



基于缺失数据 BN 参数学习的电信流失客户预测算法

赵宇翔, 卢光跃, 王航龙, 李四维

(西安邮电大学陕西省信息通信网络及安全重点实验室, 陕西 西安 710121)

摘要: 针对电信客户流失预测问题, 在数据缺失情况下, 基于贝叶斯网络 (Bayesian network, BN), 用最近邻算法填补缺失数据, 并将两类定性约束融入贝叶斯网络参数学习过程, 用以提高流失客户预测精度。仿真及实际数据分析结果表明, 所提算法较经典的期望最大化 (expectation maximization, EM) 算法有明显优势, 在牺牲代价较小的忠诚客户预测精度的情况下, 得到了更高的流失客户预测精度。

关键词: 贝叶斯网络; 参数学习; 数据缺失; 最近邻算法; 定性约束

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018018

A prediction algorithm of telecom customer churn based on Bayesian network parameters learning under incomplete data

ZHAO Yuxiang, LU Guangyue, WANG Hanglong, LI Siwei

Shaanxi Key Laboratory of Information Communication Network and Security,
Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract: Aiming at prediction of telecom customer churn, a novel method was proposed to increase the prediction accuracy with the missing data based on the Bayesian network. This method used k -nearest neighbor algorithm to fill the missing data and adds two types of monotonic influence constraints into the process of learning Bayesian network parameter. Simulations and actual data analysis demonstrate that the proposed algorithm obtains higher prediction accuracy of churn customers with the loss of less cost prediction accuracy of loyal customers, outperforms the classic expectation maximization algorithm.

Key words: Bayesian network, parameter learning, data missing, nearest neighbor algorithm, qualitative constraint

1 引言

随着移动网络的快速发展, 手机成为人与人之间沟通的重要方式。与此同时, 未入网的新用户

数量也在急剧下降。据估计, 吸引一个新用户所花费的成本是挽留一个老客户所花费成本的 5~6 倍。所以, 为了获取更大的利润, 电信企业越来越重视潜在流失客户^[1]。企业希望尽早发现

收稿日期: 2017-07-11; 修回日期: 2017-09-26

基金项目: 陕西省工业科技攻关项目 (No.2015GY-013); 陕西省工业科技攻关项目 (No.2016GY-113)

Foundation Items: Industrial Research Project of Science and Technology Department of Shaanxi Province (No.2015GY-013), Industrial Research Project of Science and Technology Department of Shaanxi Province (No.2016GY-113)

这部分人群,及时做出有针对性的调整并成功挽留这部分客户。为了准确预测流失客户,很多学者采用贝叶斯网络(Bayesian network, BN)进行分类。

贝叶斯网络由有向无环图和条件概率表两部分组成^[2],经过大约 30 年的发展,BN 已经广泛应用于人工智能与数据挖掘等热门领域^[3-5]。本文在电信客户流失网络结构已知情况下,研究离散贝叶斯网络参数学习的问题。

传统电信客户流失参数学习采用最大似然估计(maximum likelihood estimation, MLE)方法,但由于电信数据时常伴随有信息缺失,无法直接利用 MLE 方法逼近 BN 真实参数。对缺失数据进行参数估计的经典算法是期望最大化(expectation maximization, EM)算法^[6],但 EM 算法对初始点的选取较为敏感、容易陷入局部最优解,而且在数据量较小时参数学习精度低。目前,很多学者通过引入先验知识来提高含缺失数据的 BN 参数学习精度。例如,参考文献[7]提出将 EM 算法与定性约束相结合,用定性的先验知识构造惩罚函数、用样本数据构造似然函数;然后将惩罚函数与似然函数结合构造出目标函数;最后利用 CEM(constrained EM)算法求解 BN 参数。参考文献[8]则将保序回归算法融入 EM 算法每次迭代的 M 步中,并在此基础上,提出一种复杂度更低的快速保序最大期望(qirEM)算法。参考文献[9]将参数学习问题视为凸优化模型,把最大熵函数与 EM 算法相结合构造凸优化问题。通过分析上述参考文献发现:参考文献[7]方法中惩罚函数随约束类型的变化而改变,并且其权重是人为赋予的,故该方法结果受人为影响;参考文献[8]的方法并不能保证在有限的参数空间中收敛到局部最优解;参考文献[9]方法可行域比较精确,但估计参数大多分布在可行域边缘位置,这样便浪费了不等式约束中所包含的信息。因此,如何在有缺失数据的样本中进行参数学

