DOI: 10.11992/tis.202209042
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230801.1521.006

基于 Hellinger 距离的正态云相似性度量方法及应用研究

许昌林[1,2], 徐浩[1]

（1. 北方民族大学 数学与信息科学学院, 宁夏 银川 750021; 2. 北方民族大学 宁夏智能信息与大数据处理重点实验室, 宁夏 银川 750021）

摘　要: 针对现有正态云相似性度量计算复杂度较高且区分度不强等问题, 本文首先从正态云的特征曲线出发, 利用 Hellinger 距离刻画两个概率分布相似性的特点, 提出一种基于 Hellinger 距离的正态云相似性度量方法, 该方法不仅考虑了云概念的数字特征且兼顾了其分布特性, 并对相似度量具有的数学性质进行了研究。其次, 根据给出的相似度量方法, 设计了两种正态云概念的相似度算法。最后, 通过数值模拟仿真实验和时间序列数据分类实验对所提出算法的性能进行对比分析, 结果表明该算法具有较好的相似度区分能力且分类错误率和 CPU 时间代价都较低。同时, 将本文方法应用于协同过滤推荐系统中, 并在 MovieLens100k 影评数据集上进行了实验, 实验结果表明本文方法在用户评分数据极端稀疏的情况下, 仍能取得较理想的推荐质量。

关键词: 知识表示; 正态云; 不确定性; Hellinger 距离; 特征曲线; 相似性度量; 协同过滤; 推荐系统
中图分类号: TP311　　**文献标志码**: A　　**文章编号**: 1673−4785(2023)06−1312−10

中文引用格式: 许昌林, 徐浩. 基于 Hellinger 距离的正态云相似性度量方法及应用研究 [J]. 智能系统学报, 2023, 18(6): 1312–1321.
英文引用格式: XU Changlin, XU Hao. Similarity measurement method for normal cloud based on Hellinger distance and its application[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1312–1321.

Similarity measurement method for normal cloud based on Hellinger distance and its application

XU Changlin[1,2], XU Hao[1]

(1. School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Intelligent Information and Big Data Processing of NingXia Province, North Minzu University, Yinchuan, 750021, China)

Abstract: To address the problems of high computational complexity and weak discrimination of existing normal cloud model similarity measurement methods, a similarity measurement method for normal clouds based on Hellinger distance is proposed according to the characteristic curve of the normal cloud by taking inspiration from the similarity of two probability distributions described by Hellinger distance. The digital and distribution characteristics of the cloud concept were considered in the proposed method. Furthermore, the mathematical properties of the proposed similarity measurement were studied. Two similarity algorithms were then designed for the normal cloud concept on the basis of the given similarity measurement method. Finally, the performance of the proposed algorithms was compared and analyzed using numerical simulation and classification experiments on time-series data. Results showed that the proposed algorithms have good similarity discrimination capability, and their classification error rate and CPU time cost are low. Moreover, these algorithms were applied to the collaborative filtering recommendation system, and experiments were conducted on the MovieLens100k film review dataset. The experimental results revealed that the proposed methods can continue to achieve ideal recommendation quality even when the user rating data were extremely sparse.

Keywords: knowledge representation; normal cloud; uncertainty; Hellinger distance; characteristic curve; similarity measurement; collaborative filtering; recommendation system

随着信息技术不断进步, 信息过载问题日益突出推荐系统是解决过载问题的一种手段, 由 Goldberg 等[1] 提出的协同过滤推荐系统是应用最广泛的一种, 已被阿里巴巴、亚马逊等电商平台

收稿日期: 2022−09−20.　**网络出版日期**: 2023−08−01.
基金项目: 宁夏自然科学基金项目（2022AAC03238, 2023-AAC05046）; 国家自然科学基金项目（62066001）; 宁夏高等教育一流学科建设基金项目（NXYLXK-2017B09）.
通信作者: 许昌林. E-mail: xu_changlin@nun.edu.cn.

广泛应用。云模型作为研究不确定性的一种工具, 能有效处理推荐系统中的不确定信息[2-4], 同时云模型在用户识别[5]、多属性决策与优化[6-7]、综合评价[8] 等领域也得到广泛应用, 其中云概念相似度扮演重要角色。因此, 构造合适的相似度不仅能够降低计算复杂度而且能够提升运行效率。如张光卫等[3] 将云概念数字特征作为向量构造夹角余弦得到云概念相似性比较方法 (likeness comparing method based on cloud model, LICM), 并将其应用于协同过滤推荐。但 LICM 将各数字特征赋予相同权重, 而数字特征中期望往往大于熵和超熵, 导致 LICM 区分能力较弱。李海林等[9] 利用云概念几何特征提出了基于期望曲线的云模型 (expectation based cloud model, ECM) 相似度和基于最大边界曲线的云模型 (maximum boundary based cloud model, MCM) 相似度, 区分度较好但当云概念数量增加时, ECM 和 MCM 计算复杂度会急剧增加。汪军等[10] 将云概念形状相似性和距离相似性结合构建了云概念综合相似度量 PDCM (shape and distance based on cloud model), 并将其应用到分类问题中取得了一定效果, 而参数拟合和选择会影响精度。此外, 有学者从贴近度、概念跃升、散度和多粒度等方面给出云概念相似度[11-14], 取得一定效果。Li 等[15] 从区分性、有效性、稳定性和可解释性方面分析了以上相似度方法优缺点。

基于此, 本文主要工作: 1) 从正态云的特征曲线 (如期望曲线、内外包络曲线等) 出发, 融合正态云的分布特性, 利用 Hellinger 距离刻画概率分布间相似性的特点, 提出了一种基于 Hellinger 距离的正态云相似性度量方法, 该方法兼顾了云概念的数字特征和分布特性, 并讨论了所提方法的性质; 2) 设计了两种正态云相似度算法, 即基于 Hellinger 距离及期望曲线的正态云相似性度量方法 (Hellinger distance based expectation curve of cloud model, HECM) 和基于 Hellinger 距离及特征曲线的正态云相似性度量方法 (Hellinger distance based characteristic curve of cloud model, HCCM), 并将这 2 种方法与已有 LICM、ECM、MCM 和 PDCM 方法从 3 个方面进行对比分析。首先进行数值模拟仿真实验, 利用云概念差异度指标验证了本文方法具有较好的区分能力和可行性; 其次, 在时间序列数据集上进行分类实验, 结合分类错误率和 CPU 时间代价进行对比分析, 结果表明本文方法具有较好的分类性能且时间代价较低; 最后, 将本文方法应用于协同过滤推荐系统, 在电影数据集 MovieLens 100k 上进行实验分析, 采用平均绝对偏差和均方根误差指标进行精度度量, 实验结果表明本文方法在用户评分数据极端稀疏的情况下, 仍能取得较理想的推荐质量。

1　正态云及现有云概念相似度方法

1.1　正态云及正态云变换

云模型由数字特征描述不确定性概念整体特性。不同概率分布的云构成不同云模型, 鉴于正态分布的重要性和钟型隶属函数的普适性[16], 正态云模型及其相关应用得到了广泛研究, 相关定义如下。

定义 1[2,17]　设 U 是一个用精确数值表示的定量论域, C 是论域 U 上的定性概念, 若定量值 $x\in U$, 且 x 是定性概念 C 的一次随机实现, x 对 C 的确定度为 $\mu(x)\in[0,1]$ 是具有稳定倾向的随机数, 则 x 在论域 U 上的分布称为云, 每个 x 称为一个云滴。

