第 46 卷 第 5 期 计 算 机 工 程 2020 年 5 月
Vol. 46 No. 5 Computer Engineering May 2020

·人工智能与模式识别· 文章编号：1000-3428(2020)05-0054-09 文献标志码：A 中图分类号：TP393

基于深度信念网与隐变量模型的用户偏好建模

潘良辰，吴鑫然，岳 昆

（云南大学 信息学院，昆明 650500）

摘 要： 从高维、稀疏的用户评分数据中构建用户偏好模型，存在迭代计算复杂度高、中间结果规模大和难以实现有效推理等问题。为此，提出一种基于深度信念网（DBN）和贝叶斯网（BN）的用户偏好建模方法。采用 DBN 对评分数据进行分类，用隐变量表示不能直接观测到的用户偏好，利用含隐变量的 BN 描述评分数据中蕴含的相关属性间的依赖关系及其不确定性。在 MovieLens 和大众点评数据集上的实验结果表明，该方法能够有效描述评分数据中与用户偏好相关的各属性间的依赖关系，其精确率和执行效率均高于隐变量模型。

关键词： 贝叶斯网；用户偏好；评分数据；隐变量模型；深度信念网

开放科学（资源服务）标志码（OSID）：

中文引用格式： 潘良辰，吴鑫然，岳昆. 基于深度信念网与隐变量模型的用户偏好建模[J]. 计算机工程，2020，46(5)：54-62.

英文引用格式： PAN Liangchen，WU Xinran，YUE Kun. User preference modeling based on deep belief network and latent variable model[J]. Computer Engineering，2020，46(5)：54-62.

User Preference Modeling Based on Deep Belief Network and Latent Variable Model

PAN Liangchen，WU Xinran，YUE Kun

（School of Information Science and Engineering，Yunnan University，Kunming 650500，China）

【Abstract】 To address complex iterative computations，large-scale intermediate results and ineffective inference of user preference modeling from high dimensional and sparse user rating data，this paper proposes a user preference modeling method based on Deep Belief Network（DBN）and Bayesian Network（BN）. The DBN is used to classify rating data，and the latent variables are used to represent user preferences that cannot be directly observed. Then，the BN with latent variables is used to describe the uncertain dependences among related attributes in rating data. Experimental results on MovieLens and DianPing datasets show that the proposed method can effectively describe the dependences relationship between attributes related to user preferences in rating data，and its precision and execution efficiency are higher than that of Latent Variable Model（LVM）.

【Key words】 Bayesian Network（BN）；user preference；rating data；Latent Variable Model（LVM）；Deep Belief Network（DBN）

DOI： 10.19678/j.issn.1000-3428.0054183

0 概述

随着 Web2.0、电子商务和社交网络的快速发展，越来越多的用户通过各种 Web 平台参与到互联网活动中，因此，产生了大量的用户行为数据。用户对电影或商品的评分是目前 Web2.0 应用中具有代表性的一类用户行为数据，这些数据一般包含用户属性、评分对象属性和用户评分等可以直接观测到的信息，也蕴含着一些无法被直接观测到的信息，如表示用户喜好或选择倾向性的偏好信息。用户评分数据反映了用户偏好，用户偏好决定了用户评分。例如，MovieLens 数据集[1]包括用户信息、电影信息和评分，电影信息包括年代、类型、语言等多种属性，用户信息包括年龄、性别、职业等属性信息，不同用户对不同电影属性有相应的喜好，用户、电影以及评分等属性间存在依赖关系并具有不确定性。因此，从评分数据中建立描述相关属性之间依赖关系及其不确定性的用户偏好模型，为个性化推荐和用户行

基金项目： 国家自然科学基金（U1802271）；云南省应用基础研究计划重点项目（2017FA032）；云南大学科研项目（2017YDJQ06）。

作者简介： 潘良辰（1994—），男，硕士，主研方向为数据与知识工程；吴鑫然，硕士；岳 昆（通信作者），教授、博士、博士生导师。

收稿日期： 2019-03-11 **修回日期：** 2019-05-09 **E-mail：** kyue@ynu.edu.cn

为建模等应用提供知识模型和计算框架,具有重要的现实意义。

贝叶斯网(Bayesian Network,BN)[2]是一种重要的概率图模型,其是由一组节点组成的有向无环图(Directed Acyclic Graph,DAG),每个节点都有一个条件概率表(Conditional Probability Table,CPT)[3]。BN 可定量地描述属性间的依赖关系,因此,其被广泛应用于智能分析和推断决策等领域。含隐变量的 BN 称为隐变量模型(Latent Variable Model,LVM)。通过 LVM 可有效描述评分数据中能直接观测到和不能直接观测到的属性之间的依赖关系,并可进行有效的推理计算,从而为用户偏好模型的建立提供支持。例如,对于 MovieLens 数据而言,可使用隐变量表示用户对电影的偏好,基于 LVM 学习方法构建包括用户、电影、评分和偏好属性的模型。

本文提出一种基于深度信念网(Deep Belief Network,DBN)与隐变量模型的用户偏好建模方法。采用深度信念网对评分数据进行分类,利用类别变量扩展隐变量模型,同时基于评分数据的特点和隐变量模型构建的关键步骤,给出模型构建时需要满足的约束条件以及该约束条件下模型的参数学习和结构学习方法。

1　问题分析

由于隐变量的取值无法被直接观测到,可认为其数据缺失。期望最大(Expectation Maximization,EM)算法[4]是在不完整数据情况下对数据进行填充并用于模型参数最大似然估计的一种有效方法。结构期望最大(Structure Expectation Maximization,SEM)[5]算法是一种结合了 EM 算法和打分搜索的结构学习方法。EM 算法的运行涉及大量迭代计算[6],计算复杂度较高。从 BN 的结构及特点来看,CPT 中概率参数的规模由其父节点组合的取值数量决定,不同组合会带来较高的时间和空间复杂度。在实际应用中,用户评分数据具有海量、高维、稀疏、内部结构复杂等特征,从评分数据中构建 LVM 以有效描述用户偏好具有较高的难度,这也是本文拟解决的一个问题。

