

基于马尔可夫毯的贝叶斯网络结构学习算法

赵建喆¹, 吴辰妮¹, 王兴伟¹, 裴丽亚²

(1. 东北大学 软件学院, 辽宁 沈阳 110169; 2. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 贝叶斯网络图结构的自动学习是机器学习中的一个挑战, 针对传统算法学习效率低、难于去除冗余边及确定结构中边的方向等问题, 提出了一种基于马尔可夫毯的贝叶斯网络结构学习算法. 该算法改进了经典的马尔可夫毯学习算法, 使之减少条件独立检验次数, 并在后续确定有向结构方面更适应贝叶斯网络结构学习, 同时给出了两种有向边方向确定的一般性解决方案, 有效提高了学习算法的学习效率. 最后建立了基于贝叶斯网络的互联云 QoE 评价模型, 并进行了仿真实验, 结果表明改进后的学习算法在预测准确率、学习效率上均优于传统算法.

关 键 词: 贝叶斯网络; 结构学习; 马尔可夫毯; 互联云; QoE 评价

中图分类号: TP 18 **文献标志码:** A **文章编号:** 1005-3026(2020)04-0464-07

Structure Learning Algorithm of Bayesian Networks Based on Markov Blanket

ZHAO Jian-zhe¹, WU Chen-ni¹, WANG Xing-wei¹, PEI Li-ya²

(1. School of Software, Northeastern University, Shenyang 110169, China; 2. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: ZHAO Jian-zhe, E-mail: zhaojz@swc.neu.edu.cn)

Abstract: The automatic learning of Bayesian network graph structure is a challenge in machine learning. Aiming at the problems of low learning efficiency of traditional algorithm, difficulty in removing redundant edges and determining the direction of the edges in the structure, a Bayesian network structure learning algorithm based on Markov blanket was proposed. The proposed algorithm improves the classical Markov blanket learning algorithm, reduces the number of conditional independent inspections, and is more suitable for Bayesian network structure learning in the subsequent determination of directed structures. At the same time, a general solution for determining the direction of two directed edges was given, which effectively improves the learning efficiency of the learning algorithm. Finally, the Bayesian network-based interconnected cloud QoE evaluation model was established, and the simulation experiment was carried out. The results showed that the improved learning algorithm is superior to the traditional algorithm in prediction accuracy and learning efficiency.

Key words: Bayesian networks; structure learning; Markov blanket; intercloud; QoE evaluation

目前, 贝叶斯网络作为数据处理和分析工具被广泛应用, 特别是用于解决不确定性环境下的决策问题^[1]. 在大数据环境下, 如何提高复杂贝叶斯网络学习和推理的时间效率是贝叶斯网络在新的应用环境下的挑战.

目前, 贝叶斯网络结构学习存在大量研究, 其中具有代表性的算法分为基于评分搜索的算

法^[2-4]和基于依赖分析的算法^[5-7]. 基于评分搜索的方法以一个评分函数作为评定准则, 利用搜索算法在网络结构空间中寻找与训练数据最为匹配的最优网络结构. 但是目前已证明寻找最优网络结构是一个 NP 难问题^[8]. 基于依赖分析的方法是将贝叶斯网络视为表示独立节点间关系的图模型, 通过对其进行条件依赖及独立性检验, 获取

对这些依赖和独立性关系给出最好解释的某个图模型的马尔可夫等价类,经典算法包括 SGS 算法、PC 算法^[9]等。

一个节点的马尔可夫毯 (Markov blanket, MB) 是指图形模型中除去该节点的剩余节点中与之有信息传递的所有节点变量,即该节点所有的父子节点和配偶节点。通过发现图结构中所有节点的马尔可夫毯可以高效地构建贝叶斯网络无向依赖图。现有的马尔可夫毯发现算法分为基于全局条件独立信息的发现算法和基于局部条件独立信息的发现算法。全局算法中候选的马尔可夫毯节点个数往往较多,由于搜索策略的局限性,数据效率普遍不高。基于局部条件独立信息的发现算法是通过检查单个节点 $X \notin \text{MB}$ 与目标节点的条件独立性来搜寻目标节点的马尔可夫毯,而非直接对目标节点的马尔可夫毯进行寻找。这种方法所需的条件集合个数通常比节点的马尔可夫毯集合个数少,因此学习效率相对较高。典型算法有 PCMB (parents and children based Markov boundary) 算法^[10] 和 IPC - MB (iterative parent-child based search of Markov blanket) 算法^[11] 等。

