文章编号: 1001-0920(2007)06-0626-06

一种基于 Bayesian 信念网络的客户行为预测方法

何 蓓, 吴 敏
(中南大学 信息科学与工程学院, 长沙 410083)

摘 要: 提出一种基于 Bayesian 信念网络(BN)的客户行为预测方法. 通过知识学习构建客户行为 Bayesian 网络(CBN),根据 CBN 对预实例计算联合分布概率,准确预测了一对一营销优化中的客户行为. CBN 学习算法包括连线和定向部分,复杂度为 $O(N^4)$ 条件相关测试. 在零售行业一对一营销实际应用表明,CBN 学习算法较现有 BN 学习算法更快构建 CBN,预测精度高于朴素 Bayesina 分类法.

关键词: Bayesian 信念网络; 一对一营销; 数据挖掘; 客户行为预测
中图分类号: TP311 **文献标识码**: A

A approach for customer behavior precdiction based on Bayesian belief network

HE Bei, WU Min
(School of Information Science and Engineering, Central South University, Changsha 410083, China. Correspondent: WU Min, E-mail: min@csu.edu.cn)

Abstract: A new customer behavior prediction approach based on Bayesian belief network is presented. The customer behavior Bayesian network (CBN) is constructed through knowledge study, and joint probabilities are calculated with this network to precdict the customer behavior. The CBN learning algorithm is composed of connecting and directing parts, and complexity is $O(N^4)$ conditional dependence test. The empirical applications in retail one-to-one marketing show that CBN is constructed more quickly by using this approach than other existing BN learning algorithm, and the accuracy is better than that of naive Bayesian classificiation.

Key words: Bayesian belief network; One-to-one marketing; Data mining; Customer behavior prediction

1 引 言

在过去的十多年中,随着商业活动和 IT 技术的迅速发展,企业同客户之间的交互方式已经发生了显著的变化. 交易周期的不断缩短,市场的逐渐扩大,商品种类和物流方式的显著增加,都使得企业和客户之间的关系日益复杂. 面对日益激烈的市场竞争,为尽可能多地保持和发展客户,企业需要更多地了解客户,并对其行为进行准确地预测,以便及时采取相应的措施.

在客户关系管理(CRM)系统中,客户行为预测主要涉及客户保持、客户获取及营销优化等领域. 它能预先估计客户的行为和需求,帮助企业提供最佳营销方案,从而提高客户忠诚度,增加企业利润. 目前,在客户行为预测领域除采用统计、最近邻以及聚类分析等一些经典的数据挖掘技术外,一些较新型的数据挖掘技术也逐渐应用于该领域,如决策树、神经网络及规则归纳等方法. 其中比较典型的算法有 CHAID(X^2 自动交感侦察器)、CART(分类回归树)以及类神经网络分析等[1]. 另外,随着该领域研究的日益增加,近几年也相继提出了一些新的研究方法,如 Chen 等[2]针对零售市场利用类似性和意外性两种方法对不同时期客户改变模式的类同程度进行分析,从而预测客户行为的改变; Kim 等[3]基于遗传算法将投票法、行为-知识空间法以及神经网络等多种分类方法进行结合,对电子商务中的客户购买行为进行预测.

尽管这些方法能在一定程度上为企业制订正确的营销方案提供参考,但在预测范围和精确性方面仍存在局限性. 在预测范围方面,由于目前大多数的行为预测方法在预测过程中并未涉及客户的所有关

收稿日期: 2006-03-29; **修回日期**: 2006-06-01.
基金项目: 国家杰出青年科学基金项目(60425310); 国家 863 计划项目(2006AA04Z172).
作者简介: 何蓓(1979—),女,湖南邵阳人,博士生,从事一对一营销优化、最优化计算等研究; 吴敏(1963—),男,广东化州人,教授,博士,从事先进控制、过程控制等研究.

