

基于混合贝叶斯网络数据挖掘及研究生升学预测模型的研究

叶善文¹, 赵杰煜²

(1. 宁波大学 研究生院, 浙江 宁波 315211; 2. 宁波大学 计算机科学技术研究所, 浙江 宁波 315211)

摘要: 给出了基于 MAP 和 MDL 混合机制的贝叶斯网络结构学习算法, 新算法吸取了两种方法各自的特点, 具有计算简单、收敛速度快且能综合利用先验知识及专家知识的优点. 并结合 A 校研究生的海量数据进行实验, 结果表明: 新的预测模型准确率可达 84%, 且推理高效合理.

关键词: 贝叶斯网; 混合模型; 数据挖掘; 升学预测

中图分类号: TP302

文献标志码: A

文章编号: 1001-5132 (2013) 03-0040-05

我国学位与研究生教育规模增长迅速, 研究生报考人数也逐年大幅上升, 2002 年考生为 62.4 万人, 到 2012 年已经增长至 165.6 万人, 年均增长达 16.5%. 招生考试每年带来了数百万条的信息记录, 科学有效地利用这些海量信息, 挖掘其中有益的知识, 为高校研究生招生决策提供支持, 具有非常重要的现实意义.

1 模型选择

贝叶斯网络(BN)是一个带有概率注释的有向无环图, 此图模型能表达大变量集合的联合概率分布和随机变量间复杂的不确定性, 便于发现数据间的潜在关系, 在海量数据的知识发现方面具有重要研究意义与广阔的应用前景. 研究人员利用贝叶斯网络的这种特性, 研究从海量数据中学习并生成贝叶斯网络的方法, 提出了包括 Bayesian 方法、类 Bayesian 方法和非 Bayesian 方法、混合 Bayesian 方法等, 为 Bayesian 网络用于数据挖掘和知识发现开辟了道路, 在一些数据建模问题中已取得令人瞩目的效果, 而新方法和技术还在不断被引入到 Bayesian 结构学习及应用之中^[1-17].

笔者结合 2008~2012 年间 A 校研究生报考数据, 建立基于 MAP-MDL (the Maximum Aposterior Probability and Minimum Description Length) 的混合贝叶斯网络模型, 挖掘数据变量间可能存在的相互关系, 学习存在海量数据中的知识信息, 并架

构预测模型, 推理变量之间的因果分析, 为研究生招生决策提供支持, 有着一定的现实意义.

2 基于 MAP-MDL 的混合贝叶斯网络模型

2.1 极大后验概率(MAP)

机器学习的任务就是在给定样本数据 D 的情况下, 确立最佳网络图模型 S . 用 $P(D)$ 表示样本数据 D 的先验概率, $P(D|S)$ 表示假设图模型 S 成立时 D 的先验概率. 然而在机器学习中, 我们关心的是 $P(S|D)$, 即给定 D 时 S 的成立概率, 称为 S 的后验概率. 而后验概率极大的模型就是最优模型^[9,11], 此模型可表示为: $S_{\text{best}} = \arg \max P(S|D)$, 而 S 即被称为极大后验假设模型(MAP). 贝叶斯公式提供了从先验概率 $P(S)$ 、 $P(D)$ 和 $P(D|S)$ 来计算后验概率 $P(S|D)$ 的方法, 相关公式如下:

$$\begin{aligned} S_D &= \arg \max_{S \in S} [P(S|D)] = \\ &= \arg \max_{S \in S} [P(D|S)P(S)] / P(D) = \\ &= \arg \max_{S \in S} [P(D|S)P(S)], \end{aligned} \quad (1)$$

由于 $P(D)$ 是不依赖于 S 的常量, 因此可略去.

假设作为(1)式的 MAP 模型虽可以综合考虑和利用先验知识 S 与观测数据 D , 但也易受主观片面性及数据噪音的影响.

2.2 最小描述长度(MDL)

最小描述长度(MDL)是 Rissanen 在研究通用编码时提出的. 根据该理论, 如果将 Bayesian S 作为

对样本数据 D 进行压缩编码的模型, 而模型描述长度等于描述网络结构 S 的长度和在给定结构 S 下描述样本数据集 D 的长度之和. 且 S 的描述长度随模型复杂度的增加而增长, 而 D 的描述长度随模型复杂度的增加而减短. 因此, 可构建一个评分函数, 发现其中评分最大的结构, 即为最优结构, 此时结构 S 的总描述长度最小(MDL). 评分函数为:

$$S_D = \arg \max_{S \in \mathcal{S}} [L(D|S) + L(S)]. \quad (2)$$

该学习方法基于迭代的优化搜索方法: (1) 优点是无需先验知识, 实验结果易于接受, 网络简洁性及数据拟合度之间也可以很好地达到平衡; (2) 缺点是不适用于不完整样本集的学习, 且计算量大, 速度较慢; 也缺乏对先验知识处理的能力; 不能很好地利用网络结构的继承性; 学习结果受样本数据量的制约, 一般要求海量的样本数据^[7,9,11].

