は以下の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を扱うためのものである.
提案手法の処理チャートは図4.1である.

1. 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 (4.1節)
2. カーネル密度推定を利用した複雑な分布への対応 (4.2節)
3. 鈴木らの関心度式からノイズ部分を除去して,離散値特徴量へ対応

(4.3節)

4. 許容範囲フィルターにより,特徴量値の特定区間にこだわりをもつような特異な嗜好ケースに対応 (4.4節)

4.1 離散値特徴量への数値マッピング

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.

このように嗜好順を考慮したマッピングをするために人気度pplを以下のように定義した. 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(4.1.1)で定義される.

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

$$ppl_i(v) = \frac{|S_{ScoreUser}(i,v)|}{|S_{ScoreAll}(i,v)|} \tag{4.1.1}$$

$score_i(item)$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利用した複雑な分布への対応

鈴木らのシステム[10]が特徴量分布の推定に正規分布モデルを用いていたのに対し,本研究ではカーネル密度推定を用いる. これにより,特徴量の分布が正

規分布にならない場合にも対応できることが期待できる. 式 (2.3.1) のパラメータであるカーネル関数式には式 (4.2.1) を用いた.

$$K(x) = \begin{cases} tophat(x) & (\text{特徴量}\, x \,\text{が離散値}) \\ gaussian(x) & (\text{特徴量}\, x \,\text{が連続値}) \end{cases} \tag{4.2.1}$$

$$tophat(x) = \begin{cases} \frac{1}{2} & (-1 \leq x \leq 1) \\ 0 & (otherwise) \end{cases} \tag{4.2.2}$$

$$gaussian(x) = \frac{1}{2\sqrt{\pi}\sigma} exp\left(-\frac{(x-u)^2}{2\sigma^2}\right) \tag{4.2.3}$$

4.2.1 バンド幅の決定

カーネル密度推定式 (式 (2.3.1)) のパラメータであるバンド幅 h の選択には式 (4.2.4) 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 2.3 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.

$$h_{opt} = \begin{cases} 0.1 & (K = tophat) \\ h_{G\_opt} & (K = gaussian) \end{cases} \tag{4.2.4}$$

$$S_{lct} = \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, ..., 5 \cdot 10^{-1}\} \tag{4.2.5}$$

$$S_{opt} = \{silverman, scott\} \tag{4.2.6}$$

$$S_{Grd} = S_{lct} \cup S_{opt} \tag{4.2.7}$$

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} から, 未評価の要素 h をとりだす.
3. 式 (4.2.9) のようにして S_{scored_i} で h の評価値を求める.
4. 式 (4.2.10) 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.

$$
\begin{aligned}
S_{scr\_i} &= S_{trn} \backslash S_{Nsrc\_i} && (4.2.8)\\
score(h,i) &= \sum_{x \in S_{scored\_i}} \log f(x) && (4.2.9)\\
score(h) &= \frac{1}{k} \sum_{i=1}^{k} score(h,i) && (4.2.10)
\end{aligned}
$$

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 以下になればよい. 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 ノイズを除去した関心度式

既存の鈴木らの関心度式 Att_i(式 (3.3.1)) 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したとき, 離散値の場合,$tophat$ を用いるためノイズが 0 になる. よって既存の関心度式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 (式 (4.3.1)) を Att_{i_Shr} として提案する.

$$
\begin{aligned}
Att_{i\_Shr} &= D_{\mathrm{KL}}(User_i \| ALL_i)\\
&= (\int_0^1 pdf_{user}(x_i) \log \frac{pdf_{user}(x_i)}{pdf_{all}(x_i)}\, dx) && (4.3.1)
\end{aligned}
$$

式 (4.3.1) 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 $pdf_{all}(x) \leq pdf_{user}(x)$ の部分をより反映させるために鈴木らが用いた 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, 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

TABLE 4.1: 特異な分布例

ROLE 番号	ROLE の説明
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.

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 4.1 の ROLE7 や ROLE14 のような状況において, 特徴量の数値が高いがユーザが関心を示していないアイテムが上位に推薦されてしまうという問題が生じため tlr の提案を行う.

tlr_i は特徴量 i と, 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, 式 (4.4.1) のように定義される.

$$tlr_i = \{x | pdf_{all,i}(x) \leq pdf_{user,i}(x)\} \tag{4.4.1}$$

tlr_i は, 特徴量 i について全アイテムの pdf である $pdf_{all,i}$ よりも, ユーザが関心を示したアイテムの pdf である $pdf_{user,i}$ が上回るような区間であり, ユーザが特に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. 許容範囲フィルターを特異な分布をとる特徴量 i の値が,tlr_i に含まれるアイテムを優先的に選択する手法として定義する. この手法を用いれば, 特定の特徴量値が特定の範囲にあるアイテムに強い嗜好を示すようなケースに対応できることが期待できる.

しかし, この許容範囲フィルターを有効にする特徴量をどのようにして選択するかという問題がある. 許容範囲フィルターを有効にする特徴量が多いほど, アイテムが推薦対象から外される可能性が高くなりフィルターが有効に機能しにくくなる.