定义 2[17]　设 U 是一个用精确数值表示的定量论域, C 是 U 上用数字特征 (E_x,E_n,H_e) 表示的定性概念。若定量值 $x\in U$, 且 x 是定性概念 C 的一次随机实现, 若 x 满足: $x=R_N(E_x,|y|)$, 其中, $y=R_N(E_n,H_e)$, 且 x 对 C 的确定度满足:

$$\mu(x)=\exp\left(-\frac{(x-E_x)^2}{2y^2}\right) \tag{1}$$

则 x 在论域 U 上的分布称为二阶正态云。这里 $y=R_N(E_n,H_e)$ 表示以 E_n 为期望, 以 H_e 为标准差的正态随机数。

正态云主要通过正态云变换实现定性概念与定量数值间的相互转换, 其中正向正态云变换将表征概念内涵的数字特征 $C(E_x,E_n,H_e)$ 转化为定量数值。根据定义 2, 二阶正向正态云变换 (the 2nd-order forward normal cloud transformation, 2nd-FNCT) 见算法 1。比如用数字特征 $C(25, 3, 0.5)$ 表征定性概念“年轻人”的内涵[17], $E_x=25$ 表示对“年轻人”的总体期望年龄, 由算法 1 可得“年轻人”的云图如图 1 所示。

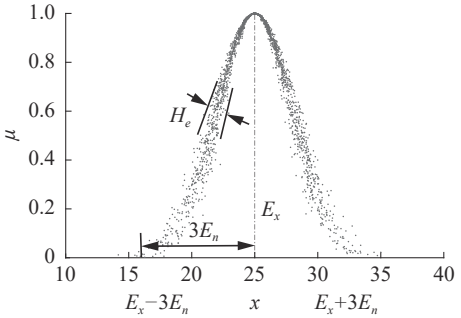


图 1　概念“年轻人”$C(25, 3, 0.5)$ 的云图

Fig. 1　Cloud map of concept “young people” $C(25, 3, 0.5)$

算法 1[17]　2nd-FNCT 算法

输入　三个数字特征 (E_x, E_n, H_e) 和云滴个数 n

输出　n 样本点 (云滴) x_i 和 $\mu(x_i)(i=1,2,\cdots,n)$

1) 以 E_n 为期望, H_e 为标准差, 生成一个正态随机数 $y_i = R_N(E_n, H_e)$;

2) 以 E_x 为期望, $|y_i|$ 为标准差, 生成一个正态随机数 $x_i = R_N(E_x, |y_i|)$;

3) 计算确定度 $\mu(x_i) = \exp\left(-\dfrac{(x_i - E_x)^2}{2y_i^2}\right)$;

4) 具有确定度 $\mu(x_i)$ 的 x_i 成为数域中的一个云滴, 重复步骤 1) ~3), 直至产生要求的 n 个云滴 x_i 为止。

逆向云变换是将定量数值有效转换为由数字特征 $C(E_x, E_n, H_e)$ 表示的定性概念。目前已有多种逆向云变换算法[17], 本文使用基于样本一阶绝对中心矩和样本方差的逆向云变换算法 (single-step backward cloud transformation algorithm based on the first-order absolutely center moment, SBCT-1stM), 如算法 2 所示。

算法 2[17]　SBCT-1stM 算法

输入　样本点 $x_i(i=1,2,\cdots,n)$

输出　反映定性概念数字特征的估计值 $\hat{E}_x$, $\hat{E}_n, \hat{H}_e$

1) 根据样本点 x_i 计算样本均值 $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, 一阶样本绝对中心矩 $\dfrac{1}{n}\sum_{i=1}^{n}|x_i - \bar{X}|$ 和样本方差 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$;

2) 分别计算期望、熵和超熵的估计值:

$$\hat{E}_x = \bar{X}, \hat{E}_n = \sqrt{\frac{\pi}{2}} \times \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{E}_x|, \hat{H}_e = \sqrt{S^2 - \hat{E}_n^2}。$$

1.2　二阶正态云的特征曲线

特征曲线能够在一定程度上反映云概念的几何特征。由定义 2 和正态分布的“3σ”原则知:

$$P(E_n - 3H_e \leqslant y \leqslant E_n + 3H_e) = 0.997\,4$$

当 $0 < H_e < E_n/3$ 时, 有 99.74% 的云滴确定度 $\mu(x)$ 处于曲线 $\mu_{\text{Out}}(x)$ 与 $\mu_{\text{In}}(x)$ 之间的区域[17-19], 其中

$$\mu_{\text{Out}}(x) = \exp\left\{-\frac{(x - E_x)^2}{2(E_n + 3H_e)^2}\right\} \tag{2}$$

$$\mu_{\text{In}}(x) = \exp\left\{-\frac{(x - E_x)^2}{2(E_n - 3H_e)^2}\right\} \tag{3}$$

则称 $\mu_{\text{Out}}(x)$ 和 $\mu_{\text{In}}(x)$ 分别为二阶正态云的外包络曲线和内包络曲线 (如图 2)。当超熵 H_e=0 时, 云滴确定度聚集分布在曲线上, 称 $\mu_{\text{Exp}}(x)$ 为二阶正态云的期望曲线 (如图 2)。因此, 对定性概念有贡献的云滴 99.74% 都落在区间 $[E_n - 3H_e, E_n + 3H_e]$ 中, 本文正是基于这一特点来构建云概念相似度量。

$$\mu_{\text{Exp}}(x) = \exp\left\{-\frac{(x - E_x)^2}{2E_n^2}\right\} \tag{4}$$

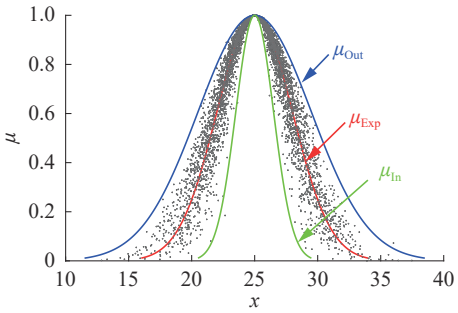


图 2　二阶正态云概念 C(25, 3, 0.5) 的特征曲线

Fig. 2　Characteristic curve of 2nd- order normal cloud concept C (25, 3, 0.5)

1.3　现有正态云概念相似度算法

根据前文所述, 下面对已有正态云概念相似度算法 LICM、ECM、MCM 和 PDCM 进行简要介绍。

算法 3[3]　LICM 算法

输入　数字特征 $C_1(E_{x_1}, E_{n_1}, H_{e_1})$ 和 $C_2(E_{x_2}, E_{n_2}, H_{e_2})$

输出　相似度 $S_{\text{LICM}}(C_1, C_2)$

1) 令 $\boldsymbol{U}_1 = (E_{x_1}, E_{n_1}, H_{e_1})$, $\boldsymbol{U}_2 = (E_{x_2}, E_{n_2}, H_{e_2})$

2) 计算 C_1, C_2 之间的相似度:

$$S_{\text{LICM}}(C_1, C_2) = \cos\langle \boldsymbol{U}_1, \boldsymbol{U}_2\rangle = \frac{\boldsymbol{U}_1 \cdot \boldsymbol{U}_2}{\|\boldsymbol{U}_1\|\|\boldsymbol{U}_2\|}$$