从基于偏好模型的用户偏好或评分估算的角度来看,可利用基于 BN 的概率推理算法,根据观测到的用户或对象属性以及评分来估计用户偏好,或根据用户偏好来估算可能的评分。但是,针对 CPT 中并未包括的新用户或新对象信息,传统的 BN 模型难以进行有效的概率推理。例如,若利用基于样本集{(男,<18,Educator,R),(男,25~34,Admin,R),(女,>34,Other,R)}构建的 BN 来估算用户(男,>34,Farmer)的评分,由于该用户信息并未包含于模型的 CPT 中,因此无法进行有效的概率推理。综上,如何使得模型能有效支持新用户或新对象的偏好估计及评分估算,是用户偏好模型构建时面临的又一挑战。

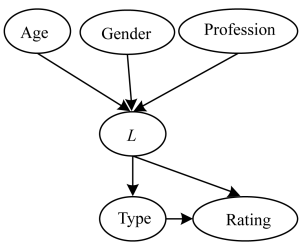
图 1 所示为一个简单的隐变量模型,在实际情形中,Age 和 Profession 等取值的个数远多于该例中变量取值的个数,因此,CPT 存在组合爆炸的情况,从而导致 LVM 构建效率较低、所占存储空间较大等问题。

Age	P(Age)
<18	0.2
18~24	0.3
25~34	0.3
>34	0.2

Gender	P(Gender)
male	0.5
female	0.5

Profession	P(Profession)
other	0.5
educator	0.2
admin	0.2
farmer	0.1

Age	Gender	Profession	$P(L\|\text{Age,Gender,Profession})$		
			$L=L_1$	$L=L_2$	$L=L_3$
<18	male	other	0.20	0.10	0.7
<18	female	other	0.20	0.50	0.3
18~24	male	farmer	0.10	0.30	0.6
18~24	female	educator	0.15	0.35	0.5
25~34	male	admin	0.05	0.65	0.3
25~34	female	farmer	0.55	0.25	0.2
>34	male	admin	0.40	0.50	0.1
>34	female	educator	0.60	0.10	0.3
⋮	⋮	⋮	⋮	⋮	⋮



L	$P(\text{Type}\|L)$		
	Type=1	Type=2	Type=3
L_1	0.35	0.55	0.1
L_2	0.55	0.25	0.2
L_3	0.05	0.35	0.6

L	Type	$P(\text{Rating}\|L,\text{Type})$				
		Rating=1	Rating=2	Rating=3	Rating=4	Rating=5
L_1	1	0.02	0.18	0.30	0.05	0.45
L_1	2	0.01	0.09	0.10	0.50	0.30
L_1	3	0.05	0.25	0.20	0.45	0.05
L_2	1	0.15	0.05	0.35	0.05	0.40
L_2	2	0.02	0.08	0.20	0.10	0.60
L_2	3	0.14	0.06	0.40	0.20	0.20
L_3	1	0.25	0.02	0.43	0.10	0.20
L_3	2	0.05	0.13	0.42	0.10	0.30
L_3	3	0.06	0.24	0.10	0.50	0.10

图 1　简单的隐变量模型

Fig. 1　Simple latent variable model

本文在基于 LVM 的用户偏好建模方面,研究如何提升模型构建效率、克服模型构建过程中 CPT 组合爆炸等问题;在基于传统 LVM 的概率推理方面,研究对新用户或新对象信息进行概率推理的问题。具体而言,本文首先利用隐变量表示用户偏好,建立一种基于隐变量模型的用户偏好模型;然后通过深度信念网对评分数据进行分类,利用类别变量扩展隐变量模型,得到类别简化的贝叶斯网(Class Simplified BN,CSBN);接着给出模型构建时需满足的约束条件及该约束条件下的模型参数学习和结构学习方法;最后在 MovieLens 和大众点评数据集上进行实验,以验证本文方法的可行性和高效性。

2 相关工作

目前,从评分数据中构建用户偏好模型,有基于评分矩阵和项目流行度的推荐方法[7]、建立商品服务评估模型[8]和使用向量模型或主题模型来表示用户偏好[9-10]等方式。例如,结合 LDA(Latent Dirichlet Allocation)[10]的主题分布描述用户偏好,利用 SVD(Singular Value Decomposition)[11]等矩阵因式分解模型描述用户偏好,这些方法能够表达预先给定的依赖关系。但 LDA 是一种线性模型,SVD 的分解矩阵可解释性较差,因此,两者均无法表述评分数据中属性间的依赖关系及其不确定性。

研究人员在基于 BN 或 LVM 的用户偏好建模方面进行了较多研究。文献[12]用隐变量表示用户的评价行为,文献[13-14]用隐变量描述用户对电影类型的偏好,文献[15]依据旅游的专家知识构建 BN 并估计用户的旅游偏好,文献[16]使用隐变量刻画用户兴趣并提出一种用以描述用户点击行为的动态贝叶斯网模型。但是,上述模型构建效率较低,难以适用于高维、稀疏的评分数据,因此,在评分数据上以隐变量模型为基础构建用户偏好模型,成为亟需解决的问题。

在提高模型构建效率方面,文献[17]采用决策树对观测数据进行分类,用分类后的变量构建 BN,将观测值的类别作为证据变量进行概率推理,但该方法难以处理海量、稀疏、高维的评分数据集。文献[18-19]提出 DBN 以对数据进行分类和降维,DBN 在很大程度上保存和还原了原始信息,可适用于海量、多维、内部结构复杂的评分数据。本文将 DBN 和隐变量模型进行结合,以构建用户偏好模型。

3 符号定义

将包含用户属性、对象属性和评分值等信息的评分数据记为 D,$U=(U_1,U_2,\cdots,U_{|U|})$表示用户属性集合,$I\in\{i_1,i_2,\cdots,i_{|I|}\}$表示对象属性,$R$ 表示用户评分。用户偏好由用户对评分对象各个属性的喜好构成,表示为 $L\in\{l_1,l_2,\cdots,l_{|L|}\}$,其中,$i_x=l_x(1\leqslant x\leqslant|I|)$,$l_x$ 称为第 x 维的偏好。例如,某一评分数据为 U_1(性别)="男性"、U_2(年龄)="18 岁 ~ 24 岁"、U_3(职业)="学生"、i_1(电影类别)="动作片"、R(评分)="4",表示该用户对某动作片的评分为 4,$l_1=i_l$ 表示用户喜好"动作片"。