习,仍然值得探讨。

本文首先用改进的 k 近邻算法解决缺失数据的填充问题,然后在此基础上将先验知识融入 BN 参数学习过程中,用两类定性约束来调整估计参数。实验结果表明:在样本缺失率不同时,与传统的 EM 算法相比,本文算法学习得到精度更高的 BN 参数。

2 相关理论

2.1 贝叶斯网络

BN 是由有向无环图 $G=(U, E)$ 和网络中各节点的条件概率表组成,其中 $U=\{X_1, X_2, \dots, X_n\}$ 表示节点变量集合,而 E 表示节点间的依赖关系。如果有一条弧 $a_{ji} \in E$ 从 X_j 指向 X_i ,则称 X_j 为 X_i 的父节点。一个节点条件概率表包含在给定父节点值时子节点取每个值的概率,记作 $P(X_i|\pi(X_i))$,其中 $\pi(X_i)$ 表示 U 中 X_i 的父节点集合。

贝叶斯网络参数学习指的是已知网络结构,从样本数据中拟合网络中各节点的条件概率。若一个 n 节点贝叶斯网络中节点 X_i 共有 r_i 种可能状态,其父节点集合 $\pi(X_i)$ 共有 q_i 种可能状态。那么,该网络节点参数可表示为:

$$\theta_{ijk} = P(X_i = k | \pi(X_i) = j) \quad (1)$$

$$1 \leq i \leq n, 1 \leq j \leq q_i, 1 \leq k \leq r_i \quad i, j, k \in Z$$

网络中所有 θ_{ijk} 组成的向量记为 θ 。为了找到一组与给定样本数据 $D=(D_1, D_2, \dots, D_m)$ 拟合程度最高的参数,构造关于 θ 的对数似然函数:

$$L(\theta|D) = \log \prod_{l=1}^m p(D_l|\theta) = \sum_{l=1}^m \log p(D_l|\theta) \quad (2)$$

经推导^[10],参数 θ_{ijk} 最大似然估计为:

$$\theta_{ijk}^* = \begin{cases} \frac{N_{ijk}}{\sum_{k=1}^{r_i} N_{ijk}}, & \sum_{k=1}^{r_i} N_{ijk} > 0 \\ \frac{1}{r_i}, & \text{其他} \end{cases} \quad (3)$$



其中, N_{ijk} 是数据中满足 $X_i = k$ 和 $\pi(X_i) = j$ 的样本数量。

本文将网络中的所有参数分为两类: 第一类是父节点组合状态相同而子节点状态不同的参数; 第二类是子节点状态相同而父节点组合状态不同的参数, 如图 1 所示。

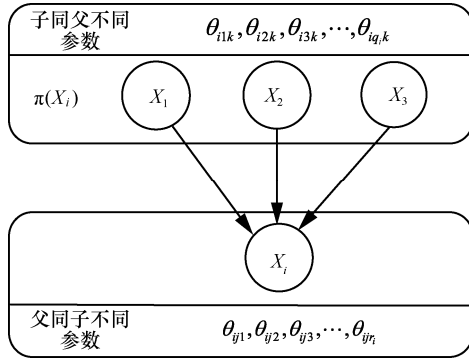


图 1 BN 节点参数分类示意

2.2 定性约束

对贝叶斯网络参数学习时, 其应用领域的专家通常能够提供网络中相关节点之间的定性关系^[11], 即约束信息。例如: 范围约束, 定义了参数的上下界, 如 $b \leq \theta_{ijk} \leq B$; 不等式约束, 定义了一对参数的相对大小关系, 如 $\theta_{ijk} \leq \theta_{i'j'k'}, i \neq i' \text{ 或 } j \neq j' \text{ 或 } k \neq k'$; 近似相等约束, 定义了两个参数的近似相等的关系, 如 $\theta_{ijk} \approx \theta_{i'j'k'}, i \neq i' \text{ 或 } j \neq j' \text{ 或 } k \neq k'$ 。

根据被约束参数父节点的不同状态, 参数约束可以分为内部约束和外部约束^[11]。内部约束是限制父节点组合状态相同而子节点状态不同时参数之间的关系; 外部约束则是限制子节点状态相同而父节点组合状态不同时参数之间的关系。

