利用人工智能案例推动概率统计课程教学

李超群[1]， 张玉洁[1]， 蒋良孝[2]
(1. 中国地质大学(武汉) 数学与物理学院, 武汉 430074； 2. 中国地质大学(武汉) 计算机学院, 武汉 430074)

[摘　要] 人工智能的迅猛发展离不开概率统计的支持,在概率统计的日常教学中,融入人工智能案例的介绍,能够促进学生的学习兴趣,提高教学质量. 以举例的方式说明了概率统计的理论知识和人工智能技术的结合,为利用人工智能案例推动概率统计的课程教学提供了示例.

[关键词] 概率统计；熵；分类；决策树
[中图分类号] O211.1　[文献标识码] C　[文章编号] 1672-1454(2020)04-0043-05

1 引　　言

近年来,人工智能的迅猛发展正在深刻地改变人类社会生活,改变世界. 为了贯彻落实《国务院关于印发新一代人工智能发展规划的通知》(国发[2017]35 号),教育部最新印发了《高等学校人工智能创新行动计划》,要求高校不断推动人工智能与教育深度融合,引领我国人工智能科技创新、人才培养和技术应用示范,带动我国人工智能总体实力的提升.

为了配合教育部提出的"高校不断推动人工智能与教育深度融合"这一目标,作为数学基础课部所要做的,就是推动与人工智能密切相关的基础课教育的改革和实践. 人工智能离不开数学学科,特别是概率论与数理统计[1]的支持. 人工智能的重要并非只面向信息、计算机等专业的学生. 非信息专业的学生,也要面临对数据的处理、应用和挖掘,不可避免地要接触和应用到人工智能技术. 概率论与数理统计是高等院校理工科学生的必修课. 然而,因为大量的理论知识与相对落后的应用实例,使得学生在学习该课程时仍然感到枯燥无味,动力不足. 比如概率论与数理统计的教材中,采用的仍旧是"摸球问题"、"产品检验问题"、"扔硬币问题"等作为应用实例. 以这些问题作为应用背景,学生很难窥探到这门课程与人工智能的紧密联系.

关于概率论与数理统计课程的教学改革,一些学者已经展开过研究,包括在概率统计课程中研究性学习方法的探讨[2],在概率论课程中引入建模思想[3],基于学生创新能力培养的概率统计课程教学改革与实践[4],利用 R 软件进行辅助性教学[5]等. 笔者长期面向本科生讲授概率论与数理统计课程,同时从事人工智能方向的科研工作十余年之久,对概率论与数理统计在人工智能方向的应用有较深的理解和认识. 因此笔者试图在课程教学中,结合人工智能的一些关键技术,介绍概率论与数理统计在人工智能中的应用,借助人工智能的热潮推动课程的改革,使学生认清"枯燥"的理论知识如何推动了炙手可热的人工智能的发展.

下面笔者就利用人工智能中的一个重要算法:决策树算法,展示概率论与数理统计知识在这个算法模型中的作用. 通过对这种算法的介绍,会让学生在概率统计的课堂上,一方面能认识到概率统计的理论知识与人工智能技术的紧密结合,调动学生的学习热情;另一方面又能从侧面了解一些人工智能技

[收稿日期] 2019-09-02； [修改日期] 2019-11-11
[基金项目] 中国地质大学(武汉)教学研究基金资助项目(2018A55, 2018A30);中央高校基本科研业务费(CUG2018JM18)
[作者简介] 李超群(1981—),女,博士,副教授,从事大学数学教学与研究. Email: chqli@cug.edu.cn

术，丰富了他们的知识结构和知识面.

2 一个分类问题

分类是人工智能的一个常见任务. 比如人工智能在辅助医疗中的一个应用就是，利用人工智能技术对病人是否患有某种疾病进行分类. 再比如借助人工智能技术对网络新闻进行分类（分为政治新闻、财经新闻、娱乐新闻等），在分类的基础上再推荐给不同的读者.

下面举一个简单的分类问题的例子. 表 1 是一个隐形眼镜的镜片推荐的数据集[6]. 每个实例由四个特征属性和一个类属性来描述，四个特征属性是“年龄”、“视力诊断”、“散光”和“泪流量”，这四个特征属性用于描述一个人的视力状况. 类属性为“推荐镜片”，类属性也称为目标属性，这个数据集中“推荐镜片”取值为 none，soft 和 hard 三种情况. 分类问题所需要做的，就是在这样一个类属性取值已知的训练数据集上建立一个分类器，然后用这个分类器来预测类属性未知的实例，即对下一位顾客，直接利用这个分类器来决定这位顾客应该配哪种类型的隐形眼镜.

