文章编号：2095-2163(2020)02-0048-09 中图分类号：TP391 文献标志码：A

基于滑动窗口的 CP-nets 增量式学习研究

何新新，朱 阳
（烟台大学 计算机与控制工程学院，山东 烟台 264005）

摘 要：偏好信息挖掘是人工智能领域数据挖掘中一个重要的研究内容，近年来得到了广泛的研究.随着当前社会发展和数字数据的急剧增长，实时应用中的偏好数据是以数据流的形式快速生成。然而，挖掘偏好的动态特性越来越需要适应流式数据变化的解决方案。造成这种情况的主要原因是用户的偏好不是静态的，可以随着时间的推移发生变化，而传统的偏好求取方法大多集中应用在静态数据集中，不能高效地处理数据流。本文主要采用基于窗口的 CP-nets 增量式学习算法处理偏好数据流。该算法在合成数据集和真实数据集上的实验结果表明本文提出的算法能够根据用户的上下文偏好信息得到与传统学习算法大体一致的较准确的用户 CP-nets 模型，并且与传统算法比较，该算法的时间复杂度低，算法效率更高。

关键词：数据挖掘；数据流；增量式算法；CP-nets

CP-nets incremental learning algorithms from preference data streams based on sliding windows form

HE Xinxin, ZHU Yang
(School of Computer and Control Engineering, Yantai University, Yantai Shandong 264005, China)

【Abstract】 Mining preference information is an important research content in artificial intelligence field data mining, and has been widely studied in recent years. However, with the rapidly growth of digital data in the current society, more and more data are generated in new applications, and preference data in real-time applications are generated sharply in the form of data streams. Mining the dynamic characteristics of preferences increasingly requires solutions that quickly adapt to change. The main reason for this situation is that the user preferences are not static, and can change over time. Traditional preference extraction methods are mostly concentrated on static scenes and cannot process data streams efficiently. In this paper, incremental Conditional preference networks (CP-nets) based on sliding window algorithm is introduced to deal with preference data streams . The experimental results of the algorithm on synthetic data sets and real data sets show that the proposed algorithm could obtain user CP - nets model more accurately consistent with traditional learning algorithms according to the context of user preference information . Compared with traditional algorithm, this algorithm has low time complexity and higher efficiency.

【Key words】 data mining; data stream; incremental algorithm; CP-nets

0 引 言

在现代社会各专业领域的发展中，实时数据的研究与处理已经比较成熟，其中包括推荐系统中数据流的偏好信息挖掘和决策的生成等方面的数据处理[1]。多数的偏好数据处理工作都是在静态时刻场景中进行，并不能排除随着时间逐渐推移甚至其他因素的不确定性对用户的数据偏好会产生不容忽视的影响[2]。某个静态时刻场景的数据库只能判断用户静态的偏好信息。并不能有效地处理动态性比较强的数据流挖掘，例如在金融市场的银行交易[3]。

本篇论文主要集中研究的是时间因素对用户偏好演化的影响。在现实社会的应用中，用户偏好大多是动态性的，也就是说能够随着时间的推移而产生偏好的变化[4]。传感网络发展潜力下的网页随时间推移能够动态地批量变换，商务网站提供的产品广告因为客户需求和企业计划的改变而做出动态的调整，这都属于动态的数据流偏好模式。要得到用户动态的偏好信息帮助判断用户的购买需要和模拟用户的购买过程，就要运用动态的数据流处理模型、即 CP-nets 的增量式学习研究确定偏好信息[5]。这种模型下的推荐系统具有的优点就是动态地获取用户的偏好信息，然后根据用户的偏好推荐合适的个性化产品信息[6]。

用户的某个静态时刻场景中的偏好在现实社会中的合理性和实用性都是较低的外界环境中的各种

基金项目：山东省重点研发计划(2015GSF115009)；国家自然科学基金(61403328, 61572419)。

作者简介：何新新(1994-)，女，硕士研究生，主要研究方向：人工智能领域的 CP-nets 增量式学习；朱 阳(1995-)，男，硕士研究生，主要研究方向：人工智能领域的 CP-nets 启发式学习。

通讯作者：何新新 Email：hexinxin_a@ sina.com

收稿日期：2019-12-04

因素随着时间前进的改变导致用户的偏好是动态变动的，例如在头条的新闻网站或者是社交网站中的热门新闻和热门搜索都是随着时间的变化在不断地改变，这是因为用户的偏好在不断地改变，兴趣在不断地更新，因此数据挖掘技术就要提出新的偏好挖掘模型来应对这种基于时间的动态数据流[7]。

动态的偏好数据挖掘在人工智能领域更具有挑战性，虽然动态的偏好数据挖掘在数据存储和数据流的海量增加两方面都具有优势[8-9]，但也存在以下两方面困难：

(1)数据没有储存，在需要时也无法获得，每个数据元组在到达时必须被接受，一旦被丢弃，不可能再次被检查。

(2)在静态数据处理中，连续增加的数据将成为海量数据，由于存储空间的增加使得 CPU 负荷过高。

在本文中提出的基于滑动窗口的 CP-nets 增量式学习方法将数据流进行动态挖掘处理，本文的主要贡献如下：

(1)基于滑动窗口的 CP-nets 增量式学习方法解决了静态数据存储以及运算上存在的存储空间占用大，算法耗时长等困难。

(2) 基于滑动窗口的 CP-nets 增量式学习方法不仅可以处理可扩展的累积偏好数据，还可以处理不断增加流式偏好数据。

(3)实验证明，CP-nets 增量式学习方法可以学习得到一个准确的 CP-net。对于同一个数据集，该方法得到的实验结果与传统学习方法得到的结果大体一致，支持任意的结果比较。

1　相关工作

随着社会的发展，数据的产生逐渐趋于流式化，数据流中偏好数据挖掘研究在近年得到广泛发展。Papini 等人[10-11] 提出 2 种从数据流中挖掘偏好关系的算法，允许指定一些特定属性取值在所有数据中优先可取，即提前定义不同属性的重要性，而不是计算属性之间的依赖关系。从现实意义上说一个数据模式里面属性的偏好关系更体现在属性的依赖关系，即同一属性不同取值在约束相同的条件下对其他属性取值偏好产生的重要影响。本文的工作是从偏好问题的紧凑关系定义角度看，属性之间的依赖关系大于属性之间的优先关系。与 Papini 等人的研究不同的是，本文提出的算法不是研究属性之间的优先级，而是主要研究属性之间的依赖关系，并且是在动态的偏好数据流中学习 CP-nets。