根据以上定义, 用电信数据中由入网时长 X_T 和是否 VIP 用户 X_V 组成的 BN 为例子 (如图 2 所示), 此时, BN 中的内部约束为: $P(X_V = \text{false} | X_T = \text{long}) < P(X_V = \text{true} | X_T = \text{long})$; 外部约束为: $P(X_V = \text{true} | X_T = \text{long}) > P(X_V = \text{true} | X_T = \text{medium}) > P(X_V = \text{true} | X_T = \text{short})$ 。

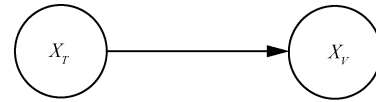


图 2 两节点贝叶斯网络结构

3 含缺失数据的 BN 参数学习算法

在电信数据中, 由于机械或人为原因会造成部分属性缺失, 若直接用这类数据集进行 BN 参数学习将难以保证参数的精度。为此, 本节首先利用最近邻 (k -nearest neighbor, k NN) 算法填补缺失数据; 接着用贝叶斯 MAP 思想进行参数的内部约束学习; 最后用保序回归模型^[12]进行参数的外部约束学习。

3.1 缺失数据填充

k NN 算法起初是用于解决分类问题, 它为了找到与目标样本相似度最高的 k 个样本, 也就是距离目标样本最近的 k 个样本, 因此距离的表示形式很重要。当属性之间具有联系或量纲不同时, 马氏距离比欧氏距离更适合度量邻近性, 所以本文选取马氏距离作为 k NN 算法中的距离度量形式^[13]。

两个样本 D_1 和 D_2 之间的马氏距离定义为:

$$d(D_1, D_2) = \sqrt{(D_1 - D_2)^T \Sigma^{-1} (D_1 - D_2)} \quad (4)$$

其中, Σ^{-1} 是数据协方差矩阵的逆, 协方差矩阵的第 i 行第 j 列元素是第 i 个属性和第 j 个属性的协方差, 定义为:

$$\sum_{ij} = \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \quad (5)$$

下面给出 k NN 算法填补缺失数据步骤。

步骤 1 数据预处理, 将初始数据分为完整数据与缺失数据两部分, 同时把缺失数据作为目标数据。

步骤 2 根据式 (4) 计算目标数据与完整数据中每个样本的马氏距离 d 。

步骤 3 对所有 d 进行排序, 选取其中 k 个马氏距离最小的样本作为目标数据的 k 近邻。

步骤 4 取 k 个近邻中对应缺失属性出现最

多的值来填补缺失处。

步骤 5 重复步骤 2~4, 直到所有数据完整。

3.2 内部约束学习

如果网络中父节点组合状态相同而子节点状态不同的参数呈单调性, 即满足: $\theta_{ij1} < \theta_{ij2} < \dots < \theta_{ijk} < \theta_{ijr_i}$ 。这时, 由参数的单调性和归一性可得:

$$\begin{cases} \theta_{ij1} + \theta_{ij2} + \dots + \theta_{ijr_i} = 1 \\ \theta_{ij1} < \theta_{ij2} < \dots < \theta_{ijr_i} \end{cases} \Rightarrow \begin{cases} 0 < \theta_{ij1} < \frac{1}{r_i} \\ \theta_{ij1} < \theta_{ij2} < \frac{1 - \theta_{ij1}}{r_i - 1} \\ \vdots \\ \theta_{ijr_i} = 1 - \sum_{k=1}^{r_i-1} \theta_{ijk} \end{cases} \quad (6)$$

假定这些参数在区间范围内服从均匀分布, 即参数 $\theta_{ijk} \sim U(\theta_1, \theta_2)$, 则:

$$\begin{cases} E(\theta_{ijk}) = \frac{\theta_1 + \theta_2}{2} \\ D(\theta_{ijk}) = \frac{(\theta_1 - \theta_2)^2}{12} \end{cases} \quad (7)$$

其中, 一般假设每个参数 θ_{ijk} 的先验分布服从 Beta 分布^[10]。为了将定性的参数范围约束有效转化为定量的数学计算中, 用 Beta 分布 $B(\alpha_1, \alpha_2)$ 来逼近均匀分布, 以获取超参数 α_1 、 α_2 的取值。即:

$$\min_{\alpha_1, \alpha_2} \int_0^1 (U(\theta_1, \theta_2) - B(\alpha_1, \alpha_2))^2 d\theta \quad (8)$$