算法 4[9]　ECM 算法

输入　数字特征 $C_1(E_{x_1}, E_{n_1}, H_{e_1})$ 和 $C_2(E_{x_2}, E_{n_2}, H_{e_2})$

输出　相似度 $S_{\text{ECM}}(C_1, C_2)$

1) 若 $E_{x_1} \leqslant E_{x_2}$ 且初始设置 $S=0$, 计算两云概念期望曲线 $\mu_{\text{Exp}}(x_1)$ 与 $\mu_{\text{Exp}}(x_2)$ 的交点 $x_0^{(1)}$ 与 $x_0^{(2)}$, 设 $x_0^{(1)} \leqslant x_0^{(2)}$;

2) 若 $x_0^{(1)} \leqslant \min(E_{x_1} - 3E_{n_1}, E_{x_2} - 3E_{n_2})$

且 $x_0^{(2)} > \max(E_{x_1} + 3E_{n_1}, E_{x_2} + 3E_{n_2})$, 则 $S_{\text{ECM}}(C_1, C_2) = 0$; 否则, 执行 3);

3) 若 $x_0^{(1)} \geqslant \max(E_{x_1} - 3E_{n_1}, E_{x_2} - 3E_{n_2})$

且 $x_0^{(2)} \leqslant \min(E_{x_1} + 3E_{n_1}, E_{x_2} + 3E_{n_2})$, 由 E_{n_1} 和 E_{n_2} 大小计算面积:

$$S = S_1 + S_2 + S_3$$

$$S_1 = \sqrt{2\pi}E_{n_1}\int_{-\infty}^{x_0^{(1)}} \mu_{\text{Exp}}(x_1)\mathrm{d}x$$

$$S_2 = \sqrt{2\pi}E_{n_2}\int_{x_0^{(1)}}^{x_0^{(2)}} \mu_{\text{Exp}}(x_2)\mathrm{d}x$$

$$S_3 = \sqrt{2\pi}E_{n_1}\int_{x_0^{(2)}}^{\infty} \mu_{\text{Exp}}(x_1)\mathrm{d}x$$

否则执行 4);

4) 在其他情况下, $x_0^{(1)}$ 或 $x_0^{(2)}$ 会落在区间 $[E_{x_2} -$

$3E_{n_2}, E_{x_1}+3E_{n_1}]$中, 即$S=S_1+S_2$;

5) 计算$S_{\mathrm{ECM}}(C_1,C_2)=\dfrac{2S}{\sqrt{2\pi}(E_{n_1}+E_{n_2})}$。

算法 5[9]　MCM 算法

输入　数字特征$C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$

输出　相似度$S_{\mathrm{MCM}}(C_1,C_2)$

1) 若$E_{x_1}\leqslant E_{x_2}$且初始设置$S=0$, 计算两云概念外包络曲线$\mu_{\mathrm{Out}}(x_1)$与$\mu_{\mathrm{Out}}(x_2)$的交点$x_0^{(1)}$与$x_0^{(2)}$, 设$x_0^{(1)}\leqslant x_0^{(2)}$, 令$E_{H_1}=E_{n_1}+3H_{e_1}$, $E_{H_2}=E_{n_2}+3H_{e_2}$。

2) 若$x_0^{(1)}\leqslant\min(E_{x_1}-3E_{H_1},E_{x_2}-3E_{H_2})$且$x_0^{(2)}>\max(E_{x_1}+3E_{H_1},E_{x_2}+3E_{H_2})$时, $S_{\mathrm{MCM}}(C_1,C_2)=0$; 否则, 执行 3)。

3) 若$x_0^{(1)}\geqslant\max(E_{x_1}-3E_{H_1},E_{x_2}-3E_{H_2})$且$x_0^{(2)}\leqslant\min(E_{x_1}+3E_{H_1},E_{x_2}+3E_{H_2})$, 由$E_{H_1}$和$E_{H_2}$大小计算面积:

$$S=S_1+S_2+S_3$$

$$S_1=\sqrt{2\pi}E_{H_1}\int_{-\infty}^{x_0^{(1)}}\mu_{\mathrm{Out}}(x_1)\mathrm{d}x$$

$$S_2=\sqrt{2\pi}E_{H_2}\int_{x_0^{(1)}}^{x_0^{(2)}}\mu_{\mathrm{Out}}(x_2)\mathrm{d}x$$

$$S_3=\sqrt{2\pi}E_{H_1}\int_{x_0^{(2)}}^{\infty}\mu_{\mathrm{Out}}(x_1)\mathrm{d}x$$

否则, 执行 4)。

4) 在其他情况下, $x_0^{(1)}$或$x_0^{(2)}$会落在区间$[E_{x_2}-3E_{H_2},E_{x_1}+3E_{H_1}]$中, 即$S=S_1+S_2$。

5) 计算$S_{\mathrm{MCM}}(C_1,C_2)=\dfrac{2S}{\sqrt{2\pi}(E_{H_1}+E_{H_2})}$。

算法 6[10]　PDCM 算法

输入　数字特征$C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$

输出　相似度$S_{\mathrm{PDCM}}(C_1,C_2)$

1) 根据熵E_n和超熵H_e, 计算形状相似度:

$$S^{\mathrm{p}}_{\mathrm{PDCM}}(C_1,C_2)=\frac{\min(\sqrt{E_{n_1}^2+H_{e_1}^2},\sqrt{E_{n_2}^2+H_{e_2}^2})}{\max(\sqrt{E_{n_1}^2+H_{e_1}^2},\sqrt{E_{n_2}^2+H_{e_2}^2})}$$

2) 令$\theta_0=(3|E_{n_1}-E_{n_2}|)/(3(E_{n_1}+E_{n_2}))$由拟合参数表[10], 选取接近$\theta_0$值作为替代, 并查找相应拟合参数$a$、$b$、$c$。

3) 将查找到的拟合参数a、b、c代入距离相似度:

$$S^{\mathrm{d}}_{\mathrm{PDCM}}(C_1,C_2)=a\times\exp(-((\theta-b)/c)^2)$$

4) 计算综合相似度:

$$S_{\mathrm{PDCM}}(C_1,C_2)=S^{\mathrm{p}}_{\mathrm{PDCM}}(C_1,C_2)\times S^{\mathrm{d}}_{\mathrm{PDCM}}(C_1,C_2)$$

上述算法中, 由于 LICM 算法直接由数字特征通过夹角余弦计算相似度, 所以复杂度较低且在协同过滤实验中有一定的效果, 但大多数情况下, 由于数字特征的期望值或绝对值远大于熵和超熵, 此时对数字特征仍采用相同权重, 会导致该方法区分能力较弱, 后续实验仿真也得到了验证。在 ECM 和 MCM 算法中, 当云概念数量增加时, 期望曲线和外包络曲线交叠区域较复杂, 从而使这两种算法时间复杂度较高。在 PDCM 算法中, $S^{\mathrm{d}}_{\mathrm{PDCM}}$与$\theta$近似正态分布关系, 对参数$\theta$、$a$、$b$、$c$进行拟合, 选取合适拟合参数值计算距离相似度$S^{\mathrm{d}}_{\mathrm{PDCM}}$, 虽然参数拟合选取方法降低了时间复杂度, 但参数近似选取以及参数与距离相似度$S^{\mathrm{d}}_{\mathrm{PDCM}}$的拟合过程会导致计算误差增大, 从而使 PDCM 算法精度不够高。

2　基于 Hellinger 距离的正态云相似性度量方法

针对现有云概念相似度计算方法不足, 基于正态云特征曲线从整体上表征正态云概念的分布和 Hellinger 距离刻画概率分布间相似程度的特点[20], 本文构造了正态云相似性度量方法。