由于偏好信息在推荐系统等方面具有巨大的应用价值，从 CP-nets 这一偏好模型的提出，一系列学习该模型的算法被先后提出。学习 CP-nets 方法可以分为：主动学习、被动学习、启发式学习和精确学习。

Koriche 等人[12] 提出的主动学习算法，引导用户通过一系列查询来识别具有二进制值的 CP-nets 的偏好排序。Aggarawal 等人[13] 通过启发式学习方法提出了一种不确定数据流的聚类方法，创建了一个不确定性模型。Liu 等人[14-15] 通过启发式学习方法从含噪声数据样本中学习 CP-nets。大多数现有的学习方法忽略了对噪声数据样本的处理，文献[14]介绍了一种从噪声样本中学习 CP-nets 的新模型，提出了一种多项式时间内求解偏好问题的算法。研究中还提出了一种从不一致例子中学习 CP-nets 的被动学习方法[15]，该方法利用分支搜索和界搜索将学习偏好图形问题转化为一个 0-1 规划问题，然后将得到的偏好图形等价地转化为 CP-nets。Mengin 等人[16-17] 通过对依赖结构的假设来研究解决了多属性域上的 CP-nets 学习问题。研究所考虑的假设是属性可分离(属性值之间不依赖，每个属性值的偏好独立于其他属性值)然后采用依赖结构无圈图的形式，求取一组局部偏好关系(CP-nets)。Guerin 等人[18] 提出了一种不限于交换比较的启发式在线算法，该算法是一种新的智能体学习用户偏好的算法。通过算法生成一系列在线查询学习，通过创建节点和初始化 CPT 来为用户构建一个 CP-net。Liu 等人[19] 从不一致的例子中得到二元和多值变量上一致的 CP-net。该方法不能解决数据量增加用户首选项可能会随着新数据的变化而改变的问题。

以上工作都是针对静态时刻场景中的静态数据提出的 CP-nets 学习算法，本文主要针对动态的数据流中属性之间的依赖关系进行 CP-nets 学习。

2　CP-nets 相关概念

在本节中，主要介绍了 CP-nets 的表示和性质，并且举例说明了 CP-nets 在表示属性偏好中的应用。其中，2.1 节给出了 CP-nets 的基本定义和表示，2.2 节描述了 CP-nets 的性质。

2.1　CP-nets 的基本定义

定义 1　条件偏好关系　偏好成对存在，共有 3 种关系($\succ$、$\prec$、$\approx$)，用 $r(v,\ v')$ 来记录，$r(\succ_v, |v'|)$ 表示属性 v' 取值不变的情况下属性 v 取值的偏好。在属性集合 $Dom(V)$ 中，如果存在关系 $v \succ v', v \prec v', v \approx v'$ 分别表示相对于属性 v 用户更偏好

于属性 v'；相对于 v 用户更偏好于 v'；用户在 v 和 v' 之间没有偏好。同一属性取值下的所有偏好对的全序关系称为偏好关系。

定义 2 父子关系 属性之间的父子关系即依赖关系用 $pare(v_i, v_j)$ 表示，v_i 为父属性，v_j 为子属性，记为 $pa(v_j) = v_i$. $Dom(v_i)$ 表示 v_i 的有限定义域为 $Dom(v_i) = \{o_1, o_2, o_3, \cdots, o_m\}$，$(A,B) \in V$，$(a_1, a_2) \in Dom(A)$，$(b_1, b_2) \in Dom(B)$，属性 A,B 父子关系程度为：

$$Pare(A,B) = \frac{sum(r(\succ_B \mid A = a_1)) + sum(r(\succ_B \mid A = a_2))}{sum(r(A, B))}. \tag{1}$$

定义 3 条件偏好网络(CP-nets)[17] CP-nets 由偏好网络结构和条件偏好表(conditional preference table，CPT)两部分构成，CPT 是一个集合 (V,E) 的图模型 G，其中 $V = \{v_1, v_2, v_3, \ldots, v_n\}$ 是构成网络中节点的一组属性变量，而 E 是一组连接一对属性变量的有向边集合 $E = \{(v_i, v_j) \mid E(v_i, v_j), v_i, v_j \in V\}$。其中，$E(v_i, v_j)$ 表示属性 v_i 是属性 v_j 的父亲。每个节点都有一个条件偏好表，表示在其他属性取值约束下该节点属性取值的偏好。在 CP-nets 中，每个 $DOM(v_i)$ 表示 v_i 的有限定义域。对于每个属性 v_i 的父属性集合 $pa(v_i)$ 会影响对 v_i 值的偏好。这定义了一个依赖图，其中每个节点 v_i 与 $pa(v_i)$ 中的每个属性都有一条有向边。

为了更形象地说明 CP-nets 结构，图 1 是以客人选择晚宴菜单的 CP-net 结构图。集合 $V = \{W, M, S\}$ 分别代表菜单中的饮品、点心和主菜。$DOM(W) = \{w_1, w_2\}$ 表示饮品，包括红酒和茶，$DOM(M) = \{m_1, m_2\}$ 表示点心，包括酥饼和蛋糕，$DOM(S) = \{s_1, s_2\}$ 表示主菜，包括牛排和鱼。这位客人更喜欢红酒和蛋糕，而对于主菜的选择取决于饮品和点心的选择。

