# 基于大模型智能助教助学的数据挖掘实验设计与探讨*

李波**

中南民族大学计算机学院，武汉 430074

**摘 要** 针对大模型如何智能辅助实验教学的问题，提出了必须引导学生积极正确地使用大模型赋能学习过程，在编程学习过程中应分阶段、有限制地使用大模型。以数据挖掘课程中回归分析一节为例，设计了一个教学实验，基于大模型进行概念理解、数学推导和个性化学习路径推荐；结合大模型进行算法的比较和总结，并对实验进行分析和讨论。课程问卷调查反映出 78.7%的学生对大模型“辅助而不完全依赖”的实验模式的认同。基于大模型智能助教助学的实验教学模式，有助于节省教师对常规知识点的答疑时间，以便让教师留出更多时间来思考更高层次的教学问题；也有助于提高学生的学习和实验效率，更方便扩展学生的知识视野。

**关键词** 实验设计，数据挖掘，回归分析，大模型，智能助教助学

# Design and Exploration of Data Mining Experiments Based on Intelligent Teaching and Learning Assistance with Large Language Models *

Bo Li

School of Computer Science
South-Central Minzu University,
Wuhan 430074, China;
libo@mail.scuec.edu.cn

**Abstract—To address the problem of how large language models (LLMs) can intelligently assist experimental teaching, this paper proposes that students must be guided to actively and correctly use LLMs to empower the learning process, and that their use in programming learning should be phased and restricted. Taking the regression analysis section of a Data Mining course as an example, a teaching experiment was designed. In this experiment, the LLM was employed for concept understanding, mathematical derivation, and personalized learning path recommendation, and it was also used for algorithm comparison and summarization. Then the experiment was analyzed and discussed. The course questionnaire survey reveals that 78.7% of the students expressed their recognition of the experimental model characterized by “assistance without complete dependence” on LLMs. The experimental teaching model that employs LLMs as intelligent teaching and learning assistants helps save teachers’ time in answering routine knowledge-point questions, allowing teachers to devote more time to higher-level pedagogical issues. It also enhances students’ learning and experimental efficiency and facilitates the broadening of their knowledge horizons.**

**Keywords—experimental design; data mining; regression analysis; large language models; intelligent teaching assistant**

## 1 引 言

随着大模型的迅速发展，大模型辅助高校教育教学已经成为一个不可回避的讨论热点。如何正确积极地引入大模型到教学中，大模型应该扮演什么角色？大模型的使用在教师教学和学生学习中已经成为一种趋势，如何合理利用好大模型，是一个值得深入探讨的问题。

近年来，很多研究者在数据挖掘教学及大模型教学应用方面，开展了深入的研究。例如，陈美等[1]提出从教学内容的先进性、理论教学模式的多元性、实验项目设计的多层次性以及考核评价方式的全面性四个维度开展教学改革。牟谷芳等[2]提出探讨了案例驱动式教学模式和产教融合项目的实施。张鹏等[3]在课程教学中充分发挥大模型在教学内容自动更新、智能交互支持等方面的优势，有效提升了教学效果。刘雪峰等[4]提出了一种基于大模型辅助的项目式学习实验教学体系，并介绍如何借助大模型辅助开展人机协作教学。康雁等[5]结合大模型进行专业课程创新实践评价目标改革，并有效地提升了学生的创新实践及人机协

*基金资助：本文得到东软教育科技集团有限公司产学合作协同育人项目-新工科建设项目（241002108121937）支持

*通讯作者：李波 libo@mail.scuec.edu.cn。

作等多项能力。大模型赋能教育教学，助力学生学习，已经成为一个不可回避的现实问题。目前大模型已经具备了比较强的“知识”和“能力”，在很多方面能为高校师生提供有价值的帮助。在教学过程和学生学习过程中，如果完全限制学生使用大模型是不现实的；但是如果不加以一定的引导，则学生会出现过多依赖大模型完成学习任务的情况，这对正常教学和学生学习是很不利的。

计算机类课程，涉及到编程和项目的组织和实施，大模型应该如何发挥应有的“角色”，下文以数据挖掘课程中的“回归分析”内容为例，设计一个回归分析的教学实验，并开展一些探讨。

## 2 实验设计思想

在本课程的实验教学中，如何紧密跟进当前人工智能技术的快速发展趋势，合理引入大模型技术，既发挥其“智能助教”的积极作用，又不影响对学生的实践能力的锻炼培养，是本课程实验教学改革迫切需要解决的问题。

认知负荷理论[6]指出，教学设计应合理分配学习者的内在负荷（由学习任务复杂度决定）、外在负荷（由信息呈现方式决定）与关联负荷（学习者主动进行的信息加工）。大模型的合理使用可有效降低外在负荷（如基本概念和编程语法查询），帮助学生聚焦于内在负荷（如理解算法和模型原理）与关联负荷（知识体系整合及应用），从而提升学习的效果与效率。