よって, 特異な分布をとる特徴量を含むできるだけ小さい特徴量集合で, フィルターを有効にするのが望ましい. 特異な分布をとる特徴量の場合, ユーザはその特徴量に強い関心をもっていることが想定される. よって, 鈴木らの手法を用いて重要次元 S_{emp}(式 (3.3.4)) を抽出して,S_{emp} でフィルターを有効にするという

手法が候補に上がる. しかし,関心度が一部の特徴量で拮抗しているため重要な特徴量が外れ値として検出できないような場合に,特定の特徴量が特異な分布をとるにもかかわらずフィルターが有効にならないというケースを考慮して,フィルターを有効にする特徴量集合$S_{fitered}$を(式(4.4.2))のように定義する.

$$S_{filtered} = \begin{cases} S_{sorted}(n) & (S_{emp} = \phi) \\ S_{emp} & (otherwise) \end{cases} \tag{4.4.2}$$

(式(4.4.2))は重要次元S_{emp}が要素をもつ場合はS_{emp}の特徴量でフィルターを有効にし,S_{emp}が空の場合は鈴木らの関心度式で算出した関心度の降順でn番目までの特徴量からなる集合$S_{soreted}(n)$の特徴量でフィルターを有効にすることを意味する. nには,評価実験でU字型の分布が確認できた全特徴量を抽出できる整数のうち,最小の値である$n = 2$を用いた.

この$S_{filtered}$の各特徴量iの値がtlr_iに含まれるという条件を満たすようなアイテムを優先的に推薦することで,特異な分布をとらない特徴量による影響を抑えつつ許容範囲フィルターを用いて適切な推薦ができることが期待できる.

第5章

評価実験

提案手法が,既存の手法では扱えないような問題ケースで有効に機能するかを検証するために,離散値の特徴量を含む新たなデータセットを用いて評価実験を行う.また,問題ケースに限定せず既存の手法で扱えた通常のケースでも有効に機能するかを検証するために,旧データセットを用いた評価実験も行う.

5.1 実験条件

データセットには,鈴木らが用いた旧データセットと旧データセットに変更を加えた新データセットを用いた. 旧データセットは連続値のみを特徴量にもつデータセットである.これは,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, 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したものを特徴量としている.

新データセットは,この旧データセットに 9 個目の新たな特徴量として喫煙値を加えたものである. 喫煙値には喫煙可と禁煙の 2 値を定義し,各ホテルに各離散値を 50% の確率で割り当てた.

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 5.1 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.
3. ホテル評価時の喫煙値への立場として極度に悪い,悪い,中立,良い,極度に良い,の 5 段階で最も近いものを選択する.

TABLE 5.1: 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 比較手法

- **嗜好回答情報を利用した重み付け線形和**

$$LIN(X) = \sum_{i=1}^{n} w_i x_i \tag{5.2.1}$$

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その総和である100で割ったものである.

- **コピュラを用いた統合式**

$$method = C_{A,B,C}$$

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

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

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

鈴木らの$C_{kl-emp-prod}$(式(3.3.7))に,4章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,$C_{A,B,C}$のように表す.フィルターを用いる場合,Cの文字列は$'Tl'$であり,フィルターを用いない場合Cは$NULL$で空文字であることに注意する.

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て,関心度式には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,関心度式には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} * [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,最も精度が高いものを用いることにした.よって,旧データセットでは2^3,新データセットでは,2^4を用いた.

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ものを採用した.よって,旧データセットでは,鈴木らの手法で最高の精度を示したC_{gumbel}を用い,新データセットでは,提案手法で最高の精度を示した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ものを採用した.よって,旧データセットでは,鈴木らの手法で最高の精度を示した5を用い,新データセットでは,提案手法で,最高の精度を示した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ものを採用した.よって,旧データセットでは,鈴木らのシステムで最高の精度を示した2.5を採用し,新データセットでは,提案手法で最高の精度を示した1.5を採用した.

*https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

TABLE 5.2: ROLE別の被験者数と禁煙アイテム選択率

想定する分布	ROLE	被験者数	禁煙アイテム選択率
なし	ROLE1	2	0.55,0.56
	ROLE2	4	0.52,0.55,0.70,0.99
	ROLE3	1	0.62
	ROLE4	2	0.51,0.51
喫煙家または嫌煙家	ROLE5	2	0.43,0.55
	ROLE6	1	0.07
	ROLE9	1	0.89
	ROLE10	1	0.12
	ROLE11	1	0.9
	ROLE12	1	0.2
	ROLE13	2	1.0,1.0
特異な分布	ROLE7	6	0.48,0.53,0.57,0.65,0.66,0.96
	ROLE14	9	0.45, 0.47, 0.5, 0.53, 0.53, 0.55, 0.57, 0.77, 1.0

- *GridSearch*(式 (4.2.10)) 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.

コピュラモデル, クラスタ数, 正定数については, 新データセットでの候補比較の結果を付録 A に掲載した.

5.4 実験結果

5.4.1 ROLE別離散値選択

評価実験で用いた ROLE(表 5.1) に対して,ROLE 別の被験者数, 禁煙アイテム選択率は表 5.2 の通りである. 禁煙アイテム選択率は被験者が関心を示したアイテムに占める, 喫煙値が禁煙となるアイテムの割合と定義しているので, 禁煙アイテム選択率が高いほど禁煙のアイテムをより優先的に選択したことを意味する. 禁煙アイテムを優先的に選択するケース, 喫煙可能アイテムを優先的に選択するケース, アイテムの喫煙値に偏りがないケースの 3 つが少なくとも一つはサンプルを持つことが確認できる.