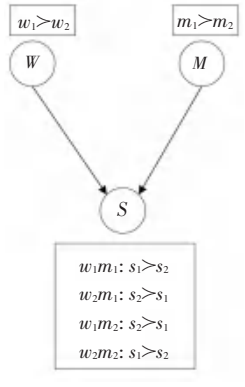


图 1 晚宴菜单的 CP-net

Fig. 1 A CP-net for dinner menu

2.2 CP-nets 的性质

设 N 为任意一个 CP-net，其所能表达的条件偏好关系为 $\succ$，$o_1, o_2, \ldots, o_m \in O$ 为决策空间 O 中的任意 m 个结果.如果在 N 的导出图中，总存在包含一个或多个边的路径，从任意顶点 o_i 出发均能到达 o_j，则说明 o_i 和 o_j 之间的占优测试成立，即 $o_j \succ o_i$。若结果序列 $(o_1, o_2, \ldots, o_m)$ 的如下关系：$o_1 \succ o_2 \succ o_3$，$o_{i-1} \succ o_i, \ldots, o_m \succ o_1$ 不成立，即 $o_1, o_2, \ldots, o_m$ 不构成环形序列，则称 CP-nets 是一致的。任意一个无环的 CP-net 都是一致的。有环 CP-nets 的一致性不确定，当导出的结构图 CPT 没有出现环路时，则其满足一致性，当导出的结构图 CPT 出现环路时则不满足一致性。

若结果集 O 上任何两个结果之间的占优关系都能被 $\succ$ 所表达，总有 $o_i \succ o_j$ 或 $o_j \succ o_i$ 成立，则称 CP-nets 满足完备性条件。如果 N 的属性集的大小为 n，属性阈值 $Dom(X_i) = k$，则属性结果集 O 的个数为 k^n，若结果之间的偏好关系个数满足 $k^{n-1}(k^n - 1)$，则 N 为完备的 CP-net[20]。

3 基于滑动窗口的增量式学习方法

为了提高数据处理的效率和速度，本文提出的基于滑动窗口的增量式学习方法主要是将偏好数据以部分滑动的形式输入到数据处理窗口中，将数据进行一种位序列形式的计数表示，再进行属性之间的依赖关系学习、即 CP-nets 的学习。

3.1 基本定义

定义 4 偏好数据流 P 被称为偏好数据流，表示为 $P = (p_1, p_2, p_3, \ldots, p_v)$，偏好数据流目前的偏好表达式个数记为 $|P| = v$。设属性 $V = \{v_1, v_2, v_3, \ldots, v_n\}$，属性个数记为 $|V| = n$。属性 v_i 取值的有限定义域表示为 $Dom(v_i) = \{o_1, o_2, o_3, \ldots, o_m\}$，属性 v_i 的一个取值表示为 $Dom(v_i) = o_i$；属性 v_i 的取值个数记为 $|O_i| = m$。偏好表达式表示为：

$$p_i = r(Dom(v_1) \times Dom(v_2) \times Dom(v_3) \times, \ldots, \times Dom(v_n)), \tag{2}$$

定义 5 偏好表达项式 x_j 被称为偏好表达项式，表示为：

$$x_j = r(Dom(v_1) \times Dom(v_1) \times, \ldots, Dom(v_z)), z \leqslant n, \tag{3}$$

偏好表达项式 x_j 与数据流中原有的偏好表达式 p_i 存在被包含关系 $x_j \subset p_i$(其中 i, j 为任意取值)。

定义 6 偏好事务序列 TDS 被称为是一个连续的偏好事务序列，表示为 $TDS = (T_1, T_2, T_3, \ldots, T_n)$。其中，$n$ 表示最新传入偏好数据流 T_n。事务

$T_i=(p_1,p_2,p_3,\dots,p_w)$ 包含 w 个偏好表达式，事务 T_i 表示为 $|T_i|=w$。每窗口 $TransSW_i$ 包含固定的事务数量 $|TransSW_i|=m$。滑动粒度大小表示为 s。滑动窗口表示为：

$$TransSW_i=[T_{N-m+s},T_{N-m+2s},T_{N-m+3s},\dots,T_N]. \quad (4)$$

滑动窗口示例如图 2 所示。

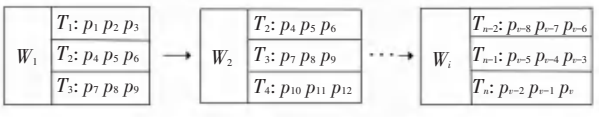


图 2　滑动窗口示例

Fig. 2　An example of sliding windows

举例说明：偏好数据库见表 1，属性个数 $|V|=3$，分别为 $V=\{A,B,C\}$，属性取值 $|O|=2$，分别为：$A=(a_1,a_2)$，$B=(b_1,b_2)$，$C=(c_1,c_2)$。$TDS=T_1,T_2,T_3,T_4$ 是当前时间一个连续的偏好事务序列，当 $w=3$ 时 T_i 表示为 $T_i=(p_1,p_2,p_3)$。窗口大小表示为 $|m|=3$，滑动粒度 $S=1$ 的滑动窗口表示为：$TransSW_1=[T_1,T_2,T_3]$，$TransSW_2=[T_2,T_3,T_4]$。

表 1　偏好数据库表

Tab. 1　Preference database table

Transaction	Preference information		
T_1	$a_1b_1c_1 \succ a_1b_2c_1$	$a_2b_2c_1 \succ a_2b_1c_2$	$a_1b_2c_1 \succ a_2b_1c_1$
T_2	$a_1b_1c_1 \succ a_1b_2c_1$	$a_2b_2c_1 \succ a_2b_1c_2$	$a_1b_2c_2 \succ a_2b_2c_1$
T_3	$a_1b_1c_2 \succ a_2b_2c_2$	$a_2b_1c_1 \succ a_1b_1c_2$	$a_1b_2c_2 \succ a_1b_2c_1$
T_4	$a_2b_1c_1 \succ a_1b_1c_2$	$a_1b_1c_2 \succ a_2b_1c_2$	$a_2b_1c_2 \succ a_1b_2c_1$

3.2　偏好关系的位序列表示

在基于滑动窗口的增量式 CP-nets 学习算法中，对于当前偏好事务序列 TSD 的滑动窗口 $TransSW_i$ 中偏好关系 p_i 的子集 x_j，构造了一个二进制位序列，表示为 $Bit(x_j)$。如果偏好表达项式 x_j 在传输中对于当前的窗口 $TransSW_n$ 存在，第 i 位 x_j 被设置为 1，否则，将被设置为 0。