本课程实验教学提出“分阶段、有限制”的大模型应用策略，在实验的不同阶段对学生使用大模型的范围和方式设置不同的要求，以达成大模型“辅助但不完全依赖”的实验改革目标。实验设计中分阶段使用大模型：在实验准备和实验结果分析比较等环节，可以充分发挥大模型的优势；而在实验编程环节，要求学生自己独立完成。本课程目前将大模型智能助教助学的作用主要定位于以下三个层面：

（1）智能答疑：学生在实验过程中遇到的环境配置、概念及算法理解、算法原理的数学推导等具体问题时，利用大模型可提供帮助，有效缓解教师一对多辅导的资源不足的问题。

（2）个性化学习指导：基于学生在超星学习通平台的学习行为数据，大模型为不同学习进度的学生推荐差异化的学习路径与实验任务难度，实现因材施教。

（3）实验结果比较分析：大模型可分别从实验算法的原理进行差异对比、优缺点对比，对算法的典型应用场景、算法选择建议、创新改进方向可以进行深入的分析和回答，帮助学生深入理解。

## 3 实验过程设计

### 3.1 实验目标

在数据挖掘课程中，回归分析是重要教学章节之一，其应用非常广泛。回归分析属于一种统计分析方法，用于确定两个或两个以上变量之间的定量依赖关系。一般通过建立一个数学方程来概化一组数据。本实验目标是对一组数据样本点进行一元线性回归分析，确定回归方程的系数。

实验拟通过最小二乘法[7]、BP 人工神经网络[7]（基于误差反向传播学习算法）、梯度下降法[8]三种实现思路来分别实现一元线性回归分析。需要用到 Python 的 Numpy 库、TensorFlow 深度学习框架 (编程集成环境可采用 Pycharm，Jupyter Notebook 等环境)。总体实验框图如图 1 所示。

### 3.2 基于大模型的概念理解与数学推导

学生在课堂虽然学习了相关算法内容，但是并不一定完全理解了。因此可以让学生通过大模型（如 DeepSeek）进一步学习这些算法的概念，可以通过大模型提供相关算法的概念解释和数学推导。通过大模型来对不同算法进行数学原理的推导和讲解，对教师教学和学生学习是很有帮助的。

目前有些企业开发的课程平台，在课程中可以内置一个课程智能体助教，也可以发挥类似的作用。

### 3.3 大模型个性化学习路径推荐

本步骤由教师完成，可以事先由实验指导老师通过调查问卷等形式，确定好学生的水平和能力分级和可能的个性化学习要求，然后通过大模型来给出个性化的学习路径推荐。分别对不同分级和提出个性化学习要求的学生提供 1-2 个个性化学习任务。基于学生表现推荐个性化学习路径，如对表现优秀的学生可以鼓励他采用其他的多种不同方法完成同样的实验目标，实现更高阶的任务。

教师在实验准备阶段，可以设计好大模型的提示词，让学生去查询资料，通过大模型推荐多种算法，但不能让学生直接通过大模型生成代码。

本步骤中，学生通过大模型的回答，可以获得有哪些方法可以实现回归分析，如最小二乘法、梯度下降法等。因为课堂上也讲述了人工神经网络，学生也可以询问人工神经网络是否可以实现回归分析，大模型的回答是肯定的。本实验可以要求学生结合课程所学内容，至少采用 3 种不同方法分别实现线性回归方程系数的求解。