本文提出一种基于马尔可夫毯的贝叶斯网络结构学习算法 (an improved structure learning algorithm of Bayesian network based on Markov blanket, IBN - MB)。首先,对经典马尔可夫毯发现算法 (IPC - MB) 进行改进,提出一种改进的马尔可夫毯学习算法,该算法能够高效、准确地发现节点的父子节点与配偶节点,减少条件独立检验次数,迅速建立贝叶斯网络结构的无向图。然后,结合贝叶斯网络中 D - 分割特性和基于马尔可夫毯发现算法得到的局部拓扑结构即 V 结构,以及 BIC 评分函数,快速确定无向图中边的方向,得到最优贝叶斯网络结构。最后,以互联云的 QoE 评价为仿真实验背景,建立模型并进行对比与分析。实验结果表明,本文提出的 IBN - MB 算法能够高效确定贝叶斯网络结构,排除网络结构的马尔可夫等价类,在机器学习效率和预测精度方面均优于传统方法。

1 IBN - MB 算法的设计

本文提出的基于马尔可夫毯的贝叶斯网络结构学习算法分为三部分,即改进的马尔可夫毯发现算法、基于拓扑信息确定有向边的方向和基于 BIC 评分函数确定有向边的方向。

1.1 改进的马尔可夫毯发现算法

本文对已有的 IPC-MB 算法进行如下改进,并提出了适合于贝叶斯网络结构学习的马尔可夫毯发现算法。

1) IPC-MB 算法通过调用 GenPC 算法发现单个节点的马尔可夫毯,但基于马尔可夫毯发现算法进行贝叶斯网络结构的学习时,需要发现所有节点的马尔可夫毯。为防止对相同节点重复调用 GenPC 算法,本文算法对所有节点调用 GenPC 算法来发现对应父子节点集合,并将发现结果存储下来,以方便后续配偶节点的发现。

2) 因为配偶节点只出现在贝叶斯网络结构 D - 分割中的汇连情况中,目标节点 T 的 V 结构如图 1 所示。为了后续节点快速发现马尔可夫毯,本文算法对已发现节点的拓扑信息即 V 结构进行存储,就可以从目标节点的父子集合中删除部分已知父节点,从而更有效率地发现配偶节点。

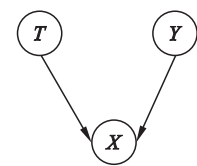


图 1 V 结构
Fig. 1 V structure

3) 本文算法利用条件独立测试发现目标节点 T 的配偶节点,寻找并存储的 V 结构经过了多次条件独立性测试,该结构是可信的、不变的,因此存储的 V 结构能够在后续算法中高效确定边的方向。

4) IPC-MB 算法的复杂度主要体现在 GenPC 算法的两层 FOR 循环上,本文算法分别对两层循环进行优化,提高算法效率。

原算法外层循环是从候选集合 $\text{ADJ}(T)$ 中随机选择一个节点,在目标节点 T 与该节点间执行条件独立性测试,若两节点条件独立,则将该节点从 $\text{ADJ}(T)$ 中删除。改进算法根据与目标节点的依赖程度对候选节点集合中的节点进行排序,首先对与目标节点依赖程度低的节点进行条件独立性测试,这样能更快地从 $\text{ADJ}(T)$ 中删除与目标节点条件独立的节点。依赖程度大小是通过 G^2 测试中的检验统计量进行衡量的,具体使用函数 p 的值,如式(1)所示。 p 的值越大表明节点 X 与目标节点条件独立的可能性越大。每次外层循环时,选择当前候选节点集合中 p 最大的节点与目标节点进行条件独立性测试。

$$p = \text{Ind}(X, T | S) . \tag{1}$$

5) IPC-MB 算法的内层循环为从候选节点集合中产生条件集 S , 在该条件集下执行条件独立性测试, 判断是否将节点 X 加入目标节点的父子节点集合中, 该方法不加区分地对待候选节点集合中的所有节点. 本文研究发现, 如果一个节点在之前的条件独立性测试中更多地参与条件集并有效分割路径, 那么在之后的测试中, 该节点被选中进入条件集的可能性越大, 因此本文提出如下定义作为节点被选为条件集的衡量标准. 改进算法对给定变量集合 S 排序, 以便尽量选择可以有效分割目标节点与候选节点的条件集 S .