例如根据表 1 的偏好数据库可得到事务 T_1 中偏好关系 p_1 存在偏好表达项式 $x:a_1b_1 \succ a_1b_2$。于是在窗口 $TransSW_1$ 中 $Bit(a_1b_1 \succ a_1b_2)=(100)$。采用位序列的表示方法得到属性 A,B 的偏好关系的位序列表见表 2。

表 2　属性 *A*，*B* 的偏好关系的位序列表

Tab. 2　The bit-sequence of the preference between *A* and *B*

Windows	Transaction	The bit-sequence of A and B
$TransSW_1$	T_1	$Bit(a_1b_1 \succ a_1b_2)=(100)$ $Bit(a_2b_2 \succ a_2b_1)=(010)$
	T_2	$Bit(a_1b_1 \succ a_1b_2)=(100)$ $Bit(a_2b_2 \succ a_2b_1)=(010)$ $Bit(a_1b_2 \succ a_2b_2)=(001)$
	T_3	$Bit(a_2b_1 \succ a_1b_1)=(010)$
$TransSW_2$	T_2	$Bit(a_1b_1 \succ a_1b_2)=(100)$ $Bit(a_2b_1 \succ a_1b_1)=(010)$ $Bit(a_1b_2 \succ a_2b_2)=(001)$
	T_3	$Bit(a_2b_1 \succ a_1b_1)=(010)$
	T_4	$Bit(a_2b_1 \succ a_1b_1)=(100)$

3.3　初始窗口和滑动窗口

由表 2 可知，$TransSW_1$ 表示初始窗口，$TransSW_1=[T_1,T_2,T_3]$。$TransSW_1$ 中将偏好关系 $x_i(x_i \subset p)$ 的二进制位序列相加，计算偏好表达项式 x_j 的支持数，即窗口 $TransSW_i$ 中偏好表达项式 x_j 的个数记为 $sup(x_j)$。例如表 3 中：

$Bit(a_2b_2 \succ a_1b_2)=(100100001)$，偏好关系的计数值 $sup(a_2b_2 \succ a_1b_2)=3$。由 $TransSW_1$ 在滑动粒度 $s=1$ 时滑动得到 $TransSW_2$。根据表 2 中每个事务 T_i 中偏好表达项式 x_j 的支持数 $sup(x_j)|T_i$ 将表 3 中初始窗口 $TransSW_1$ 中的 $sup(x_j)|T_1$ 删除并加上事务 T_4 中偏好表达项式 x_j 的支持数 $sup(x_j)|T_4$ 得到滑动窗口 $TransSW_2$。属性 A,B 偏好关系的位序列见表 4。

表 3　初始窗口 *TransSW*$_1$ 中属性 *A*，*B* 偏好关系的位序列表

Tab. 3　The bit-sequence of the preference relation between *A* and *B* in the initial window *TransSW*$_1$

Windows	The bit-sequence of A and B	Count value
$TransSW_1$	$Bit(a_1b_1 \succ a_1b_2)=(100100)$	$sup(a_1b_1 \succ a_1b_2)=2$
	$Bit(a_2b_2 \succ a_2b_1)=(010010)$	$sup(a_2b_2 \succ a_2b_1)=2$
	$Bit(a_1b_2 \succ a_2b_2)=(001)$	$sup(a_1b_2 \succ a_2b_2)=1$
	$Bit(a_2b_1 \succ a_1b_1)=(010)$	$sup(a_2b_2 \succ a_2b_1)=1$

表 4 $TransSW_2$ 中属性 A, B 偏好关系的位序列表

Tab. 4 The bit-sequence of the preference relation between A and B in the initial window $TransSW_2$

Windows	The bit-sequence of A and B	Count value
$TransSW_2$	$Bit(a_1b_1 \succ a_1b_2) = (100)$	$sup(a_1b_1 \succ a_1b_2) = 1$
	$Bit(a_2b_1 \succ a_1b_1) = (010)$	$sup(a_2b_1 \succ a_1b_1) = 1$
	$Bit(a_1b_2 \succ a_2b_2) = (001)$	$sup(a_1b_2 \succ a_2b_2) = 1$
	$Bit(a_2b_1 \succ a_1b_1) = (010100)$	$sup(a_2b_1 \succ a_1b_1) = 2$

4 CP-nets 学习算法

针对属性偏好的本质,本文提出通过设置阈值变量对属性偏好关系进行定量判断后近似为定性偏好信息,该定性信息体现在 CP-nets 中能够描述属性之间的偏好关系。

4.1 基本定理

定理 1 比较 2 个属性之间的偏好关系时,可以理论上完全产生 $S = |O|^{|O|} \times (|O|^{|O|} - 1)$ 对偏好关系,只有一个属性固定取值,另一个属性取值存在偏好关系时,即满足 $r(\succ_{v_n} | v_m = o_i)$ 或者 $r(\succ_{v_m} | v_n = o_j)$ 才是有效偏好关系。理论上产生有效偏好关系的个数为 $S' = |O|^3 \times (|O| - 1)$。在偏好数据流中有效偏好关系的数量占总数量一定比例,才能判定属性之间是否可能存在父子系。公式(5)和公式(6)可分别表示为:

$$S_1 = S_{(v_n|v_m)} = \frac{sum(r(\succ_{v_m} | v_n = o_i)) + sum(r(\succ_{v_m} | v_n = o_j))}{sum(sup(r(v_n, v_m)))} \geqslant \alpha, \frac{1}{2} \leqslant \alpha \leqslant 1, \tag{5}$$

$$S_2 = S_{(v_m|v_n)} = \frac{sum(r(\succ_{v_n} | v_m = o_i)) + sum(r(\succ_{v_n} | v_m = o_j))}{sum(sup(r(v_n, v_m)))} \geqslant \alpha, \frac{1}{2} \leqslant \alpha \leqslant 1, \tag{6}$$

公式(5)、(6)中设置一个用户自定义的阈值 α 来约束存在父子关系的可能性。阈值 α 的取值根据窗口数据流的大小、偏好数据量以及用户对偏好关系的精确度确定。

举例说明:假设属性 A 与 B 取值分别是 (a_1, a_2), (b_1, b_2),产生 $2^2 \times (2^2 - 1) = 12$ 对偏好关系。理论上有效的偏好关系有 $|2|^3 \times (|2| - 1) = 8$ 对。实际数据中有效偏好的关系随着数据的改变而改变。