利用 DBN 分别对 U 和 I 数据进行分类,得到分别表示用户属性和对象属性的类别变量 U_C 和 I_C,分类后的评分数据 D_C 包含 U_C、I_C 和 R 的信息。

定义 1 BN 是有向无环图(Directed Acyclic Graph,DAG),记为 $G=(V,E,\theta)$,其满足如下 4 个性质:

1)V 是一组多维随机变量集合,其构成了 G 中的节点,每一个节点对应一个变量。

2)含隐变量的 BN 简称隐变量模型,记为 $G=(V,L,E,\theta)$,其中,L 是描述用户偏好的隐变量节点。

3)E 是连接各节点有向边的集合,表示各节点间的依赖关系。若存在从节点 u 指向节点 v 的有向边 $u\rightarrow v$(u、$v\in V$,$u\neq v$),则称 u 是 v 的父节点。每个节点 v 在给定其父节点集 $\pi(v)$时独立于其非子孙节点。

4)θ 为各节点条件概率参数的集合,表示为 $P(v|\pi(v))$。

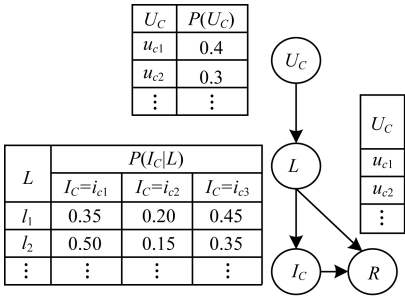
定义 2 类别简化的贝叶斯网(CSBN)记为$G=(V,L,E,\theta)$,其满足如下 3 个性质:

1)$V=\{U_C,I_C,R\}$是包括用户类别、对象类别和评分的变量集合。

2)L 为隐变量,$L=l_j$ 表示用户对评分对象属性第 j 个类别的偏好,其中,$1\leqslant j\leqslant|I_C|$。

3)E 和 θ 分别为 G 中有向边和条件概率参数的集合。

图 2 所示为一个简单的 CSBN 模型,在采用分类变量 U_C 替代图 1 中的 Age、Gender、Profession 变量构建隐变量模型时,模型中的依赖关系得以简化,且 CPT 的组合数量也大幅降低,便于计算和存储。因此,采用分类变量构建隐变量模型,可以简化模型并提高模型构建效率。



U_C	$P(U_C)$
u_{c1}	0.4
u_{c2}	0.3
⋮	⋮

L	$P(I_C\|L)$		
	$I_C=i_{c1}$	$I_C=i_{c2}$	$I_C=i_{c3}$
l_1	0.35	0.20	0.45
l_2	0.50	0.15	0.35
⋮	⋮	⋮	⋮

U_C	$P(L\|U_C)$		
	$P(L=l_1)$	$P(L=l_2)$	$P(L=l_3)$
u_{c1}	0.2	0.1	0.7
u_{c2}	0.2	0.5	0.3
⋮	⋮	⋮	⋮

L	I_C	$P(R\|L,I_C)$				
		$R=1$	$R=2$	$R=3$	$R=4$	$R=5$
l_1	i_{c1}	0.02	0.18	0.3	0.05	0.45
l_2	i_{c2}	0.01	0.09	0.1	0.50	0.30
l_3	i_{c3}	0.05	0.25	0.2	0.45	0.05
⋮	⋮	⋮	⋮	⋮	⋮	⋮

图 2 CSBN 模型示例

Fig. 2 Example of CSBN model

4　基于 DBN 的评分数据分类

根据定义 2,本文使用 DBN 分别对用户属性数据和评分对象属性数据进行分类。分类后的评分数据维度降低,CPT 中不同取值的组合也相应减少,从而提高了模型构建效率。此外,可以对训练集中未出现的变量组合进行分类,从而实现对这类取值的概率推理。DBN 分类算法[18-19]将 DBN 看作一个特殊的多层感知器,是由多个受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)[20]叠加构成的深度学习结构,预训练阶段通过逐层训练得到各个 RBM 的权值,下一层的输出作为更高一层的输入。DBN 训练分为预训练和微调 2 个步骤,最后将 DBN 模型输入 softmax 回归分类器[21]得到用户属性类别数据集。以用户属性数据的分类为例,基于 DBN 的评分数据分类算法 R-DBN-Classification 步骤如下:

算法 1　R-DBN-Classification 算法

输入　用户属性数据 D,迭代次数上限 g_n,学习率 η

输出　带有分类标签的评分数据集 D_c

1. $\eta \leftarrow 0.01, g_n \leftarrow 2\ 000$

(学习率、迭代次数由经验值确定)

2. 初始化 DBN 分类器模型,RBMs 层数为 3,隐层单元分别设置为 18、36、18//最后一层隐层单元个数为输出的类别数

3. 令 n 为输入层神经元个数 //例如,针对性别、年龄和//职业,设置 n = 3

4. 初始化 DBN 网络的权值 $W = [W_1', W_2', W_3']$ 为 0 矩阵,随机初始化偏重 b 和 c

5. W,b,c←DBN_train(n,v,g_n)

(使用文献[18-19]中的 DBN_train 算法)

6. $U_c \leftarrow$softmax(W,b,c,D)

算法 1 采用非监督贪婪逐层训练的方法,即对比散度(Contrastive Divergence,CD)算法[20]获取权值,只需单个步骤就可以接近最大似然学习,因此,可显著缩短训练时间,提高收敛速度。