键属性(如年龄、性别、婚姻状态、收入情况、学历、职业、家庭小孩数等),它虽然适用于基于市场细分的管理模式,但不能较好地满足一对一营销考虑每个客户均有不同需求的特性.另外,在精确性方面,现有预测方法虽然可反映出客户行为的大致趋势,但不能完全保证其预测结果的准确性,而预测结果的准确性又直接涉及企业营销方案的成功率问题.

Bayesian 分类方法可以较好地解决以上问题,它通过预测类成员关系的可能性,给出给定样本属于某个特定类的概率.从理论上讲,与其他现有分类算法相比,具有最小的出错率[4].Bayesian 分类方法大致上分为朴素 Bayesian 分类和 Bayesian 信念网络(BN)两种.

朴素 Bayesian 分类在假定类条件独立,即给定样本的类标号、属性的值相互独立的情况下进行概率分析,这一假定简化了计算. Arne 等[5]利用朴素 Bayesian 和最大熵方法来解决客户行为预测问题, Lian 等在文献[6]中也涉及基于该方法解决通讯行业的客户行为预测问题. Baysian 信念网络则说明联合条件概率的分布情况,它允许在变量的子集间定义类的条件独立性,其预测过程由两部分组成:信念网络学习和推理.

考虑现实生活中,客户属性之间可能存在依赖关系(如客户的职业与其教育程度和收入情况等),本文基于 Bayesian 信念网络方法,对一对一营销中客户行为知识的网络学习机理进行研究,提出一种更为精确的客户行为预测方法.

2　问题描述

定义 1　客户行为知识集 $BK = (C, Q, I)$,其中:属性集 $C = B \cup K$, B 为客户行为属性集,包括客户购买行为、客户保持行为、客户转移行为等; K 为可能与客户行为 B 具有依赖关系的环境属性集,包括相关客户属性、推销渠道、产品型号等; Q 为专家经验集,给出某些属性之间的依赖关系; I 为 C 的样本实例集,其各属性的值均为离散值.

定义 2　客户行为 Bayesian 网络(CBN) $CN = (G, T)$,其中: $G = (V, E)$ 为有向无环图(DAG), T 为条件概率表(CPT)集.节点集 V 由属性集 C 构成, $V = C$,节点之间的有向弧线 E 代表节点间的概率依赖.设节点 $c_i, c_j \in C$, $i \neq j$, $i, j = 1, 2, \cdots, N$, $N = |C| + |B|$ 为 G 中的节点数,其中 $|\cdot|$ 表示集合所包含的元素个数.在图 G 中,如果有一条弧线由节点 c_i 指向节点 c_j,则 c_i 和 c_j 之间存在概率依赖,它们互为邻居(用 S_{ai} 表示 c_i 在 G 中的所有邻居节点集合), c_i 是 c_j 的双亲,而 c_j 为 c_i 的子女. c_j 的子女则称为 c_i 的后代, c_i 的双亲称为 c_j 的祖先.节点 c_j 的条件概率表 T_j 给出条件分布 $P(c_j \mid \text{Parents}(c_j))$,其中 $\text{Parents}(c_j)$ 为 c_j 的双亲.

定义 3　预测实例 $PI = (D, KC, KI)$,其中: D 为待预测的客户行为, KC 为已知的环境属性集, KI 为已知的环境属性实例.

根据以上定义,基于 Bayesian 网络的客户行为预测可描述为首先通过学习知识集 BK 构建相应 Bayesian 网络 CN;然后利用 CN 对预测实例 PI 进行推理,计算其联合分布概率 $P(D \mid KC = KI)$,确定在 KI 环境下发生客户行为 D 的可能性.

3　CBN 学习相关定义

客户行为知识集 BK 的建立是根据企业历史营销数据构建的,然而一个拥有数十万乃至数百万客户的大型企业,很显然其规模相当庞大.根据文献[7]的分析,即使是相对简单的、基于评分准则的 Bayesian 网络学习算法,从大规模样本中获取一个较为精确的 Bayesian 网络也是 NP-hard 问题.为避免算法出现指数复杂性,本文对基于评分准则的 Bayesian 网络学习算法进行改进,并结合一对一营销专家经验,提出了一种有效的 CBN 网络学习算法.为方便算法的描述,给出相关定义.