2.1　两正态分布间的 Hellinger 距离

Hellinger 距离是两个统计样本或总体之间重叠量的度量, 在概率统计理论中, Hellinger 距离常被用于度量两个概率分布的相似度。具体来说, 连续型随机变量概率分布 P 和 Q 的 Hellinger 距离[20]定义为

$$D_{\mathrm{H}}(P,Q)=\sqrt{1-\int_{-\infty}^{+\infty}\sqrt{p(x)q(x)}\mathrm{d}x}\tag{5}$$

其中, $p(x)$、$q(x)$分别为分布 P、Q 的概率密度函数典型。情况下, P 表示数据真实分布, Q 表示数据理论分布、模型分布或 P 的近似分布。$D_{\mathrm{H}}(P,Q)$越大表示两分布差异性越大。根据式 (5) 易得如下结论。

定理 1　设$P\sim N(\mu_1,\sigma_1^2),Q\sim N(\mu_2,\sigma_2^2)$, 则 P 和 Q 的 Hellinger 距离为

$$D^N_{\mathrm{H}}(P,Q)=\sqrt{1-\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\cdot\exp\left\{-\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right\}}\tag{6}$$

由定理 1 知, 对任意两正态分布, 其 Hellinger 距离都可转为由期望与方差的代数运算, 无需进行积分运算, 这一特点会将大大降低计算复杂度, 而且进一步可得到$D^N_{\mathrm{H}}(P,Q)$满足如下性质。

性质 1　设$P\sim N(\mu_1,\sigma_1^2),Q\sim N(\mu_2,\sigma_2^2)$, 则

1) $D^N_{\mathrm{H}}(P,Q)=D^N_{\mathrm{H}}(Q,P)$;

2) $0\leqslant D^N_{\mathrm{H}}(P,Q)<1$;

3) 若 P 和 Q 同分布于正态分布, 即$\mu_1=\mu_2,\sigma_1^2=\sigma_2^2$, 当且仅当$D^N_{\mathrm{H}}(P,Q)=0$。

证明　1) 由距离对称性知$D^N_{\mathrm{H}}(P,Q)=D^N_{\mathrm{H}}(Q,P)$。

2) 由基本不等式$a+b\geqslant2\sqrt{ab},(a>0,b>0)$知,

$$0<\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\leqslant1\ (\sigma_1>0,\sigma_2>0)$$

由于$0<\exp\left\{-\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right\}\leqslant 1$, 故$0\leqslant D_{\mathrm{H}}^{N}(P,Q)<1$。

3) 若$\mu_1=\mu_2,\sigma_1^2=\sigma_2^2$, 显然有$D_{\mathrm{H}}^{N}(P,Q)=0$; 反过来, 若$D_{\mathrm{H}}^{N}(P,Q)=0$, 则有

$$\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\cdot\exp\left\{-\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right\}=1$$

化简得:

$$\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}=\exp\left\{\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right\} \tag{7}$$

由于$0<\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\leqslant 1$, 且$\exp\left\{\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right\}\geqslant 1$, 所以要使式 (7) 成立, 只有

$$\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}=\exp\left\{\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right\}=1$$

从而有$\mu_1=\mu_2,\sigma_1=\sigma_2$。

2.2 两正态云概念间的 Hellinger 距离与相似度

由文献 [21] 知, 二阶正态云概率密度不存在解析解, 故直接利用概率密度无法得到$D_{\mathrm{H}}(P,Q)$的解析式。而正态云还可由特征曲线刻画其整体分布, 并且将特征曲线按其不确定性特征 (熵、超熵) 进行缩放时不会改变原云概念几何性质, 故本文间接采用正态云特征曲线计算$D_{\mathrm{H}}(P,Q)$。首先将特征曲线$\mu_{\mathrm{Exp}}(x)$、$\mu_{\mathrm{In}}(x)$与$\mu_{\mathrm{Out}}(x)$分别乘相应系数正态化, 得到对应特征曲线的密度函数, 分别为

$$p_{\mathrm{Exp}}(x)=\frac{1}{\sqrt{2\pi}E_n}\exp\left\{-\frac{(x-E_x)^2}{2E_n^2}\right\} \tag{8}$$

$$p_{\mathrm{In}}(x)=\frac{1}{\sqrt{2\pi}|E_n-3H_e|}\exp\left\{-\frac{(x-E_x)^2}{2(E_n-3H_e)^2}\right\} \tag{9}$$

$$p_{\mathrm{Out}}(x)=\frac{1}{\sqrt{2\pi}(E_n+3H_e)}\exp\left\{-\frac{(x-E_x)^2}{2(E_n+3H_e)^2}\right\} \tag{10}$$

根据定理 1, 由式 (8)~(10), 容易得到基于期望曲线$\mu_{\mathrm{Exp}}(x)$、内包络曲线$\mu_{\mathrm{In}}(x)$和外包络曲线$\mu_{\mathrm{Out}}(x)$的 Hellinger 距离。

定理 2 设 U 是用精确数值表示的定量论域, $C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$是$U$上的两个二阶正态云概念, 则基于期望曲线、内包络曲线和外包络曲线的 Hellinger 距离分别为

$$D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)=\sqrt{1-\sqrt{\frac{2E_{n_1}E_{n_2}}{E_{n1}^{\ 2}+E_{n2}^{\ 2}}}\cdot\exp\left\{-\frac{(E_{x_1}-E_{x_2})^2}{4(E_{n1}^{\ 2}+E_{n2}^{\ 2})}\right\}} \tag{11}$$

$$D_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)=\sqrt{1-\sqrt{\frac{2\sigma_{\mathrm{In1}}\sigma_{\mathrm{In2}}}{\sigma_{\mathrm{In1}}^2+\sigma_{\mathrm{In2}}^2}}\cdot\exp\left\{-\frac{(E_{x_1}-E_{x_2})^2}{4(\sigma_{\mathrm{In1}}^2+\sigma_{\mathrm{In2}}^2)}\right\}} \tag{12}$$

其中: $\sigma_{\mathrm{In1}}=E_{n_1}-3H_{e_1},\sigma_{\mathrm{In2}}=E_{n_2}-3H_{e_2}$。

$$D_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)=\sqrt{1-\sqrt{\frac{2\sigma_{\mathrm{Out1}}\sigma_{\mathrm{Out2}}}{\sigma_{\mathrm{Out1}}^2+\sigma_{\mathrm{Out2}}^2}}\cdot\exp\left\{-\frac{(E_{x_1}-E_{x_2})^2}{4(\sigma_{\mathrm{Out1}}^2+\sigma_{\mathrm{Out2}}^2)}\right\}} \tag{13}$$

其中: $\sigma_{\mathrm{Out1}}=E_{n_1}+3H_{e_1},\sigma_{\mathrm{Out2}}=E_{n_2}+3H_{e_2}$。

根据距离和相似度转换关系, 由此得到两二阶正态云概念的相似度度量如下。

定理 3 设 U 是用精确数值表示的定量论域, $C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$是$U$上的两个二阶正态云概念, 那么基于期望曲线、内包络曲线和外包络曲线的相似度分别为

$$S_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)=1-D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)$$

$$S_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)=1-D_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)$$

$$S_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)=1-D_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)$$

此外, 相似度量$S_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)$、$S_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)$和$S_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)$还满足如下性质。

性质 2 设$C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$是论域$U$上的两个二阶正态云概念, 则