以用户属性信息为例设计的 DBN 分类器模型结构如图 3 所示,其中,输入为用户属性,包括性别(Gender)、年龄(Age)和职业(Profession)。Gender 取值为 0、1,分别代表男、女,Age 取值为年龄段,Profession 取值为职业对应的编号。虽然输入神经元个数为 3,输出神经元个数为 18,但输入神经元的 3 个节点取值组合数为 $2 \times 7 \times 21 = 294$,相当于经过 DBN 分类器后用户属性组合从 294 降到了 18,大幅降低了 EM 算法中间结果规模以及 CSBN 中节点的 CPT 规模,进而提高了 BN 的学习效率。

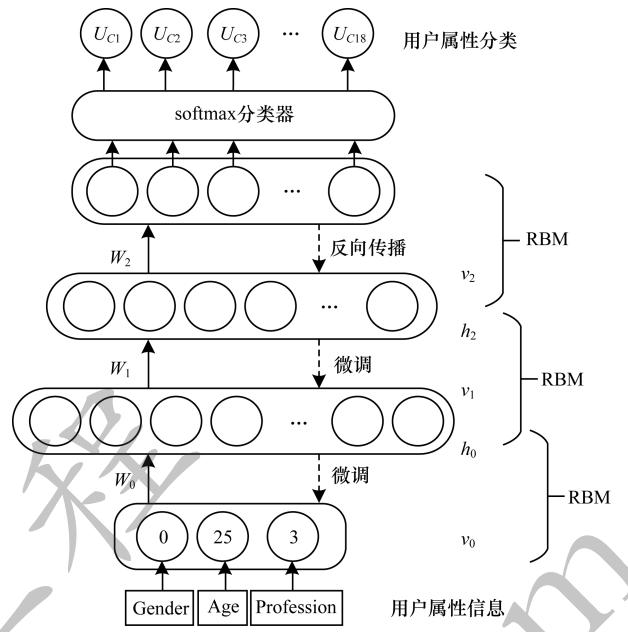


图 3　DBN 分类器模型

Fig. 3　DBN classifier model

5　带分类变量的隐变量模型构建

5.1　约束条件

为保证模型构建的有效性,根据用户偏好、隐变量在评分数据中的特定含义,本文给出如下约束:

约束 1　第 j 维的偏好 l_j 表达用户对评分对象属性类别节点 I_C 的第 j 个取值 i_{cj} 的倾向($1 \leqslant j \leqslant |I_C|$),用户属性 U_C 不依赖于其他变量。

约束 2　$P(i_{c1}|l_1) > P(i_{c2}|l_1)$,当用户偏好为 l_1 时,该用户偏好 i_{c1} 的概率大于 i_{c2};$P(R = r_2 | i_{c1}, l_1) > P(R = r_1 | i_{c1}, l_1)$,用户更可能对倾向或喜好的对象评高分($r_2 > r_1$),反之评低分($r_1 > r_2$)。

5.2　基于约束的 CSBN 参数学习

将引入分类变量后的用户属性类别数据集、评分对象属性类别数据集以及评分数据,处理为带有分类标签的评分数据集 D_c,D_c 中一次用户评分记录为一个样本 $D_{cy} \in \{D_{c1}, D_{c2}, \cdots, D_{c|D_c|}\}$,包括 ID、用户属性类别、评分属性类别以及评分,用户偏好 L 的取值个数为 $|I|$,$L \in \{l_1, l_2, \cdots, l_{|I|}\}$。然后,随机生成一组满足约束条件的初始参数,对数据集中隐变量的值进行填充,再计算并更新参数,直至收敛或达到迭代次数上限,所得参数优化结果即为最终参数学习结果。基于第 t 次迭代得到的参数估计 θ^t,第 $t+1$ 次迭代过程如下:

E 步:根据当前参数 θ^t,利用式(1)为每个样本 D_{cy} 计算不同用户偏好取值 L_j 的后验概率 $P(L = L_j |$

D_{cy},θ^t)($1\leqslant j\leqslant c$),使数据集修补为带权完整数据集 D_c^t。然后根据 D_c^t,利用式(2)计算期望统计量 m_{ijk}^t。

$$P(L=L_j\mid D_{cy},\theta^t)=\frac{P(L=L_j,D_{cy}\mid\theta^t)}{\sum_{j=1}^{c}P(L=L_j,D_{cy}\mid\theta^t)} \tag{1}$$

$$m_{ijk}^t=\sum_{l=1}^{m}P(C_i=k,\pi(C_i)=j\mid D_{cl}^t) \tag{2}$$

M 步:根据 m_{ijk}^t,利用式(3)计算参数的最大似然估计 θ_{ijk}^{t+1}。

$$\theta_{ijk}^{t+1}=\frac{m_{ijk}^t}{\sum_{k=1}^{r_i}m_{ijk}^t} \tag{3}$$

为了保证算法执行的收敛效率,本文使用式(4)来度量参数更新的程度,$\ln P(D_c^{t+1}\mid\theta^{t+1})$和 $\ln P(D_c^t\mid\theta^t)$分别是第 $t+1$ 次和第 t 次迭代所需参数的对数似然函数,若 $\mathrm{sim}(\theta^t,\theta^{t+1})<\delta$($\delta$ 为根据经验值设置的收敛阈值,如 0.000 01),则认为参数已经收敛,迭代计算结束。

$$\mathrm{sim}(\theta^t,\theta^{t+1})=\left|\ln P(D_c^{t+1}\mid\theta^{t+1})-\ln P(D_c^t\mid\theta^t)\right| \tag{4}$$

EM 算法不断迭代直至收敛,基于约束的 CSBN 参数学习算法 Parameter_learning 描述如下:

算法 2 Parameter_learning 算法

输入 带有分类标签的评分数据集 D_c,收敛阈值 δ,迭代次数上限 T

输出 CSBN 模型参数 θ

1. 随机产生一组满足约束 2 的初始参数 θ′

2. for t←0 to T do

E 步:

3. 根据式(1)修补 D_c 中的每个样本,得到带权完整数据集 D_c^t,并利用式(2)计算 m_{ijk}^t

M 步:

4. 根据 m_{ijk}^t,利用式(3)计算 θ^{t+1},利用 $\mathrm{sim}(\theta^t,\theta^{t+1})$判断是否收敛

5. end for

例如,表 1 给出了利用 DBN 分类的用户评分类别数据的片段示例,以图 2 结构作为当前模型结构,执行算法 2 迭代一次修补后的样本数据如表 2 所示。假设数据集大小为 $|D_c|$,L 的取值个数为 $|I|$,则每次 EM 过程修补后的数据有 $|I|\times|D_c|$条,计算最大似然估计 $|I|\times|D_c|$次,EM 迭代次数最多为 T 次,因此,需要计算 $T\times|I|\times|D_c|$次最大似然估计,算法 2 的时间复杂度为 $O(T\times|I|\times|D_c|)$。可见,EM 算法的执行效率与修补后的数据量、输入数据集以及隐变量取值个数呈负相关[22],即采用类别变量构建 CSBN 后数据量下降,从而提高了算法效率。

表 1 数据集片段

Table 1 Fragment of the dataset

ID	c_1	L	c_2	R
0	0	—	3	1
1	1	—	0	2
⋮	⋮	⋮	⋮	⋮

表 2 修补后的数据

Table 2 Repaired date

ID	c_1	L	c_2	R
0-1	0	0	3	1
0-2	0	1	3	1
0-3	0	2	3	1
0-4	0	3	3	1

5.3 基于约束的 CSBN 结构学习

贝叶斯信息准则(Bayesian Information Criterion,BIC)是一种常用的打分标准,能在缺值样本前提下对结构进行打分,为本文学习 CSBN 结构提供了模型选择基准。模型结构 ζ 的 BIC 评分计算公式如下:

$$\lg P(D\mid\zeta)\approx\lg P(D\mid\zeta,\theta^*)-\frac{d}{2}\lg m \tag{5}$$

式(5)右侧第 1 项是模型 ζ 的优参对数似然度,其度量结构 ζ 数据 D 的拟合程度;第 2 项是一个关于模型复杂度的罚项,其能够有效避免依据优参似然度选择模型导致的过拟合现象。本文使用 SEM 算法结合 BIC 打分准则作为 CSBN 结构学习方法的基础。首先根据约束 1 和约束 2,随机生成一组满足约束 1 的初始结构和一组满足约束 2 的初始参数,以生成的初始结构与初始参数作为 SEM 算法的初始值进行参数学习;然后计算初始结构的 BIC 评分,通过当前结构边的变化得出一系列的候选结构,根据当前修补的数据集学习候选结构的参数并进行 BIC 打分,选出局部最优候选模型并与当前模型作对比,选其中评分较高者作为当前模型继续进行参数学习,重复迭代上述步骤。

从初始 CSBN 模型出发进行参数学习,基于约束的 CSBN 结构学习算法 Structure_learning 描述如算法 3 所示。

算法 3 Structure_learning 算法

输入 带有分类标签的评分数据集 D_c,收敛阈值 δ,迭代次数上限 T,节点个数 C_num

输出 CSBN 模型结构 G,CSBN 模型参数 θ

1. 随机生成满足约束 1 的 BN 结构 G′和满足约束 2 的初始参数 θ′

2. oldscore←BIC(G,θ|D′),newscore←－∞//计算当前//BIC评分
3. for i←0 to (C_num－1) do
4. 令 G_set 为当前节点加、减或转边得到的候选结构
5. $\theta^i, D^i \leftarrow EM(G^i, D'^i, \theta^i, \delta, 1)$//参数计算
6. temp←θ^i, D^i,tempBIC←BIC(G,θ|D′)
7. if tempBIC > newscore then
8. $\theta^i, D^i \leftarrow$temp,newscore←tempBIC
9. end if
10. if newscore > oldscore then
11. $\theta^{i+1}, D^{i+1} \leftarrow EM(G^i, D'^i, \theta^i, \delta, T)$
12. $(G,\theta) \leftarrow (G^{i+1}, \theta^{i+1})$
13. oldscore←$BIC(G,\theta|D^{i+1})$
14. else
return (G,θ)
15. end if
16. end for

假设模型的节点数为 C_num,一次结构学习过程产生的候选结构个数为 z,对所有候选结构执行一次 EM 算法,得出当前最优候选结构并对其执行 EM 算法直至收敛或达到迭代次数上限 T,候选结构选择的时间开销远低于 EM 算法的执行时间开销,则 SEM 算法需要执行 $z \times C$_num 次 EM 算法,C_num 在算法开始时就给定,z 由初始结构决定,则算法 3 的时间复杂度为 $O(T \times |I| \times |D_c|)$。

6　实验结果与分析

6.1　实验设置

本文实验采用 GroupLens 提供的 MovieLens-1M 数据集[1],其包括 6 040 条用户属性数据、3 952 条电影属性数据、1 000 209 条电影评分数据。此外,在大众点评网利用爬虫爬取 20 个城市各 100 个用户的评分数据,该数据集包括 2 000 条用户属性数据、5 162家餐厅属性数据以及 114 023 条评分数据。2 个数据集经过预处理后,每行数据对应一次用户评分记录,每个记录分别由 3 个用户属性信息、1 个电影/餐厅属性信息和 1 个评分值组成。

实验环境如下:Intel Core i7-6700 @ 3.40 GHz 处理器,12 GB DDR4 内存,Nvidia GeForce GTX 750 Ti 显卡,Windows 10(64 位)操作系统,Python 作为开发语言。

本文主要针对模型构建效率、所构建模型有效性等方面进行实验分析。首先,分别在 MovieLens 和大众点评数据集上对 LVM 和 CSBN 的模型构建时间进行对比,经测试 DBN 分类算法执行时间在 CSBN 模型构建时间中占比较小,因此,本文效率测试部分忽略 DBN 分类算法带来的时间开销。然后,基于 CSBN 模型的结构和参数,计算条件概率 $P(Q|e)$,其中,e 是由用户属性构成的证据变量取值,Q 是电影类型的查询变量,计算具有最大后验概率的电影类型并作为用户偏好。本文对计算出的用户偏好电影类型和统计出的真实用户偏好电影类型进行对比,估计召回率(Recall)、准确率(Precision)以及F 值。三者计算公式分别如下:

$$R_{\text{recall}} = \frac{\text{num(ture)}}{\text{num(sample)}}$$

$$P_{\text{precision}} = \frac{\text{num(ture)}}{\text{num(inference)}}$$

$$F = \frac{(\alpha^2+1) \cdot R_{\text{recall}} \cdot P_{\text{precision}}}{\alpha^2 \cdot (R_{\text{recall}} + P_{\text{precision}})}$$