定义 1 如果集合 S 为节点 X 曾经参与过的条件集, 且该条件集曾有效分割路径, 则称为有效条件集, $|S|$ 为集合的大小. 节点 X 被选进条件集的概率为

$$\text{Freq}(X) = \sum_s \frac{1}{|S|} . \tag{2}$$

对两节点执行条件独立性测试时, 条件集的大小不止为 1, 因此需考虑一个集合被选作条件集的概率.

定义 2 一个集合 S 被选作条件集的概率等于集合 S 中所有节点 $X(X \in S)$ 被选入条件集的概率之和, 集合 S 被选作条件集的概率为

$$\text{Freq}(S) = \sum_{X \in S} \text{Freq}(X) . \tag{3}$$

6) IPC-MB 算法得到的目标节点 T 的父子节点集合可能包含假正节点.

定义 3 如果在候选节点集合 $\text{ADJ}(T)$ 的任何子集下, 两个节点的条件独立性测试均显示两节点是条件依赖的, 那么这两个节点是邻接的, 且两个节点互为对方的父子节点集合中一员, 则该节点称为对方节点的假正节点(false positive).

如图 2 所示, 若求目标节点 T 的父子节点集合, 当条件集的大小为 0 时, 节点 B 与 T 独立, 此时从 $\text{ADJ}(T)$ 中将节点 B 删除, 但节点 T, C 在条件集合 $\{A, B\}$ 下条件独立, 因此过早删除节点 B 将会导致在后续的条件独立性测试中, 无法测试出节点 T 与 C 条件独立, 导致节点 C 在 GenPC 过程结束后依然存在于节点 T 的父子节点集合中.

为解决上述问题, 改进算法使用两个节点间的对称关系, 即两个节点相互在对方节点的父子节点集合中删除多余的假正节点. 在图 2 中, 节点 T 经过 GenPC 算法后得到父子节点集合 $\text{CPC}(T) = \{A, C\}$, 对 C 调用 GenPC 过程得到父子节点集合 $\text{CPC}(C) = \{A, B\}$, $C \in \text{CPC}(T)$ 而

$T \notin \text{CPC}(C)$, 不满足节点间的对称关系, 因此 C 是 T 的假正节点, 可将节点 C 从 T 的父子节点集合中删除.

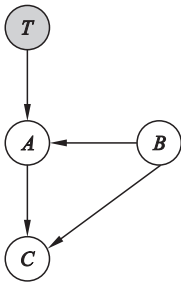


图 2 对称原则示例

Fig. 2 Example of symmetry principle

基于以上改进, 本文提出了改进的马尔可夫毯发现算法(算法 1), 该算法通过调用改进的 GenPC 算法(算法 2)和 GenSET 算法(算法 3)完成了贝叶斯网络所有节点的马尔可夫毯的发现并存储了局部拓扑信息即 V 结构用于后续边的方向确定.

算法 1 改进的马尔可夫毯发现算法

```
输入: 数据集  $D$ 、贝叶斯网络节点集合  $U$ 、阈值  $\alpha$ 
输出: 所有节点的马尔可夫毯集合 MB
1. FOR  $T \in U$ 
2.   设置  $\text{CanADJ}(T)$  为网络中除节点  $T$  外所有节点;
3.   初始化节点  $T$  的父子节点集合  $\text{CPC}(T)$  为 NULL;
4.   初始化节点  $T$  的父节点集合  $\text{Pa}(T)$  与配偶节点集合  $\text{Sp}(T)$  为 NULL;
5.   令  $\text{CPC}(T) = \text{GenPC}(T, \text{CanADJ}(T), D, \alpha)$  存储结果;
6. END FOR
7. FOR  $T \in U$ 
8.   设置节点  $T$  的马尔可夫毯  $\text{MB}(T) = \text{CPC}(T)$ ;
9.   FOR  $X \in (\text{CPC}(T) - \text{Pa}(T))$ 
10.    设置  $\text{CanSP} = \text{CPC}(X)$ ;
11.    IF  $T \notin \text{CanSP}$ 
12.       $\text{MB}(T) = \text{MB}(T) - \{X\}$ ;
13.    END IF
14.    FOR  $Y \in \text{CanSP}$  且  $Y \notin \text{MB}(T)$ 
15.      IF  $\text{Ind}(T, Y | \text{Sepset}_{T,Y} \cup \{X\}) > \alpha$ 
16.         $\text{MB}(T) = \text{MB}(T) \cup \{Y\}$ ;
17.         $\text{Pa}(X) = \text{Pa}(X) \cup \{T, Y\}$ ;
18.         $\text{Sp}(T) = \text{Sp}(T) \cup \{Y\}$ ;
19.         $\text{Sp}(Y) = \text{Sp}(Y) \cup \{T\}$ ;
20.      END IF
21.    END FOR
22.  END FOR
23. END FOR
24. RETURN MB
```