2.3 MAP-MDL 的混合贝叶斯网络模型

笔者拟合 MAP 与 MDL 各自的优点, 构建基于 MAP-MDL 的混合贝叶斯网络模型, 从而提高网络学习效率, 减少主观片面性及噪音所带来的影响.

根据香农信息论, 1 个离散信源可以看成 1 个马尔可夫过程来建模. 把 1 个变量 X 初始化为 1 个值集 $\{x_1, x_2, \dots, x_i, \dots\}$, 如果 1 个值 x 对应于信源的概率 $P(X=x)=P(x)$, 则 x 的信息量可定义为:

$$I(X=x)=I(x)=-\log_2 P(x). \quad (3)$$

变量 X 的最小描述长度则是该信源的最优描述语言 $\mathcal{L}$. 在这种最优描述语言中, 描述长度 $L(x)$ 与信息量 $I(x)$ 相等, 即:

$$L(X=x)=L(x)=I(x)=-\log_2 P(x). \quad (4)$$

由(1)~(2)式和(4)式可以建立用于选取最优结构 S_D 的一种复合成准则, 即 MAP 与 MDL 的集成准则:

$$S_D = \arg \min_{S \in \mathcal{S}} [-\log_2 P(D|S) + L(S)]. \quad (5)$$

2.3.1 计算 $P(D|S)$

在无约束多项分布、参数独立、采用狄利特里先验和数据完整的前提下, 数据的结构似然等于每个结点对于结构似然的乘积^[11,13,15], 即:

$$P(D|S) = \prod_{i=1}^n \prod_{j=1}^{q_i} \left(\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right) = \prod_{i=1}^n \prod_{j=1}^{q_i} C \cdot \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk} + N_{ijk} - 1}, \quad (6)$$

其中, C 为狄利特里分布的常数部分; n 为变量数目; r_i 为变量 X_i 的状态数; q_i 为变量 X_i 父节点的状态数; N_{ij} 为变量 X_i 父结点在第 j 个状态时, 其数据集 D 的案例数量; N_{ijk} 为变量 X_i 取状态为 k 时, 而其父节点取状态为 j 时, 在数据集 D 中的案例数量. N'_{ij} 和 N'_{ijk} 则表示在相应先验数据集中出现的案例数量, 如果没有先验信息, 则可用贝叶斯假设, 设变量各个取值的概率相等, 则有 $N'_{ijk}=1, N'_{ij}=r_i$.

根据狄利特里分布:

$$\hat{\theta}_{ijk} = \frac{N'_{ijk} + N_{ijk}}{N'_{ij} + N_{ij}}, N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}, N_{ij} = \sum_{k=1}^{r_i} N_{ijk}. \quad (7)$$

2.3.2 计算 $L(S)$

MDL 评分是常用的评分函数之一, 可形式化定义为^[11,13]:

$$L(S) = \sum_{i=1}^n [\text{par}(X_i) | \ln(n) + \frac{(r_i-1)}{d} \prod_{X_j \in \text{Par}(X_i)} r_j], \quad (8)$$

其中, $\text{par}(X_i)$ 为变量 X_i 的父节点集; d 为表示存储每个变量值所需的存储单元(常量); r_i 为变量 X_i 的状态数; r_j 为变量 X_j 的状态数.

将(6)~(8)式代入(5)式, 即可求得 S_D 的值, 最终求得最佳网络结构模型如下:

$$S_D = -C \prod_{i=1}^n \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (N_{ijk} + N'_{ijk} - 1) \log_2 \left(\frac{N'_{ijk} + N_{ijk}}{N'_{ij} + N_{ij}} \right) + \sum_{i=1}^n [\text{par}(X_i) | \ln(n) + \frac{(r_i-1)}{d} \prod_{X_j \in \text{Par}(X_i)} r_j]. \quad (9)$$

3 构建模型

3.1 数据选择及离散化

笔者选择 2006~2012 年报考 A 校研究生的 14999 条考生记录作为学习总样本, 并随机抽取 2008 年报考库中的 700 条记录作为抽样验证样本. 根据实际能采集到的样本数据, 选取学生性别、毕业学校、学历层次、报考专业与所学专业的相关度、工作经历等方面对学生的综合能力进行评价. 考虑到统考科目的比重及统考科目中的数学学科在研究生的升学率中扮演着重要的角色, 因此也引入至统考数学及统考课比重的评价指标. 由于统考英语也是影响研究生升学的一个重要因素, 但考虑到英语是每个学生的必考科目, 因此不具有可比性, 所以没有列入指标体系中. 最终变量的定

义及其离散化处理见表 1.