其中,num(inference)是推理出用户有倾向观看的电影类型数目,num(ture)是推理出有倾向且实际也有倾向观看的电影类型数目,num(sample)是实际用户有倾向观看的电影类型数目。

6.2　效率测试

本文在 MovieLens 数据集上选取不同用户数,分别测试 CSBN 和 LVM 的模型构建执行时间,结果如图 4 所示。从图 4 可以看出,随着数据量的增加,CSBN 和 LVM 的执行时间均增加,在同样大小的数据集下,CSBN 的执行效率高于 LVM,并且随着数据量的增加,CSBN 更能有效地提升用户偏好模型构建效率。

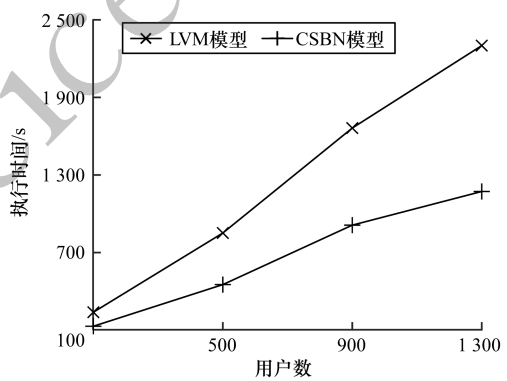


图 4　MovieLens 数据集上模型构建执行时间对比

Fig. 4　Comparison of model construction execution time on MovieLens dataset

图 5 所示为大众点评数据集上 LVM 和 CSBN 的模型构建时间对比,由于大众点评数据集在每个城市均爬取了 100 个用户的数据,从中随机选取 30%、50%、80% 的数据来测试算法 3 的效率。从图 5可以看出,随着数据集的增大,LVM 和 CSBN 的执行时间均增加,在相同比例的数据集下,CSBN 的执行时间远低于 LVM,原因在于数据量增加时,LVM 模型中间结果规模以 $|I|$ 倍增长,导致其执行时间增长更快。

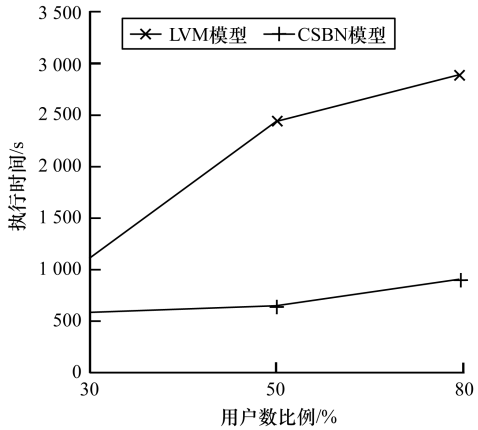


图 5 大众点评数据集上模型构建时间对比

Fig. 5 Comparison of model construction time on DianPing dataset

此外,本文进一步比较不同数据集下 LVM 和 CSBN 模型构建时的中间结果规模,以反映算法执行过程中所需要的内存空间大小。表 3 和表 4 分别为 MovieLens 和大众点评数据集上构建 LVM 和 CSBN 模型时一次迭代计算的中间结果规模,可以看出,在相同数据量时,CSBN 构建时中间结果规模比 LVM 模型低 1 个数量级,随着数据量的增加,LVM 构建时中间结果规模以接近 60% 的比例快速增长,而 CSBN 模型增长较为平缓,这说明本文通过对评分数据进行分类再构建隐变量模型的方法,大幅减少了模型构建中的中间结果数量,且保证了模型构建的高效性。

表 3 MovieLens 数据集上模型构建的中间结果规模比较

Table 3 Comparison of intermediate result size in model construction on MovieLens dataset

用户数	LVM 模型	CSBN 模型
100	5.49×10^{7}	3.36×10^{6}
500	3.13×10^{8}	1.19×10^{7}
900	5.95×10^{8}	3.64×10^{7}
1 300	9.07×10^{8}	5.55×10^{7}

表 4 大众点评数据集上模型构建的中间结果规模比较

Table 4 Comparison of intermediate result size in model construction on DianPing dataset

用户比例/%	LVM 模型	CSBN 模型
30	2.01×10^{8}	1.23×10^{7}
50	3.02×10^{8}	1.85×10^{7}
80	4.83×10^{8}	2.96×10^{7}

6.3 有效性测试

为了测试模型的有效性,本文在 1 300 个用户的 MovieLens 数据集和 80% 大众点评数据集上,测试基于 CSBN 和 LVM 模型估计的不同 top-k 用户偏好的召回率、准确率和 F 值,结果分别如图 6 和图 7 所示。

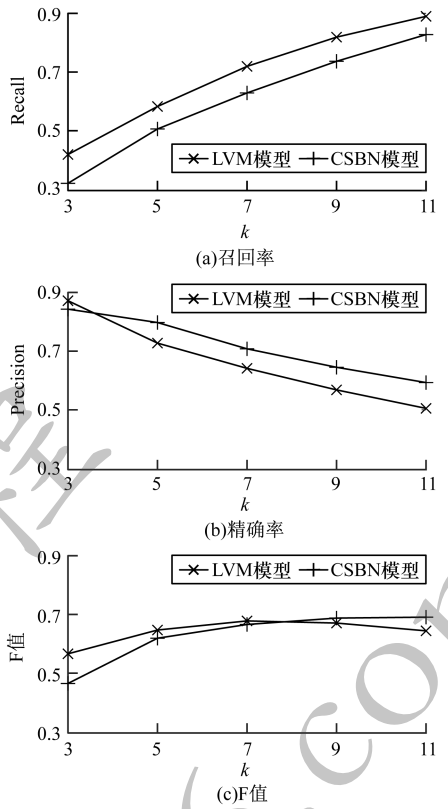


图 6 MovieLens 数据集上基于 LVM 和 CSBN 模型的用户偏好对比

Fig. 6 Comparison of user preferences based on LVM and CSBN models on MovieLens dataset

