朴素贝叶斯分类算法在大学生体质分析中的应用

杜云梅，刘东
（广州商学院 信息技术与工程学院，广东 广州 510363）

摘 要：基于大数据对大学生体质进行分类预测，有助于大学体育治理体系的建设，朴素贝叶斯模型是一种操作简单且性能较好的机器学习分类算法。基于朴素贝叶斯分类算法，采用广州商学院 2014、2015 年学生体测数据及其评分结果作为源数据，构建大学生体质分类器。应用此分类器可对大学生的体质状况实现一定概率意义上正确的判断，从而可以对体质存在隐患概率比较大的学生给出主动性预警，以便大学体育对学生进行群体性的体质判断、进行个性化的有效干预，从而促进学生健康发展，提高大学生整体体质水平。分类器模型用 Python 编码实现，最后用与训练数据不重叠的历史体质数据检测分类器的准确率，结果显示，基于朴素贝叶斯算法的体质分类器达到了 78%的正确率。

关 键 词：学校体育；大学生体质分析；运动干预；朴素贝叶斯分类算法；大数据
中图分类号：G80-05 **文献标志码**：A **文章编号**：1006-7116(2018)01-0117-05

The application of Naive Bayes classification algorithm in university student fitness analysis

DU Yun-mei，LIU Dong
（Department of Information Technology and Engineering，Guangzhou Business College，Guangzhou 510363，China)

Abstract: Based on big data, the authors carried out classification and prediction on university student fitness, which is conducive to university sports governance system construction; the Naïve Bayes model is a machine learning classification algorithm that is simple to operate and provided with good performance. Based on Naive Bayes classification algorithm, and using the physical test data of classes 2014 and 2015 students of Guangzhou Business College and their score results as source data, the authors established a university student fitness classifier. By applying such a classifier, researchers can, in a certain sense of probability, correctly determine newly or previously enrolled university students' fitness condition, thus give a proactive early warning to those students whose fitness has a relatively high probability of hidden troubles, so that university physical education can carry out group fitness determination and individualized effective intervention on the students, thus promoting student healthy development and improving university students' overall fitness level. The classifier mode was realized by using Python coding, in the end, the classifier's accuracy rate was verified by using historical fitness data that did not overlap with training data, and the result showed that the fitness classifier based on naïve Bayes algorithm reached a correct rate of 78%.

Key words: school physical education; university student fitness analysis; sports intervention; Naive Bayes classifier algorithm; big data

2007 年中共中央国务院《关于加强青少年体育增强青少年体质的意见》[1]印发实施，2012 年教育部等出台《关于教育部加强学校体育工作的若干意见》[2]，2014 年重新修订了《国家学生体质健康标准》[3]，2016 年《“健康中国 2030”规划纲要》更将青少年体质问题上升到国家战略层面[4]。各级政府、各类学校和社会各界凝共识、聚合力、谋发展，协同加强学校体育治理体系建设。

收稿日期：2017-08-05
作者简介：杜云梅(1975–)，女，副教授，研究方向：软件工程、大数据与人工智能。E-mail：3947653@qq.com

但根据 1985 年开始的每 5 年一次的学生体质调研数据，大学生体质健康下滑趋势依然未得到遏制，甚至在很多指标上不如中学生[5–7]。各大学有必要结合新技术新理论推进体育教学改革，加强体育干预体系建设。

最近 10 年来，数据积累的急剧增加和针对数据的全链条技术整体成熟，催生了大数据以及接踵而来的人工智能的热潮。利用体质数据监测与人工智能分析技术，对疾病预防和健康趋势分析都具有积极的意义。国家也将健康医疗大数据应用发展纳入了国家大数据战略布局[8–9]。

本研究正是尝试应用大数据与人工智能技术，对体质监测数据进行建模与分析。基于朴素贝叶斯分类算法，构建大学生体质分类器，应用此分类器可对大学生的体质状况实现一定概率意义上正确的判断，从而对体质存在隐患概率比较大的学生给出主动性预警，以便大学体育对学生进行群体性的体质判断，为促进大学生体质健康发展提供数据与决策支撑。