表 1 变量定义及其离散化处理

变量名称	变量定义及离散化处理
政治面貌(X_1)	党员或预备党员为 1, 其它为 2
考生来源(X_2)	应届生为 1, 历届生为 2
性别(X_3)	男为 1, 女为 2
最高学历(X_4)	本科及以上为 1, 本科以下为 2
专业相关度(X_5)	一级学科内相近或相同为 1, 不同为 2
是否考数学(X_6)	统考数学为 1, 不统考数学为 2
统考课比重(X_7)	二门及以下为 1, 三门及以上为 2
毕业学校(X_8)	研究生院及 211 学校为 1, 普通本科为 2
上线情况(X_9)	合格为 1, 不合格为 2

3.2 模型的结构学习

根据前文提出的混合贝叶斯网络学习机制——(9)式, 最终得到的 Bayesian 网络图模型如图 1 所示. 根据模型建立所有节点的参数表(Condition Parameter Table, CPT), 由于对于已知贝叶斯网络结构、样本数据完整下的 Bayesian 网络参数学习方法已经有非常成熟的计算方法^[7-8,12-13], 因此能方便地求得网络的概率参数表(由于求得的 CPT 表数据过长, 在此省略).

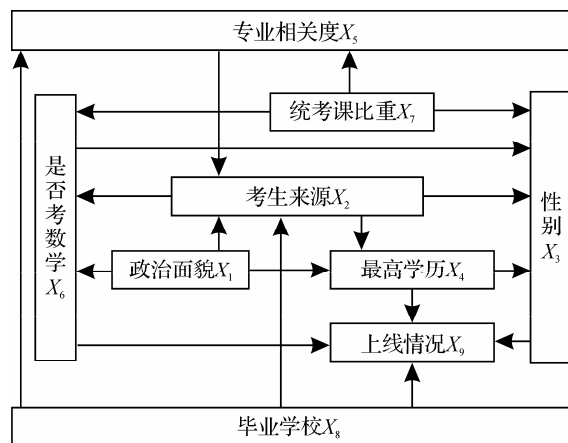


图 1 研究生入学升学预测的 Bayesian 网络模型

4 模型推理及分析

4.1 平均升学率

由图模型以及 CPT 表, 可计算出平均升学率 $P(X_9 = 1) = 0.178$, 其值低于全国平均水平(全国近年来研究生的平均录取率在 0.33 左右). 实际上, A 校报考生源总体质量不高, 每年招生都需要大量

调剂才能满足实际需要, 显得非常被动, 也显然不利于优秀生源的筛选. 究其原因, 一方面 A 校是一个新兴高校, 社会认知度不是太高, 对优秀考生缺乏吸引力; 另一方面也说明 A 校在招生宣传、相关配套政策及措施上还有进一步提升的空间.

4.2 升学率预测

以计算所得的平均升学率作为考生是否上线的判别阈值. 假定考生上线概率 $M > 0.178$, 则认定考生上线可能性很大, 即判定为上线. 表 2 则是模型对 700 条测试样本进行升学率预测的结果统计, 可见模型预测准确率 $= (55 + 532) / 700 = 83.9\%$, 预测效果比较理想, 达到预期目标, 认为可以为 A 校研究生招生提供决策参考.

表 2 升学率结果统计表

结果分类	上线情况	不上线情况
预测值	73	627
实际值	55	532

4.3 决策分析

4.3.1 考生毕业学校对升学率的影响

根据模型及 CPT 表, 可求得 $P(\text{升学}|\text{重点院校毕业}) = 0.171$, $P(\text{上线}|\text{一般院校}) = 0.178$. 从概率上看, 毕业于一般院校的考生总体升学率要略高于重点院校毕业的考生. 这是个非常有趣的结果, 与人们的日常认知不相吻合, 属于一个异常分布. 但其实它与 A 校实际招生情况相一致. 虽然 A 校近年来发展迅猛, 但对优秀生源的吸引力不强, 重点院校的优秀考生一般不会直接报考 A 校, 只能是一些成绩相对一般的考生, 因此上线率也就不高. 而对于一般院校毕业的考生, 却是一些相对优秀的考生前来报考, 而且准备也较为充分, 上线率高也就必然.