定义 4　在有向无环图 G 中:

1) 不考虑弧线 E 的方向.如果存在一条路径将两个节点相连,则这条路径为邻接路径,简称为路径,称没有方向的弧线为无向边.

2) 考虑弧线 E 的方向.如果存在一条路径将两个节点相连,则该节点对是有向相连的,称这条路径为有向邻接路径,或有向路径,称具有方向的弧线为有向边.

3) 两节点间存在邻接路径,若它不是有向邻接路径,则称该邻接路径为无向邻接路径,或无向路径.

定义 5　在邻接路径上的任意一个节点,如果在该路径上的两个弧线均指向该节点,则称该节点为碰撞点,否则称为非碰撞点.

定义 6　在 $G = (C, E)$ 中,设 $x, y \in C$, $x \neq y$, $Z \in C$ 中子集 $\{x, y\}$ 的补集表示为 $C \setminus \{x, y\}$.如果 x 和 y 之间有一条邻接路径 P, P 上所有的节点均属于以下情况之一,则称为 x 和 y 在给定 Z 下是 d-连通的:1) 为非碰撞点且不属于 Z;2) 为碰撞点或者本身或其子女属于 Z.如果两个节点间不存在 d-连通路径,则称它们为在给定 Z 下是 d-分离的, Z 称为条件集.

定义 7　在 G 中,若通过改变两节点无向邻接路径中一个或多个节点的状态(将碰撞点改为非碰撞点或将非碰撞点改为碰撞点),使得该节点对为有

向相连,则称此行为为打开路径,否则称为关闭路径.

定义 8 在Bayesian网络中如果两个节点存在依赖关系,则知道其中一个节点的值将会对获取另一节点的值提供一定的信息.根据信息传递理论,可由其相关信息进行测量.下式表示节点 x 和 y 之间的相关信息:

$$I(x,y) = \sum_{x,y} P(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}. \quad (1)$$

类似地,对与两个 d- 连通的节点而言,它们之间的信息传递值可由条件相关信息进行测量,即

$$\mathrm{CI}(x,y\mid Z) = \sum_{x,y,z} P(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)\,p(y\mid z)}, \quad (2)$$

其中 $z \in Z$ 表示 Z 中一个给定节点.当 $I(x,y)$ 小于某个特定的阈值 ε,则称 x 和 y 是无关的,否则称为相关.同样,当 $\mathrm{CI}(x,y\mid Z)$ 小于 ε,则称 x 和 y 是条件无关的,否则称为条件相关.将式(1)和式(2)计算分别称为 I 测试和 CI 测试.

4 CBN学习算法

CBN 学习算法主要包括连线和定向两部分.连线部分确定图 G 中各客户属性之间、客户属性与客户行为以及各客户行为之间是否存在依赖,根据依赖关系连接各节点;定向部分则确定各个依赖关系的方向,即弧线的方向.

4.1 连线部分

连线部分由草拟和完善两个阶段组成.草拟阶段计算每对节点之间的相关信息,并基于该信息描绘一个草图.完善阶段首先对图进行添补,在条件相关的两个节点间添加边;然后再测量所有边两端节点间的 CI 测试,去除多余边.

4.1.1 草拟阶段

草拟阶段利用 I 测试来描绘Bayesian网络的草图,各节点之间最多只有一条邻接路径,确保不会形成环结构.

Step1: 初始化图 $G(V, E)$,链表 L 和指针 p,其中 $V = C, E = L = \{\}, p = \mathrm{NULL}$.

Step2: 对每对节点 (c_i, c_j),其中 $c_i, c_j \in C, i \neq j, j, i = 1, \cdots, N$.根据式(1)进行 I 测试,计算相关信息 $I(c_i, c_j)$.去除所有 $I(c_i, c_j)$ 值小于阈值 ε 的节点对,并对剩下的每对节点按 $I(c_i, c_j)$ 从大到小排序,将其存储在 L 中.令指针 p 指向 L 的第 1 个节点对.