下文则以最小二乘法、梯度下降法、BP 人工神经网络算法为例，分别求解回归方程的系数。

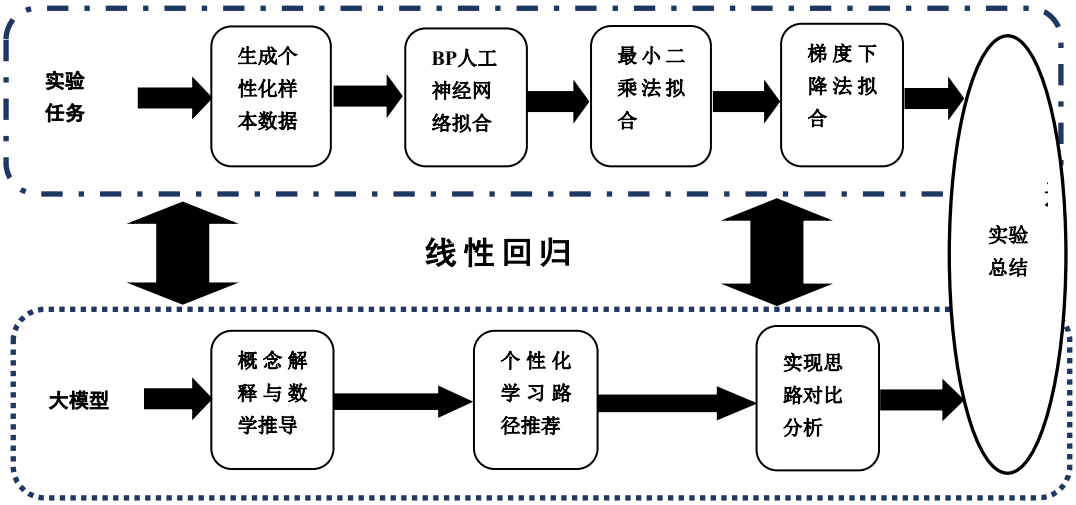



图 1 大模型赋能的实验设计框图

### 3.4 实验任务设计

实验步骤可以参考如下：

(1)随机生成一批样本数据(x_sample,y_sample)，在 pycharm 环境采用 python 语言调用 numpy 等库实现，将这批数据保存为一个 CSV 数据文件，以便在后续步骤中使用；

（2）在 pycharm 环境中，基于 TensorFlow 深度学习框架（以 1.0 版本为例），针对（1）的数据，采用 BP 反向传播学习算法实现对回归方程系数的迭代求解；

（3）在 pycharm 环境中，调用在第（1）步中建立的数据文件，根据最小二乘法原理，推导求解回归方程系数的联立方程组，采用 python 语言调用 numpy 库计算方程参数；

（4）在 pycharm 环境中，调用在第（1）步中建立的数据文件，根据梯度下降法原理，采用 python 语言调用 numpy 库计算方程参数；

（5）基于大模型对以上实验结果进行比较分析，总结算法的异同点和优劣点。

## 4 实验过程

### 4.1 随机生成样本点

本实验首先从自动构造样本数据开始。可以采用个性化生成数据，拟合方程的系数和学生在班级的序号相关(如可以设方程的斜率 k=0.2*学生班级序号)。以下以序号为 1 的学生为例。假设待拟合的线性方程为 $y = f(x) = kx + c = 0.2 * 1 * x + 0.5$，通过随机函数生成 100 个 float 类型 x，而 y 值可以由该方程直接求取。这样产生了 100 个实例点$(xi, yi)$，而这 100 个实例点，客观上一定是满足线性方程 $y = 0.2x + 0.5$ 的。学生生成的实验数据是不同的。

```
x_sample=np.random.rand(100).astype(np.float32) #随机生成100个实例点
y_sample=0.2*x_sample+0.5
#定义方程式y=k*x_sample+c
```

这批数据生成后（如图 2），将它们转为一个 CSV 文件，用作后续步骤的实验数据。实验需要求取方程系数的线性方程 $y = kx + c$，$k$ 和 $c$ 未知。