1 朴素贝叶斯分类器

大学生的体质属于什么类别，其实就是一个分类问题，从数学角度来说，分类问题可做如下定义：已知集合：$\boldsymbol{C}=\{y_1, y_2, \cdots, y_n\}$和 $\boldsymbol{I}=\{x_1, x_2, \cdots, x_m, \cdots\}$，确定映射规则 $y=f(x)$，使得任意 $x_i \in I$ 有且仅有一个 $y_j \in \boldsymbol{C}$ 使得 $y_j=f(x_i)$成立。朴素贝叶斯(Naive Bayes)是一种基于贝叶斯定理与特征条件独立假设的机器学习分类算法。它的思想基础是对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

朴素贝叶斯模型是流行的十大挖掘算法之一，之所以备受人们关注，是因为它操作简单且性能较好，由于计算的高效性和高精度，朴素贝叶斯分类模型在文本分类领域得到了广泛的应用[10–13]。

$P(A|B)$表示事件 B 已经发生的前提下事件 A 发生的概率，叫做事件 B 发生下事件 A 的条件概率。其基本求解公式为：$P(A|B)=\frac{P(AB)}{P(B)}$。现实中经常遇到这种情况：$P(A|B)$可以很容易直接得出，而 $P(B|A)$则很难直接得出，但我们更关心 $P(B|A)$，贝叶斯定理便是基于条件概率，通过 $P(A|B)$来求 $P(B|A)$。贝叶斯定理即：$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$，其中，分母 $P(A)$，可以根据全概率公式分解为：$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$。

给定训练数据集(X, Y)，其中每个样本 X 都包括 n 维特征，即 $\boldsymbol{X}=(x_1, x_2, x_3, \cdots, x_n)$，类标记集合含有 k 种类别，即 $\boldsymbol{Y}=(y_1, y_2, \cdots, y_k)$。如果现在来了一个新样本 x，要判断它的类别，从概率的角度来看，这个问题就是给定 x，它属于哪个类别的概率最大。那么问题就转化为求解 $P(y_1|x)$，$P(y_2|x)$，$\cdots$，$P(y_k|x)$中最大的那个，即求后验概率最大的输出：$\text{argmax}_{yk}P(y_k|x)$。根据贝叶斯定理 $P(y_k|x)=\frac{P(x|y_k)P(y_k)}{P(x)}$，根据全概率公式，可以进一步地分解上式中的分母：

$$P(y_k|x)=\frac{P(x|y_k)P(y_k)}{\sum_k P(x|y_k)P(y_k)} \quad (1)$$

分子中的 $P(y_k)$是先验概率，根据训练集就可以简单地计算出来。而条件概率 $P(x|y_k)=P(x_1, x_2, \cdots, x_n|y_k)$，朴素贝叶斯算法对条件概率分布作出了独立性的假设，也就是说假设各个维度的特征 x_1，x_2，$\cdots$，x_n 互相独立，因此条件概率可以转化为：

$$P(x|y_k)=P(x_1, x_2, \cdots, x_n|y_k)=\prod_{i=1}^{n}P(x_i|y_k) \quad (2)$$

将公式(2)代入公式(1)得到：$P(y_k|x)=\frac{P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}$，于是朴素贝叶斯分类器可表示为：

$f(x)=\text{argmax}_{yk}P(y_k|x)=\text{argmax}_{yk}\frac{P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}$，因为对所有的 y_k，上式中的分母的值都是一样的，所以可以忽略分母部分，朴素贝叶斯分类器最终表示为：

$$f(x)=\text{argmax}P(y_k)\prod_{i=1}^{n}P(x_i|y_k)$$

可以看出，朴素贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。根据上述分析，构造朴素贝叶斯分类器主要可以分为 4 个步骤：

1)确定特征属性与类别集合：设 $X(x_1, x_2, \cdots\cdots, x_m)$为一个待分类处理项，而每个 x_i 为 X 的一个特征属性向量。类别集合 $Y=(y_1, y_2, \cdots\cdots, y_n)$，每个 y_k 为一个分类项，该集合是预先已得到的。

2)获取训练集：收集并准备训练数据，对连续型变量要进行离散化或分布处理。另外，朴素贝叶斯是有监督的机器学习算法，需要有属性标记。

3)分类模型训练：输入特征属性和训练样本，计算 $P(y_k)$，$P(x_i|y_k)$，即计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，生成分类器。