采用二阶矩法拟合, 计算过程如下:

$$\begin{aligned} m_1 &= \int_0^1 U(\theta) \theta d\theta = E_U(\theta) \\ m_2 &= \int_0^1 U(\theta) \theta^2 d\theta = D_U(\theta) + E_U^2(\theta) \end{aligned} \quad (9)$$

$$\begin{aligned} M_1 &= \int_0^1 B(\theta) \theta d\theta = \frac{\alpha_1}{\alpha_1 + \alpha_2} \\ M_2 &= \int_0^1 B(\theta) \theta^2 d\theta = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)} + M_1^2 \end{aligned} \quad (10)$$

$$\begin{cases} M_1 = m_1 \\ M_2 = m_2 \end{cases} \Rightarrow \begin{cases} \alpha_1 = \frac{m_1^2 - m_1 m_2}{m_2 - m_1^2} \\ \alpha_2 = \frac{m_1 - m_2}{m_2 - m_1^2} (1 - m_1) \end{cases} \quad (11)$$

通过式(8)~式(11)可以求得超参数 α_1 和 α_2 , 至此完成了将参数先验信息转化为虚拟样本数据的工作。然后将其代入:

$$\theta_{ijk}^* = \frac{N_{ijk} + \alpha_1}{\sum_{k=1}^{r_i} N_{ijk} + \alpha_1 + \alpha_2} \quad (12)$$

获得修正后的参数。内部约束学习方法步骤如下。

步骤 1 根据先验知识判断是否内部约束满足单调性, 如果满足单调性, 则转入步骤 2。

步骤 2 根据依据式(6)求出各个参数取值的上下界, 然后用 Monte Carlo 法基于均匀分布进行抽样。

步骤 3 依据式(8)用 Beta 分布拟合有效先验参数并得到超参数 α_1 和 α_2 。

步骤 4 利用式(12)得到校正后的参数。

3.3 外部约束学习

当明确父节点对子节点的定性影响时, 即知道网络中不同父节点组合状态值优先顺序满足:

$$X_{\pi(i)}^1 \preceq X_{\pi(i)}^2 \preceq X_{\pi(i)}^3 \preceq \dots \preceq X_{\pi(i)}^{q_i} \quad (13)$$

其中, $X_{\pi(i)}^j, j \in [1, q_i]$ 表示节点 i 的父节点组合状态值。由此可以得到参数的相对大小关系:

$$\theta_{i1k} < \theta_{i2k} < \theta_{i3k} < \dots < \theta_{iq_i k} \quad (14)$$

本文用 MLS (minimum lower set) 算法^[12]实现该约束条件下的 BN 参数学习, 首先根据优先关系构造出所有父节点组合状态值的下限集合 $\ell = \{L_1, L_2, L_3, \dots, L_s\}$, 然后根据式(15)计算每个下限集合的加权平均值^[12]:



$$\text{Average}(L_i) = \begin{cases} \frac{\sum_{j \in L_i} N_{ij} \theta_{ijk}}{\sum_{j \in L_i} N_{ij}}, & \sum_{j \in L_i} N_{ij} > 0 \\ 0.5, & \sum_{j \in L_i} N_{ij} = 0 \end{cases} \quad (15)$$

其中, N_{ij} 指父节点组合状态为 j 时出现的频次。

然后令加权平均值最小的下限集合为 B , 即:

$$B = \min_{L_i \in \ell} \text{Average}(L_i) \quad (16)$$

此时令集合 B 中所有参数值都为 $\text{Average}(B)$, 即: $\theta_{ijk} = \text{Average}(B), X_{\pi(i)}^j \in B$ 。

然后删除所有下限集合中与集合 B 相交的部分, 并从节点 i 的全部父节点组合状态中删除集合 B 中的组合状态, 即:

$$L_i = L_i \setminus B \quad (17)$$

$$X_{\pi(i)} = X_{\pi(i)} \setminus B \quad (18)$$

重复该过程直到 $X_{\pi(i)}$ 为空集, 结束计算。外部约束学习方法总结如下。

步骤 1 根据先验知识确定不同父节点组合状态值的优先顺序并构造出所有的下限集合。

步骤 2 依据式 (15) 计算每个下限集合的加权平均值, 然后根据式 (16) 得到集合 B 并令集合 B 中所有参数取值为 $\text{Average}(B)$ 。

步骤 3 依据式 (17) 和式 (18) 进行处理。

步骤 4 判断集合 $X_{\pi(i)}$ 是否为空集, 若是, 结束计算; 若否, 返回步骤 2。

3.4 新算法流程

鉴于新算法用到了最近邻算法以及内、外部约束条件, 本文称之为基于最近邻—双重约束的参数估计 (k -nearest neighbor-dual constrained estimation, k NN-DCE) 方法。总结前文的参数求解过程, 新算法的总流程如下。