1) $S_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)=S_{\mathrm{H}}^{\mathrm{Exp}}(C_2,C_1)$, $S_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)=S_{\mathrm{H}}^{\mathrm{In}}(C_2,C_1)$, $S_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)=S_{\mathrm{H}}^{\mathrm{Out}}(C_2,C_1)$;

2) $0<S_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2),S_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2),S_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)\leqslant 1$;

3) 若$C_1=C_2$, 当且仅当$S_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)=S_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)=S_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)=1$。

证明 由定义 5 和性质 1 容易得证 (略)。

2.3 基于 Hellinger 距离和特征曲线的相似度算法

根据具体应用领域, 由期望曲线、内/外包络曲线的不同组合, 通过加权求和形式计算其相似度, 这种方法体现了云概念整体的分布特性。基于此, 设计了两种相似度算法, 分别见算法 7 和算法 8。

算法 7 HECM 算法

输入 数字特征$C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$

输出 相似度$S_{\mathrm{HECM}}(C_1,C_2)$

1) 计算基于期望曲线的 Hellinger 距离$D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)$。

2) 计算相似度$S_{\mathrm{HECM}}(C_1,C_2)=1-D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)$。

算法 8 HCCM 算法

输入 数字特征$C_1(E_{x_1},E_{n_1},H_{e_1})$和$C_2(E_{x_2},E_{n_2},H_{e_2})$

输出 相似度$S_{\mathrm{HCCM}}(C_1,C_2)$

1) 分别计算基于期望曲线的 Hellinger 距离$D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)$、基于内包络曲线的 Hellinger 距离$D_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)$、基于外包络曲线 Hellinger 距离$D_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)$, 并令

$$D_{\mathrm{HCCM}}(C_1,C_2)=\frac{1}{3}\left(D_{\mathrm{H}}^{\mathrm{Out}}(C_1,C_2)+D_{\mathrm{H}}^{\mathrm{In}}(C_1,C_2)+D_{\mathrm{H}}^{\mathrm{Exp}}(C_1,C_2)\right)$$

2) 计算相似度 $S_{\mathrm{HCCM}}(C_1,C_2)=1-D_{\mathrm{HCCM}}(C_1,C_2)$。

3　实验对比分析

为说明算法 HECM 和 HCCM 有效性和可行性, 1) 通过数值仿真实验验证 HECM 和 HCCM 算法的可行性; 2) 在 UCI 数据库时间序列数据集上检验算法的分类性能和计算时间代价; 3) 将算法应用于协同过滤推荐系统中, 并在电影数据集上进行实验对比分析。开发工具为 Python3.8, 运行环境为 Windows 10-64 位操作系统, CPU 为 AMD Ryzen 54600U with Radeon Graphics 2.10 GHz, 16 GB 内存。

3.1　数值仿真实验

本文在文献 [3,9-10] 给出的 4 个正态云概念上进行数值仿真实验, 并将所提出的 HECM 和 HCCM 算法与算法 LICM[3]、ECM[9]、MCM[9] 和 PDCM[10] 进行比较, 其中正态云概念分别为: C_1(1.5, 0.626 66, 0.339), C_2(4.6, 0.601 59, 0.308 62), C_3(4.4, 0.751 99, 0.276 76) 和 C_4(1.6, 0.601 59, 0.308 62), 对应云图如图 3 所示, 不同算法计算结果如表 1 所示。

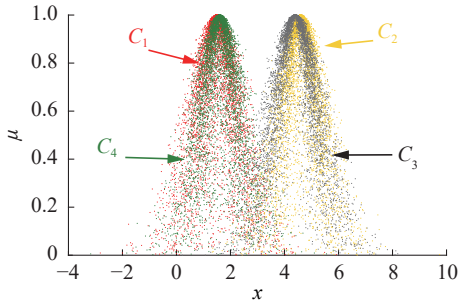


图 3　二阶正态云概念C_1、C_2、C_3、C_4云图

Fig. 3　Cloud map of 2nd- order normal cloud concept C_1, C_2, C_3, C_4

表 1　不同相似度算法下云概念$C_i(i=1,2,3,4)$的相似度

Table 1　Cloud concept $C_i(i=1,2,3,4)$similarity under different similarity algorithms

相似度	LICM	ECM	MCM	PDCM	HECM	HCCM
$S(C_1, C_2)$	0.96	0.01	0.33	0.01	0.04	0.22
$S(C_1, C_3)$	0.97	0.04	0.37	0.03	0.11	0.26
$S(C_1, C_4)$	0.99	0.94	0.96	0.89	0.99	0.99
$S(C_2, C_3)$	0.99	0.86	0.95	0.80	0.97	0.86
$S(C_2, C_4)$	0.97	0.01	0.38	0.01	0.04	0.22
$S(C_3, C_4)$	0.98	0.04	0.37	0.03	0.11	0.26

由表 1 看出, HECM、HCCM 与 ECM、MCM 和 PDCM 算法都得到 C_1 与 C_4 最相似, C_2 与 C_3 最相似, $S(C_1, C_4)$ 和 $S(C_2, C_3)$ 远大于其他任意两概念的相似度, 且 $S(C_1, C_4)>S(C_2, C_3)$, 而 LICM 得到的这 4 个云概念间的相似度都较接近, 均在 0.95 以上。若将这 4 个云概念进行二分类, 那么可以认为概念 C_1、C_4 属于同一类, 概念 C_2、C_3 属于同一类。

为比较各方法区分能力, 借鉴文献 [12] 云概念差异度思想, 即对某个云概念来说, 若与它属同类的云概念相似度越大, 而与它属不同类的云概念相似度越小, 说明该度量方法能有效区分不同类的云概念云概念, C_i差异度定义为

$$\delta_{C_i}=\sum_{j,k}\left|S(C_i,C_j)-S(C_i,C_k)\right|, \tag{14}$$

其中: C_j代表与C_i属同类的云概念, C_k代表与C_i属不同类的云概念。例如云概念C_1的差异度为

$$\delta_{C_1}=|S(C_1,C_4)-S(C_1,C_2)|+|S(C_1,C_4)-S(C_1,C_3)|$$

根据式 (14), 各云概念在不同相似度算法下的差异度如表 2。由表 2 可看出, HECM 算法得到的概念差异度均高于其他算法, 这说明 HECM 算法的区分能力较强, 而 LICM 算法得到的概念差异度都最小, 相似度区分能力最差。与 HECM 算法一样, ECM 算法得到概念差异度均高于 LICM、MCM、PDCM、HCCM 算法得到的概念差异度, 说明基于期望曲线得到的概念相似度对这 4 个云概念区分能力较强, 但期望曲线中没有体现 H_e 的作用。在同时考虑数字特征 E_x、E_n、H_e 的相似度算法中, PDCM 和 HCCM 得到的概念差异度均高于 MCM 和 LICM 算法得到的概念差异度。若从计算复杂角度分析, HECM 与 HCCM 算法只需进行代数运算, 计算复杂度远小于 ECM、MCM 以及 PDCM。所以综合对比看, HECM 和 HCCM 具有较好地性能, 在度量云概念相似度方面具有可行性, 且计算复杂度较低。

表 2　不同相似度算法下云概念$C_i(i=1,2,3,4)$的差异度δ_{C_i}

Table 2　Cloud concept $C_i(i=1,2,3,4)$difference degree δ_{C_i} under different similarity algorithms