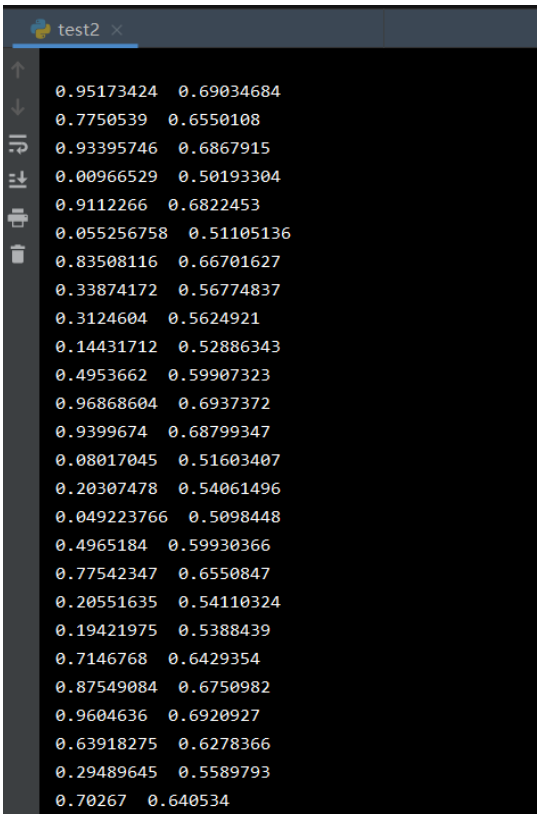



图 2 随机生成样本点部分数据

## 4.2 基于 BP 神经网络拟合方程系数

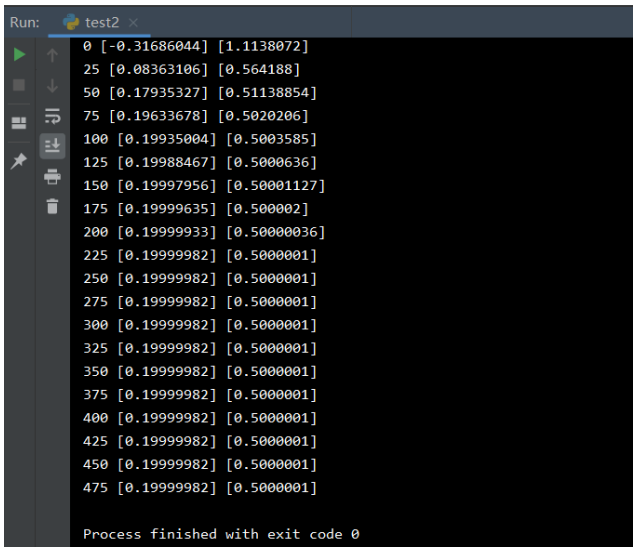



图 3 BP 神经网络拟合结果

学生在基于前面对算法原理充分理解的基础上，在 pycharm 平台下，采用 python 语言、调用 TensorFlow 深度学习框架实现。以下为部分程序语句实现示例：

```
#构建网络结构
Weight=tf.Variable(tf.random_uniform([1],-1.0,1.0))
Biases=tf.Variable(tf.zeros([1]))
y=Weight*x_sample+Biases    #网络的计算输出
Loss=tf.reduce_mean(tf.square(y-y_sample)) #计算输出值与目标值的误差
optimizer=tf.train.GradientDescentOptimizer(0.5) #梯度下降
train=optimizer.minimize(Loss) #根据误差通过反向传播调整网络参数
```

然后建立一个 session 来加载运行数据。

这里，初始自动生成的样本点 $(x_sample, y_sample)$ 可以认为是目标值，而神经网络输出 $y=Weight*x_sampel+Biases$ 为计算值。BP 神经网络拟合方程系数的关键是需要抓住目标值和计算值之间的差异，即通过神经网络的目标输出与网络计算输出的均方误差 tf.square(y-y_sample) 来反向调节网络的连接权值和阈值。该过程也是神经网络的训练过程。

当达到最小误差阈值时，方程系数的拟合值已和目标值（0.2 和 0.5）非常接近了（如图 3）。

## 4.3 基于最小二乘法计算方程系数

最小二乘法拟合方程系数实际上是根据计算值 f(xi) 和样本点实际值 yi 之间的差异，基本思路就是计算两者的均方误差 MSE，并让其最小化。这里，MSE 是 $k$ 和 $c$ 的凸函数，当 MSE 对 $k$ 和 $c$ 的偏导数为 0 时，可获得 $k$ 和 $c$ 的最优解。而通过求偏导数并令其为 0，得到一个联立的方程组，从而可求得最后的 $k$ 和 $c$ 的最优解（如图 4 所示）。