步骤 1 将样本数据分为完整数据和缺失数据两部分。

步骤 2 应用最近邻算法找出与缺失数据相似度高的完整数据, 并填补对应的缺失属性。

步骤 3 统计样本, 获得样本统计值并根据

式 (3) 求出初始参数。

步骤 4 用内部约束学习方法修正父节点组合状态相同而子节点状态不同的参数。

步骤 5 用外部约束学习方法校正父节点组合状态不同而子节点组合状态相同的参数。

4 算法仿真

本次仿真试验平台为 MATLAB R2010b, 运行环境为 Windows 7。仿真实验由 2 部分组成:

第 4.1 节从理论层面验证本文算法的优势; 第 4.2 节将该参数学习算法应用到电信客户流失预测中, 用真实数据证实算法的可行性。

4.1 经典模型仿真

采用草坪湿润模型 (如图 3 所示), 该模型已被广泛应用于 BN 参数学习算法的评估^[14], 先验知识见表 1。

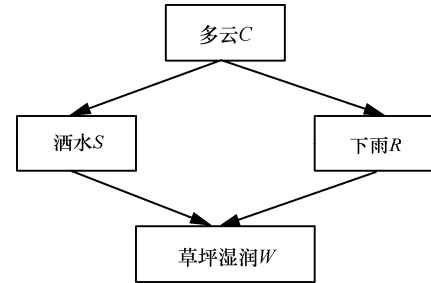


图 3 草坪湿润 BN 模型

首先依原始网络真实参数随机生成一个数据集; 然后使完整样本数据随机缺失, 分别构造样本缺失率为 20%、30%、40% 3 种情况; 最后采用 KL 散度^[15]衡量算法准确性。KL 散度值越小, 表明估计参数与真实参数越接近, 其计算式如下:

$$\text{KL}(\hat{\theta} | \theta) = \sum_x p_{\hat{\theta}}(X) \ln \frac{p_{\hat{\theta}}(X)}{p_{\theta}(X)} \quad (19)$$

其中, $p_{\theta}(X)$ 代表真实参数的概率分布, $p_{\hat{\theta}}(X)$ 代表估计参数的概率分布。为了保证仿真实验的准确性, 分别对每种算法运行 20 次, 求其 KL 散度的平均值。

表 1 试验约束

内部约束	外部约束
$P(S=1 C=0) > P(S=0 C=0)$	$P(S=1 C=0) > P(S=1 C=1)$
$P(S=0 C=1) > P(S=1 C=1)$	$P(R=1 C=1) > P(R=1 C=0)$
$P(R=1 C=1) > P(R=0 C=1)$	$P(W=1 S=1, R=1) > P(W=1 S=1, R=0)$
$P(R=0 C=0) > P(R=1 C=0)$	$P(W=1 S=1, R=1) > P(W=1 S=0, R=1)$
$P(W=1 S=1, R=1) > P(W=0 S=1, R=1)$	$P(W=1 S=1, R=1) > P(W=1 S=0, R=0)$
$P(W=1 S=0, R=1) > P(W=0 S=0, R=1)$	$P(W=1 S=0, R=1) > P(W=1 S=0, R=0)$
$P(W=1 S=1, R=0) > P(W=0 S=1, R=0)$	$P(W=1 S=1, R=0) > P(W=1 S=0, R=0)$

实验结果如图 4~图 6 所示, 图 4~图 6 中给出了 KL 散度值随样本大小的变化情况。从结果可以看出: 在样本量相同时, 传统的 EM 算法 KL 散度小于 kNN -MLE 算法, qirEM 算法 KL 散度小于传统的 EM 算法, kNN -DCE 算法 KL 散度小于 qirEM 算法。反映出数据缺失情况下 EM 算法比简单地进行数据填补然后求参数的最大似然估计要更加优越, 融入单一约束的 qirEM 算法比 EM 算法性能更好, 而融入双重约束的 kNN -DCE 算法性能最好。综上所述, 应用 kNN -DCE 算法可以很好地将先验知识融入参数学习, 获得精度更高的参数。