决策树算法面对这样一个分类问题，就是要构造一颗树状的分类器. 比如决策树算法中著名的 C4.5 算法[7]，构造出的决策树如图 1 所示.

表 1 隐形眼镜数据

年龄	视力诊断	散光	泪流量	推荐镜片
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

在图 1 中，特征属性“泪流量”所处的最顶层的结点称为树的根结点. 最下面一层的以椭圆形作标识的结点为叶子结点，其余的称为内部结点. 对需要预测的每一个目标属性未知的实例，这个实例将根据它的特征属性的取值，沿着决策树的某一个分支落入对应的叶子结点，将该叶子结点的目标值赋给该实例.

决策树算法因为这样一颗树的结构，因此有非常清晰的分类规则，可解释性很强，是应用最为广泛的归纳推理算法之一. 为了生成这样的一棵决策树，一个最基本的问题是，哪一个属性在树的根结点被测试？为了回答这个问题，我们必须要先学习概率论中关于熵的概念.

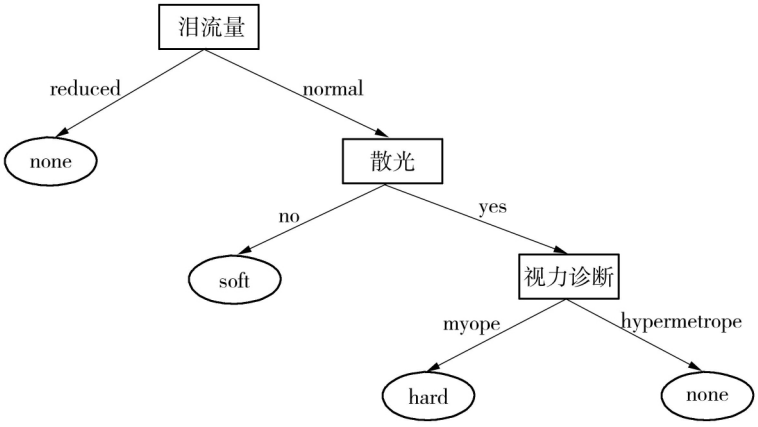

图 1　目标“推荐镜片”的 C4.5 决策树

3　熵的定义

熵的提出是为了度量随机试验的不确定性程度. 随机试验的主要特征是在试验之前无法确定地知道哪一个结果将会出现，即随机试验具有一种不确定性程度. 比如有一个问题，甲、乙、丙三个射手进行射击，甲每次射击，命中和未命中的概率各为 0.5；乙每次射击，命中和未命中的概率分别为 0.99 和 0.01；丙每次射击，命中和未命中的概率分别为 0.7 和 0.3. 显然在这三个试验中，甲的不确定性程度最大，乙的不确定性程度最小，丙的介于两者之间. 现在要用一个量来从数值上刻画这三个随机试验的不确定性程度，而这样的一个量就是美国数学家申农找到的“熵”[7].

定义 1（熵）　一个随机试验 α 有有限个不相容的结果 $A_1, A_2, \cdots, A_n$，它们的相应概率为 $\mathrm{P}(A_1)$，$\mathrm{P}(A_2), \cdots, \mathrm{P}(A_n)$，满足 $\sum_{i=1}^{n}\mathrm{P}(A_i) = 1$. 则试验 α 的熵定义为

$$Entropy(\alpha) = -\sum_{i=1}^{n}\mathrm{P}(A_i)\log\,\mathrm{P}(A_i).$$

按照定义 1 进行计算，上面射击例子中，对数计算以 10 为底，可得甲、乙、丙三个射手进行射击的随机试验的熵分别为 0.3010，0.0243 和 0.2653. 这个结果与直观完全吻合.

4　决策树构建中熵的应用

4.1　熵的第一次应用

现在，回到决策树算法的构建上来，哪一个属性在树的根结点被测试？或者说选用什么样的准则来度量属性的分类能力？为此，就需要用到熵的概念. 设 S 为训练样例集，它的类属性有 k 个取值，则这个样例集的熵定义为

$$Entropy(S) = -\sum_{i=1}^{k} P_i \log_2 P_i,$$

其中，P_i 为第 i 类样本所占的比例. 这里的对数取以 2 为底数.

