贝叶斯网络在学生成绩预测中的应用

黄建明
(昆明理工大学计算中心 昆明 650504)

摘 要 学生成绩是高校教学质量评价的重要依据。以学生课程成绩为数据样本,提出一种主干课程贝叶斯网络模型的构建方法。在网络参数学习完备后,对学生相关课程成绩进行了推理预测。结果表明,模型能有效揭示影响课程成绩的关键因素,为决策提供依据。

关键词 贝叶斯网络,结构学习,概率推理,成绩预测
中图法分类号 TP18 **文献标识码** A

Application of Bayesian Network to Predicting Students' Achievement

HUANG Jian-ming
(Computer Center, Kunming University of Science and Technology, Kunming 650504, China)

Abstract Students' achievement is an important basis for the evaluation of teaching quality in universities. This paper presented a construction method for main courses Bayesian networks, which use course score as data sample. After the network parameter learning, we have inference and prediction to course score. The results show that the model can reveal effectively the key factors of which affect the course score, and to provide a basis for decision-making.

Keywords Bayesian network, Structure learning, Probability inference, Achievement prediction

1 引言

贝叶斯网络是一种基于概率论和图论的不确定知识表示和推理模型[1],它是一种有向无环图。经过多年来的研究发展,贝叶斯网络已被广泛应用于模式识别、决策支持、医疗诊断、数据挖掘等领域[2,3]。

贝叶斯网络可以从数据中学习,或者利用专家知识构造。从数据中学习贝叶斯网络可以排除人为因素,真实自然地反映网络节点间的依赖关系。

学生成绩是高校教学质量评价的重要依据。在专业课程体系中,前导课程对后续课程有着重要和直接的影响。本文以往届学生课程成绩作为数据样本,提出一种专业课程体系的贝叶斯网络模型的构建方法。该模型直观地揭示了主干课程间的相互联系,特别是前导课程对后续课程的重要影响。基于该模型对学生后续课程成绩进行了推理预测,结果表明,模型是有效的,有助于找出影响学生课程成绩的关键因素及规律,为教学管理部门提供决策支持信息,促进教学质量的提高。

2 基于贝叶斯网络的学生成绩预测模型

贝叶斯网络的学习包括结构学习和参数学习,而结构学习是贝叶斯网络研究中的热点和难点。贝叶斯网络结构学习方法可分为两大类:基于打分搜索的方法[4]和基于依赖分析的方法[5]。基于打分搜索的结构学习方法就是使用打分函数寻找一个与数据拟合得最好的网络结构,而基于依赖分析的贝叶斯网络结构学习方法是通过节点间的条件互信息来确定是否有边存在。

本文首先以基于信息论的边删除算法[1,6]构造无向图(马尔科夫网络),然后以课程开设的时间顺序给边定向,得到课程依赖关系的贝叶斯网络。学习马尔科夫网络不需要学习其边的方向,效率更高。数据样本采自于学生真实的课程考试成绩。

2.1 学习无向图的边删除算法

设 U 是有限变量集,X,Y,Z 是不相交的子集。算法根据信息论中检验信息独立的重要结论,利用条件互信息来测试条件独立 $I(X,Z,Y)$,即在已知变量 Z 条件下定义变量 X 和 Y 之间的条件互信息 $I(X,Y|Z)$ 为:

$$I(X,Y|Z)=\sum_{I=1}^{I}\sum_{J=1}^{J}\sum_{K=1}^{K}p(x_i,y_j,z_k)\log\frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}$$

给定一个阈值 ε(一个小的正实数),如果 $I(X,Y|Z)\leqslant\varepsilon$,则给定 Z,X 和 Y 条件独立,即 $I(X,Z,Y)$ 成立。否则,给定 Z,X 和 Y 条件不独立。

条件互信息定义中的概率 $p(x_i,y_j,z_k)$,$p(x_i,y_j|z_k)$,$p(x_i|z_k)$,$p(y_j|z_k)$ 可从样本集中经过统计计算得到。

算法描述如下:

1)输入包含 N 个属性的数据样本及阈值 ε;

2)在 N 个属性中,依次测试所有任意两个属性 X 和 Y 之间的条件互信息 $I(X,Y|Z)$,其中 $Z=U-X-Y$;