5.4.2 特異な分布

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5.1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5.1と通常の分布であるROLE1の分布図5.2を比較する. これらの図からは許容範囲tlrに関して通常の分布ではscoreの右端まで連続しているのに対して,ROLE7では右端付近で途切れているのが確認できる. よって, これに許容範囲フィルターを適用することで, ユーザが強い嗜好を示す範囲のアイテムを優先的に選択することができるため,ROLE7の場合でも適切な選択ができることが期待できる.

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り, 価格特徴量の分布が低価格帯と高価格帯で峰になるU字型の分布になることを想定したものである.

6,000円以下の低価格帯のホテルの場合, ホテルのレビュー値が低いというデメリットと個室利用できるというメリットがある. 一方,12,000円付近の高価格帯のホテルの場合, 個室利用ができないというデメリットと, ホテルのレビュー値が高いものが多いというメリットがある. 両価格帯の中間に位置する中価格帯の場合, 個室利用ができず, レビュー値も高価格帯のホテルより劣るものが多いため, 選択するメリットが小さい.

よって, ユーザはメリットの薄い中価格帯では選択を控え, メリットとデメリットが等しく存在する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5.3) 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.

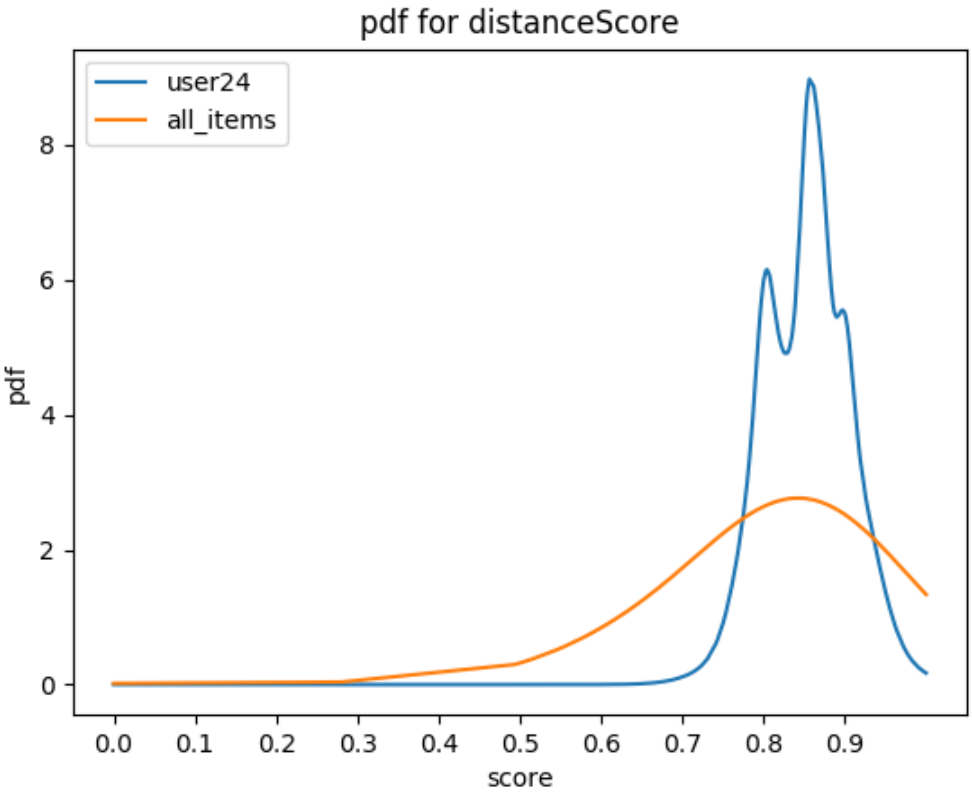

Fig. 5.1: ROLE7の距離特徴量分布

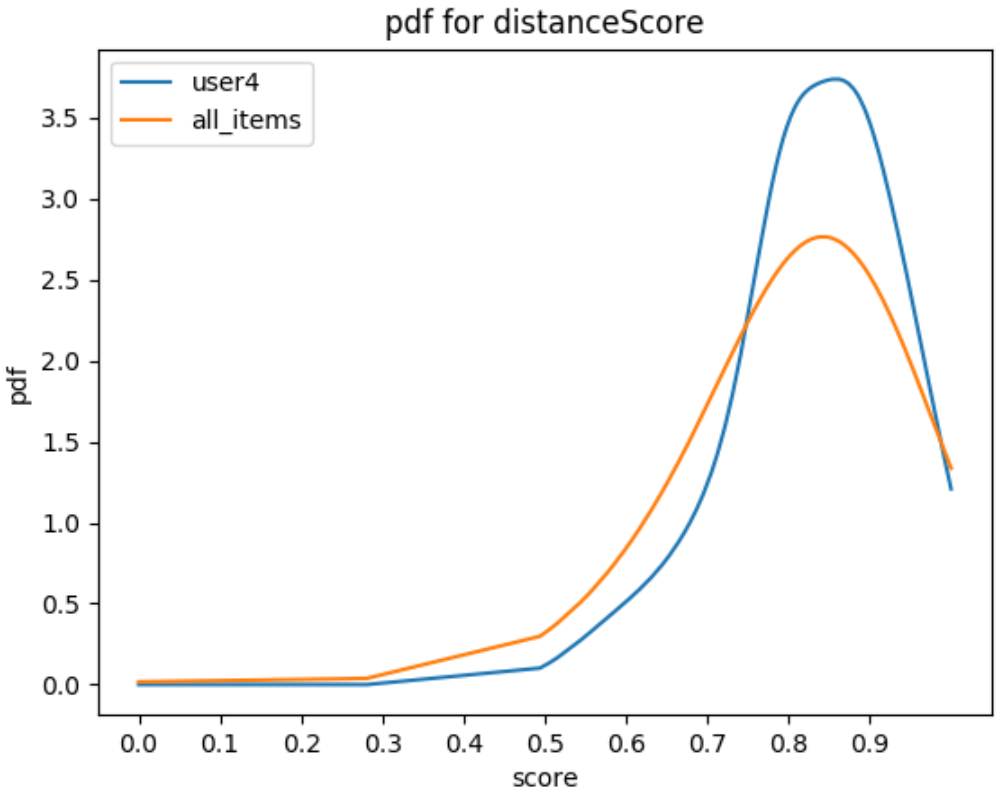

Fig. 5.2: ROLE1 の距離特徴量分布

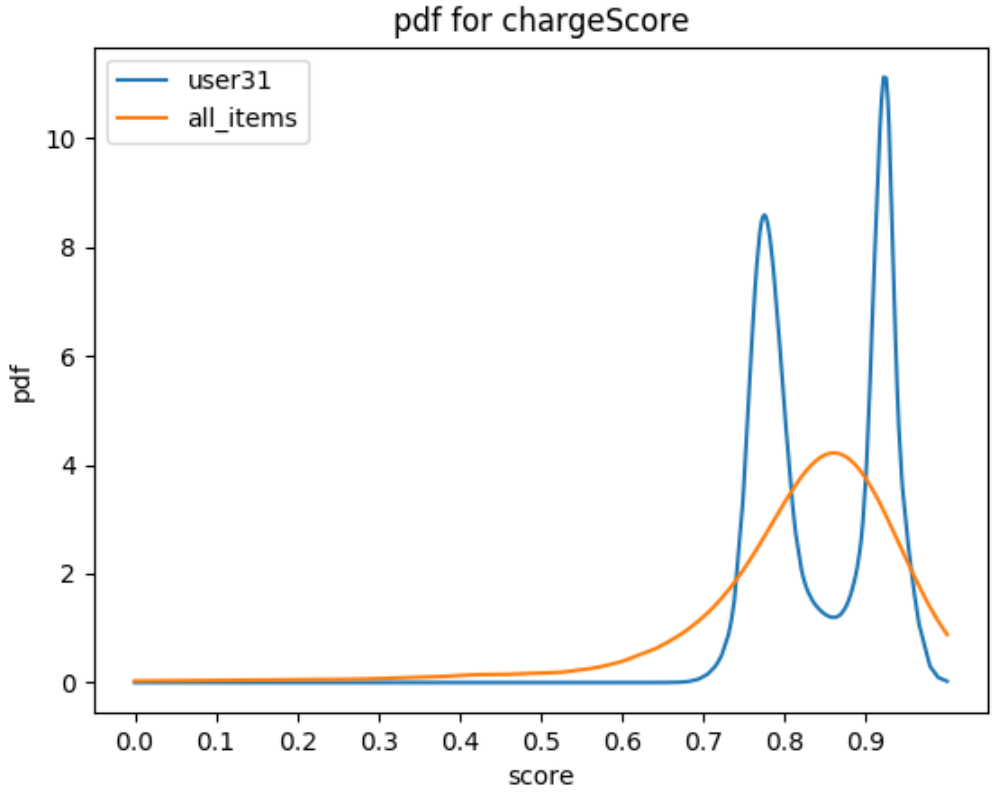

Fig. 5.3: ROLE14の価格特徴量分布

5.4.3 旧データセットでの実験結果