4)验证与应用：使用分类器对待分类项进行分类，

对于待分类项 X，如果存在 $P(y_k|X)=\max(P(y_k)\prod P(x_i|y_k)$，则 $X\in y_k$。

算法的核心部分就是训练集的准备和模型的学习训练过程，训练之后所形成的分类器可直接应用。

2　体质分类器的构造

参照上述的构造步骤，针对大学生体质问题，应用朴素贝叶斯算法构造体质分类器的具体过程如下：

2.1　属性定义

参照国家学生体质健康标准，设定了 12 项体质特征：性别、年龄、年级、籍贯、身高、体质量、身体质量指数(BMI)、肺活量、速度素质、爆发力素质、柔韧性素质、耐力素质、力量素质。

设定分类集合为：优秀、良好、及格、不及格。

2.2　数据预处理

以广州商学院 2014 年和 2015 年学生的真实体测数据作为源数据。

首先，按照《国家学生体质健康标准(2014 年修订)》中的评分标准，编写计算机程序算出每个学生的单项评分、学年总分并评定等级，去除有缺失值的数据条目，最后得到 21 664 条有效记录，形成有体质分类结果的完整数据集。广州商学院学生的体质分布如图 1 所示，其中不及格占 10.97%，及格占 78.36%，良好占 10.50%，优秀的只有 0.2%左右，学生体质状态不容乐观，虽然绝大部分学生的体质处于及格线上，但达到优秀等级的非常少。

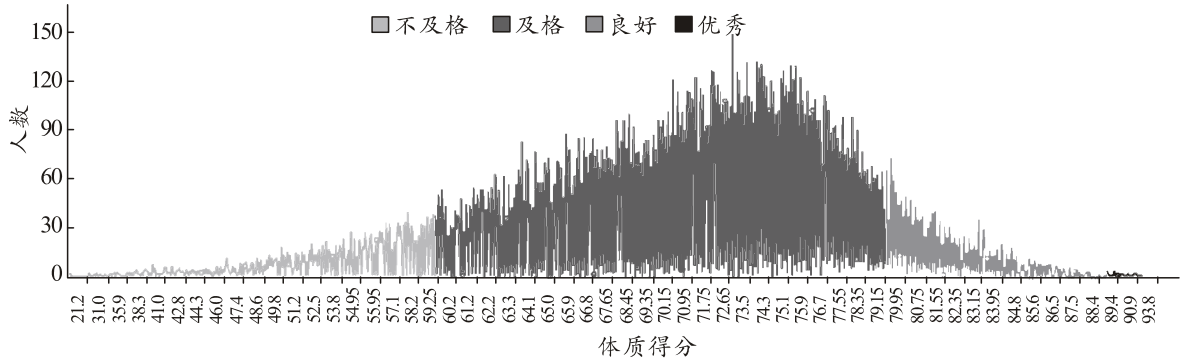

图 1　广州商学院学生体质分布图

接着，为后面分类器运算的方便，进一步将体质特征中性别的“男/女”分别转换为数值 1/2，将体质指数的“优秀/良好/及格/不及格”分别转换为数值 1/2/3/4，将身高、体质量两个数据项换算合并为 BMI 一个数据项。

除了年龄、年级和籍贯属性是离散型数据不需进一步处理，其他属性都是连续型变量，需要进行离散化处理，本研究采用的办法是参考国家体质评分标准，划分特征的取值区间，在分类器的训练过程中，计算的是区间概率。

另外，采用了 Laplace 平滑处理来解决零概率问题。在计算实例的概率时，如果某个量在观察样本库(训练集)中没有出现过，会导致整个实例的概率结果是 0，在体质分类的问题中，当一个特征取值区间没有在训练样本中出现，该取值区间的概率就为 0，使用连乘计算体质概率时也为 0，这是不合理的，不能因为一个事件没有观察到就武断地认为该事件的概率是 0。在计算实例的概率时用加 1 的方法估计没有出现过的现象的概率。

2.3　训练集

在 Python 中编码实现 2.2 节所述的数据预处理，得到的数据集存储为 csv 文件，第一行为索引行，包括 12 个体质特征和体质等级，后面每一行是每个同学的体质特征和等级取值，值之间以逗号分隔。该文件就是接下来分类器训练的数据输入，为保证模型检验的客观性，本研究采用切片法，将其中的 80%作为训练数据，另外的 20%留作检验数据。