另外, 求得 $P(\text{重点院校毕业}|\text{上线}) = 0.071$, $P(\text{一般院校毕业}|\text{上线}) = 0.929$. 即 A 校上线考生中来自于重点院校的考生仅占 7.1%, 而来自普通院校的考生占了 92.9%. 因此, 就目前而言, A 校招生的主体还是普通院校的优秀考生, 学校招生宣传在今后的相应一段时间内, 应该侧重于这个群体, 应出台政策来稳定及吸引这个群体的优秀学生前来报考, 这样才能起到事半功倍的效果.

4.3.2 专业相关度对升学率的影响

同理, 可求得 $P(\text{升学}|\text{专业不相关}) = 0.133$,

$P(\text{升学}|\text{专业相关})=0.191$. 扣除一些需要较强专业背景的专业,如轮机、医学、化学、建筑等以外,专业不相关的考生升学率与专业相关的考生实际升学率从结果来看应该是差距不大.这与A校的大部分招生专业都不需较强的专业背景,且这些专业招生学生又占大多数,因此出现了上述异常概率分布情况.当然这与现今大多数高校实施的“平台+模块”教学,实行大学科教学,淡化专业界限也有一定的关系.反之,我们可以分析上线考生中的专业相关程度,求得 $P(\text{专业相关}|\text{升学})=0.83$,从结果看,上线考生中83%的考生所学专业与所报考的专业相关,只有不到17%的考生所学专业与报考专业不相关.因此在招生中没有必要过多对考生所学专业作限制,反而可以鼓励更多的考生跨专业、跨学科报考,这样可以有利于培养复合型的人才,有利于学科间的融合.

4.3.3 学历与性别对升学率的影响

考察学历对研究生升学的影响,可求得 $P(\text{升学}|\text{学历高})=0.192$, $P(\text{升学}|\text{学历低})=0.067$.学历高的考生升学率明显优于学历低的考生.另外,分析上线考中两者所占比例关系,可求得, $P(\text{学历高}|\text{升学})=0.96$, $P(\text{学历低}|\text{升学})=0.04$,即学历高的考生占升学考生中的96%左右,占绝对多数.以此来看,A校招生政策中也没有必要对考生学历作过多的限制,少数低学历考生的报考不会影响A校研究生招生的总体质量.

另外,可求得概率 $P(\text{升学}|\text{男})=0.168$, $P(\text{升学}|\text{女})=0.188$.从这个概率结果上看,女生升学率要高于男生.而在上线考生中分析, $P(\text{男}|\text{升学})=0.47$, $P(\text{女}|\text{升学})=0.53$.即实际上线的考生中男女生基本保持平衡,女生比例略微高于男生.

4.3.4 其它分析

分析政治面貌对考生升学率的影响,求得 $P(\text{升学}|\text{党员})=0.186$, $P(\text{升学}|\text{非党员})=0.175$,即党员考生的上线率高于非党员上线率.另外,可求得 $P(\text{党员}|\text{升学})=0.34$, $P(\text{非党员}|\text{升学})=0.66$,即上线考生中非党员的考生要比党员考生大32个百分点.A校可以考虑出台相关政策措施吸引学生党员前来报考.

分析统考数学对考生升学率影响,求得 $P(\text{升学}|\text{统考数学})=0.135$, $P(\text{升学}|\text{不统考数学})=0.204$.

$P(\text{统考数学}|\text{升学})=0.28$, $P(\text{不统考数学}|\text{升学})=0.72$.结果发现统考数学对学生的升学率影响较大,这一结果反映了部分教师对统考数学对考生升学率的担忧不无道理,同时也说明了国内大学的数学教学总体质量有待提高.上线考生中统考数学的考生比例不到30%,也说明统考数学迫使一部考生向不统考数学的学科专业转移.

分析统考课比重对升学率的影响,求得 $P(\text{升学}|\text{统考课比重大})=0.203$, $P(\text{升学}|\text{统考课比重小})=0.164$.结果发现统考课比重大的升学率要高于统考课比重小的.这从另一方面说明A校专业课命题偏难,也可能是A校专业课的命题不是非常科学规范,因此,A校应该及时对专业课命题进行全面审视,查验具体问题之所在.

笔者还对其它众多因素进行综合考察,得到了较多的分析结果,限于篇幅不一一例举了.