実験結果は表5.3である．鈴木らの既存手法は$C_{Nrm,Inf}$で,提案手法は$C_{Kd,Shr,Tl}$であることに注意する．提案手法と各比較手法で平均差がない,という仮説におけるマンホイットニーのU検定ときのp値の結果が表5.4である．

表5.4から,全ての評価指標で既存手法と提案手法でp値が30%以上であることがわかる

よって既存手法と提案手法に有意な差はないと判断できるため,旧データセットに対して提案手法は,既存手法と同等の性能を示すといえる．

TABLE 5.3: 旧データセットでの実験結果

measure	$C_{Kd,Shr,Tl}$(提)	$C_{Nrm,Inf}$(既)	LIN	SVM
iP@0	1.0	0.995	0.98	0.991
iP@0.1	0.994	0.992	0.959	0.986
iP@0.2	0.976	0.99	0.94	0.981
iP@0.3	0.954	0.955	0.918	0.965
iP@0.4	0.92	0.934	0.884	0.958
iP@0.5	0.899	0.922	0.849	0.945
iP@0.6	0.848	0.888	0.811	0.913
iP@0.7	0.808	0.845	0.772	0.863
iP@0.8	0.755	0.756	0.709	0.814
iP@0.9	0.696	0.64	0.633	0.741
iP@1.0	0.568	0.503	0.531	0.634
MAiP	0.856	0.856	0.817	0.89
nDCG@5	0.995	0.992	0.96	0.981
nDCG@10	0.987	0.987	0.957	0.98
nDCG@15	0.981	0.981	0.952	0.977
nDCG@20	0.975	0.978	0.949	0.974
nDCG@30	0.967	0.972	0.942	0.969
P@5	0.954	0.954	0.887	0.946
P@10	0.888	0.9	0.862	0.923
P@15	0.81	0.847	0.799	0.857
P@20	0.764	0.775	0.732	0.794
P@30	0.648	0.638	0.617	0.669

5.4.4 新データセットでの実験結果

実験結果は表5.5である.