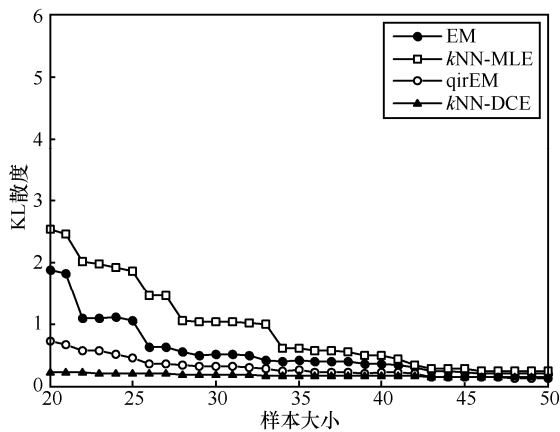


图 4 样本缺失 20% 时算法 KL 散度

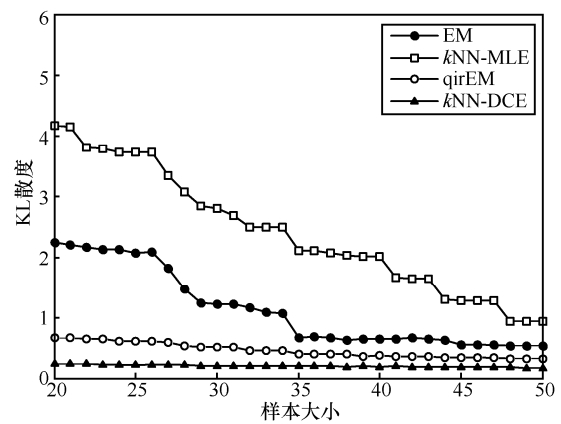


图 5 样本缺失 30% 时算法 KL 散度

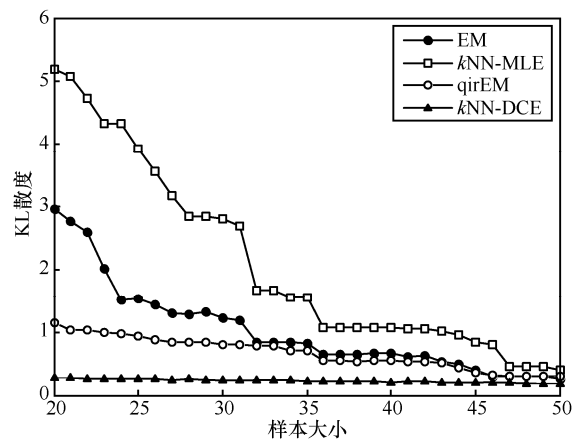


图 6 样本缺失 40% 时算法 KL 散度

4.2 电信客户流失预测

4.2.1 前期工作

本次试验以某运营商 2015 年客户数据为基础, 从电信原始数据所有属性中挑选出 9 个属性

后构建贝叶斯网络结构模型 (如图 7 所示)。为了方便描述, 用字母 $A \sim I$ 依次代替属性入网时长、用户属性、是否 VIP、客户是否流失、月通话次数、本月所耗流量、计费收入、计费时长、欠费状态。入网时长 $A = \{1, 2, 3\}$ 表示 $A = \{\text{短}, \text{中}, \text{长}\}$;