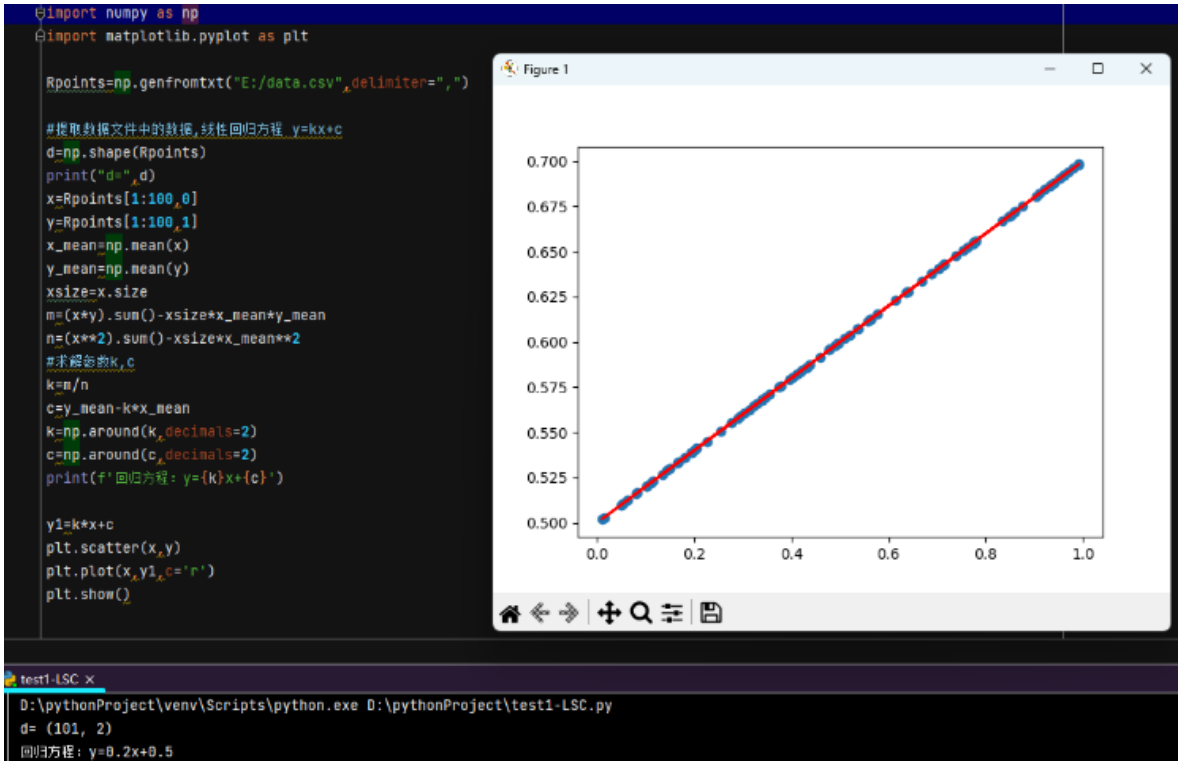



图 4 基于最小二乘法和 python Numpy 库编程求解

## 4.4 基于梯度下降法求解方程系数

梯度下降法的目标是通过迭代优化，找到使损失函数（均方误差 MSE）最小的线性方程参数（斜率 $k$ 和

截距 $c$）。首先需要初始化参数 $k$ 和 $c$；然后计算梯度，即对损失函数求偏导，得到参数更新方向；接着迭代更新参数，即沿梯度反方向调整参数学习率 α 。重复以上步骤，直到损失函数变化小于阈值或达到最大迭代次数（如图 5 所示）。

```python
#读取数据
Rpoints=np.genfromtxt("E:/data.csv",delimiter=",")
d=np.shape(Rpoints)
x=Rpoints[1:100,0]
y=Rpoints[1:100,1]

def gradient_descent(x, y, learning_rate, epochs):
    n = len(x)
    k, c = 0.0, 0.0  # 初始化参数

    for epoch in range(epochs):
        # 预测值
        y_pred = k * x + c
        # 计算梯度
        dk = (-2 / n) * np.sum(x * (y - y_pred))
        dc = (-2 / n) * np.sum(y - y_pred)
        # 更新参数
        k -= learning_rate * dk
        c -= learning_rate * dc
        # 每100次迭代打印损失
        if epoch % 100 == 0:
            loss = np.mean((y - y_pred) ** 2)
            print(f"Epoch {epoch}: loss={loss:.6f}, k={k:.6f}, c={c:.6f}")

    return k, c

# 使用梯度下降法求解
k_gd, c_gd = gradient_descent(x, y, learning_rate=0.01, epochs=2000)
```

（a）算法实现代码

```
Epoch 1600: loss=0.000004, k=0.206689, c=0.496335
Epoch 1700: loss=0.000003, k=0.205852, c=0.496793
Epoch 1800: loss=0.000002, k=0.205119, c=0.497195
Epoch 1900: loss=0.000002, k=0.204479, c=0.497546

梯度下降法结果：
k（斜率）: 0.203923
c（截距）: 0.497850

Process finished with exit code 0
```

（b）计算结果

图 5 基于梯度下降法和 python Numpy 库编程求解

## 5 结合大模型的实验比较与分析

本次实验目的是线性回归方程的参数求解，实验涉及到 3 种实现思路和 2 种实现平台（基于最小二乘法在 pycharm 下采用 Numpy 库编程求解、基于 BP 神经网络在 pycharm 下应用 TensorFlow 框架编程实现、基于梯度下降法采用 Numpy 库在 pycharm 下编程求解）。学生可以通过大模型 DeepSeek 提问，“请比较下采用 BP 神经网络、最小二乘法、梯度下降法实现线性回归的原理差异、各自特点及优劣点”。大模型分别从原理优缺点对比、数据量适应性等进行了比较（如图 6 所示）。

| 维度 | 最小二乘法 | 梯度下降法 | BP神经网络 |
|---|---|---|---|
| 优点 | 1. 解析解，计算速度快<br>2. 无超参数调优<br>3. 数学可解释性强 | 1. 适合大规模数据<br>2. 可扩展至其他模型<br>3. 内存效率高 | 1. 可拟合复杂关系<br>2. 自动特征学习<br>3. 对非线性问题泛化性好 |
| 缺点 | 1. 数据需满足严格假设<br>2. 大数据集矩阵求逆困难<br>3. 无法处理非线性 | 1. 需调学习率<br>2. 可能陷入局部最优<br>3. 收敛速度依赖参数 | 1. 计算成本高<br>2. 易过拟合<br>3. 黑箱性，解释性差 |
| 数据量适应性 | 小数据（n<10^4） | 大数据（n>10^6） | 中大数据（10^3~10^5） |
| 噪声鲁棒性 | 对异常值敏感 | 可通过正则化提升 | 依赖网络结构和正则化 |
| 实现复杂度 | 最低（几行代码） | 中等（需手动实现迭代） | 最高（需设计网络结构） |