3)如果 $I(X,Y|Z)>\varepsilon$,则两属性(节点)X 和 Y 之间有边

黄建明(1972-),男,硕士,实验师,主要研究方向为智能信息处理,E-mail:hjmkm@163.com。

存在，反之它们无直接边连接；

4)输出包含所有边的无向图(马尔科夫网络)。

马尔科夫网络是无向图，并不能反应节点间的依赖关系。我们只需要给无向图的边定向，即可得到对应的贝叶斯网络。通常贝叶斯网络的弧定向方法主要有碰撞识别和打分 & 搜索两种[7]。碰撞识别是指依靠 V 结构来对弧进行定向。所谓 V 结构，指的是形如 $A \rightarrow C \leftarrow B$ 的网络局部结构。碰撞识别方法的基本思想是：如果不邻接的两个节点 A 和 B，它们都与节点 C 之间有无向边，且 C 不属于 A 和 B 的切割集，则 C 为碰撞节点，这时确定边的方向为 $A \rightarrow C \leftarrow B$，这种结构就是贝叶斯网络中的 V 结构。

本文以课程开设的时间顺序来给无向边定向，得到课程依赖关系的贝叶斯网络。

2.2 学生成绩预测模型的构造

本文以收集到的学生课程考试成绩作为数据样本，经过清洗、离散化处理及统计分析后，以边删除算法构造无向图，然后以课程开设的先后顺序给边定向，得到用于学生成绩预测的贝叶斯网络模型。

本文的数据样本采自某高校“机械设计制造与自动化”专业 2000 至 2004 级，共 5 个年级 1600 多名学生的课程考试成绩。其中我们挑选了具有专业代表性的 7 门主要课程(《高等数学》、《大学物理》、《理论力学》、《机械原理》、《材料力学》、《机械设计》、《机械制造》)的考试成绩作为数据样本。

表 1 是本文收集到的 1600 多名学生的 7 门课程的考试成绩。

表 1 学生课程考试成绩表

学号	课程名称						
	高等数学	大学物理	理论力学	材料力学	机械原理	机械设计	机械制造
0403108	81	82	82	64	74	90	74
0403109	88	97	82	98	82	94	81
0403110	85	96	78	92	72	84	78
0403111	50	69	48	67	55	68	78
0403112	74	67	35	61	60	70	71
…	…	…	…	…	…	…	…

接着对表 1 的成绩数据进行离散化处理。方法是把每门课程学生的成绩分为两个成绩段：大于等于 70 分的成绩，把它记为“high”；小于 70 分的成绩，把它记为“low”；然后对离散化后的数据记录进行投影处理，获得元组的重复次数，如表 2 所列。

表 2 学生成绩样本统计表

课程名称							人数
高等数学	大学物理	理论力学	材料力学	机械原理	机械设计	机械制造	
low	low	low	low	low	low	low	213
low	low	low	low	low	low	high	30
low	low	low	low	low	high	low	85
low	low	low	low	low	high	high	59
…	…	…	…	…	…	…	…
high	high	high	high	high	high	high	229

从表 2 可以看出，在总共 1604 名学生中，7 门课程成绩都考评为“low”的有 213 人，而 7 门课程成绩都考评为“high”的有 229 人(篇幅所限，表 2 并未列出全面数据)。

下面以表 2 结果作为数据样本，运用边删除算法生成无向图，其中设阈值 $\varepsilon = 0.02$，得到含 9 条边的马尔科夫网络。紧接着以 7 门课程开设的时间顺序给无向边定向，即如果网络中任意两个节点(课程)有无向边相连，则由先开设的课程指向后开设的课程。无向图所有边都定向后，得到其对应的贝叶斯网络。图 1 即是本文构建的课程依赖关系的贝叶斯网络模型。