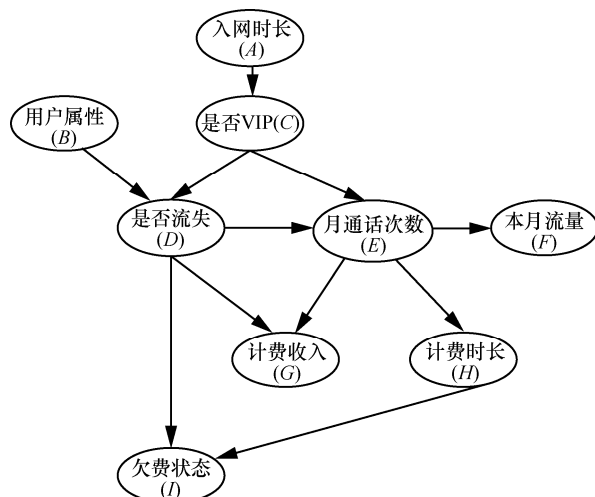


图7 电信客户流失预测贝叶斯网络结构

用户属性 $B = \{1, 2\}$ 表示 $B = \{\text{集团, 公众}\}$; 是否VIP $C = \{1, 2\}$ 表示 $C = \{\text{否, 是}\}$; 客户是否流失 $D = \{1, 2\}$ 表示 $D = \{\text{否, 是}\}$; 月通话次数 $E = \{1, 2, 3\}$ 表示 $E = \{\text{低, 中, 高}\}$; 本月所耗流量 $F = \{1, 2, 3\}$ 表示 $F = \{\text{低, 中, 高}\}$; 计费收入 $G = \{1, 2, 3\}$ 表示 $G = \{\text{低, 中, 高}\}$; 计费时长 $H = \{1, 2, 3\}$ 表示 $H = \{\text{短, 中, 长}\}$; 欠费状态 $I = \{1, 2, 3\}$ 表示 $I = \{\text{不欠, 欠少, 欠多}\}$ 。根据上面的约定, 结合专家的经验, 可以形成系列约束条件。例如, 内部约束 $P(D=1|B=1, C=2) > P(D=2|B=1, C=2) > P(D=1|B=2, C=2)$ 说明在用户属性同为集团且是VIP用户时, 客户不流失的概率大于客户流失

的概率; 而外部约束 $P(C=2|A=3) > P(C=2|A=2) > P(C=2|A=1)$ 则说明入网时长越长是VIP用户的概率越大。专家给出的参数约束条件见表4。

4.2.2 预测结果

试验用 550 个样本构成训练集, 样本缺失率 (missing completely at random, MCAR) 为 25%, 应用 kNN-DCE 算法学习电信客户流失预测网络的参数。为了验证算法的有效性, 用 561 个测试样本来检验该网络的预测精度, 贝叶斯网络推理采用联合树算法^[16]。电信客户流失预测是一个不平衡数据集分类问题, 一般定义少数类为 P , 多数类为 N , FP 是指将多数类错分成少数类的样本总数, FN 是指将少数类错分成多数类总数, TN 表示分类正确的多数类样本数, TP 表示分类正确的少数类样本数。用以下指标来评价预测精度:

少数类预测精度:

$$SE = \frac{TP}{TP + FN} \quad (21)$$

多数类预测精度:

$$SP = \frac{TN}{TN + FP} \quad (22)$$

表4 客户流失预测参数约束条件

约束类别	约束条件
内部约束	$P(D=1 B=1, C=2) > P(D=2 B=1, C=2)$
	$P(D=1 B=2, C=2) > P(D=2 B=2, C=2)$
	$P(E=3 C=2, D=1) > P(E=2 C=2, D=1) > P(E=1 C=2, D=1)$
	$P(F=1 E=1) > P(F=2 E=1) > P(F=3 E=1)$
	$P(H=1 E=1) > P(H=2 E=1) > P(H=3 E=1)$
	$P(G=1 D=1, E=1) > P(G=2 D=1, E=1) > P(G=3 D=1, E=1)$
	$P(G=3 D=1, E=3) > P(G=2 D=1, E=3) > P(G=1 D=1, E=3)$
	$P(C=2 A=3) > P(C=2 A=2) > P(C=2 A=1)$
	$P(D=1 B=2, C=2) > P(D=1 B=1, C=2) > P(D=1 B=2, C=1) > P(D=1 B=1, C=1)$
	$P(F=1 E=1) > P(F=1 E=2) > P(F=1 E=3)$
外部约束	$P(I=1 D=1, H=3) > P(I=1 D=1, H=2) > P(I=1 D=1, H=1)$
	$P(G=1 D=1, E=1) > P(G=1 D=1, E=2) > P(G=1 D=1, E=3)$
	$P(G=3 D=1, E=3) > P(G=3 D=1, E=2) > P(G=3 D=1, E=1)$

总体分类精度:

$$Gm = \sqrt{SE \times SP} \quad (23)$$

仿真结果由表 5 给出。从表 5 中可以看出,应用本文算法总体的预测精度略高于 EM 算法和 qirEM 算法,最为关注的潜在流失客户的预测精度明显高于 EM 算法及 qirEM 算法,所以本文算法能够很好地应用到电信数据贝叶斯网络中进行参数学习。

表 5 电信客户流失预测仿真结果

电信数据	kNN-DCE	EM	qirEM
SE	84.56%	70.49%	75.4%
SP	82%	90%	88%
Gm	83.27%	79.37%	81.45%

5 结束语

围绕电信流失客户预测中 BN 参数学习问题,本文提出 kNN-DCE 参数学习算法。本文采用改进的 kNN 算法填充缺失数据,然后在完整的数据基础上融入两类约束条件完成 BN 参数学习。仿真结果表明:所提算法性能优于 EM 算法,能有效提高电信流失客户预测精度。由于先验知识是否可靠会直接影响约束的正确性,进而影响参数学习的结果。在今后的工作中,将考虑先验知识的不确定性,研究如何缓解错误的先验知识对参数学习结果造成的负面影响。