5 结论与展望

Bayesian网络具有语义清晰,概率推理能力强又易于理解的特点,又可有效地利用先验知识发现数据中隐藏的依赖及因果关系,具有独特的优越性.而基于MAP和MDL混合贝叶斯网络的学习机制更是综合了MAP及MDL的特点,具有计算简单,收敛速度快,能综合利用先验知识及专家知识的优点,具有一定的推广价值.现被成功地应用到研究生招生考试海量数据的挖掘及入学预测模型构建中,验证了算法的可行性和正确性.

另外,该模型的构建也为高校或研究生教育管理部门提供一种良好的分析决策工具,具有一定的实用价值.当然,模型还有待进一步完善及提高之处,如模型变量的定义、选取由于受采集变量的限制还比较粗糙,数据的规模也还不够大,一些变量指标有待更进一步的细化.

今后的工作将进一步扩展指标及分析规模,也可以尝试多种方法建立贝叶斯网络,进行模型间的比较分析,以便选取更加可靠、快速的学习方法,为研究生招生考试海量数据的挖掘、研究生的招生科学决策提供全面的信息支持.

参考文献:

- [1] Geiger D, Verma T, Pearl J. Identifying independence in

- Bayesian networks[J]. Communications of the Association for Computing, 1990, 20:507-534.
- [2] Cooper G. Computational complexity of probabilistic using Bayesian belief Network[J]. Artificial Intelligence, 1990, 42:393-405.
- [3] Cooper G, Herskovits E. A Bayesian method for the induction of Bayesian networks from data[J]. Machine Learning, 1992(9):309-347.
- [4] Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL Principle[J]. Computational Intelligence, 1994, 10(4):269-293.
- [5] Heckerman D, Mamdani A, Wellman M. Real-world applications of Bayesian networks[J]. Communications of ACM, 1995, 38(3):24-26.
- [6] Heckerman D. Bayesian network for data mining[J]. Data Mining and Knowledge Discovery, 1997(1):79-119.
- [7] 林士敏, 田凤占, 陆玉昌. 贝叶斯网络的建造及其在数据挖掘中的应用[J]. 清华大学学报: 自然科学版, 2001, 41(1):49-52.
- [8] Wong M L, Lee S Y, Leung K S. A hybrid approach to discover Bayesian networks from databases using evolutionary programming[C]. Proceedings on the second International Conference Data Mining, 2002:498-505.
- [9] 黄解军. 贝叶斯网络结构学习及其在数据挖掘中的应用研究[D]. 武汉: 武汉大学, 2005.
- [10] Samardinos T I, Brown E, Aleferis C F. The max-min Hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2006, 65(1):31-78.
- [11] 胡学钢, 胡春玲. 一种基于依赖分析的贝叶斯网络结构学习算法[J]. 模式识别与人工智能, 2006, 19(4):445-449.
- [12] 冀俊忠, 刘椿年, 阎静. 一种快速的贝叶斯网络结构学习算法[J]. 计算机研究与发展, 2007, 44(3):412-419.
- [13] 范敏. 基于贝叶斯网络的学习与决策方法研究及应用[D]. 重庆: 重庆大学, 2008.
- [14] 关著华. 基于贝叶斯网数据挖掘若干问题研究[D]. 吉林: 吉林大学, 2009.
- [15] 王越, 谭暑秋, 刘亚辉. 基于互信息的贝叶斯网络结构学习算法[J]. 计算机工程, 2011, 37(7):62-64.
- [16] 王洪泊, 涂序彦. 一种面向最经济服务流的可视化动态贝叶斯网络模型[J]. 电子学报, 2011, 39(6):1331-1335.
- [17] 朱明敏, 刘三阳, 杨有龙. 基于混合方式的贝叶斯网络等价类学习算法[J]. 电子学报, 2013(1):98-104.
- [18] 胡春玲. 贝叶斯网络研究综述[J]. 合肥学院学报: 自然科学版, 2013, 23(1):33-39.

Mass Data Mining and Graduate Prediction Model Construction Based on Cooperative Bayesian Network

YE Shan-wen¹, ZHAO Jie-yu²

(1.Graduate School, Ningbo University, Ningbo 315211, China;

2.Research Institute of Computer Science and Technology, Ningbo University, Ningbo 315211, China)

Abstract: Based on the Maximum Aposterior Probability (MAP) and Minimum Description Length (MDL), this paper presents an algorithm of Bayesian networks structure learning. The algorithm effectively combines the characteristics of two methods, and takes the full advantage of simplistic computation, rapid convergence and priori knowledge and expert system. The experiment with mass data collected from a University graduate entrance examination shows that the model-based algorithm can efficiently be used for prediction and the success rate reaches 84%.

Key words: bayesian networks; cooperative model; data mining; examination predicting

(责任编辑 章踐立)