2.4　分类器训练

在特征选取和训练数据基础上，可以构造多种不同目的的分类器。这里有代表性地列举了两个分类器。第 1 个分类器是朴素贝叶斯算法的正向应用，即已知部分属性来预测体质分类。为了演示贝叶斯的工作过程，这个分类器只选取了 4 个特征以方便演示朴素贝叶斯的分类原理。第 2 个分类器反过来把体质分类结果作为一个特征项，来预测属性的取值区间。

1)分类器 1。

特征：性别，身高，体质量，肺活量。其中，性别(x_1)有两个取值(男，女)；年级(x_2)有两个取值(大一大二，大三大四)；用身高、体质量换算成 BMI(x_3)，分成 4 个取值区间(≤17.1，17.2~23.9，24.0~27.9，≥28.0)；肺活量(x_4)，按以下值(3 400，3 350，3 300，3 150，3 000，2 900，2 800，2 700，2 600，2 500，2 400，

2 300，2 200，2 100，2 000，1 960，1 920，1 880，1 840，1 800)分成 20 个取值区间。

分类：体质级别(优秀 y_1，良好 y_2，及格 y_3，不及格 y_4)。

待分类项：例如身高 160 cm、体质量 48 kg、肺活量 2 400 mL 的大一女生，体质最可能是什么级别？这个问题即是给定条件 X=(女，1，18.75，2 400)，条件概率 $P(y_1|X)P(y_2|X)P(y_3|X)P(y_4|X)$中最大的那个，就是分类器预测那个类别。根据特征条件独立的假设，$P(y_1|X)=P(y_1|x_1，x_2，x_3，x_4)=P(y_1)P(x_1|y_1)P(x_2|y_1)P(x_3|y_1)P(x_4|y_1)/P(x_1，x_2，x_3，x_4)$。这些都可以通过训练集中数据计算出来。

2)分类器 2。

特征：性别，年级，身高，体质量，肺活量，速度，爆发力，体质等级。

分类：耐力级别。

待分类项：如一个身高 160 cm、体质量 48 kg、肺活量 2 700 mL，50 m 跑成绩 10.2 s 的大一女生想要得到优秀体质级别，800 m 跑要达到什么水平？

分类器的原理不再赘述，都能在 Python 中编码实现，用到 Pandas、Sklearn 和 Numpy 等外部库，采用 GaussianNB 实现模型。

2.5 分类器检验

将 2.2 节中得到的数据集用切片法切出数据总量的另外 20%作为检验数据，采用了 Precision、Recall、Fb-score 和 Accuracy 四个评价指标，其中 Precision(精度)是精确性的度量，表示被分为正例的示例中实际为正例的比例；Recall(召回率)是覆盖面的度量，度量有多个正例被分为正例，Fb-score 是准确率和召回率的调和平均：$Fb=[(1+b_2)\times P\times R]/(b_2\times P+R)$。Accuracy(正确率)表示被分为正例的条目数与检验数据条目数的比例。检测结果表 1 所示。从检验结果可以看出，分类器的综合正确率达到 77.98%。

表 1 分类器正确率检测结果

体质分类	精度/%	召回率/%	调和平均/%	样本
1(优秀)	0	0	0	7
2(良好)	0	0	0	312
3(及格)	0.78	1	0.87	1 841
4(不及格)	0.78	0.11	0.20	224
平均/%	0.68	0.78	0.69	2 384
正确率/%	0.779 8			

2.6 体质分类器在体育教学实践中的应用

用训练数据训练得到的分类器可以直接使用，输入学生的几项体质特征值，就可以得到相应的分类结果，可以作为对学生体质状况的预测。

分类器 1：

给定条件 X=(女，1，160，45，2 400)

给出的结果是 y_3 即身高 160 cm、体质量 45 kg、肺活量 2 400 mL 的大一女生，历史数据显示如果不加干预的话，其体质检测结果最可能是“不及格”。

可以将全部学生的体质进行分类预测，按照分类结果将学生分成不同的组别，对于体质检测结果较大可能为“不及格”的那部分同学，可以制定特别的干预计划，加强体质锻炼。

分类器 2：

给定条件：X=(女，1，160，48，2 700，10.2)