图 6 大模型对三种算法的比较

有些学生提出了问题“以上三种方法都用到了求导，有什么区别吗？”，这个问题也可以通过大模型解答。大模型给出了简明的比较：

①最小二乘法：解析求导，一次性得到精确解；

②梯度下降法：数值求导，迭代逼近最优解；

③BP 神经网络：链式求导，分层计算复杂梯度。

还需要让学生理解，BP 算法的两个核心组成部分，包括了反向传播（计算梯度的数学方法（链式法则））和梯度下降（使用梯度更新参数的优化策略）。反向传播是计算工具，梯度下降是优化工具。BP 神经网络更适合和擅长去处理更复杂的非线性的问题。

此外，也可以让大模型对各个算法的应用场景进行比较分析，比较结果如图 7 所示。

| 场景 | 推荐方法 | 理由 |
| --- | --- | --- |
| 小数据集+线性关系 | 最小二乘法 | 计算快且精确 |
| 大数据集+内存限制 | 梯度下降（SGD） | 分批训练，内存高效 |
| 特征间存在非线性 | BP神经网络 | 可捕捉复杂模式 |
| 需要模型解释性 | 最小二乘法 | 参数直接对应物理意义 |
| 实时更新模型 | 梯度下降（在线学习） | 支持逐样本更新 |

图 7 三种算法适用场景推荐

BP 神经网络属于有监督学习算法，一般是用于目标分类和识别等问题，但本实验用该算法来解决线性回归问题中的参数拟合求解。让学生认识到同一个问题往往可以采用多种方法来解决；BP 神经网络属于数据挖掘课程中的一个重点学习内容，通过本实验也可以让学生认识到神经网络的功能和作用，不仅是用于分类，还可以用于其他应用场景，以开阔他们的思路。

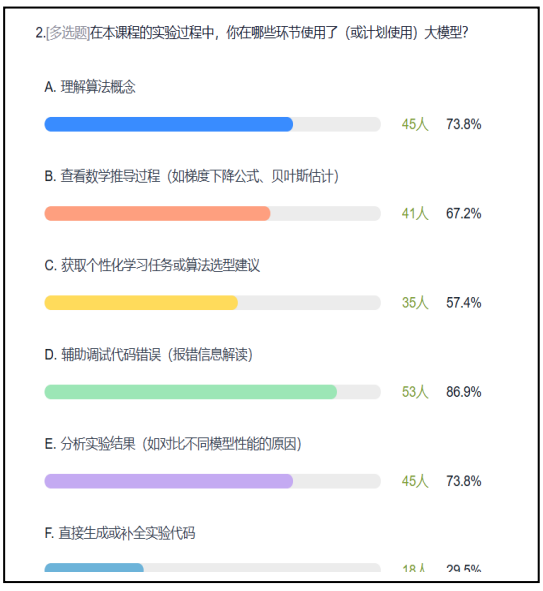
## 6 大模型赋能实验教学的分析与思考

大模型是一把“双刃剑”。大模型的合理利用可以在教学中发挥它积极的一面，应该在教学中“分阶段、有限制”地使用大模型。在实验准备和实验结果分析比较环节，可以充分使用大模型，发挥大模型的优势。在实验的编程环节，需要严格要求学生自己独立完成。在本科期间，至少应该保证学生有独立编程的一个学习阶段。笔者认为，如果学生依靠大模型直接生成代码，那么会失去了自己独立思考和编程的机会，实验教学的目标就失败了；而如果学生依靠大模型来修改代码的 Bug，那么就会削弱学生利用编程平台调试 Bug 和独立分析解决问题的能力，也从而降低实验教学目标的达成度。如果学生利用大模型来调试错误，前提条件是学生经过了充分的独立思考，仍不能解决问题。