为什么在决策树的算法构建中要介绍熵的概念？因为要用熵来刻画训练样例集的纯度. 举个简单的例子，当类属性只有两个取值时，即样例只分为正样例和负样例两类时，样例集的熵为

$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-,$$

其中，P_+ 和 P_- 分别是正样例和负样例所占的比例. 当 S 中只有正样例或者只有负样例时（即样例集是

纯的),则 $Entropy(S)=0$. 当正负样例各占一半时(即样例集的不纯度最高),则 $Entropy(S)=1$. 当正负样例的数量不相等时,熵介于 0 和 1 之间. 显然可以这样理解,即随机地从样例集中抽出一个样例,这是一个随机试验,P_+ 就表示正样例被抽到的概率,P_- 表示负样例被抽到的概率. 样例集中某一类样例的比例越大(即样例集越纯),则随机试验的不确定性程度越小,相应的熵值越小. 反之,两类样例的比例越接近(即样例集越不纯),则随机试验的不确定性程度越大,相应的熵值越大. 因此熵的大小就直接反应了样例集的纯度. 熵越小,样例集越纯.

利用熵来刻画训练样例集的纯度,基于熵就有了"信息增益"(Information gain)这个标准,信息增益被用来度量属性分裂训练数据的能力. 信息增益如下定义.

定义 2(信息增益) 一个属性 A 相对样例集合 S 的信息增益 $Gain(S,A)$ 为

$$Gain(S,A)=Entropy(S)-\sum_{v\in Values(A)}\frac{|S_v|}{|S|}Entropy(S_v),$$

其中,$Values(A)$ 为属性 A 的所有可能取值的集合,S_v 为 S 中属性 A 的值为 v 的子集. 那么上式的右边第一项就是原集合 S 的熵,第二项是用属性 A 分类样例集 S 之后的熵的期望值,显然,属性 A 分类样例集的能力越高,则经属性 A 的不同取值分裂之后的子集 S_v 的纯度就越高,因此等式右边第二项的值就会越小. 因此 $Gain(S,A)$ 是由于知道了 A 的值而导致的期望熵的减少. 这种减少越多越好,因此信息增益最大的那个属性将被用在根结点用于分裂数据集.

以表 1 的眼镜数据集为例,样例集 S 包含了 24 个样例,其中 15 个样例的目标函数为 none,5 个的目标函数为 soft,4 个的目标函数为 hard,记 $S=[15\text{none},\ 5\text{soft},\ 4\text{hard}]$. 而特征属性"视力诊断"有两个取值,一个是 myope,一个是 hypermetrope. "视力诊断"为 myope 的样例一共有 12 个,对应的目标函数 7 个为 none,2 个为 soft,3 个为 hard,记 $S_1=[7\text{none},\ 2\text{soft},\ 3\text{hard}]$. "视力诊断"为 hypermetrope 的样例也有 12 个,对应的目标函数 8 个为 none,3 个为 soft,1 个为 hard,记 $S_2=[8\text{none},\ 3\text{soft},\ 1\text{hard}]$. 则 S,S_1,S_2 的熵为

$$Entropy(S)=-\frac{15}{24}\log_2\frac{15}{24}-\frac{5}{24}\log_2\frac{5}{24}-\frac{4}{24}\log_2\frac{4}{24}=1.326;$$

$$Entropy(S_1)=-\frac{7}{12}\log_2\frac{7}{12}-\frac{2}{12}\log_2\frac{2}{12}-\frac{3}{12}\log_2\frac{3}{12}=1.384;$$

$$Entropy(S_2)=-\frac{8}{12}\log_2\frac{8}{12}-\frac{3}{12}\log_2\frac{3}{12}-\frac{1}{12}\log_2\frac{1}{12}=1.189.$$

按照属性"视力诊断"分类 24 个样例得到的信息增益计算如下:

$$\begin{aligned}Gain(S,\text{视力诊断})&=Entropy(S)-\frac{|S_1|}{|S|}Entropy(S_1)-\frac{|S_2|}{|S|}Entropy(S_2)\\&=1.326-(1/2)1.384-(1/2)1.189=0.040.\end{aligned}$$

同样可以计算出另外三个属性的信息增益为

$$Gain(S,\text{年龄})=0.039,\quad Gain(S,\text{散光})=0.377,\quad Gain(S,\text{泪流量})=0.548.$$

4.2 熵的第二次应用

信息增益可以作为决策树的一个属性分裂的度量标准,但是信息增益有一个内在偏置,它偏袒有较多值的属性. 就是说一个属性的取值个数越多,相应的用这个属性对训练样例进行分类,得到的信息增益越高. 但是用这样的属性生成的决策树往往对后来数据的预测性能会很差,原因在于取值很多的属性会把训练样例分割成非常小的空间,因此对于训练样例会有很高的信息增益,但对于未见的实例却是一个非常差的目标函数预测器.