Step3: 根据 p 的位置从 L 取出一个节点对,如果该节点对间没有邻接路径,则添加相应的边到 E,去除该节点对,将 p 移至下一个节点对.

Step4: 如果 E 已包含 $N - 1$ 条边,或 L 已经为空,则该阶段结束,否则重复 Step3.

4.1.2 完善阶段

完善阶段首先检查所有相关信息值大于 ε 却并没有被直接连接的节点对,调用程序 seperate 检测这些节点对是否条件相关.如果是则在该节点对间添加边;然后利用 CI 测试识别草拟阶段中添加的多余边,将它们从 E 中去除.

Step1: 对 L 中所有剩余的节点对,调用程序 separate (current_graph, node1, node2) 判断该节点对是否应该建立连接.如果是则连接节点对,加入相应边到 E,否则不做任何操作.

Step2: 对 E 中的每条边,若其两端的节点间存在其他的邻接路径,则暂时从 E 中去除该边.调用程序 separate (current_graph, node1, node2).如果这对节点不能被分离,则将该边返回到 E 中,否则永远去除该边.

```
separate (current_graph, node1, node2) {
  N1 = S_node1 ∩{node1 和 node2 邻接路径上的节点};
  N2 = S_node2 ∩{node1 和 node2 邻接路径上的节点};
  N1′= S_N1 ∩{node1 和 node2 邻接路径上的节点};
  N2′= S_N2 ∩{node1 和 node2 邻接路径上的节点};
  if |N1 ∪ N1′| <|N2 ∪ N2′|,
     Z = N1 ∪ N1′;
  else Z = N2 ∪ N2′;
  v = CI(node1,node2 | Z); //CI 测试
  if v < ε return 'seperated'
  else if |Z| = 1 return 'connected'
  while |Z′| <|Z| {
     Z′= Z;
     Z_i = Z\{Z 的第 i 个节点} //i = 1, …, |Z|
     v_i = CI(node1,node2 | Z_i)
     if v < ε return 'seperated'
     else if v_i ≤v + ε, Z′= Z\{Z 的第 i 个节点}
  }
  return 'connected'
  }
```

假设 node2 不是 node1 的祖先节点,根据定义 6 可知如果节点 node1 和 node2 不是 d- 连通的,则它们可被在邻接路径上 node2 的双亲结点 d- 分离.但目前图 G 中的弧线仍为无向边,所以无法直接判断 node2 的邻居节点是否为其双亲节点.根据Bayesian

网络的性质,如果转变同一路径中两个连续节点的状态,则必将关闭该路径(因为在同一路径上不可能存在两个连续的碰撞点),那么通过改变 $N1 \cup N1'$ 或 $N2 \cup N2$ 的状态就可以关闭所有通过两个或两个以上节点将 node1 和 node2 有向相连的路径. 不过在这种情况下,通过一个碰撞点将 node1 和 node2 相连的路径仍有可能被打开. 因此,只要保证在不打开两节点间任何相对路径的前提下,将所有碰撞点从 $N1 \cup N1$ 或 $N2 \cup N2$ 去除,就可保证在潜在模型中 node1 和 node2 间的所有路径均能被关闭.

设这些双亲节点构成条件集 Z,则 Z 为 $N1 \cup N1$ 或 $N2 \cup N2$ 的子集. 在假设从 Z 中去除 node2 的一个双亲节点将不会减少 node1 和 node2 间相关信息的前提下,程序 separate 通过比较两个节点在不同条件集下的条件相关信息,去除 $N1 \cup N1'$ 和 $N2 \cup N2'$ 中 node2 的子女节点和其他无关节点.

4.2　定向部分

定向部分主要通过识别 G 中的碰撞点并结合专家经验确定 E 中弧线的方向.