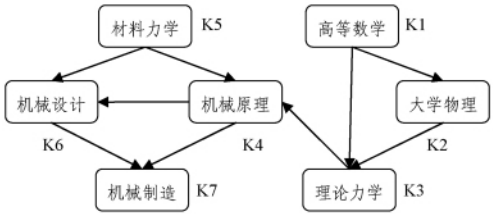

图 1 课程依赖关系的贝叶斯网络

从图 1 中可以看出，7 门课程分成了两大模块，右边部分为基础课程模块，包括《高等数学》、《大学物理》、《理论力学》等课程；而左边部分主要是专业课程模块，包括了《材料力学》、《机械原理》、《机械设计》、《机械制造》等课程。整个网络包括 9 条边。有些课程间没有边连接，是因为它们之间联系的“信息”偏弱，没有前面 9 条边的联系紧密。当然，要在网络中改变边数的多少，只要把阈值 ε 做适当调节即可。

2.3 网络参数的学习

一个完整的贝叶斯网络除了具有网络结构外，还应当包括它的网络参数，即条件概率表(Conditional Probability Table，CPT)。本文根据表 2 的样本数据，以数理统计的方法获得条件概率表。

P(K1=high)	P(K1=low)
0.38	0.62

P(K5=high)	P(K5=low)
0.5	0.5

K1	K2	P(K3=high)	P(K3=low)
high	high	0.71	0.29
high	low	0.23	0.77
low	high	0.35	0.65
low	low	0.11	0.89

K3	K5	P(K4=high)	P(K4=low)
high	high	0.88	0.12
high	low	0.56	0.44
low	high	0.59	0.41
low	low	0.27	0.73

K1	P(K2=high)	P(K2=low)
high	0.75	0.25
low	0.22	0.78

K5	K4	P(K6=high)	P(K6=low)
high	high	0.94	0.06
high	low	0.64	0.36
low	high	0.75	0.25
low	low	0.41	0.59

K4	K6	P(K7=high)	P(K7=low)
high	high	0.75	0.25
high	low	0.34	0.66
low	high	0.45	0.55
low	low	0.14	0.86

图 2 条件概率表

如图 1 所示，为了表述方便，本文用 K1－K7 来依次表示图中的 7 门课程。

例如，图 1 中《材料力学》(K5)和《高等数学》(K1)一样无父节点，那么它的边缘概率就等于《材料力学》成绩为“high”的人数除以总人数。

$$P(K5=\text{high})=\frac{\text{“}K5=\text{high”的人数}}{\text{总人数}}=\frac{802}{1604}=0.5$$

这样，$P(K5=\text{low})=1-P(K5=\text{high})=1-0.5=0.5$。

而对于节点《机械制造》(K7)，其父节点为《机械原理》(K4)和《机械设计》(K6)，那么它的条件概率：

$$P(K7=\text{high}\mid K4=\text{high},K6=\text{high})=\frac{\text{“}K7=\text{high”且“}K4、K6\text{ 都为 high”的人数}}{\text{“}K4、K6\text{ 都为 high”的人数}}=\frac{556}{742}\approx 0.75$$

同理可求得，$P(K7=\text{high}\mid K4=\text{low},K6=\text{high})\approx 0.45$；$P(K7=\text{high}\mid K4=\text{low},K6=\text{low})\approx 0.14$；$P(K7=\text{high}\mid K4=\text{high},K6=\text{low})\approx 0.34$。

以同样的方法，可依次求出其它节点的条件概率。图 2 就是本文学习到的图 1 网络模型的条件概率表(CPT)。

3 基于贝叶斯网络推理的学生成绩预测

本文以贝叶斯网络推理的似然加权算法[8]来进行学生成绩的仿真估计与预测。

3.1 似然加权算法

似然加权算法根据贝叶斯网络的拓扑顺序对网络中的节点逐个进行采样：当 X 是证据变量时，不采样直接取证据值；当 X 是非证据变量时，如果它没有父节点，则按先验概率 $P(X)$进行采样；如果它有父节点，则根据其父节点的采样结果和该节点的条件概率 $P(X\mid parents(X))$进行采样。似然加权算法产生的每一个样本都与证据值一致，不会造成样本的浪费。这样产生的样本都是有效的，避免了逻辑抽样算法的不足。