其中, 第 1~6 行是对所有节点调用改进的 GenPC 算法发现所有节点的父子节点集合, 通过第 2 行将 $\text{CanADJ}(T)$ 设置为网络结构中除节点

T 外所有的节点. 第 7 ~ 23 行发现所有节点的马尔可夫毯集合. 其中, 第 8 行将节点 T 的马尔可夫毯初始化为其父子节点集合. 第 10 行中 CanSP 为节点 X 的父子节点集合, 表示目标节点将从其子节点的父子节点集合中寻找配偶节点. 第 11 ~ 13 行采用了对称原则. 第 14 ~ 21 行通过条件独立性测试来寻找目标节点的配偶节点, 这里检验条件独立性使用的测试方法为 G^2 测试. 第 16 ~ 19 行存储在目标节点 T 的马尔可夫毯发现过程中得到 V 结构.

改进的 GenPC 算法发现父子节点的基本原理如下: 在一个可信网络理论中, 如果节点 X, Y 满足对于所有的 $Z (Z \subseteq U - \{X, Y\})$, 均不存在 $X \perp Y | Z$, 那么节点 X, Y 在贝叶斯网络中是邻接的. 因此, 只要证明对任一集合 $Z (Z \subseteq U - \{X, Y\})$ 的条件下, 存在 $X \perp Y | Z$, 那么节点 X, Y 是不邻接的, 它们之间不存在边. 在候选节点集合的所有子集条件下, 检验节点 T 与候选节点 X 间的条件独立性, 一旦存在一个条件集 S , 使得 $X \perp T | S$, 则将候选节点 X 从 $CPC(T)$ 中删除.

算法 2 改进的 GenPC 算法

输入: 目标节点 T 、节点 T 的相邻节点候选集合 $ADJ(T)$ 、数据集 D 、阈值 α
输出: 节点 T 的父子节点集合 $CPC(T)$

1. 设置条件集大小 $setsize = 0$;
2. WHILE $|ADJ(T)| > setsize$
3. 基于 $Ind(T, Y)$ 对 $ADJ(T)$ 中的节点进行排序;
4. FOR $X \in ADJ(T)$
5. 利用 GenSet 算法得到条件集构成的集合
 $SET = GenSet(ADJ(T) - \{X\}, setsize)$;
6. FOR $S \in SET$
7. IF $Ind(X, T | S) \leq \alpha$ THEN
8. 从 $ADJ(T)$ 中将节点 X 删除;
9. 存储条件集 $Sepset_{T,X} = S$;
10. FOR $Y \in S$
11. 更新 $Freq(Y)$;
12. END FOR
13. BREAK
14. END IF
15. END FOR
16. END WHILE
17. $setsize = setsize + 1$;
18. END WHILE
19. $CPC(T) = ADJ(T)$;
20. RETURN $CPC(T)$;

其中, 第 7 ~ 14 行进行目标节点与候选节点的条件独立性测试, 若测试通过, 则在第 8 行将候选节点从候选集合中删除; 第 9 行存储两节点条件独立的条件集; 第 10 ~ 12 行更新条件集中所有

节点被选入条件集的概率, 第 19 行将 $ADJ(T)$ 赋值给父子节点集合 $CPC(T)$; 第 20 行返回目标节点的父子节点集合.

GenSET 算法的主要目的是返回排序后的、给定大小的条件集.

算法 3 GenSET 算法

输入: 待排序节点集合 V 、产生的集合大小 $setsize$
输出: 排序产生的条件集集合 SET

1. 从集合 V 中生成大小为 $setsize$ 的子集放入集合 SET 中
2. FOR $S \in SET$
3. 计算 S 被选作条件集的概率 $Freq(S)$
4. END FOR
5. 根据 $Freq(S)$ 对集合 SET 中的元素进行排序
6. RETURN SET ;

1.2 基于 V 结构确定边的方向

本文提出的 IBN-MB 算法设计两种贝叶斯网络结构中边的方向的确定方法. 其中, 基于 V 结构的边的方向确定方法具体过程分为以下两个步骤: 一是直接根据马尔可夫毯发现算法中得到并存储的 V 结构确定边的方向; 二是根据 V 结构进行一定推理来确定边的方向.