根据定义 5,在 G 中所有的节点可分为两种:碰撞点和非碰撞点. 若节点 b 为碰撞点,则它与邻居节点 a,c 的连接情况为 $a \rightarrow b \leftarrow c$;若为非碰撞点,则其可能的连接情况为 $a \leftarrow b \rightarrow c$ 或 $a \rightarrow b \rightarrow c$. 根据 Charniak[8] 的分析,对 Bayesian 网络而言,只有当 b 为碰撞点时,才能在 b 已知的情况下 a 和 c 之间可以传递信息. 换言之,在此情况下,a 和 c 在给定条件集 b 下是条件相关的. 因此可利用 CI 测试识别碰撞点,决定 G 中弧线的方向.

Step1: 在 G 中寻找至少有一个共同邻居的两节点 $v1$ 和 $v2$,$N1 = S_{v1} \cap \{v1$ 和 $v2$ 邻接路径上的节点$\}$,$N2 = S_{v2} \cap \{v1$ 和 $v2$ 邻接路径上的节点$\}$.

Step2: $N1' = S_{N1} \cap \{v1$ 和 $v2$ 邻接路径上的节点$\}$,$N2' = S_{n2} \cap \{v1$ 和 $v2$ 邻接路径上的节点$\}$.

Step3: 如果 $|N1 \cup N1'| < |N2 \cup N2'|$,令 $Z = N1 \cup N1'$,否则 $Z = N2 \cup N2'$.

Setp4:设 $v = CI(v1, v2 | Z)$,如果 $v < \varepsilon$,转向 Step7,否则,如果 Z 仅包含一个节点,令 $v1$ 和 $v2$ 为 Z 中节点的双亲节点.

Step5: 设 $Z' = Z$,$Z_i = Z \setminus \{Z$ 的第 i 个节点$\}$,$i = 1, \cdots, |Z|$,计算 $v_i = CI(node1, node2 | Z_i)$,如果 $v_i < \varepsilon$,转向 Step7,否则,如果 $v_i \leq v + \varepsilon$,$Z' = Z' \setminus \{Z$ 的第 i 个节点$\}$,若 Z 的第 i 个节点 $\in S_{v1,v2}$,则令 $v1$ 和 $v2$ 为其双亲节点.

Step6: 如果 $|Z'| < |Z|$,设 $Z = Z'$,如果 $|Z| > 0$,转到 Step4.

Step7: 转到 Step1,直到所有的节点对都被检测.

Step8: 对任意 3 个节点,如果 a 是 b 的双亲节点,b 和 c 存在无向边,且 a 和 c 没有连接,令 b 为 c 的双亲节点.

Step9: 如果任意边 (a, b) 无定向,且专家经验集 Q 指定 b 是依赖 a 的,则令 a 为 b 的双亲节点.

Step10: 转到 Step8,直到没有边可定向为止.

Step4 ～ Step6 利用 CI 检测对两节点间的碰撞点进行识别,Setp8 根据识别到碰撞点确定其他弧线的方向. 虽然,这种靠识别碰撞点来确定弧线方向的方法并不能确保可以定向所有弧线(在最坏情况下,如果 Bayesian 网络中不存在碰撞点,那么将无法对任何弧线定向),但所幸的是这种情况在实际应用中很少发生,在绝大多数情况下,它可以确定绝大多数甚至全部的弧线方向. 为保证 CBN 中所有弧线均为有向边,Step9 引入专家判断机制,根据专家经验集 Q 对 CBN 学习算法未能定向的弧线进行定向. 考虑专家经验有可能是错误的,因此算法以实际计算的依赖关系为准,仅在无法定向的情况下采用其建议.

5　复杂度分析

在 CBN 学习算法中,其计算大部分集中在 CI 测试上,因此用 CI 测试作为衡量本算法时间复杂度的指标(I 测试可以视为当条件集为空时 CI 测试的特殊情况). 实际上它也是在相关性分析算法中广泛采用的一个衡量指标[9].