似然加权算法给证据事件的发生赋予一个权值，这个权值由每个证据变量在给定其父节点取值下的条件概率的乘积得到。

$$w=\prod_{i=1}^{m}P(E=e\mid parents(E))$$

把每个样本的权值 w 累加到 w_e 中，而把符合查询变量值 $Q=q$ 的样本的权值 w 累加到 w_{qe} 中。则：

$$P(Q=q\mid E=e)\approx\frac{w_{qe}}{w_e}$$

算法描述如下：

1)输入证据变量 E 及其取值 e、查询变量 Q 及其取值 q，样本大小 m；

2)权值 $w\leftarrow 1$；

3)按网络的拓扑顺序对每个节点 x 依次采样：如果该节点是证据节点，则直接取证据值 e，且 $w=w\times P(E=e\mid parents(E))$；如果不是证据节点，则根据 $P(X\mid parents(X))$进行采样；

4)$w_e\leftarrow w_e+w$，如果样本符合 $Q=q$，则 $w_{qe}\leftarrow w_{qe}+w$；

5)返回 2)开始下一轮的采样，直至样本数等于 m；

输出 $P(Q=q\mid E=e)=\frac{w_{qe}}{w_e}$。

3.2 成绩预测实例

这里以预测 $P(K7=\text{high}\mid K1=\text{low})$的值为例，分析成绩预测的过程和结果。实验中采样次数 M 分别取 2000、5000、10000 3 种情况生成样本，不同 M 取值下分别重复执行 10 次，得到 10 组样本。

图 3 为 3 种不同采样次数下 10 组样本中概率值的变化情况。从图中可以看出，当采样次数 M 取 2000 次时，概率值波动很大，而当采样次数逐步取大时(如 M 取 10000 次)，概率值的变化趋于稳定，并最终收敛于 48.7%，即 $P(K7=\text{high}\mid K1=\text{low})\approx 0.487$。

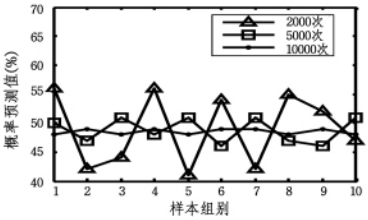

图 3 不同采样次数下概率值变化图

这一结果表示：如果已知一个学生的《高等数学》分数不高(小于 70 分)，则他的《机械制造》成绩为高分(大于等于 70 分)的概率为 48.7%，而分数为低分的概率为 51.3%。

同样的方法可对图 1 中的任意一门或几门课程成绩进行预测。例如已知某学生《机械原理》得分不高，同时已知他的《机械设计》得到了高分，那么他的《大学物理》得高分的概率为：$P(K2=\text{high}\mid K4=\text{low},K6=\text{high})\approx 0.32$，即 32%。

结束语 本文提出一种以学生课程成绩为数据样本，构造用于学生成绩预测的贝叶斯网络模型的方法。采用贝叶斯网络推理的似然加权算法，举例说明了成绩预测的方法和过程。结果表明该网络模型及其构造方法是有效的，为教学管理部门进行科学决策提供了依据，具有较好的应用价值和指导意义。需要说明的是，本文收集的数据样本不是足够大，而为了保证网络的准确性和不失真，本文中的模型并没有包涵更多的课程。当然，只要数据样本足够大，涵盖大部分课程的贝叶斯网络模型是可以构建的，这样更多门课程的成绩预测即可实现。

参 考 文 献

[1] Pear J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. San Mateo: Morgan Kaufmann Publishers, 1988

[2] Heckerman D. Bayesian networks for data mining [J]. Data Mining and Knowledge Discovery, 1997, 1: 79-119

[3] 赵进晓，肖飞. 一种基于贝叶斯网络的模型诊断方法[J]. 计算机科学，2009，36(1)：291-295

[4] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data [J]. Machine Learning, 1992, 9: 309-347

[5] Chen J, Bell D, Liu W. Learning Bayesian networks from data: An efficient approach based on information theory[J]. Artificial Intelligence, 2002, 137(1/2): 43-90

[6] 何盈捷，刘惟一. 由 Markov 网到 Bayesian 网[J]. 计算机研究与发展，2002，39(1)：87-99

[7] 王双成，苑森淼. 具有丢失数据的贝叶斯网络结构学习研究[J]. 软件学报，2004，15(7)：1042-1048

[8] Russell S, Norvig P. Artificial Intelligence: A Modern Approach (2nd Edition)[M]. New Jersey: Prentice Hall, 2003