本文提出的改进马尔可夫毯发现算法中采取的是后向策略, 即删除已确认的不邻接节点, 因此已发现的目标节点的父子节点集合必然包含最终网络结构中所有节点的父子节点集合, 根据父子节点集合进一步推导出的 V 结构也必定包含最终网络结构中所有的 V 结构. 由此可知, 在下一步确定边的方向时不会再产生 V 结构. 根据 D - 分割理论, 贝叶斯网络结构中 3 个节点之间只存在顺连、分连和汇连三种结构. 而汇连结构就是 V 结构, 由于网络结构中不会再产生该结构, 因此在确定边的方向时只可能产生顺连和分连两种结构, 通过以上理论进行推理, 确定剩余边的方向.

1.3 基于 BIC 评分函数确定边的方向

基于本文算法中得到的 V 结构基本可以确定贝叶斯网络结构中边的方向, 但是也有特殊情况, 即不能通过 V 结构确定所有边的方向, IBN - MB 算法给出了基于 BIC 评分函数的解决方案.

BIC 评分函数是在样本满足独立同分布假设的前提下, 用对数似然度量结构与数据之间的拟合程度. 具体评分函数式为

$$BIC(G|D) = \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \lg \theta_{ijk} - \frac{1}{2} \sum_{i=1}^n q_i (r_i - 1) \lg m. \tag{4}$$

式中: G 是由 n 个节点构成的贝叶斯网络结构; D 为给定的数据集, 样本集数目为 m ; r_i 为第 i 个节

点 X_i 的取值空间大小, X_i 的父节点集合 $\pi(X_i)$ 中所有可能的取值组合个数 $q_i = \prod_{X_j \in \pi(X_i)} r_j$; m_{ijk} 为 X_i 的父节点取第 j 个节点($j=1,2,\dots,q_i$), X_i 取第 k 个值($k=1,2,\dots,k_i$)时的样本个数; $\theta_{ijk} = m_{ijk}/m_{ij}$ 为似然条件概率,且有 $0 \leq \theta_{ijk} \leq 1$, $\sum_k \theta_{ijk} = 1$.

利用 BIC 评分函数对剩余边的方向进行确认时,需要将评分函数进行分解,使其变为局部结构的评分函数:

$$\text{BIC}(\langle X_i, \pi(X_i) \rangle | D) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \lg \theta_{ijk} - \frac{1}{2} q_i (r_i - 1) \lg m. \quad (5)$$

式中:第 1 项为度量结构与数据之间的拟合程度;第 2 项为对模型复杂度的罚项,防止过度拟合情况的出现.

当节点 X 与 Y 之间存在一条未确定方向的边时,首先假定这条边为 $X \rightarrow Y$,即 Y 为 X 的父节点,这种情况下局部结构的 BIC 评分为 S_1 ,如式(6)所示:

$$S_1 = \text{BIC}(\langle X, \pi(X) \cup \{Y\} \rangle | D) + \text{BIC}(\langle Y, \pi(Y) \rangle | D). \quad (6)$$

然后假设这条边为 $X \leftarrow Y$,即 X 为 Y 的父节点,这种情况下局部结构的 BIC 评分 S_2 为

$$S_2 = \text{BIC}(\langle X, \pi(X) \rangle | D) + \text{BIC}(\langle Y, \pi(Y) \cup \{X\} \rangle | D). \quad (7)$$

比较 S_1 与 S_2 的大小,若 $S_1 < S_2$,说明 S_1 代表的网络结构即为最优结构,即有向边方向为 $X \rightarrow Y$;反之,则 $X \leftarrow Y$.

贝叶斯网络为有向无环图,但得到的贝叶斯网络结构可能有环路存在,需要进行解环操作.具体方案为:若环与 V 结构并存,首先保证涉及 V 结构的边的方向不作改变,之后将其余边中任意一条边的方向改变,达到解环的效果;若两者不并存,则直接将三条边中任意一条边的方向改变解环.