在连线部分,草图阶段对图 G 中任意两个节点分别进行 I 测试,测试次数为 $N(N-1)/2$,时间复杂度为 $O(N^2)$. 在完善阶段首先在草图阶段的基础上添加边,最坏情况下它将添加 $N(N-1)/2-(N-1)$ 条边,因此需调用 $N(N-1)/2-(N-1)$ 次程序 seperate. 在去除多余边时,最坏情况下它调用 $N(N-1)/2$ 次程序 seperate. 程序 seperate 最坏情况下 $N1 \cup N1$ 和 $N2 \cup N2$ 包含 $N-2$ 个节点,在 while 循环中共需进行 $(N-2)+(N-3)+\cdots+2+1=(N-2)(N-3)/2$ 次 CI 测试. 该阶段的时间复杂度为 $O(N^4)$.

在定向部分对每对节点进行检测,判断它们之间是否存在碰撞点. 判断的次数为 $N(N-1)/2$,每次判断的时间复杂度为 $O(N^2)$(其分析方法类似于程序 seperate),定向阶段的时间复杂度为 $O(N^4)$.

综合以上分析,CBN 学习算法的复杂性为 $O(N^4)$. 需要说明的是虽然在最坏情况下,该算法的复杂性与 Cheng 等[9] 提出的 Bayesian 网络学习方法相同,但在实际应用中本文的 CBN 学习算法速度更快,这主要是因为 CBN 学习算法利用两个连续节

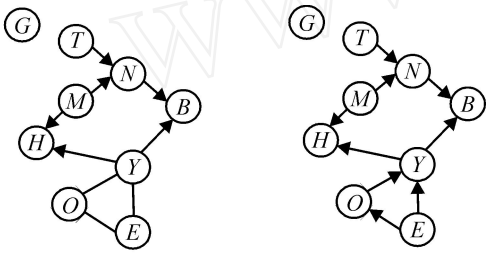
点不可能同时为碰撞点的特点,对边的添加进行了更严格的判断.

6 客户行为预测实例

本节以零售行业客户会员卡申请行为 B 为例,给出基于Bayesian网络方法进行客户行为预测的应用实例. 其数据信息来自在美国、墨西哥和加拿大均有销售业务的大型连锁超市 FoodMart 的数据库. 在该数据库中 customer 表属性中包括客户属性:M 为 marital_status, G 为 gender, H 为 houseowner, O 为 occupation, E 为 education, Y 为 yearly_income, T 为 total_children, N 为 num_children_at_home 等.

定义客户行为知识集 $BK = (C, Q, I)$,其中:$C = \{M, G, H, O, E, Y, T, N, B\}$; $Q = \{E \rightarrow O, E \rightarrow Y, O \rightarrow Y, T \rightarrow N\}$;实例集 I 为从历史数据库中抽取的 10^4 随机样本,并进行了相关预处理,以保证其属性值均为离散值.

根据 CBN 学习算法,得到不采用专家知识集 Q 和采用专家知识集的 BN,分别如图 1(a) 和图 1(b) 所示.



(a) 未使用Q (b) 使用Q

图 1 客户会员卡申请行为 BN

从图 1 可以看出,节点 G 与 B 之间没有路径,因此客户性别与会员卡申请行为没有依赖关系. 另外,节点 H 不为 B 的祖先,因此客户的房产情况也与会员卡申请行为无关. 图 1(a) 与图 1(b) 不同之处在于,后者根据专家知识集 Q 确定客户年收入 Y、职业 O 和教育情况 E 之间的依赖方向,从而保证了图的结构为 DAG 图. 将图 1(b) 用 CBN_G 表示.

表 1 给出了 CBN 学习法与文献[9]中的 Bayesian 网络学习算法在处理该问题上的性能比较. 根据表 1 可知,CBN 学习算法的性能提高主要体现在完善阶段,CI 测试数和时间消耗为文献[9]方法的 52 % 左右. 根据算法比较分析,在该阶段 CBN 算法调用程序 separate 采用邻接路径上两个连续节点为条件集进行 CI 测试,避免了边的不正确添加.