定理 1 中假设 $S_1 > \alpha$ (或者 $S_2 > \alpha$)成立,说明 v_n, v_m 可能存在偏好关系 $v_n \succ v_m$, (或者有偏好关系 $v_m \succ v_n$) 即 $Pa(v_n) = v_m (Pa(v_m) = v_n)$ 成立,$S_{(v_n|v_m)} \leqslant \alpha$ 说明 v_m, v_n 不可能存在偏好关系 $v_n \succ v_m$ (或者偏好关系 $v_m \succ v_n$), 即 $Pa(v_m) \neq v_n (Pa(v_n) \neq v_m)$。如果公式(5)、(6)同时成立则比较 S_1 与 S_2 的取值大小,如果 $S_1 > S_2 \geqslant \alpha$, 则 $v_n \succ v_m$ 关系成立存在最大可能性,否则 $v_m \succ v_n$。

定理 2 如果满足定理 1 即说明两属性之间可能存在一种偏好关系,确立了子属性的待定父亲属性,待定父亲属性的不同取值对于子属性取值偏好关系产生的影响大体一致,这里设计一个自定义的阈值 β 来容纳父属性对子属性的影响存在一定偏差。阈值变量 β 可以根据用户对于属性依赖度的需求自行定义。该定理的公式(7)、(8)可分别表示为:

$$\begin{cases} S_1 \geqslant \alpha; \\ N_1 = N_{(v_n|v_m)} = \dfrac{sum(sup(v_{m|o_j \succ o_i} | v_{n|o_i}))}{sum(sup(v_{m|o_i \succ o_i} | v_{n|o_j}))} = 1 \pm \beta. \end{cases} \tag{7}$$

$$\begin{cases} S_2 \geqslant \alpha; \\ N_2 = N_{(v_m|v_n)} = \dfrac{sum(sup(v_{n|o_j \succ o_i} | v_{m|o_i}))}{sum(sup(v_{n|o_i \succ o_i} | v_{m|o_j}))} = 1 \pm \beta. \end{cases} \tag{8}$$

公式(7)、(8)能够进一步确定偏好关系,确定定理 1 中筛选出的父子关系是否成立。

进一步考虑 v_n 的不同取值对 v_m 取值的偏好关系产生的影响在允许偏差内一致,当定理 1 满足表达式 $S_1 > \alpha, S_1 > S_2 \geqslant \alpha$, 定理 2 满足表达式 $N_1 = 1 \pm \beta$ 时属性 v_n 与属性 v_m 之间存在偏好关系 $v_n \succ v_m, v_n = pa(v_m)$ 成立;同理,当定理 1 满足表达式 $S_2 > \alpha, S_2 > S_1 \geqslant \alpha$, 定理 2 满足表达式 $N_2 = 1 \pm \beta$ 时属性 v_n 与属性 v_m 之间存在偏好关系 $v_m \succ v_n, v_m = pa(v_n)$ 成立。

基于以上定理分析,可以得到算法 1 所示的伪代码和图 3 所示的算法流程图。

算法 1 基于滑动窗口的 CP-nets 增量式学习算法

输入 TDS(偏好事务序列), 阈值 α, 阈值 β

输出　a CP-net

TransSW = null;

/* TransSW 由 m 个 TDS 组成 */

repeat:

for each incoming transaction *Ti* in *TransSW*

do *bit* - *sequence*(*x*)

if *bit* - *sequence*(*x*) ≠ 0, then

do *Sub*(*xi*)

for

if $S_{(v_n|v_m)} \geqslant \alpha$, then

if $S^*_{(v_n|v_m)} = 1 \pm \beta$ then

draw the dependencies $v_n \succ v_m$ in existence

else error

end if

else $S_{(v_m|v_n)} \geqslant \alpha$, then

if $S^*_{(v_n|v_m)} = 1 \pm \beta$ then

draw the dependencies $v_m \succ v_n$ in existence

else error

end if

end if

end for

end for

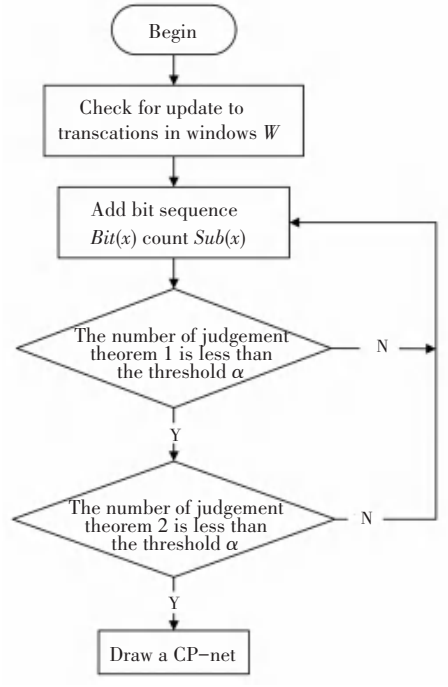


图 3　增量式算法流程图

Fig. 3　Incremental algorithm flow chart

算法 1 的计算复杂度首先从扫描数据库进行单个比较数据时算起，单个比较数据的属性个数为 n，这时的算法复杂度为 $O(n)$，一对属性之间依赖程度的计算复杂性为 $O(n^2)$。计算一个偏好事务序列的长度为 w，所以计算每个事务的计算复杂度是 $O(n^2)w$，每滑动窗口的长度为 $T = m$，所以计算每个窗口中数据偏好的计算复杂度是 $O(n^2)wm$。

定理 3　定理 1 中的 $S_{(v_n,v_m)}$ 表示父属性对子属性的影响度，$\mathrm{Min}(S_{(v_n,v_m)})$ 表示影响度最小的两属性之间的边 $E(v_n,v_m)$。