为了避免信息增益的不足,另一个度量标准增益比率(Gain Ratio)被提出. 增益比率在信息增益的基础上加入了一个分裂信息(split information)项. 分裂信息项定义如下:

定义 3(分裂信息) 一个属性 A 相对样例集合 S 的分裂信息 $SplitInformation(S,A)$ 为

$$SplitInformation(S,A)=-\sum_{i=1}^{c}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}.$$

其中,S_1 到 S_c 是有 c 个取值的属性 A 分割 S 而形成的 c 个样例子集. 可以看到,这个分裂信息其实也是

一个熵，就是 S 关于属性 A 的各个取值的熵. 就是说，样例集 S 根据属性 A 的取值分成了 S_1 到 S_c 的 c 个样例子集，随机地从样例集 S 中取一个样例，则这个样例来自于第 i 个样例子集 S_i 的概率为 $|S_i|/|S|$，显然如果一个属性的取值个数越多，越不能确定这个样例是来自于哪个样例子集，因此随机试验的不确定程度越大，由此该属性的分裂信息也越大.

在本文的眼镜数据集中，“年龄”这个属性有三个取值，每个取值对应有 8 个样例，因此年龄这个属性的分裂信息为

$$SplitInformation(S,\text{年龄}) = -3(1/3)\log_2(1/3) = 1.585.$$

剩下的三个属性，每个属性都是两个取值，每个取值均对应 12 个样例，因此这三个属性的分裂信息都是 1.

最终增益比率使用增益度量和分裂信息项共同定义得到.

定义 4（增益比率）　一个属性 A 相对样例集合 S 的增益比率 $GainRatio(S,A)$ 为

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}.$$

这里可以看到，增益比率的定义用到了两次熵的概念. 第一次是用熵的概念定义了信息增益，第二次是利用熵的概念定义了分裂信息. 在增益比率的定义中，可以充分看到熵这个概念所发挥的至关重要的作用.

最终，眼镜数据集中，年龄、视力诊断、散光和泪流量这四个属性的增益比率分别为 0.025，0.039，0.377 和 0.548. 后三个属性因为分裂信息都是 1，所以增益比率和信息增益是相同的.

使用增益比率代替信息增益来选择属性，则增益比率越大的属性越好. 但这样做也有一个缺陷，因为当样例集在某一个属性上的取值几乎相同时，即 $|S_i| \approx |S|$ 时，则分裂信息会趋于 0，因此导致的增益比率会非常高. 但这样的属性对分类也是无用的. 为了避免选择到这样的属性，C4.5 算法采用的是一个启发式的规则，即先计算每个属性的信息增益，对增益高过平均值的属性再应用增益比率来测试. 利用这样的准则，最终可以看到，泪流量这个属性，不论是信息增益，还是增益比率都是最高的. 因此被选择在树的根结点，并为它的两个可能取值（reduced，normal）在根结点下创建分支. 得到的部分决策树如图 2 所示. 因为泪流量为 reduced 的样例其目标函数值都是 none，因此这个结点成为一个叶子结点. 而对泪流量为 normal 的分支上的子样例，将在这个子样例集上重复上面的过程，选择新的属性继续对训练样例进行分裂. 最终得到了图 1 所示的决策树（这棵树的建成还有一些细节问题，在这里不再详细讨论）.

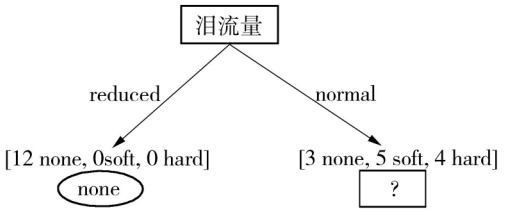

图 2　在根结点处形成的部分决策树

5　结　　论

至此，关于 C4.5 算法的构建就已经完成. 可以看到，在这个过程中，“熵”这个概念起了至关重要的作用. 可以说，没有熵，就没有决策树算法 C4.5. 本文所举的这个分类问题的例子很简单，也容易解释，并不需要占用很多的课堂时间. 但通过这样一个举例，就能够让学生了解枯燥的理论知识是如何成就了这样一个知名的分类算法，可以用于解决现实生活中的一大类分类问题. 通过这样一种将理论学习与炙手可热的人工智能技术相结合的方式，能够极大地提高学生的学习兴趣，提升教学质量. 笔者在课堂上向学生演示了这个例子，学生显示出了极大的兴趣，表示这种从理论到实际应用的展示，特别是与当前热门的人工智能的应用相结合的展示，使他们真正看到了数学的“学有所用”和“学以致用”.