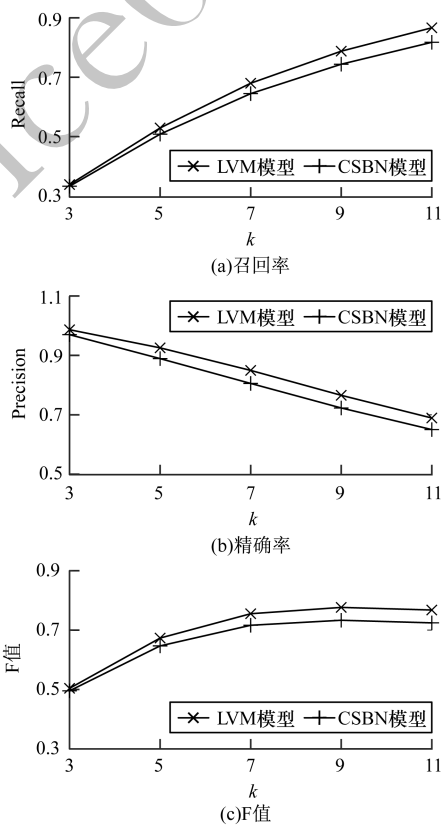


图 7 大众点评数据集上基于 LVM 和 CSBN 模型的用户偏好对比

Fig. 7 Comparison of user preferences based on LVM and CSBN models on DianPing dataset

从图 6、图 7 可以看出,在 2 种数据集上基于 CSBN 和 LVM 估计的偏好的召回率、精确率和 F 值基本相同,随着 k 值的增加,基于 CSBN 和 LVM 的召回率、F 值随之上升,精确率随之下降,并且相差不大。在 $k=7$ 和 $k=3$ 时,两者的 F 值基本相同,说明此时 CSBN 和 LVM 的召回率和精确率达到了平衡,这在一定程度上说明了本文方法有效。

本文在不同数据量的 2 种数据集上,分别测试基于 CSBN 模型估计的偏好的召回率、精确率和 F 值,结果如图 8 和图 9 所示。可以看出,在 MovieLens 数据集上,随着 k 值的增加,CSBN 的召回率上升,精确率下降,而 F 值趋于稳定,在不同用户个数下训练的模型结果几乎相同,说明了本文方法的稳定性。但在大众点评数据集上,不同数据量的召回率、精确率和 F 值差距较 MovieLens 上偏大,原因在于大众点评数据量较小,仅为 MovieLens 的 10%,导致模型的结构和参数并未达到最优,这也符合大众点评数据集的真实情况。

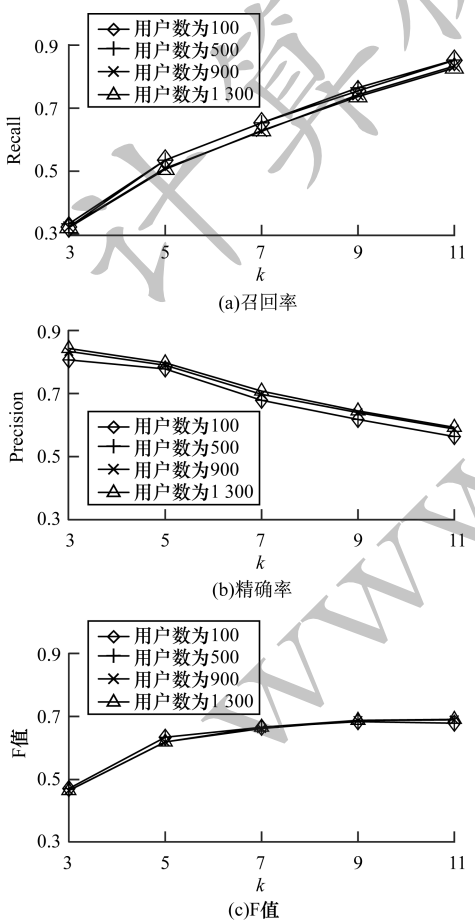


图 8　MovieLens 数据集上 CSBN 模型的用户偏好发现结果

Fig. 8　User preference discovery results of CSBN model on MovieLens dataset

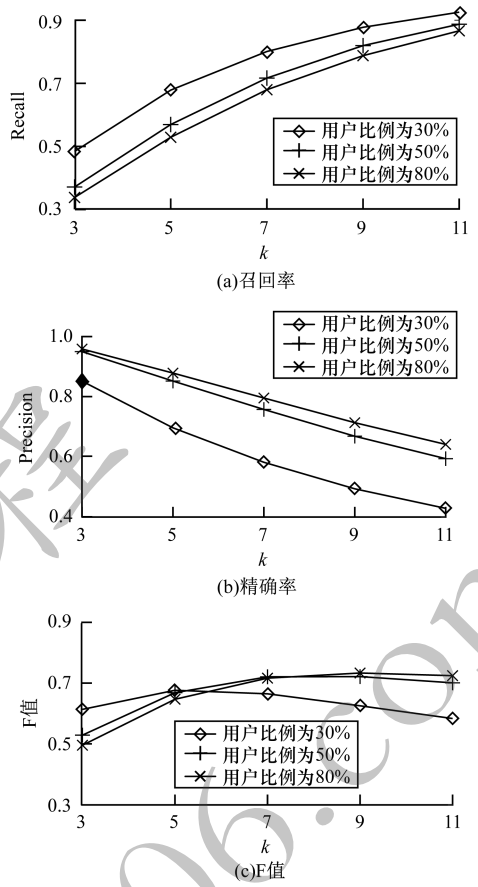


图 9　大众点评数据集上 CSBN 模型的用户偏好发现结果

Fig. 9　User preference discovery results of CSBN model on DianPing dataset

大众点评数据集上偏好估计的精确率高于 MovieLens 数据集,原因在于用户喜好的“餐饮”类型比“电影”类型更加明显。上述实验结果说明了 CSBN 模型构建方法在真实数据集上的有效性,并且没有因为提高效率而降低有效性。

7　结束语

本文提出一种基于深度信念网和隐变量模型的用户偏好发现方法,该方法既可以描述隐含的用户偏好,又可以反映用户评分数据中各属性间的任意不确定性依赖关系,且能够提高模型构建效率,克服传统隐变量模型构建过程中迭代计算多、中间结果规模大等问题。但是,本文方法是在静态数据中构建 CSBN 模型,实际应用的评分数据是动态变化且不断增加的。因此,下一步将以增量学习的方式在动态数据上构建 CSBN 模型。此外,用户的偏好也随着时间发生变化,建立动态的模型对不断变化的偏好进行预测,也是今后的研究方向。

参考文献

[1] GroupLens. MovieLens-1M dataset[EB/OL]. [2019-02-15]. https://grouplens.org/datasets/movielens/1m/.