検索結果上位に対する評価指標を含めた過半数の評価指標で,提案手法が比較手法よりも優れた結果を示している.

提案手法と各比較手法で平均差がない,という帰無仮説におけるマンホイットニーのU検定をしたときのp値の結果が表5.6である. 既存手法におけるp値は全評価指標において1%未満である. よって,提案手法は既存手法と比較して有意水準1%以下で性能がよいといえる.

以下で指摘した問題ケースごとの結果について分析する.

- **喫煙値について偏った嗜好をもつケース**

TABLE 5.4: 旧データセットでの提案手法との比較におけるU検定でのp値

measure	$C_{Nrm,Inf}$(既)	LIN	SVM
iP@0	0.359	0.079	0.166
iP@0.1	0.965	0.013	0.383
iP@0.2	0.449	0.112	0.777
iP@0.3	0.952	0.13	0.953
iP@0.4	0.93	0.224	0.485
iP@0.5	0.644	0.175	0.401
iP@0.6	0.544	0.47	0.341
iP@0.7	0.488	0.583	0.47
iP@0.8	0.931	0.544	0.402
iP@0.9	0.47	0.624	0.402
iP@1.0	0.544	0.773	0.371
MAiP	0.977	0.402	0.436
nDCG@5	0.535	0.055	0.468
nDCG@10	0.839	0.093	0.643
nDCG@15	0.954	0.088	0.794
nDCG@20	0.664	0.106	1.0
nDCG@30	0.665	0.184	0.908
P@5	0.697	0.084	0.622
P@10	0.861	0.561	0.6
P@15	0.562	0.817	0.524
P@20	0.908	0.665	0.729
P@30	0.862	0.729	0.84

結果は表5.8である.

提案手法が比較手法の過半数で,SVM に次いで,最高の結果を示しており,4章で提案した各小手法を部分的に組み合わせたどの手法よりも,優れた結果を示している. また,関心度式に Att_{Shr} を採用する手法が,Att_{Inf} を採用する手法よりも優れた結果を示す傾向が通常のケース(表5.7)より顕著である.

よって離散値特徴量を扱う場合,Att_{Shr} が Att_{Inf} よりもより有効に機能するといえる.

- **特異な分布ケース**

表5.1のROLE7やROLE14のような特異な嗜好ケースに対する結果は,表5.9である.

提案手法が比較手法の過半数で,最高の結果を示している.また,フィルターを有効にする手法が,フィルターを有効にしない手法より優れた結果を示す傾向が通常のケース(図5.7)よりも顕著である.

よって,フィルターが有効に機能しているといえる.

通常のケースだけでなく,離散値について偏った嗜好を持つケース(表5.8)や,表5.1のROLE7やROLE14のような特異な嗜好ケース(表5.9)においても,提案手法は過半数の評価指標で最高の結果あるいはSVMに次ぐ結果を示すことから,提案手法は他の比較手法と比較して評価実験で用いたような複雑な嗜好ケースにおいても安定した結果を示す手法であることがわかる.

TABLE 5.5: 新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
nDCG@30	0.945	0.728	0.918	0.898	0.879	0.852	0.928
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
P@30	0.597	0.432	0.562	0.586	0.564	0.541	0.614

TABLE 5.6: 新データセットでの提案手法との比較におけるU検定でのp値

measur	$C_{Nrm,Inf}$(既)	$C_{Kd,Inf,Tl}$	$C_{Kd,Shr}$	$C_{Kd,Inf}$	LIN	SVM
iP@0	0.0	0.251	0.064	0.011	0.0	0.458
iP@0.1	0.0	0.324	0.125	0.02	0.005	0.707
iP@0.2	0.0	0.035	0.036	0.017	0.003	0.631
iP@0.3	0.0	0.041	0.047	0.01	0.001	0.37
iP@0.4	0.0	0.027	0.055	0.006	0.0	0.089
iP@0.5	0.0	0.036	0.139	0.024	0.003	0.234
iP@0.6	0.0	0.021	0.161	0.036	0.008	0.215
iP@0.7	0.0	0.015	0.18	0.035	0.018	0.306
iP@0.8	0.0	0.04	0.308	0.084	0.059	0.564
iP@0.9	0.0	0.064	0.313	0.119	0.144	0.682
iP@1.0	0.001	0.21	0.513	0.263	0.338	0.8
MAiP	0.0	0.043	0.197	0.029	0.017	0.561
nDCG@5	0.0	0.231	0.048	0.014	0.001	0.497
nDCG@10	0.0	0.174	0.086	0.023	0.001	0.482
nDCG@15	0.0	0.125	0.093	0.024	0.001	0.487
nDCG@20	0.0	0.089	0.095	0.021	0.002	0.396
nDCG@30	0.0	0.066	0.115	0.019	0.003	0.409
P@5	0.0	0.178	0.073	0.04	0.003	0.484
P@10	0.0	0.05	0.084	0.022	0.004	0.369
P@15	0.0	0.125	0.117	0.052	0.01	0.439
P@20	0.0	0.151	0.271	0.135	0.051	0.579
P@30	0.003	0.246	0.424	0.261	0.168	0.633