但在实践中，碰到的一个最大的困难在于学时的有限. 现在高校的许多课程相较之前课时都有所削减，在较之以前更少的教学学时的情形之下，教学内容没有减少，如果还要增加人工智能方向的应用案例，则更加需要谨慎地选择和考量. 概率论的课程里面可以和人工智能的应用案例相结合的知识点很多，选取哪些有代表性的知识点，和有代表性的人工智能技术，是需要教师反复斟酌的. 比如概率论的

“全概率公式”和“贝叶斯公式”是一个重要的知识点，同时它们也与人工智能方向的“贝叶斯分类器”有紧密的联系，可以在讲完全概率公式和贝叶斯公式之后，给学生简单地介绍一下朴素贝叶斯分类器. 为了平衡教学内容和学时的矛盾，笔者的做法和建议是，在习题课上，介绍这种热点应用，一定比例地缩减传统的求解性习题，可以在不减少教学内容，保证教学质量的情况下，极大地提升学生的学习兴趣.

当然，不仅仅是概率统计知识，其实人工智能的发展离不开各种各样的数学知识. 比如著名的 BP 神经网络的训练，就需要依靠梯度下降法. 那么在高等数学课堂上，可以在学习梯度这个概念时，结合 BP 神经网络进行介绍. 而卷积神经网络(CNN)的训练，离不开矩阵的运算. 因此在线性代数的课堂上，讲矩阵的运算时，可以给学生初步介绍一下卷积神经网络. 特别是线性代数课程，许多学生觉得比较枯燥. 但是一幅图像在数学上的形式就是一个矩阵，图像处理、图像分类等问题在数学上都需要对矩阵进行操作. 因此教师也可以借助图像处理的相关内容在线性代数的课堂上简单地介绍矩阵的相关知识与图像处理的结合. 总之，现在高等院校的教师都有自己的科研方向，都可以根据自己的知识储备，将数学的知识点的讲解和各种实际问题结合起来，让学生感觉到学有所用，学以致用. 这样能极大地提到学生的学习兴趣，促进他们学习的主观能动性，提升课堂的教学质量.

致谢 作者非常感谢相关文献对本文的启发以及审稿专家提出的宝贵意见.

[参 考 文 献]

[1] 盛骤，谢式千，潘承毅. 概率论与数理统计[M]. 北京：高等教育出版社，2008.

[2] 边家文，付丽华，彭惠明，陆建华，邢婧，方秉武. 概率统计课程中研究性学习方法探讨[J]. 大学数学，2012，28(2)：11－15.

[3] 丁海峰. 建模思想在概率论教学中的应用意义研究[J]. 黑龙江教育(理论与实践)，2018，1263(11)：65－67.

[4] 陈绍刚，黄廷祝. 基于学生创新能力培养的概率统计课程教学改革与实践[J]. 大学数学，2018，34(2)：53－57.

[5] 曹丽，张莉. 基于R的概率统计直观教学展示[J]. 大学数学，2017，33(4)：86－89.

[6] Ian H W，Eibe F，Mark A H，Christopher J P. 数据挖掘：实用机器学习工具与技术[M]. 北京：机械工业出版社，2014.

[7] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

Using Artificial Intelligence Cases to Lift Probability and Statistic Course Teaching

LI Chao-qun[1]， *ZHANG Yu-jie*[1]， *JIANG Liang-xiao*[2]

(1. School of Mathematics and Physics，China University of Geosciences，Wuhan 430074，China；
2. School of Computer，China University of Geosciences，Wuhan 430074，China)

Abstract：The rapid development of artificial intelligence is inseparable from the support of probability and statistics. In the course teaching of probability and statistics，by introducing the cases of artificial intelligence，we can promote students′ interest in learning and lift the quality of the course teaching. This paper gives some examples to show how to bridge the theory knowledge of probability and statistics and the artificial intelligence cases.

Key words：probability and statistics；entropy；classification；decision trees