为了满足 CP-nets 的一致性，CP-nets 结构中不允许环的存在，在一个 CP-net 中每个属性都至少有一个父属性的存在则说明该 CP-net 是有环 CP-net，即存在环的 CPT 结构中不存在边入度（*V* - *in* - *degree*）为 0 的属性。根据定理 3，设计算法 *Dedge*(*C*,*C*′) 进行去环。如果 *C* 不存在边入度为 0 的属性节点，则说明 C 存在环路；若 C 中存在边入度为 0 的属性节点 v_i，在其副本 *C*′ 中删除当前节点以及与其相连的边，并重新更新节点的边入度值后删除新的边入度为 0 的节点。如果所有节点都被删除则说明 *C* 中不存在环，如果存在未被删除的节点则说明 *C* 中存在环，删除定理 3 中影响度最小的两属性之间的有向边 $E(v_n,v_m)$，重新进行 *Dedge*(*C*,*C*′) 算法直至结构 *C* 不存在环。至此，给出算法 2 的伪代码详见如下。

算法 2　无环 CP-nets 结构学习算法

输入　增量式学习得到的 CP-net *C*

输出　最优 CP-net

Dedge(*C*,*C*′)

/* C′表示记录结构学习的副本 */

while(existence *V* - *in* - *degree* = 0) do

for each $v_i \subset C$

If *vi* - *in* - *degree* = 0

Delete vi,*edgevi* - > *vi*,

C′← new in-degree table; /*将新的边入度信息更新到副本 C′*/

end if

end for

end while

if *C*′ ! = null

deleteMin($S_{(v_n,v_m)}$) *edge* from *C*;

Dedge(*C*,*C*′)

end if

if *C*′ =null

return *C*

end if

4.2 算法举例

举例说明基于滑动窗口的 CP-nets 学习算法：求窗口 $TransSW_1$ 和 $TransSW_2$ 中属性 A,B 存在的偏好关系。因为举例数据较少,定理中的阈值假设取值分别为 $\alpha=\frac{1}{2},\beta=\frac{1}{2}$。根据定理 1 中的公式(5)得窗口 $TransSW_1$ 中 $S_1=S_{(A|B)}=\frac{5}{6}>\frac{1}{2},S_2=S_{(B|A)}=\frac{1}{6}<\frac{1}{2}$,所以 A,B 属性之间可能存在关系 $A\succ B$, 即 $Pa(B)=A$。通过定理 2 确定这一关系是否成立,根据公式(8)可得 $N_1=N_{(A|B)}=\frac{2}{3}=1-\frac{1}{3}$,所以属性 A,B 存在偏好关系 $A\succ B$,即 $Pa(B)=A$ 成立,属性 A, B 的偏好结构如图 4 所示。

图 4 属性 *A*, *B* 的偏好网络结构图

Fig. 4 The preference structure CPT of *A* and *B*

同理,滑动窗口 $TransSW_2$ 中 $S_{(A|B)}=\frac{1}{5}<=\frac{1}{2}$, $S_{(B|A)}=\frac{3}{5}>\frac{1}{2}$, 所以 A,B 属性间可能存在关系 $B\succ A$ 即 $Pa(A)=B$。根据定理 2 可得 $N_{(B|A)}=\frac{1}{2}=1-\frac{1}{2}$,所以 A,B 属性之间存在偏好关系 $B\succ A$,即 $Pa(A)=B$。

在得到滑动窗口 $TransSW_1$ 中属性 A 与属性 B 的依赖关系后,根据表 1 偏好数据库表分别求属性 B 与属性 C,属性 A 与属性 C 之间的偏好关系。属性 B 与属性 C 偏好关系最后的位序列表,即 $TransSW_1$ 中得到的 $sup(\succ_{(Dom(B)\times Dom(C))})$ 计数如下：

$$sup(b_1c_1\succ b_1c_2)=2$$

$$sup(b_2c_2\succ b_2c_1)=2$$

$$sup(b_2c_2\succ b_1c_2)=1$$

$$sup(b_2c_1\succ b_1c_1)=1$$

根据定理可得属性 B 与属性 C 之间存在偏好关系 $B\succ C$。属性 A 与属性 C 偏好关系最后的位序列表,即 $TransSW_1$ 中得到的 $sup(\succ_{(Dom(A)\times Dom(B))})$ 计数如下：

$$sup(a_1c_1\succ a_2c_1)=1$$

$$sup(a_2c_2\succ a_1c_2)=1$$

根据定理可得属性 A 与属性 C 的关系为 $C\succ A$ 即 $Pa(A)=C$。 属性之间的偏好关系是 $A\succ B\succ C\succ A$,每个属性的入度都不为 0,$\mathrm{Min}(S_{(A,C)})=2$,所以删除 $edge(A,C)$。根据表 1 的偏好数据库表可得属性 A, B, C 的偏好网络结构图如图 5 所示。

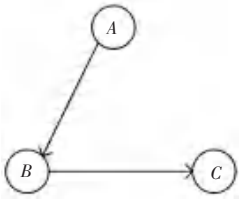


图 5 属性 *A*, *B*, *C* 的偏好网络结构图

Fig. 5 The preference structure CPT of *A*, *B* and *C*

5 实验与结果

5.1 数据来源

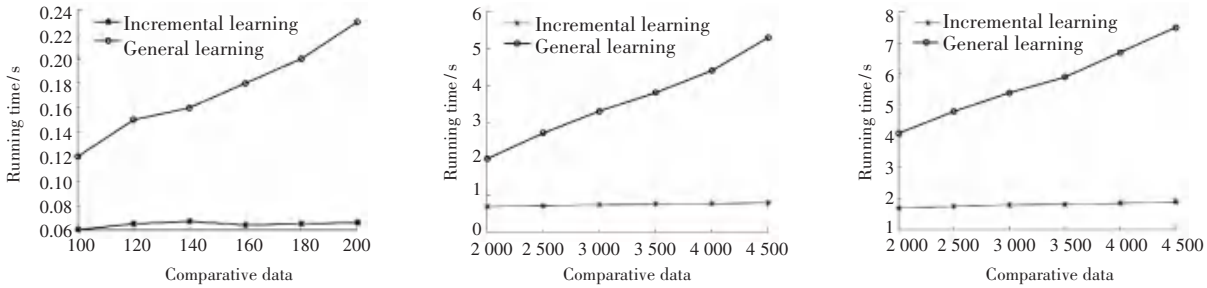
本文在 Matlab 上进行实验测试,实验数据分别是合成的随机数据集和真实数据集。对合成的随机数据集进行测试的目的是评估算法处理大数据的能力,对真实数据集进行测试是为了评估算法在真实数据集上的可应用性。其中真实数据集来自于 Kamishima 收集的寿司数据集[21]。