TABLE 5.7: 新データセットにおける通常の分布の実験結果

measure	$C_{Kd,Shr,Tl}$(提)	$C_{Nrm,Inf}$(既)	$C_{Kd,Inf,Tl}$	$C_{Kd,Shr}$	$C_{Kd,Inf}$	LIN	SVM
iP@0	0.973	0.705	0.962	0.918	0.903	0.836	0.936
iP@0.1	0.958	0.632	0.942	0.873	0.86	0.83	0.921
iP@0.2	0.947	0.578	0.905	0.833	0.823	0.806	0.902
iP@0.3	0.919	0.535	0.881	0.802	0.791	0.761	0.869
iP@0.4	0.895	0.511	0.864	0.784	0.774	0.74	0.839
iP@0.5	0.856	0.499	0.833	0.758	0.754	0.701	0.808
iP@0.6	0.818	0.477	0.776	0.719	0.72	0.652	0.771
iP@0.7	0.785	0.462	0.729	0.691	0.68	0.619	0.733
iP@0.8	0.719	0.448	0.663	0.645	0.637	0.583	0.691
iP@0.9	0.657	0.425	0.61	0.594	0.596	0.555	0.618
iP@1.0	0.559	0.412	0.54	0.535	0.532	0.509	0.544
MAiP	0.826	0.517	0.791	0.741	0.734	0.69	0.785
nDCG@5	0.953	0.684	0.947	0.863	0.864	0.733	0.92
nDCG@10	0.951	0.702	0.937	0.877	0.862	0.776	0.919
nDCG@15	0.944	0.717	0.934	0.875	0.868	0.798	0.919
nDCG@20	0.941	0.718	0.928	0.875	0.87	0.807	0.916
nDCG@30	0.935	0.727	0.921	0.874	0.869	0.821	0.911
P@5	0.869	0.46	0.85	0.712	0.725	0.627	0.829
P@10	0.819	0.472	0.784	0.697	0.699	0.611	0.779
P@15	0.758	0.447	0.728	0.654	0.672	0.584	0.726
P@20	0.697	0.449	0.673	0.627	0.628	0.555	0.683
P@30	0.609	0.441	0.584	0.573	0.567	0.514	0.601

TABLE 5.8: 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
nDCG@30	0.971	0.729	0.911	0.96	0.907	0.933	0.974
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
P@30	0.568	0.407	0.502	0.621	0.556	0.613	0.648

TABLE 5.9: 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
nDCG@30	0.907	0.671	0.884	0.809	0.799	0.733	0.868
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
P@30	0.505	0.341	0.467	0.441	0.433	0.373	0.482

第6章

結論

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式や,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,特徴量の分布に正規分布を仮定している点や,離散値の特徴量を扱えない点や,特異な分布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,正規分布モデルをカーネル密度推定に変更する手法や,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式 Att_{Shr} を用いる手法や,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,旧データセットでは同程度の性能を示し,指摘した問題ケースを含む新たなデータセットでは,有意水準1%以下でより高い性能を示すことを確認できた. 特に新たなデータセットでは,提案手法が比較手法の中でも機械学習手法である SVM よりも高い性能を示すことを確認できた. よって,提案手法は通常の嗜好ケースで既存手法と同程度の性能を維持しつつ,特異な嗜好に対して既存手法や SVM のような機械学習手法よりも高い性能を示すといえる.

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(4.1.1)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と,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

本研究をすすめるにあたり, 研究テーマを設定して頂き, 終始熱心なご指導してくださった宮崎純教授に深く感謝いたします. また, 日頃から研究について細部にわたり多くのご助言を頂きました欅惇志助教に深く感謝いたします. 事務, 書類手続き等をサポートして頂きました, 秘書の麻生さん, 事務の三浦さんに感謝いたします. ホテルのデータセットを提供して頂いた楽天株式会社に感謝いたします. 実験の協力を快く引き受けていただいた宮崎研究室の皆さんに感謝いたします. 最後に, 学生生活を支援してくださった家族に感謝いたします.

参考文献

[1] Dietmar Jannach,Markus Zanker, Alexander Felfernig,Gerhard Friendrich, (田中克己, 角谷和俊). 情報推薦システム入門,pp.13-21, 共立出版,2012.

[2] 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.

[3] 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.

[4] 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.

[5] C. Cortes and V. Vapnik. "Support vector networks". Machine Learning, 20:pp.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 1831-1834. 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] 鈴木崇弘, 欅 惇志, 宮崎 純. コピュラを用いたユーザー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ー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.