参考文献:

- [1] JAMIL S, KHAN A. Churn comprehension analysis for telecommunication industry using ALBA[C]//International Conference on Emerging Technologies(ICET 2016), Oct 1, 2016, Islamabad, Pakistan. Piscataway: IEEE Press, 2016.
- [2] PEARL J. Probabilistic reasoning in intelligent systems: networks of plausible inference[J]. Computer Science Artificial Intelligence, 1988, 70(2): 1022-1027.
- [3] WAGNER S. A Bayesian network approach to assess and predict software quality using activity-based quality models[J]. Information and Software Technology, 2016, 52(11): 1230-1241.
- [4] PENDHARKAR P C, KHOSROWPOUR M, RODGER J A. Application of Bayesian network classifiers and data envelopment analysis for mining breast cancer patterns[J]. Journal of Computer Information Systems, 2016, 40(4): 127-131.
- [5] COOK J, LEWANDOWSKY S. Rational irrationality: modeling climate change belief polarization using Bayesian networks[J]. Topics in Cognitive Science, 2016, 8(1): 160-179.
- [6] LAURITZEN S L. The EM algorithm for graphical association models with missing data[J]. Computational Statistics & Data Analysis, 1995, 19(2): 191-201.
- [7] LIAO W, JI Q. Learning Bayesian network parameters under incomplete data with domain knowledge[J]. Pattern Recognition, 2009, 42(11): 3046-3056.
- [8] MASEGOSA A R, FEELDERS A J, GAAG L C V D. Learning from incomplete data in Bayesian networks with qualitative influences[J]. International Journal of Approximate Reasoning, 2015, 69(C): 18-34.
- [9] CORANI G, CAMPOS C P D. A maximum entropy approach to learn Bayesian networks from incomplete data[M]// Interdisciplinary Bayesian Statistics. Berlin: Springer International Publishing, 2015: 69-82.
- [10] 张连文. 贝叶斯网引论[M]. 北京: 科学出版社, 2006.
- [11] ZHANG L W. Introduction of Bayesian network[M]. Beijing: Science Press, 2006.
- [12] ZHOU Y, FENTON N, ZHU C. An empirical study of Bayesian network parameter learning with monotonic influence constraints[J]. Decision Support Systems, 2016, 87(C): 69-79.
- [13] FEELDERS A, VAN D G L C. Learning Bayesian network parameters under order constraints[J]. International Journal of Approximate Reasoning, 2005, 42(1-2): 37-53.
- [14] 刘星毅. 基于马氏距离和灰色分析的缺失值填充算法[J]. 计算机应用, 2009, 29(9): 2502-2504.
- [15] LIU X Y. Improved kNN algorithm based on Mahalanobis distance and gray analysis[J]. Journal of Computer Applications, 2009, 29(9): 2502-2504.
- [16] 杨宇, 高晓光, 郭志高. 小数据集条件下基于数据再利用的 BN 参数学习[J]. 自动化学报, 2015, 41(12): 2058-2071.
- [17] YANG Y, GAO X G, GUO Z G. Learning BN parameters with small data sets based reutilization[J]. Acta Automatica Sinica, 2015, 41(12): 2058-2071.
- [18] KULLBACK S, LEIBLER R A. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(22): 79-86.
- [19] HU X J, YANG S L, MA X J. Inference structure and construction algorithms of Bayesian network based on junction tree[J]. Acta Simulata Systematica Sinica, 2004, 16(11): 2559-2558.



[作者简介]



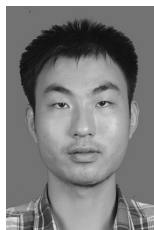
赵宇翔（1993-），男，西安邮电大学陕西省信息通信网络及安全重点实验室硕士生，主要研究方向为数据挖掘。



王航龙（1989-），男，西安邮电大学陕西省信息通信网络及安全重点实验室硕士生，主要研究方向为数据挖掘。



卢光跃（1971-），男，博士，西安邮电大学陕西省信息通信网络及安全重点实验室教授，主要研究方向为信号与信息处理、认知无线电和大数据分析。



李四维（1989-），男，西安邮电大学陕西省信息通信网络及安全重点实验室硕士生，主要研究方向为数据挖掘。