2 仿真实验与结果分析

2.1 仿真实验设计

本文构建基于贝叶斯网络的互联云 QoE 评价模型作为仿真实验的背景,验证所提出算法的准确率和效率.实验设计了互联云应用测试环境,开发了视频播放应用,实验中邀请每位用户观看在不同网络传输状况下的 9 段视频,并分别对这些视频给出 QoE 评分,汇总并存储为 QoE 评价数据集.然后随机分为训练样本和测试样本,使用 K-折交叉检验算法进行训练,其中 $K=10$,基于

本文提出的 IBN-MB 算法训练得到的贝叶斯网络互联云 QoE 评价模型如图 3 所示.模型中的节点分别代表 QoE 及其部分影响因素; X_1 表示视频类型, X_2 表示用户性别, X_3 表示网络延迟, X_4 表示网络带宽, X_5 表示视频码率, X_6 表示用户清晰度感受, X_7 表示视频分辨率, X_8 表示用户响应速度感受, X_9 表示用户流畅度感受, Y 表示用户给出的 QoE 评价.

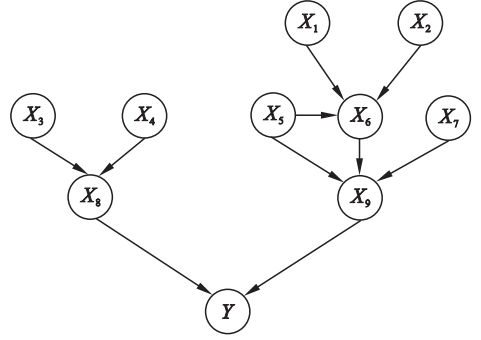


图 3 QoE 评价模型贝叶斯网络结构图

Fig. 3 Structure of QoE evaluation model based on BN

2.2 性能评价

使用决策树 C4.5 算法与 IBN-MB 算法进行性能对比与分析.为了反映本文设计的互联云 QoE 评价模型的优劣,将通过以下几个指标进行性能评价.

1) F -measure: F -measure 综合了预测准确率与召回率两方面指标,其中预测准确率见性能指标 2 的定义.召回率 Recall 定义为正确预测 QoE 值的个数占应被成功预测的 QoE 值个数的比例.其中, BNF_i 为属于第 i 类但未被正确识别的测试样例的个数.

$$\text{Recall} = \sum_{i=1}^M \text{recall}_i, \quad (8)$$

$$\text{recall}_i = \frac{\text{BNC}_i}{\text{BNC}_i + \text{BNF}_i} \times 100\%. \quad (9)$$

F -measure 综合了预测准确率与召回率两方面指标,定义为

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (10)$$

2) 预测准确率:预测准确率是指从给定的 QoE 评价指标中正确预测 QoE 值的个数占所有个数的百分比.预测准确率计算公式为

$$\text{Precision} = \frac{\sum_{i=1}^M \text{precision}_i}{M}, \quad (11)$$

$$\text{precision}_i = \frac{\text{BNC}_i}{\text{BNI}_i} \times 100\%. \quad (12)$$

式中: BNC_i 为正确识别的第 i 类测试样例的个数; BNI_i 为属于第 i 类的测试样例的个数; M 为

QoE 取值空间个数。

3) 时间开销:时间开销是衡量人工智能模型性能的重要指标。本文利用对条件独立性测试 (conditional independence test, CI) 的调用次数作为衡量模型时间开销性能的具体方法。

通过以上 3 个性能指标,对基于 IBN-MB 算法与 C4.5 算法的评价模型进行了仿真实验。实验结果分析如下:

1) 对模型的 F -measure 指标进行比较,令 k 为 10,结果如图 4 所示:随着样本数的增加,基于两种算法的评价模型的 F -measure 值均呈现平缓上升的趋势,样本数与预测准确率近似呈线性关系。基于 IBN-MB 算法的 F -measure 值略高于基于 C4.5 算法的评价模型。