结果为[103]，即 160 cm、体质量 48 kg、肺活量 2 700 mL、50 m 跑成绩 10.2 s 的大一女生 800 m 要跑到 3 min 3 s 以内，才最有可能得到“优秀”体质等级。如果现在的 800 m 跑不能达到这个成绩，为达到“优秀”体质等级，就要加强耐力训练。

随着学生各项测试数据的积累，在此分类器的辅助下，可以以目标为导向，即要让学生的体质分类结果达到“优秀”，应该让学生加强哪方面能力的锻炼；进一步，可以按学生有待加强的能力进行分组，对不同组制定不同的锻炼计划与干预措施。

3 展望

本研究用朴素贝叶斯算法，构建了大学生体质分类器，应用该分类器可以对每个在校学生的体质状态进行预测，为个性化的运动指导与干预提供依据；也可以对学生群体进行客观的体质分析，发现不同群体的体质短板。检验结果显示，本分类器能达到 78%的综合正确率，具有一定的可信度。

本研究采用了广州商学院 2 年的学生数据做试验，当加入越来越多的训练数据时，模型会变得越来越准确。而全国的学生体测数据都是依照《国家学生体质健康标准》，所以数据项与数据结构基本一致，从而可以很容易的将其他省市高校学生体测数据纳入到本分类模型的训练集中。当有了更多高校数据时，还可以按省市、按南北方等不同地域对学生体质状况进行横向的对比分析等。

另外，在此体质分类模型给出的预测与判断基础上，学校体育部门可以有针对性地对学生进行个性化的体育锻炼指导与干预，跟进采集下一年的体测数据，就可以对学生体质进行时间纵向上的体质变化分析、运动干预的有效性分析等。

因为整个数据预处理与分类器训练过程都用

Python 编码，所以扩展数据后的训练集准备与模型更新可由程序自动完成。而且在朴素贝叶斯分类下可以构造出更多结构相似、目的不同的分类器，以满足学校体育对学生体质的促进和监督的需求。

参考文献：

[1] 中共中央，国务院. 关于加强青少年体育增强青少年体质的意见[EB/OL]. [2017-07-02]. www.gov.cn/jrzg/2007-05/24/content_625090.htm.

[2] 教育部，发展改革委，财政部，等. 关于进一步加强学校体育工作的若干意见[EB/OL]. [2017-07-02]. www.gov.cn/zwgk/2012-10/29/content_2252887.htm.

[3] 教育部关于印发《国家学生体质健康标准(2014 年修订)》的通知[EB/OL]. [2017-07-02]. http://www.moe.edu.cn/s78/A17/twys_left/moe_938/moe_792/s3273/201407/t20140708_171692.html.

[4] 中共中央，国务院. “健康中国 2030”规划纲要[EB/OL]. [2017-07-02]. http://news.xinhuanet.com/ health/2016-10/25/c_1119786029.htm.

[5] 国家体育总局，教育部，科技部，等. 2014 年国民体质监测公报[EB/OL]. (2015-11-25) [2017-0702]. http://www.sport.gov.cn/n16/n1077/n1227/7328132.html.

[6] 国家体育总局，教育部，科技部，等. 2010 年国民体质监测公报[EB/OL]. （2011-09-02） [2017-07-02]. http://www.sport.gov.cn/n16/n1077/n297454/2052709.html.

[7] 教育部发布 30 年来我国学生体质与健康“大数据”[EB/OL]. [2017-07-02]. http://www.jyb.cn/china/gnxw/201407/t20140729_592098.html.

[8] 国务院印发关于促进大数据发展行动纲要[EB/OL]. [2017-07-02]. http://business.sohu.com/20150906/n420463676.shtml.

[9] 国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见[EB/OL]. [2017-07-02]. http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.

[10] 杨雷，曹翠玲，孙建国，等. 改进的朴素贝叶斯算法在垃圾邮件过滤中的研究[J]. 通信学报，2017，38(4)：140-148.

[11] 刘秋阳，林泽锋，栾青青. 基于朴素贝叶斯算法的垃圾短信智能识别系统[J]. 电脑知识与技术：学术交流，2016，12(12)：190-192.

[12] 贾志鹏. 基于朴素贝叶斯分类器的校园信息智能推荐算法[J]. 软件工程，2016，19(12)：30-32.

[13] 谢小军，陈光喜. 基于多属性联合的朴素贝叶斯分类算法[J]. 计算机技术与发展，2016，26(12)：77-81.