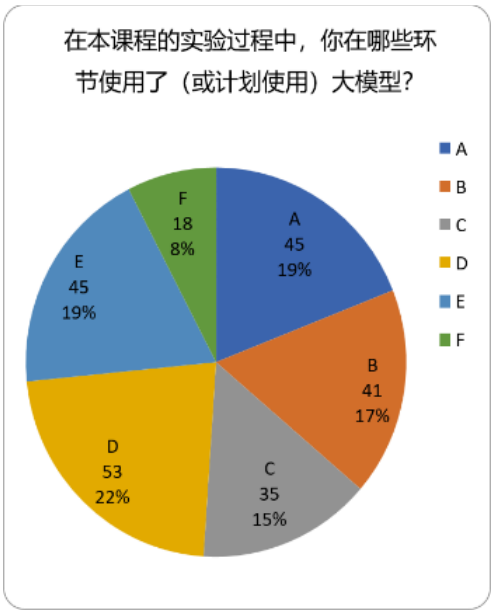
在编程实现和调试环节上，笔者认为对于在校学生，尽可能不基于大模型而应自己独立完成，需要教师从多个角度严格要求。例如：①在实验任务书强调实验的要求，提前沟通，明确告知学生实验要求，不提倡学生利用大模型编码；②提供编程指导支持，给予学生足够的技术指导，让学生有信心独立来完成编程；③强调独立思考和编程的价值和意义，说明亲手编程实践对自己能力提升和职业发展的重要性；④加强编程中间过程的考核，可要求学生提供中间调试代码过程的截图和实时编程过程记录，要求分阶段代码提交与审查提供若干中间阶段的代码版本截图；⑤实验完成后开展技术实现细节的答辩，等等。

基于大模型智能助教助学的实验教学模式，有助于节省教师对常规知识点的答疑时间，以便让教师留出更多时间来思考更高层次的教学问题，如个性化教学；同时也对教师的教学提出了更高的要求，教师必须引导好和监管好实验的全过程。基于大模型智能助教助学的实验教学模式，也有助于提高学生的学习和实验效率，方便扩展学生的知识视野，但对学生应用大模型的自控能力提出了更高要求。经过期末考试成绩分析比较，发现采用了大模型辅助教学学习的方式后，卷面不及格的学生比例和上学期比较下降了 3.5%，说明基于大模型辅助的学习方式对部分学习后进学生可能会发挥更有效的帮助作用。

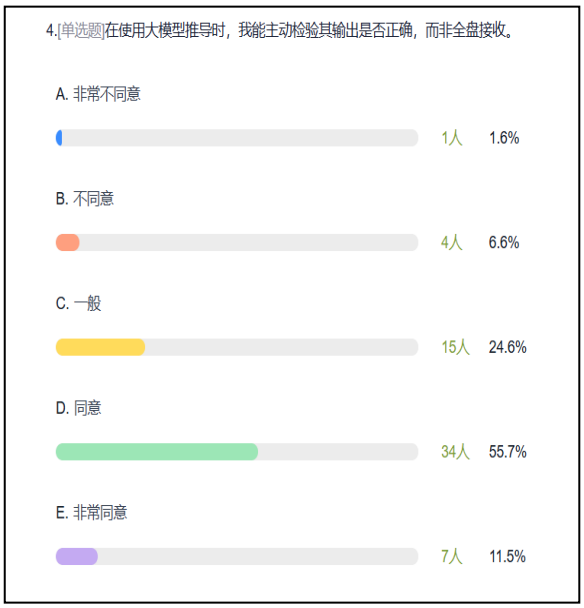
在课程结束的基于超星学习通的调查问卷中，有 12 项问题。图 8 列出了其中 3 个问题的统计结果示例。