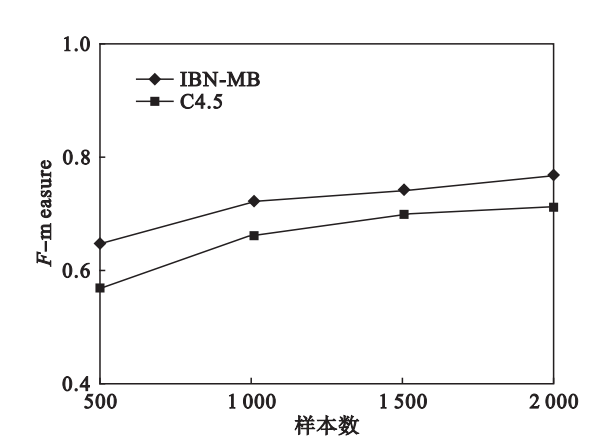


图 4 不同 IBN 与 C4.5 算法时样本数对 F -measure 的影响
Fig. 4 Effect of sample number on F -measure under different IBN and C4.5

2) 对基于 IBN-MB 算法与 C4.5 算法训练的模型预测准确率进行对比,结果如图 5 所示。

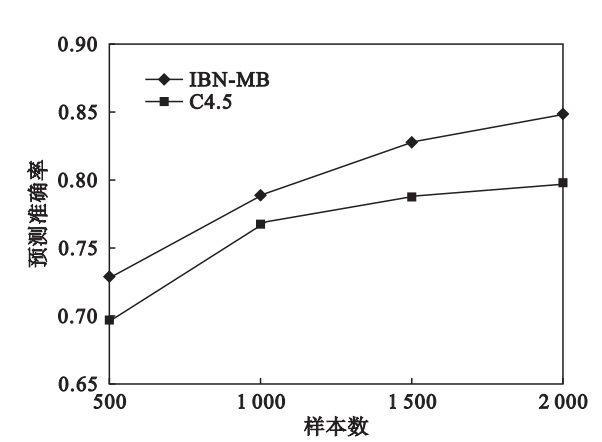


图 5 不同 IBN 与 C4.5 算法时样本数对预测准确率的影响
Fig. 5 Effect of sample number on prediction accuracy under different IBN and C4.5

算法的评价模型预测准确率均呈现稳步上升的趋势,且上升趋势趋于稳定。基于 IBN-MB 算法的评价模型预测准确率略高于基于 C4.5 算法的评价模型。

3) 分别令 k 取 3,5 和 10 时,对基于 IBN-MB 算法的评价模型预测准确率进行验证,结果如图 6 所示。可知,随着样本数的增加,基于 IBN-MB 算法的 QoE 评价模型预测准确率呈逐步上升的趋势。当 k 取 10 时,评价模型的预测准确率基本最高,当 k 取 3 时,模型的预测准确率最低。

4) 分别对评价模型的时间开销指标进行比较,基于 IBN-MB 算法的评价模型中调用 CI 测试次数明显低于基于 PC 算法。这是因为 IBN-MB 算法在构建无向图时,使用了存储网络结构拓扑信息和对候选节点排序等改进策略,因此能够尽早删除与目标节点条件独立的节点,从而降低了 CI 测试次数,减少时间开销。

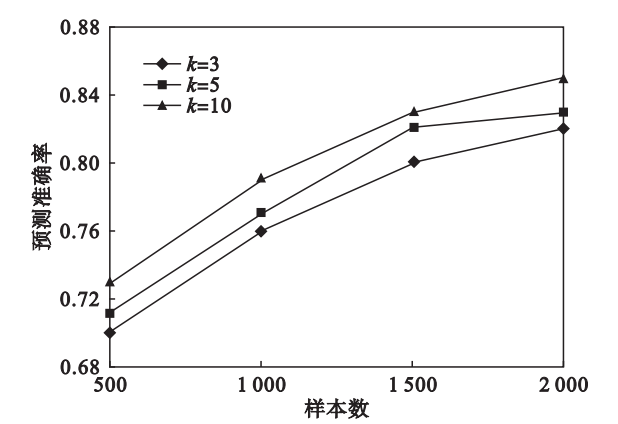


图 6 不同 k 时样本数对预测准确率的影响
Fig. 6 Effect of sample number on prediction accuracy under different k

3 结 论

1) 本文提出一种改进的基于马尔可夫毯的贝叶斯网络结构发现算法,该算法能高效、准确地发现父子节点与配偶节点,确定贝叶斯网络结构无向图。在此基础上高效确定拓扑结构中边的方向。

2) 基于本文算法对互联云 QoE 评价问题进行建模,仿真实验结果表明,新算法在准确率和性能上均优于传统算法。

参考文献:

[1] Pearl J. Probabilistic reasoning in intelligent systems [J]. Computer Science Artificial Intelligence, 1988, 70 (2): 1022-1027.