付録 A

パラメータ候補の比較

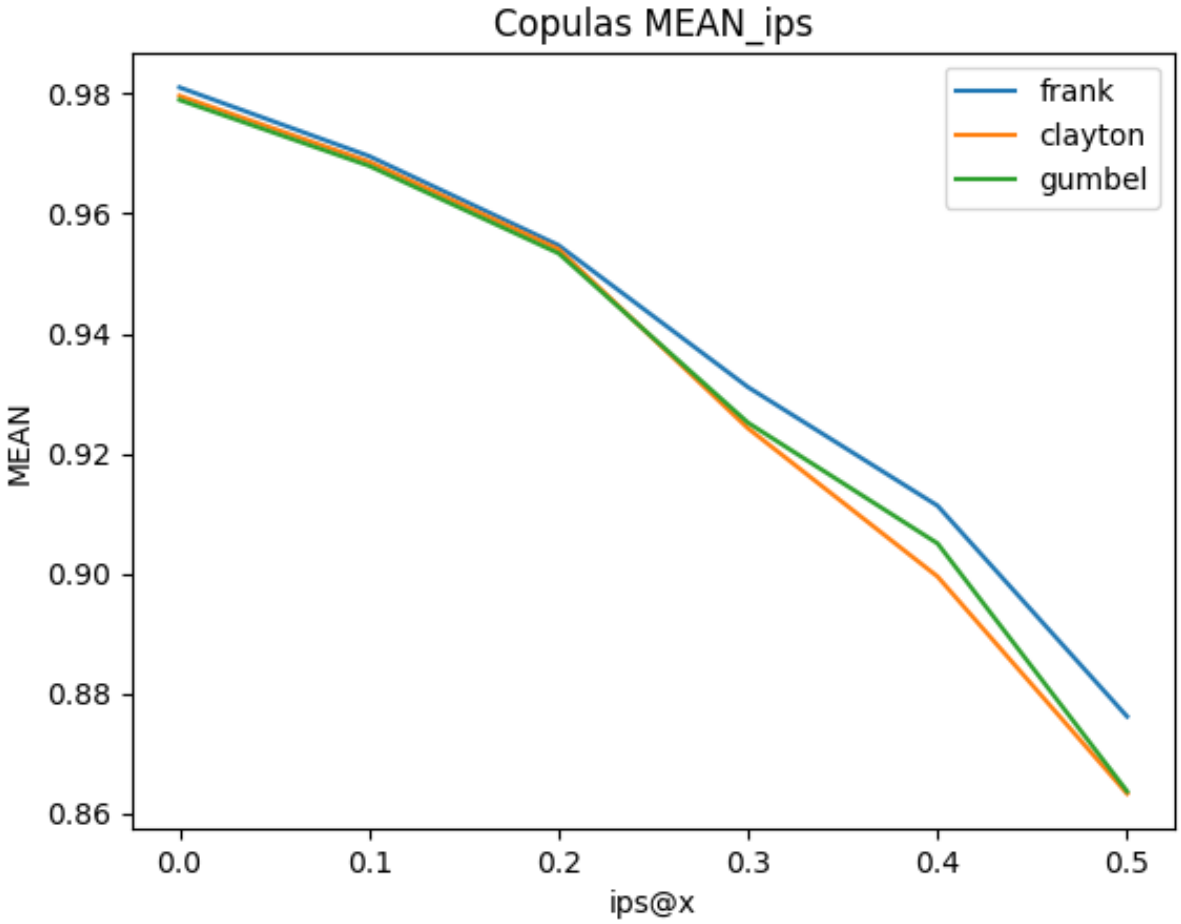

Fig. A.1: コピュラの iP 比較

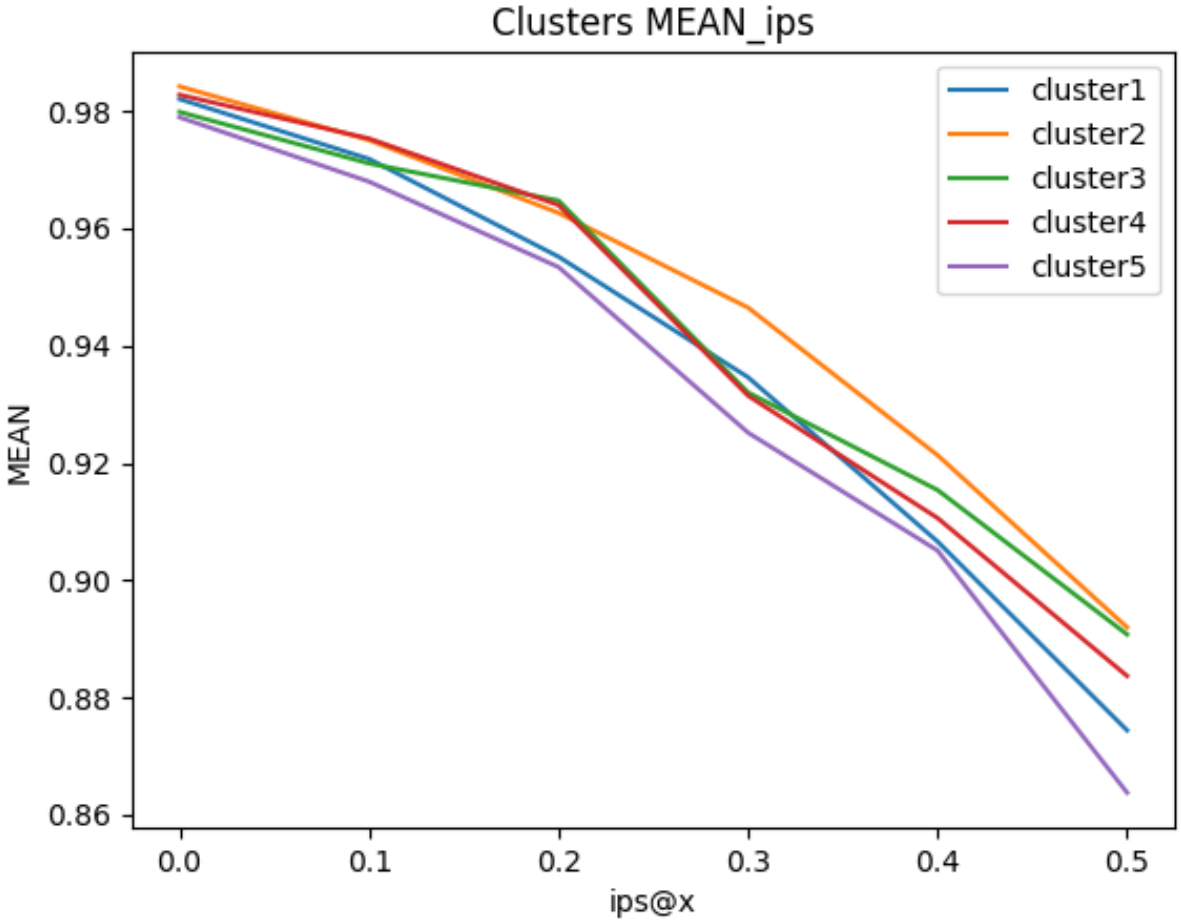