5.2 实验结果分析

本文在合成数据集上的测试目的是将 CP-nets 增量式学习与 CP-nets 传统学习[14-15]的运行时间进行对比。合成数据集包含用户的 4 500 对偏好关系,其中数据集包含 30 个属性。分别将随机用户 U_1, U_2做 CP-nets 增量式学习与 CP-nets 传统学习[14-15]运行时间的平均值进行对比。

图 6 中水平轴表示成对属性比较的次数, 垂直轴表示用户 U_1, U_2 运行时间的平均值。 图 6(a)是用户输入 15 个属性的 200 对偏好数据时 CP-nets 增量式学习与 CP-nets 传统学习[14-15]运行时间的平均值进行对比;图 6(b)是用户输入 15 个属性的 4 500对偏好数据时 CP-nets 增量式学习与 CP-nets 传统学习[14-15]运行时间的平均值进行对比;图 6(c)是用户输入 30 个属性的4 500对偏好数据时 CP-net 增量式学习与 CP-net 传统学习[14-15]运行时间的平均值进行对比。实验结果表示本文提出的 CP-nets 增量式学习算法在运行时间上少于 CP-net 传统学习的运行时间。

在真实数据集的测试目的是将 CP-nets 增量式学习与 CP-nets 传统学习[14-15]的运行结果和运行时间进行对比。研究考虑 2 个随机用户 U_1,U_2 与寿司选择相关的偏好数据集[21],寿司的 5 个属性分别表示为 A_1,A_2,A_3,A_4,A_5。文中分别求了用户 U_1,U_2 以每次 20 对偏好数据增量从 100 对数据增加到 200 对数据的偏好网络结构图。



（a）对比结果 1　（b）对比结果 2　（c）对比结果 3

（a）Comparison result 1　（b）Comparison result 2　（c）Comparison result 3

图 6　用户 U_1，U_2学习 CP-nets 平均运行时间对比图

Fig. 6　User U_1, U_2 learning CP-nets average runtime comparison

图 7 表示 CP-nets 增量式学习算法中用户 U_1 以每次 20 对偏好数据增量从 100 对数据增加到 200 对数据得到的偏好网络结构图。用户 U_1 在数据增加的每个阶段得到偏好网络结构图不同，该结果表明偏好数据增量会对用户的 CP-nets、即用户偏好产生影响。图 8 表示的是传统学习 CP-nets 算法中用户 U_1 以每次 20 对偏好数据增量从 100 对数据增加到 200 对数据得到的偏好网络结构图。对比与图 7 所得的结果图，图 8 展示的所有属性偏好包含于图 7 所示的属性偏好中，该实验结果表示 CP-nets 增量式学习与 CP-nets 传统学习得到的偏好结果大体一致，本文提到的增量式学习 CP-nets 算法得到的偏好关系是准确的。

综上所述，CP-nets 增量式学习算法能够得到用户属性的动态性偏好网络结构图，并且得到的属性之间的偏好关系是准确的。

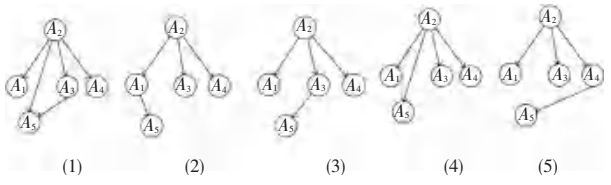


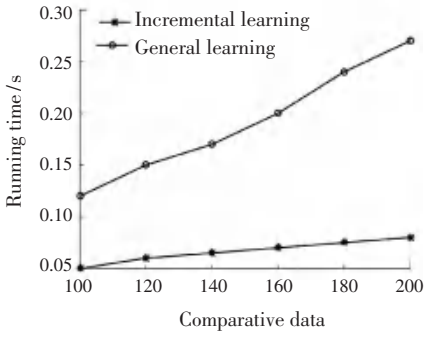
图 7　CP-nets 增量式学习算法中用户 U_1的实验结果

Fig. 7　Experimental results of user U_1 in incremental learning algorithm

图 8　传统学习 CP-nets 算法中用户 U_1的实验结果

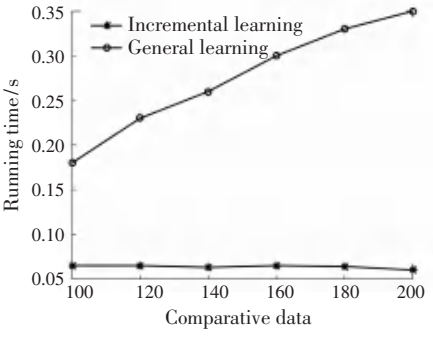
Fig. 8　Experimental results of user U_1 in traditional learning algorithm

图 8 分别表示用户 U_1，U_2 以每次 20 对的增量从 100 对偏好数据增加到 200 偏好数据，CP-nets 增量式学习与 CP-nets 传统学习[14-15]在得到某一时刻的 CP-nets 所用的确定时间进行了比较。图 9(a)表示用户 U_1 的增量学习与传统学习运行时间的比较；图 9(b)表示用户 U_2 的增量学习与传统学习运行时间的比较。图 9 中，用户 U_1，U_2 做 CP-nets 增量式学习的运行时间在数据增加过程中大致相同并且耗时随着数据增量的逐渐增加大幅度少于同时刻传统学习 CP-nets 的运行时间，而 CP-nets 传统学习的运行时间随着数据增量的增加逐渐呈线性增加。实验表明，本文提出的 CP-nets 增量式学习方法对于随机用户都是适用且高效的。