表 1 CBN学习法和文献[9] BN学习法性能比较

阶段	CI 测试次数		计算时间 / s	
	CBN 学习法	文献[9] 法	CBN 学习法	文献[9] 法
草拟	28	28	0.82	0.82
完善	**68**	**129**	**11.72**	**22.48**
定向	15	15	2.80	2.80
合计	111	172	15.34	26.10

根据图 CBN_G 提供的依赖关系,结合概率论中的条件概率公式

$$P(B \mid A) = P(AB) / P(A), \quad (3)$$

其中 A, B 为两个客户属性节点,可得到各个节点的 CPT 集 CBN_T. 根据定义 2,即可获得 CN = (CBN_G, CBN_T).

根据所得 CN,采用文献[10]中给出的 PPTC 方法进行 Bayesian 网络推理(其思路为将 Bayesian 网络转换为一个由簇和邻接节点集组成的相容连接树结构,然后根据该连接树进行全局传播和边缘化计算,从而得到所需的条件概率分布),则可根据预测实例 PI,对客户会员卡申请的预测情况进行预测. 随机抽取 10 组客户行为预测属性实例如表 2 所示,考虑到客户性别 G 和房产情况 H 对客户会员卡申请行为没有依赖关系,因此可不考虑这两个属性的取值.

表 3 给出了客户会员卡行为预测结果与实际会员卡申请概率以及基于朴素 Bayesian 的分类结果比较. 从表 3 可以看出,本文给出的基于 Bayesian 信念网络的客户行为预测法能给出具体的预测概率,且精度较高,与实际概率仅偏离 ±1 % ～ ±3 % 左右(需要说明的是在 ID = 10 的情况下,其预测偏差达

表 2 客户行为预测属性实例

ID	客户属性					
	M	T	E	O	N	Y
1	S	1	Partial high school	Professional	0	$70k ～ $90k
2	M	4	Partial high school	Skilled manual	4	$10k ～ $30k
3	S	4	Bachelors degree	Professional	0	$50k ～ $70k
4	S	1	High school degree	Skilled manual	0	$30k ～ $50k
5	M	1	Partial high school	Manual	1	$10k ～ $30k
6	M	5	Bachelors degree	Professional	3	$50k ～ $70k
7	S	5	Bachelors degree	Management	0	$50k ～ $70k
8	M	4	Partial high school	Skilled manual	3	$10k ～ $30k
9	M	2	Partial high school	Manual	2	$10k ～ $30k
10	S	0	High school degree	Management	0	$150k +

表 3　客户会员卡行为预测结果比较

ID	实例概率/%				Bayesian 信念网络预测法/% $P(B \mid M,T,E,O,N,Y)$				朴素 Bayesian 分类法 $P(M,T,E,O,N,Y \mid B)P(B)$			
	Bronze	Normal	Silver	Golden	Bronze	Normal	Silver	Golden	Bronze	Normal	Silver	Golden
1	**80**	8	10	2	**81**	6	9	5	**1.7E-04**	9.9E-06	2.8E-05	1.8E-06
2	2	**94**	4	0	4	**93**	3	0	1.0E-07	**3.2E-04**	9.0E-07	9.0E-06
3	**83**	2	9	7	**81**	4	10	5	**1.1E-03**	0.0E+00	2.0E-04	0.0E+00
4	**83**	4	8	4	**82**	4	10	4	**1.6E-03**	0.0E+00	1.0E-04	0.0E+00
5	3	**86**	9	2	5	**88**	3	4	3.9E-07	**4.5E-06**	2.8E-07	9.9E-08
6	22	0	11	**67**	10	6	14	**70**	6.3E-06	1.0E-07	8.9E-06	**1.4E-04**
7	**82**	0	9	9	**81**	4	10	5	**2.1E-04**	5.0E-07	3.8E-05	1.0E-05
8	9	**91**	0	0	5	**93**	2	0	**2.1E-04**	5.0E-07	3.8E-05	1.0E-05
9	2	**91**	2	4	5	**88**	3	4	**1.1E-03**	0.0E+00	2.0E-04	0.0E+00
10	0	0	**100**	0	5	0	**78**	16	3.0E-06	2.0E-08	**2.3E-05**	8.0E-07