Fig. A.2: クラスタ数の iP 比較

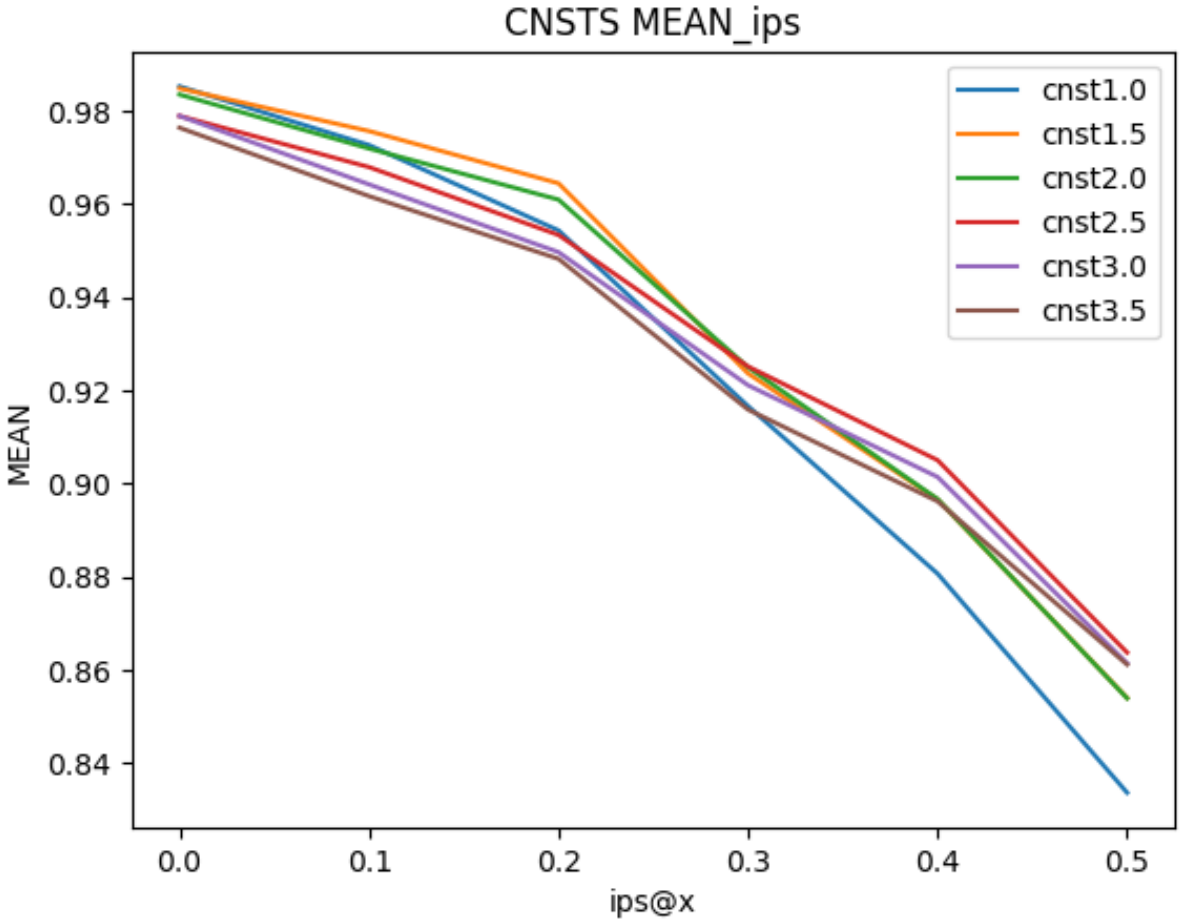

Fig. A.3: 正定数 cnst の iP 比較