差异度	LICM	ECM	MCM	PDCM	HECM	HCCM
δ_{C_1}	0.05	**1.83**	1.22	1.74	**1.83**	1.50
δ_{C_2}	0.05	1.70	1.19	1.58	**1.86**	1.28
δ_{C_3}	0.03	1.64	1.16	1.54	**1.72**	1.20
δ_{C_4}	0.03	**1.83**	1.17	1.74	**1.83**	1.50

3.2　时间序列数据分类

时间序列数据由于其高维性, 能够较好检验

分类算法的性能，采用 UCI 数据库中时间序列数据集 (synthetic control chart time series)[22]，该数据集分 6 类 (共 600 行 60 列)，每行数据代表一个时间序列，每 100 行为一类 (如表 3)，其中 Time_i 代表 600 条时间序列数据，Num_j 代表 60 个维度。实验选取每类后 10 行为测试集，前 90 行为训练集。为提高分类效率，将每个时间序列降维分段处理，训练集和测试集降维后的维数分别为 2、3、4、5、6、10、12、20 维。具体时间序列数据分类过程见算法 9。

表 3　时间序列数据集 $D_{m\times n}$

Table 3　Time series dataset $D_{m\times n}$

时间序列	Num_1	Num_2	…	Num_j	…	Num_n
Time_1	$D_{1,1}$	$D_{1,2}$	…	$D_{1,j}$	…	$D_{1,n}$
Time_2	$D_{2,1}$	$D_{2,2}$	…	$D_{2,j}$	…	$D_{2,n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Time_i	$D_{i,1}$	$D_{i,2}$	…	$D_{i,j}$	…	$D_{i,n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Time_m	$D_{m,1}$	$D_{m,2}$	…	$D_{m,j}$	…	$D_{m,n}$

算法 9　时间序列数据分类算法

输入　时间序列数据集 $D_{m\times n}$

输出　分类错误率和计算相似度 CPU 时间代价

1) 划分数据集。取每类数据前 90 行作为训练集，每类数据的后 10 行为测试集，即训练集为 540 个时间序列数据，测试集为 60 个时间序列数据，并将时间序列数据集分段降维处理，降维后维数分别为 2、3、4、5、6、10、12、20 维，即分割后数据的分段数为 2、3、4、5、6、10、12、20 段。

2) 对分割后的每一段数据按照类别进行逆向云变换，得到相应云概念数字特征。

3) 在同一维数段上云概念，分别利用 LICM、ECM、MCM、PDCM、HECM 和 HCCM 算法计算每一类训练集云概念与其他类测试集云概念的相似度，得到相似度矩阵。

4) 根据最近邻思想，在每一维度矩阵下取相似度最大的类作为分类结果 (例如 2 维时，共 2×6=12 类；3 维时，共 3×6=18 类，依此类推)，并根据分类结果计算分类错误率和计算相似度 CPU 时间代价。

由算法 9，LICM、ECM、MCM、PDCM、HECM 和 HCCM 算法在不同维数下分类错误率、分类错误率平均值和标准差分别如图 4 和表 4 所示，同时各算法相似度计算 CPU 时间代价如图 5 所示。

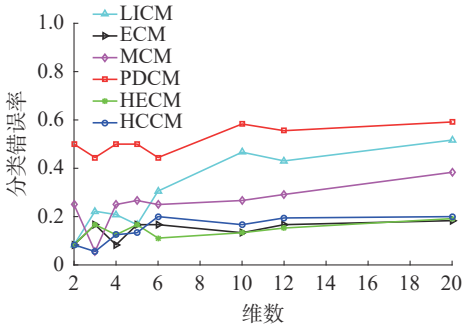


图 4　各算法时间序列数据分类错误率

Fig. 4　Classification error rate for time series data of each algorithm

表 4　不同维数下不同算法的分类错误率均值和标准差

Table 4　Mean value and standard deviation of classification error rate of different algorithms under different dimensions

指标	LICM	ECM	MCM	PDCM	HECM	HCCM
分类错误率均值	0.3000	**0.1437**	0.2517	0.5149	**0.1413**	**0.1448**
分类错误率标准差	0.1563	0.0398	0.0909	0.0571	**0.0350**	0.0552

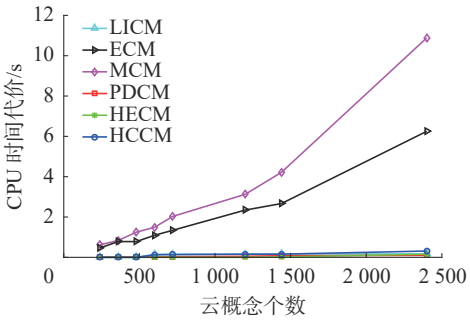


图 5　各算法相似度计算 CPU 时间代价

Fig. 5　CPU time cost of each algorithm to calculate similarity

由图 4 可知，维数为 2、3、4、5 维时，各算法的分类错误率均有波动。根据表 4 分类错误率标准差，LICM 算法稳定性较差，其他几种算法的分类错误率较稳定。从分类错误率来看，PDCM 算法在不同维数下分类错误率都较高，LICM 算法随维数增加分类错误率呈现增加趋势，ECM、HECM 和 HCCM 算法相比其他几种算法分类错误率都较低，且 HECM 算法平均分类错误率和标准差都最小，说明 HECM 算法分类性能和稳定性更好。此外，除 ECM 和 HECM 算法外，HCCM 与 LICM、MCM 和 PDCM 算法相比有更低的错分率和稳定性。尽管 ECM 和 MCM 算法的分类错误率整体低于 LICM 和 PDCM 算法，但由图 5 知，ECM 与 MCM 算法的时间复杂度远高于 HECM 与 HCCM 算法，且随云概念个数增加，ECM 与

MCM 算法 CPU 时间代价呈现增大趋势。因此，综合看, HECM 和 HCCM 算法在时间序列数据集上都具有较好的分类性能。

3.3　不同算法在协同过滤推荐中的应用

3.3.1　协同过滤推荐算法描述

协同过滤 (collaborative filtering, CF) 推荐假设相似用户可能喜欢相似项目，通过分析用户的历史行为数据对目标用户行为进行预测并进行有效推荐，详细步骤见算法 10。

算法 10　协同过滤推荐算法

输入　用户评分表

输出　目标用户 UID 对项目 IID 的推荐评分

1) 计算用户−项目矩阵$\boldsymbol{R}_{m\times n}$。根据用户评分详情，列出用户−项目评分矩阵$\boldsymbol{R}_{m\times n}$，共 m 行用户，n 列项目，则第 i 行第 j 列元素 r_{ij} 表示第 i 个用户对 j 个项目的评分，即

$$\boldsymbol{R}_{m\times n}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix}$$

其中，$r_{ij}=\begin{cases} \text{实际评分, 若用户}i\text{对项目}j\text{有评分} \\ 0,\ \text{若用户}i\text{对项目}j\text{没有评分} \end{cases}$

2) 计算用户评分频度向量。根据 1) 中用户项目矩阵$\boldsymbol{R}_{m\times n}$，统计出每个用户的评分频度向量$\boldsymbol{U}_i=[u_1\ u_2\ \cdots\ u_G]\ (1\leqslant i\leqslant m)$，其中$u_g(g=1,2,\cdots,G)$代表用户 i 对每个项目评分为 g 的频数，G 为项目评分最高分值。

3) 计算用户评分特征向量。根据用户评分频度向量$\boldsymbol{U}_i$，将用户的每一次评分视为云滴，通过逆向云变换算法计算得到每个用户的评分特征向量$\boldsymbol{V}_i=[E_{x_i}\quad E_{n_i}\quad H_{e_i}],(1\leqslant i\leqslant m)$。