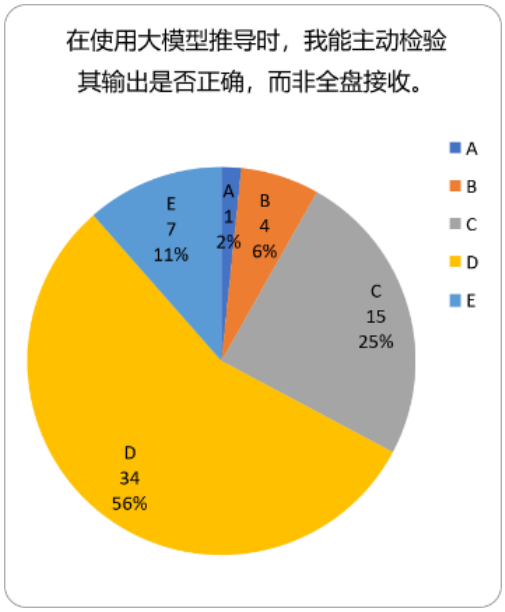

(a) 问题2回答情况



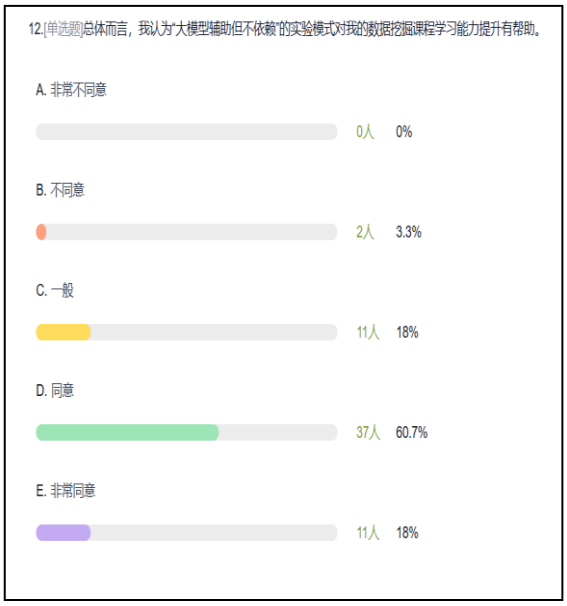

(b) 问题2统计饼状图



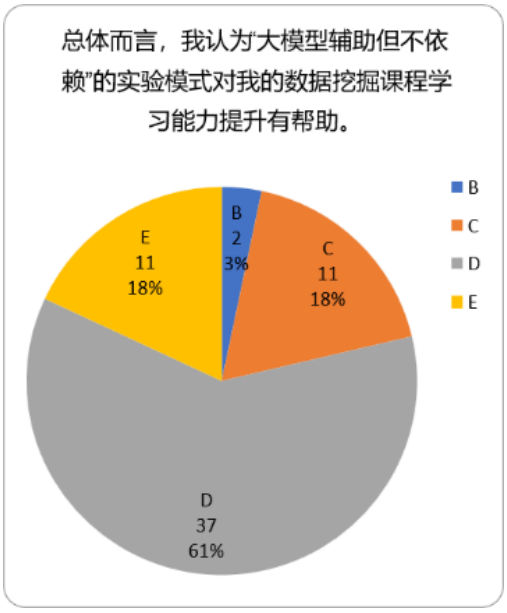

(c) 问题4回答情况




(d) 问题4统计饼状图




(e) 问题12回答情况




(f) 问题12统计饼状图

图 8 课程调查问卷的部分问答情况统计

在问到“在本课程的实验过程中，你在哪些环节使用了大模型”时，有 73.8%的学生回答了“理解算法概念”，67.2%的学生回答了“查看数学推导过程”，57.4%的学生回答了“获取个性化学习任务或算法选项建议”，86.9%的学生回答了“辅助调试代码错误（报错信息解读）”，73.8%的学生回答了“分析实验结果”，还有 29.5%的学生回答了“直接生成或补全实验代码”。在问到“总体而言，我认为‘大模型辅助但不依赖’的实验模式对我的数据挖掘课程学习能力提升有帮助”时，没有学生选“非常不同意”，3.3%的学生选“不同意”，18%的学生选“一般”，60.7%的学生选“同意”，18%的学生选“非常同意”。可见，大模型辅助而不完全依赖的实验模式，大多数学生是认同的，它是一种新的实验范式，但在今后实验中还需要有效加强对实验过程的合理引导和管理，及时了解学生现状，以让大模型助教助学能够扬长避短，充分发挥其积极的一面，减少其消极的一面。

## 7 结束语

在大模型时代，大模型赋能辅助高校教育教学已经成为一个不可回避的重要趋势，需要教师正确积极地引导学生利用大模型到学习过程中，让大模型扮演

合适和正确的“角色”，让大模型真正助力教师教学、助力学生成才，而不是让大模型成为给学生“包办一切”、从而吞噬学生逻辑训练和能力提升的工具。基于大模型助教助学，在实验教学过程中如何有效加强对学生的引导和管理，将在后续教学中继续深入研究。

## 参 考 文 献

[1]陈美，张永清，周航. 实践与理论双驱动的数据挖掘课程教学改革研究[J]. 教育教学论坛, 2025,11:77-80

[2]牟谷芳，张秋燕，李胜坤. 新工科背景下数据挖掘实践教学探索[J].高教学刊, 2025,5: 103-111

[3] 张鹏, 杜航原, 郭虎升, 张虎, 王文剑. 大模型赋能的大数据课程新型教学模式探索与实践[J]. 计算机技术与教育学报, 2025,13(4): 241-249

[4]刘雪峰，陈玮，韩旭，李牧，牛建伟，马殿富. 大模型时代计算机实验教学的挑战与机遇[J].实验技术与管理,2025(网络首发)

[5] 康雁, 张璇, 杨云, 朱锐. 大模型驱动的软件工程实践课程改革与实践[J]. 计算机技术与教育学报,2025,13(7): 57-62

[6] Sweller J，Cognitive Load During Problem Solving: Effects on Learning[J]. Cognitive science, 1988,12,257-285

[7]周志华.机器学习[M].北京:清华大学出版社,2016

[8]周杰,郭振华,张林.模式识别[M].北京:高等教育出版社,2022