“机器学习”案例化教学实践与探讨

——以贝叶斯分类器为例

范彦勤[a],覃杨森[b],史旭明[a],袁 媛[a]

(桂林航天工业学院a.理学院 b.计算机科学与工程学院,广西 桂林 541004)

[摘 要]《机器学习》教学内容理论深、算法多、难度大,难以理解,造成学习兴趣难以维持和提高,采用案例化的教学方法是改善这一困境的有益尝试。该过程可让学生从实际场景入手,由浅入深,逐步引导学生解决问题,既巩固已学理论知识,又让学生掌握新课程内容,激发学生的积极性和参与度。给出一个贝叶斯分类器案例教学过程,实践证明,该方法能够有效地帮助学生掌握贝叶斯分类器的分类过程及实际应用现状,并为他们以后的工作打下基础。

[关键词] 机器学习;案例教学;贝叶斯分类器

[基金项目] 2019年度桂林航天工业学院教改项目“大数据背景下机器学习课程建设研究”(2019JB28)

[作者简介] 范彦勤(1988—),女,河南驻马店人,硕士,桂林航天工业学院理学院助教,研究方向为贝叶斯网络及概率图模型;覃杨森(1986—),男(壮族),广西来宾人,硕士,桂林航天工业学院计算机科学与工程学院工程师,研究方向为计算机软件技术。

[中图分类号] G642 [文献标识码] A [文章编号] 1674-9324(2020)43-0109-02 [收稿日期] 2020-03-23

机器学习作为人工智能发展最快的分支之一,其理论和方法已被广泛应用于各领域。它是一门多领域交叉学科,学习算法多而抽象,不易理解掌握,容易导致教学理论与实践脱节[1-2]。同时大部分教材及参考书,包括很多的教学过程,重在抽象概念和课程理论的讲解,缺乏结合案例。

一、原理教学设计

(一) 贝叶斯分类器的信用评估原理

贝叶斯分类器[3-4]的信用评估原理是在个人信用的先验概率基础上,利用贝叶斯公式计算出类别的后验概率,将后验概率最大的类作为信用评估所属的类。

设$U=\{X,C\}$是样本数据集,其中$X=\{X_1X_2,\cdots,X_n\}$是样本数据的指标变量集,C是信用评估的类变量,分类结果为:{$c_1=0$,表示信用好;$c_2=1$,表示信用坏},x_i是属性X_i的取值。样本$x_i=(x_1,x_2,\cdots,x_n)$属于c_j,$i=1,2.$的概率,由贝叶斯定理及概率的链式法则可表示为:

$$P(C=c_j|X=x)=\frac{P(c_j)\cdot P(x_1,x_2,\cdots,x_n|c_j)}{P(x_1,x_2,\cdots x_n)}=aP(c_j)\cdot P(x_1,x_2,\cdots,x_n|c_j)=aP(c_j)\prod_{i=1}^{n}P(x_i|x_1,\cdots,x_{i-1},c_j)\quad(1)$$

其中:a是正则化因子;$P(c_j)$是类c_j的先验概率;$P(x_1,x_2,\cdots,x_n|c_j)$是类$c_j$关于$x_i$的似然。

给定信用数据样本集D,分类任务的目标是对D进行分析,根据贝叶斯网络的信用评估原理,可知贝叶斯信用评估的关键是计算出$P(x_i|x_1,\cdots,x_{i-1},c_j)$。

(二) 构造朴素贝叶斯分类器(NB)

朴素贝叶斯分类器是最简单的贝叶斯分类器,具有高效率和良好分类精度的优点。该分类器将类变量作为父节点,属性变量作为子节点,并假定子节点之间相互独立。

(三) 结合案例建立个人信用评估模型

1.采集获取个人信用评估数据,对其预处理。

2.利用训练样本在NB分类器上构建模型,具体如下:(1) 根据贝叶斯分类器结构学习算法,得到最优贝叶斯网络结构;(2) 学习贝叶斯分类器的条件概率分布。

对于朴素贝叶斯分类器(NB),由于该分类器的网络图中各指标变量间是相互独立的,则彼此之间不再需要进一步的结构学习,只需要估计出指标变量的条件概率即可。一般采用频率估计的方法对离散型指标变量进行估计;对于连续属性变量一般把正态分布设定为其分布函数。