4) 计算用户相似度矩阵。用户相似度矩阵表示为

$$\boldsymbol{S}_{m\times m}=\begin{pmatrix} S(1,1) & S(1,2) & \cdots & S(1,m) \\ S(2,1) & S(2,2) & \cdots & S(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ S(m,1) & S(m,2) & \cdots & S(m,m) \end{pmatrix}$$

其中，$S(i,l)$表示用户 i 与 l 的相似度$(1\leqslant i,l\leqslant m)$，分别由 LICM、ECM、MCM、PDCM、HECM 和 HCCM 算法计算。

5) 形成推荐。根据目标用户 UID、用户相似矩阵$\boldsymbol{S}_{m\times m}$和用户−项目矩阵$\boldsymbol{R}_{m\times n}$，在用户空间中查找对该项目有评价记录且与目标用户最接近的 k 个最相似邻居用户，得到最近邻居集$N_{\text{eih}}=\{N_{\text{eih}_1},N_{\text{eih}_2},\cdots,N_{\text{eih}_k}\}$，其中，$N_{\text{eih}_1}$与目标用户相似度最高，$N_{\text{eih}_2}$与目标用户相似度次之，依次类推。根据最近邻集合$N_{\text{eih}}$形成推荐，预测目标用户 UID 对待推荐项目 IID 的评分$P_{\text{UID}\to\text{IID}}$。本文采用加权平均策略得到预测评分$P_{\text{UID}\to\text{IID}}$[4]，计算方法如下：

$$P_{\text{UID}\to\text{IID}}=\frac{\sum\limits_{u_i\in N_{\text{eih}}}S(\text{UID},u_i)\times r_{u_i\to\text{IID}}}{\sum\limits_{u_i\in N_{\text{eih}}}S(\text{UID},u_i)} \tag{15}$$

式中：$r_{u_i\to\text{IID}}$为用户 u_i 对待推荐项目 IID 的评分，$S(\text{UID},u_i)$为目标用户 UID 对近邻用户 u_i 的相似度。

3.3.2　协同过滤推荐算法在影评数据集上的比较

MovieLens100k 数据集[23] 是收集用户对电影评分信息，并通过历史打分信息将预测评分较高的电影推荐给目标用户。数据集从 1997 年 9 月 19 日至 1998 年 4 月 22 日收集 943 个用户对 1 682 部电影的评分记录，共 100 000 条，该数据集用户评分数据稀疏等级为 1−(100 000/9 431 682)=0.937。将数据集以 80% 和 20% 比例划分训练集和测试集，推荐质量评价指标采用平均绝对偏差 (mean absolute error, MAE) 和均方根误差 (root mean squared error, RMSE)。

$$\begin{aligned} E_{\text{MA}}&=\frac{\sum\limits_{i=1}^{N}|p_i-q_i|}{N} \\ E_{\text{RMS}}&=\sqrt{\frac{\sum\limits_{i=1}^{N}(p_i-q_i)^2}{N}} \end{aligned} \tag{16}$$

其中：预测用户评分为 p_i，实际用户评分为 q_i。一般情况下，MAE 与 RMSE 越小推荐质量越高。文献 [3] 已说明 LICM 算法推荐效果优于余弦相似性、修正余弦相似性和 BP-CF(back propagation-collaborative filtering) 方法，故此次实验只将 HECM 和 HCCM 算法与 LICM、ECM、MCM 和 PDCM 算法的推荐效果进行对比。其中最近邻居数 k 分别取 10、20、30、40、50、60，各算法在 k 不断增加时推荐效果的 MAE 和 RMSE 变化分别见图 6 和图 7，不同算法在 k 取不同值时的 MAE 和 RMSE 平均值如表 5 所示。

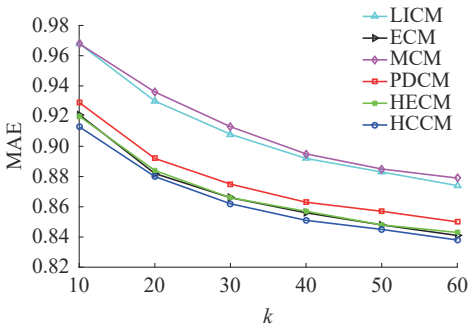


图 6　最近邻居个数 k 增加时各算法的 MAE 取值

Fig. 6　MAE value of each algorithm when the nearest neighbor number k increases

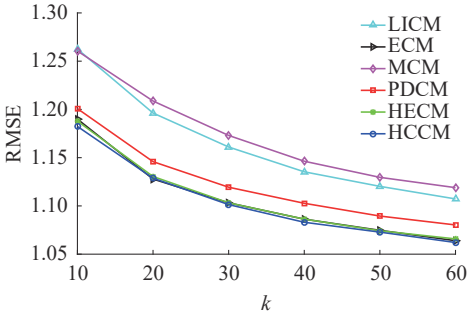


图 7　最近邻居个数 *k* 增加时各算法的 RMSE 取值

Fig. 7　RMSE value of each algorithm when the nearest neighbor number *k* increases

表 5　不同算法在最近邻居个数*k*取不同值时对应的 MAE 均值和 RMSE 均值

Table 5　MAE mean and RMSE mean corresponding to different algorithms when the nearest neighbor *k* takes different values

指标	LICM	ECM	MCM	PDCM	HECM	HCCM
MAE 均值	0.9092	0.8690	0.9127	0.8777	0.8697	0.8648
RMSE 均值	1.1638	1.1077	1.1729	1.1231	1.1080	1.1051

从图 6 和图 7 看出，随着 *k* 从 10 增加至 60，6 种相似度算法的 MAE 和 RMSE 均呈现下降趋势。结合表 5 可看出，LICM、MCM 和 PDCM 算法相比 ECM、HECM 和 HCCM 算法得到的 MAE 和 RMSE 都较高，推荐质量相对较差。相比之下，ECM、HECM 和 HCCM 算法得到的 MAE 和 RMSE 在 6 种算法中比较小且取值接近，且 HCCM 算法得到 MAE 和 RMSE 是最小的，拥有更准确的推荐效果，推荐质量最优，说明 HCCM 算法拥有一定的优越性。

根据上述实验结果，本文方法与其他方法相比有如下优势：1) 从计算角度和效果看，同时考虑云概念 3 个数字特征，且通过 3 条特征曲线研究正态云相似性，综合考虑了云概念的几何特性，并综合量化云概念间的差异，考虑了更多的信息，信息损失少，所以概念区分度和分类性能都较好；2) 从计算过程看，利用数字特征只进行简单的代数运算而无需进行较为复杂的积分运算，与 ECM、MCM、PDCM 算法相较而言更为简单，所以具有较低的计算复杂度；3) 从推广角度看，由于 Hellinger 距离是一种 f 散度且满足距离公理化定义，所以由此得到的云概念相似度具有较好的性质，容易推广至高阶正态云和高维云模型中，具有普遍适用性。

4　结束语

本文主要针对现有正态云相似性方法存在问题，结合正态云特征曲线几何特性和 Hellinger 距离刻画概率分布相似性特点，提出了基于 Hellinger 距离的正态云相似性度量方法，并构造了 2 种正态云概念相似度计算算法。通过数值仿真、时间序列数据分类实验，将本文方法与已有方法进行对比，最后将本文方法应用于协同过滤推荐，实验结果均表明本文方法拥有良好性能和推荐质量。基于 Hellinger 距离和正态云特征曲线构造的云概念相似度为云概念相似度的测量提供了一种新思路，容易将其推广至高阶正态云和高维云模型中。与此同时，结合领域问题，如何选择合适的特征曲线构造相应的 Hellinger 距离，将是下一步需要进行研究的主要工作。

参考文献：

[1] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61–70.