[2] ZHANG Lianwen, GUO Haipeng. Introduction to Bayesian networks[M]. Beijing: Science Press, 2006. (in Chinese)
张连文,郭海鹏. 贝叶斯网引论[M]. 北京:科学出版社,2006.

[3] KOLLER D, FRIEDMAN N. Probabilistic graphical models: principles and techniques[M]. WANG Feiyue, HAN Suqing, Translate. Beijing: Tsinghua University Press, 2015. (in Chinese)
KOLLER D, FRIEDMAN N. 概率图模型:原理与技术[M]. 王飞跃,韩素青,译. 北京:清华大学出版社,2015.

[4] SCHÜTZ W, SCHÄFER R. Bayesian networks for estimating the user's interests in the context of a configuration task [C]//Proceedings of UM2001 Workshop on Machine Learning for User Modeling. Washington D. C., USA: IEEE Press, 2001: 13-17.

[5] ZHANG Hongyi, WANG Liwei, CHEN Yuxi. Research progress of probabilistic graphical models: a survey[J]. Journal of Software, 2013, 24(11): 2476-2497. (in Chinese)
张宏毅,王立威,陈瑜希. 概率图模型研究进展综述[J]. 软件学报,2013,24(11):2476-2497.

[6] FRIEDMAN N. The Bayesian structural EM algorithm[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. [S. l.]: Morgan Kaufmann Publishers Inc., 1998: 129-138.

[7] HAN Yanan, CAO Han, LIU Liangliang. Collaborative filtering recommendation algorithm based on score matrix filling and user interest[J]. Computer Engineering, 2016, 42(1): 36-40. (in Chinese)
韩亚楠,曹菡,刘亮亮. 基于评分矩阵填充与用户兴趣的协同过滤推荐算法[J]. 计算机工程,2016,42(1):36-40.

[8] ZHAO Guoshuai, QIAN Xueming, XIE Xing. User-service rating prediction by exploring social users' rating behaviors[J]. IEEE Transactions on Multimedia, 2016, 18(3): 496-506.

[9] KASSAK O, KOMPAN M, BIELIKOVA M. User preference modeling by global and individual weights for personalized recommendation [J]. Acta Polytechnica Hungarica, 2015, 12(8): 27-41.

[10] WEN Junhao, YUAN Peilei, ZENG Jun, et al. Research on collaborative filtering recommendation algorithm based on topic of tags[J]. Computer Engineering, 2017, 43(1): 247-252, 258. (in Chinese)
文俊浩,袁培雷,曾骏,等. 基于标签主题的协同过滤推荐算法研究[J]. 计算机工程,2017,43(1):247-252,258.

[11] FANG Bing, NIU Xiaoting. Tag-based matrix factorization recommendation algorithm [J]. Application Research of Computers, 2017, 34(4): 1022-1025, 1031. (in Chinese)
方冰,牛晓婷. 基于标签的矩阵分解推荐算法[J]. 计算机应用研究,2017,34(4):1022-1025,1031.

[12] KIM J S, JUN C H. Ranking evaluation of institutions based on a Bayesian network having a latent variable[J]. Knowledge-Based Systems, 2013, 50: 87-99.

[13] GAO Renshang, YUE Kun, WU Hao, et al. Modeling user preference from rating data based on the Bayesian network with a latent variable [C]//Proceedings of International Conference on Web-Age Information Management. Berlin, Germany: Springer, 2016: 3-16.

[14] GAO Yan, YUE Kun, WU Hao, et al. Construction and inference of latent variable model oriented to user preference discovery [J]. Journal of Computer Applications, 2017, 37(2): 360-366. (in Chinese)
高艳,岳昆,武浩,等. 面向用户偏好发现的隐变量模型构建与推理[J]. 计算机应用,2017,37(2):360-366.

[15] HUANG Y, BIAN L. A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet [J]. Expert Systems with Applications, 2009, 36(1): 933-943.

[16] CHAPELLE O, ZHANG Y. A dynamic Bayesian network click model for Web search ranking[C]//Proceedings of International Conference on World Wide Web. New York, USA: ACM Press, 2009: 1-10.

[17] YUE Kun, WEI Mujin, TIAN Kailin, et al. Representing and inferring causalities among classes of multidimensional data[M]// LI Qing, FENG Ling, PEI Jina, et al. Advances in data and Web management. Berlin, Germany: Springer, 2009: 223-234.

[18] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[19] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[20] FISCHER A, IGEL C. An introduction to restricted Boltzmann machines [C]//Proceedings of Iberoamerican Congress on Pattern Recognition. Berlin, Germany: Springer, 2012: 14-36.

[21] LIU Gang, XU Chao, CHEN Siyi, et al. Image classification with stacked restricted Boltzmann machines and hybrid neural network[J]. Journal of Chinese Computer Systems, 2017, 38(9): 2146-2151. (in Chinese)
刘罡,徐超,陈思义,等. 结合深度置信网络与混合神经网络的图像分类方法[J]. 小型微型计算机系统,2017,38(9):2146-2151.

[22] MCLACHLAN G, KRISHNAN T. The EM algorithm and extensions [M]. New York, USA: John Wiley & Sons, 2007.

编辑 吴云芳