22%,这是因为在实际商业活动中,ID = 10 的客户属性实例较少,使其预测结果不能得以较好的评价). 而基于朴素 Bayesian 分类法的预测方法可进行简单的分类(实例被指派到 $P(M,T,E,O,N,Y \mid B)P(B)$ 值最大的类),但不能对其发生的具体概率进行估计,且存在预测结果不正确的情况. 如朴素 Bayesian 分类法预测 ID 分别为 8 和 9 的客户属性实例会员卡申请行为为 Bronze 卡,而实际客户会员卡申请活动中 Normal 卡申请概率最高达 91%.

根据表 3 中 Bayesian 信念网络预测法提供的结果,对于那些年薪较高、家中小孩数较多的客户来说,申请 Golden 的可能性较高;而对孩子数较少、年薪较低的客户来说,申请 Normal 和 Bronze 卡的概率较大. 因此,会员卡申请的客户行为是与属性 N 和 Y 密切相关的,图 1 也证实了这一点(N 和 Y 是 B 的双亲节点). 根据这个结论,也可将该预测方法应用于基于市场细分的营销活动中,针对客户家中小孩数和年收入情况进行有针对性的促销.

7　结　　语

本文提出了一种基于 BN 的客户行为预测方法,它较传统的基于决策树和朴素 Bayesian 的客户预测方法而言,具有更高的精度(BN 是目前分类方法中精度最高的算法),且应用范围更广,适用于市场细分和一对一营销等多种营销方法. 在该预测方法中提出的 CBN 学习算法,时间复杂度为 $O(N^4)$,避免了多 Bayesian 网络学习算法的指数发散问题,且在实际的客户行为预测应用中,与其他复杂度相同的 Bayesian 网络学习算法相比,学习速度更快. 零售行业的实际应用表明,该方法在解决一对一营销问题上具有预测精度高、运算速度快等特点,能有效帮助企业做出最佳决策.

参考文献(References)

[1] Alex Berson, Stephen Smith, Kurt Thearling. Building data mining applications for CRM[M]. Beijing: Posts and Telecom Press, 2001.

[2] Chen M C, Chiu A L, Chang H H. Mining changes in customer behavior in retail marketing[J]. Expert Systems with Applications, 2005, 28(4): 773-781.

[3] Kim Eunju, Kim Wooju, Lee Yillbyung. Combination of multiple classifiers for the customer's purchase behavior prediction[J]. Decision Support Systems, 2002, 34(2): 167-175.

[4] Jiawei Han, Micheline Kamber. Data mining concepts and techniques[M]. Beijing: China Machine Press, 2001: 196-200.

[5] Arne Mauser, Ilja Bezrukov, Thomas Deselaers, et al. Predicting customer behavior using naive bayes and maximum entropy — Winning the data-mining-cup 2004[C]. Proc Informatiktage 2005. Augustin, 2005.

[6] Lian Yan, Richard H Wolniewicz, Robert Dodier. Predicting customer behavior in telecommunications[J]. IEEE Intelligent Systems, 2004, 19(2): 50-58.

[7] David Maxwell Chickering, David Heckerman, Christopher Meek. Large-sample learning of bayesian networks is NP-hard[J]. J of Machine Learning Research, 2004, 5(10): 1287-1330.

[8] Eugene Charniak. Bayesian networks without tears[J]. AI Magazine, 1991, 12(4): 50-63.

[9] Cheng J, David Bell, Liu W R. Learning bayesian networks from data: An efficient approach based on information theory[J]. Artificial Intelligence, 2002, 137(1/2): 43-90.

[10] Cecil Huang, Adnan Darwiche. Inference in belief networks: A procedural guide[J]. Int J Approximate Reasoning, 1996, 15(3): 225-263.