[2] LI Deyi. Artificial intelligence with uncertainty[C]//The Fourth International Conference onComputer and Information Technology, 2004. Wuhan: IEEE, 2004: 2.

[3] 张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推荐算法 [J]. 软件学报, 2007, 18(10): 2403–2411.
ZHANG Guangwei, LI Deyi, LI Peng, et al. A collaborative filtering recommendation algorithm based on cloud model[J]. Journal of software, 2007, 18(10): 2403–2411.

[4] WANG Fan, MENG Xiangwu, ZHANG Yujie, et al. Mining user preferences of new locations on location-based social networks: a multidimensional cloud model approach[J]. Wireless networks, 2018, 24(1): 113–125.

[5] 刘苏, 黄纯, 李克明, 等. 基于自适应分段云模型的单相用户相别辨识方法 [J]. 电力系统自动化, 2022, 46(3): 42–49.
LIU Su, HUANG Chun, LI Keming, et al. Phase identification method for single-phase user based on adaptive piecewise cloud model[J]. Automation of electric power systems, 2022, 46(3): 42–49.

[6] 龚艳冰, 徐绪堪, 刘高峰. 基于正态云期望和方差距离的语言型多属性决策方法研究 [J]. 统计与信息论坛, 2021, 36(10): 12–19.
GONG Yanbing, XU Xukan, LIU Gaofeng. Research on linguistic multi-attribute decision making method based on normal cloud expectation and variance distance[J]. Journal of statistics and information, 2021, 36(10): 12–19.

[7] 狄鹏, 倪子纯, 尹东亮. 基于云模型和证据理论的多属性决策优化算法 [J]. 系统工程理论与实践, 2021, 41(4): 1061–1070.
DI Peng, NI Zichun, YIN Dongliang. A multi-attribute

decision making optimization algorithm based on cloud model and evidence theory[J]. Systems engineering-theory & practice, 2021, 41(4): 1061–1070.

[8] 马丽叶, 张涛, 卢志刚, 等. 基于变权可拓云模型的区域综合能源系统综合评价 [J]. 电工技术学报, 2022, 37(11): 2789–2799.
MA Liye, ZHANG Tao, LU Zhigang, et al. Comprehensive evaluation of regional integrated energy system based on variable weight extension cloud model[J]. Transactions of China electrotechnical society, 2022, 37(11): 2789–2799.

[9] 李海林, 郭崇慧, 邱望仁. 正态云模型相似度计算方法 [J]. 电子学报, 2011, 39(11): 2561–2567.
LI Hailin, GUO Chonghui, QIU Wangren. Similarity measurement between normal cloud models[J]. Acta electronica sinica, 2011, 39(11): 2561–2567.

[10] 汪军, 朱建军, 刘小弟. 兼顾形状–距离的正态云模型综合相似度测算 [J]. 系统工程理论与实践, 2017, 37(3): 742–751.
WANG Jun, ZHU Jianjun, LIU Xiaodi. An integrated similarity measure method for normal cloud model based on shape and distance[J]. Systems engineering-theory & practice, 2017, 37(3): 742–751.

[11] 金璐, 覃思义. 基于云模型间贴近度的相似度量法 [J]. 计算机应用研究, 2014, 31(5): 1308–1311.
JIN Lu, QIN Siyi. Similarity measurement between cloud models based on close degree[J]. Application research of computers, 2014, 31(5): 1308–1311.

[12] 查翔, 倪世宏, 谢川, 等. 云相似度的概念跃升间接计算方法 [J]. 系统工程与电子技术, 2015, 37(7): 1676–1682.
ZHA Xiang, NI Shihong, XIE Chuan, et al. Indirect computation approach of cloud model similarity based on conception skipping[J]. Systems engineering and electronics, 2015, 37(7): 1676–1682.

[13] 许昌林, 王国胤. 正态云概念的漂移性度量及分析 [J]. 计算机科学, 2014, 41(7): 9–14,51.
XU Changlin, WANG Guoyin. Excursive measurement and analysis of normal cloud concept[J]. Computer science, 2014, 41(7): 9–14,51.

[14] YANG Jie, WANG Guoyin, LI Xukun. Multi-granularity similarity measure of cloud concept[C]//International Joint Conference on Rough Sets. Cham: Springer, 2016: 318−330.

[15] LI Shuai, WANG Guoyin, YANG Jie. Survey on cloud model based similarity measure of uncertain concepts[J]. CAAI transactions on intelligence technology, 2019, 4(4): 223–230.

[16] 李德毅, 刘常昱. 论正态云模型的普适性 [J]. 中国工程科学, 2004, 6(8): 28–34.
LI Deyi, LIU Changyu. Study on the universality of the normal cloud model[J]. Engineering science, 2004, 6(8): 28–34.

[17] 许昌林. 基于云模型的双向认知计算方法研究 [D]. 成都: 西南交通大学, 2014.
XU Changlin. Method of bidirectional cognitive computing based on cloud model[D]. Chengdu: Southwest Jiaotong University, 2014.

[18] 刘常昱, 李德毅, 杜鹢, 等. 正态云模型的统计分析 [J]. 信息与控制, 2005, 34(2): 236–239,248.
LIU Changyu, LI Deyi, DU Yi, et al. Some statistical analysis of the normal cloud model[J]. Information and control, 2005, 34(2): 236–239,248.

[19] 刘禹, 李德毅. 正态云模型雾化性质统计分析 [J]. 北京航空航天大学学报, 2010, 36(11): 1320–1324.
LIU Yu, LI Deyi. Statistics on atomized feature of normal cloud model[J]. Journal of Beijing university of aeronautics and astronautics, 2010, 36(11): 1320–1324.

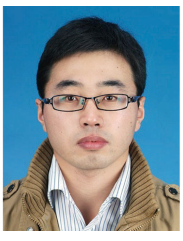
[20] ZHENG Yayun, YANG Fang, DUAN Jinqiao, et al. Quantifying model uncertainty for the observed non-Gaussian data by the Hellinger distance[J]. Communications in nonlinear science and numerical simulation, 2021, 96: 105720.

[21] 王国胤, 许昌林, 张清华, 等. 双向认知计算的 p 阶正态云模型递归定义及分析 [J]. 计算机学报, 2013, 36(11): 2316–2329.
WANG Guoyin, XU Changlin, ZHANG Qinghua, et al. P-order normal cloud model recursive definition and analysis of bidirectional cognitive computing[J]. Chinese journal of computers, 2013, 36(11): 2316–2329.

[22] Synthetic control chart time series[EB/OL]. (1999−06−07) [2022−09−20]. http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series.

[23] MovieLens 100K dataset [EB/OL]. (1999−04−07) [2022−09−20].https://grouplens.org/datasets/movielens.

作者简介:



许昌林，副教授，博士，主要研究方向为智能信息处理、云模型理论、认知计算、不确定性决策。主持在研国家自然科学基金项目 1 项、宁夏自然科学基金项目 2 项，完成宁夏自然科学基金项目 2 项；发表学术论文 20 余篇。

徐浩，硕士研究生，主要研究方向为基于云模型理论的不确定性决策、数据挖掘。