（a）用户 U_1 学习 CP-nets 确定运行时间变化

（a）User U_1 learning CP-nets runtime comparison



（b）用户 U_2 学习 CP-nets 确定运行时间对比

（b）User U_2 learning CP-nets runtime comparison

图 9　用户 U_1，U_2学习 CP-nets 确定运行时间对比图

Fig. 9　User U_1, U_2 learning CP-nets determine runtime comparison

6 结束语

本文主要设计了一种从偏好数据流中学习 CP-nets 的增量方法。通过在合成数据和真实数据上进行的一系列广泛的实验表明,本文提出的算法可以用于处理大数据集,并且可以输出一个表示对应用户属性偏好依赖关系的无环 CP-nets。对于不同时间段的偏好信息,也可以根据增量数据求出特定时刻用户的 CP-nets,有效利用偏好信息的即时性价值。同时,实验表明本文的算法相较于其它算法能够得到一个大致相同的 CP-nets,并且该算法可以有效地节省存储空间和减少运行时间[22]。

在今后的学习研究中主要有 2 部分工作。一方面是,提高 CP-nets 模型性能的学习方法是进一步研究的主要方向。并计划研究其他分类的 CP-nets 学习方法。希望能够进一步更准确高效地学习 CP-nets 模型。后续还考虑将 CP-nets 模型与模糊集学习相结合,研究模糊条件偏好网络的可行性[23]。另一方面是计划学习比 CP-nets 更具有拓展性和研究性的 TCP-nets[24] 和 PCP-nets[25]。

参考文献

[1] KRAWCZYK B, MINKU L L, GAMA J, et al. Ensemble learning for datastream analysis: A survey [J]. Information Fusion, 2017, 37:132.

[2] CHOMICKI J. Peference formulas in relational queries[J]. ACM Transaction on Database Systems, 2003, 28(4): 427.

[3] PAPINI J, AMO S D, SOARES A K S. Strategies for mining user preferences in a data stream setting[J]. Journal of Information & Data Management, 2013, 5(1): 64.

[4] JIANG T, TUZHILIN A. Dynamic micro-targeting: Fitness-based approach to predicting individual preferences [J]. Knowledge and Information Systems, 2009, 19(3): 337.

[5] CHEVALEYRE Y, KORICHE F, LANG J, et al. Learning ordinal preferences on multiattribute domains: The case of CP-nets[M]// FÜRNKRANZ J, HÜLLERMEIER E. Preference learning. Berlin/Heidelberg:Springer,2010: 273.

[6] LIU W, WU C, FENG B, et al. Conditional preference in recommender systems [J]. Expert Systems with Applications, 2014, 42(2): 774.

[7] KRAWCZYK B, MINKU L L, GAMA J, et al. Ensemble learning for data stream analysis: A survey [J]. Information Fusion, 2017, 37: 132.

[8] KORICHE F, ZANUTTINI B. Learning conditional preference networks[J]. Artificial Intelligence, 2010, 174(11): 685.

[9] RIBEIRO M R, BARIONI M C N, AMO S. Temporal conditional preference queries on streams [M] //BENSLIMANE D, DAMIANI E, GROSKY W, et al. Database and expert systems applications. DEXA 2017. Lecture Notes in Computer Science. Cham:Springer, 2017,10438:143.

[10] PAPINI J A J, AMO S D, SOARES A K S. Strategies for mining user preferences in a data stream setting [J]. Journal of Information and Data Management, 2013, 1(1): 1.

[11] PAPINI J A J, AMO S D, SOARES A K S. FPSmining: A fast algorithm for mining user preferences in data streams[J]. Journal of Information and Data Management, 2014, 5(1): 4.

[12] KORICHE F, ZANUTTINI B. Learning conditional preference networks[J]. Artificial Intelligence, 2010, 174(11): 685.

[13] AGGARWAL C C, YU P S. A framework for clustering uncertain data streams[C]//Proc of the 24th International Conference on Data Engineering. Mexico:IEEE, 2008: 150.

[14] LIU Juntao, YAO Zhijun, YI Xiong, et al. Learning conditional preference network from noisy samples using hypothesis testing [J]. Knowledge-Based Systems, 2013, 40: 7.

[15] LIU Juntao, XIONG Yi, WU Caihua, et al. Learning conditional preference networks from inconsistent examples [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 376.

[16] MENGIN J, LANG J. Learning preference relations over combinatorial domains [C]//Proc of the 49th Experimental Nuclear Magnetic Resonance Conference. Pacific Grove: ENC, 2008:207.

[17] BOUTILIER C, BRAFMAN R I, DOMSHLAK C, et al. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements[J]. Journal of Artificial Intelligence Research,2011, 21(1): 135.

[18] GUERIN J T, ALLEN T E, GOLDSMITH J. Learning CP-net preferences online from user queries[M]// PERNY P, PIRLOT M, TSOUKIÀS A. Algorithmic decision theory. Berlin/Heidelberg:Springer-Verlag, 2013:208.

[19] LIU Zhaowei, ZHONG Zhaolin, LI Ke, et al. Structure learning of conditional preference networks based on dependent degree of attributes from preference database[J]. IEEE Access, 2018, 6: 27864.

[20] 刘惊雷. CP-nets 及其表达能力研究[J]. 自动化学报, 2011, 37(3): 290.

[21] KAMISHIMA T. Nantonac collaborative filtering: Recommendation based on order responses [C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington: KDD, 2003: 583.

[22] DIMOPOULOS Y, MICHAEL L, ATHIENITOU F. Ceteris paribus preference elicitation with predictive guarantees[C]//Proc of the 21st International Joint Conference on Artificial Intelligence. Pasadena : IJCAI, 2009: 1890.

[23] ZHAO Xudong, SHI Peng, ZHENG Xiaolong, et al. Fuzzy adaptive control design and discretization for a class of nonlinear uncertain systems[J]. IEEE Transactions on Cybernetics, 2016, 46(6): 1476.

[24] ZHANG Shu, MOUHOUB M, SADAOUI S. Integrating TCP-Nets and CSPs: The constrained TCP-Net (CTCP-Net) model [M]// ALI M, KWON Y, LEE C H,et al. Current approaches in applied artificial intelligence. IEA/AIE 2015. Lecture Notes in Computer Science. Cham:Springer, 2015,9101: 201.

[25] BIGOT D, ZANUTTINI B, FARGIER H, et al. Probabilistic conditional preference networks[J]. arXiv preprint arXiv:1309.6817, 2013.