3.分类测试集。基于已训练好的贝叶斯分类模型进行分类。

4.输出分类结果。

二、案例应用

(一) 样本数据收集

给定数据为UCI[5]上德国标准个人信贷数据,该数据库主要用于评估个人信用的研究。该样本数据共1000条,信用结果分好和坏两种,其中评估结果为好客户700条,坏客户300条。同时每个样本中有21个变量,20个为属性指标变量,1个为类变量。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（二）数据预处理

样本数据20个属性变量中2-5-8-11-13-16-18为连续属性，其余为离散属性。属性变量类型不统一，使用前需将数据全部转化为离散型。处理方法：对于离散型数据，保留其实际数值；对于连续型数据，需将其转化成离散型数据。此外，需对各指标数据进行标准化处理。

三、应用分析与总结

模型实现方法比较多，大家可以使用Python，C，MATLAB等编程语言，下面我们提供具体算法如下：

1.输入训练样本，定义类变量和属性变量。

2.参数学习。观测所有训练样本：首先给定$P(C_k)=\hat{P}(C_k)=\frac{N_{C_k}}{N}$（其中$N_{C_k}$为第$C_k$类中记录的个数）(2)。

①当X_i是离散型，$P(X_i=x_i|C_k)=\hat{P}(X_i=x_i|C_k)=\frac{N_{C_k}^{(x_i)}}{N_{C_k}}$（其中$N_{C_k}^{(x_i)}$为第$C_k$类中$X_i=x_i$的情况数量）(3)。

②当X_i是连续型，假设其服从正态分布。则第C_k类的样本均值和样本方差如下：

$$P(X_i=x_i|C)=g(x_i,\mu_{C_k},\sigma_{C_k})=\frac{1}{\sqrt{2\pi\sigma_{C_k}}}e^{\frac{(x_i-\mu_{x_k})^2}{2\sigma^2}} \quad (4)$$

$$\sigma_{C_k}=\frac{\sum_{j=1}^{N}(x_{ji}(c_k)-uc_k(c_k))^2}{N_{Ck}} \quad (5)$$

3.分类测试。

我们采用五折交叉验证方法进行计算，用Matlab编制并运行了NB的分类程序，各类的先验概率按训练样本中的各类占训练样本总数的比例计算。其中定义第一类错误为将坏客户错判为好客户的比率；第二类错误为将好客户错判为坏客户的比率，总分类错误为总分类错判人数占样本总人数的比率。最终实现分类结果为：第一类错判比率0.1357；第二类错判比率0.4833；总错判比率0.2400。通过引入信用评估案例，学生学习主动性高，结合之前所学的编程及建模思想，较好的掌握了贝叶斯分类器的分类应用，可有效推广至其他应用领域中。

四、结束语

机器学习作为一门数学理论深且实践操作难的课程，如何将抽象、枯燥的理论知识简单有效地传授给学生，显得尤为重要。本文以教学目标为出发点，将实际应用案例引入机器学习教学中，可以让学生解决实际问题的过程中。

参考文献

[1] 李勇.本科机器学习课程教改实践与探索[J].计算机教育,2015(13):63-66.

[2] 闵锋,鲁统伟.《机器学习》课程教学探索与实践[J].教育教学论坛,2014(53):158-159.

[3] 周志华.机器学习[M].北京:清华大学出版社,2018:13-16.

[4] 张连文,郭海鹏.贝叶斯网引论[M].北京:科技出版社,2006:80-85.

[5] Asuncion A,Newman D J.UCI Repository of Machine Learning Databases [DB/OL] [http://www.ics.uci.edu/~mlearn/MLRepository.html].Irvine,CA:University of California,Department of Information and Computer Science,2007.

Case-based Teaching Practice and Discussion on Machine Learning:Taking Bayesian Classifier as an Example

FAN Yan-qin[a],TAN Yang-sen[b],SHI Xu-ming[a],YUAN Yuan[a]

(a.College of Science,b.College of Computer Science and Engineering,Guilin Institute of Aerospace Technology,Guilin,Guangxi 541004,China)

Abstract:The teaching content of Machine Learning course involves various deep and difficult theory,especially many algorithms,which makes it difficult for students to understand and improve learning interest.Case teaching method is a more effective way for this course.It enables students to start from the actual case,helps step by step,and gradually guides students to solve problems.It not only consolidates what they have learned,but also helps students to master new contents and stimulates students' enthusiasm and participation.A Bayesian classifier case is applied in this course teaching.It has been proved that it effectively helps students master the classification process and practical application of Bayesian classifiers,and lays a foundation for future work.

Key words:Machine Learning;teaching case;Bayesian classifier

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net