数据挖掘应用课程教学案例建设

胡皎月，侯 佳

内蒙古农业大学 职业技术学院 内蒙古 包头 014109

[摘 要]针对新工科新农科人才培养的需求，分析课程特点及学情现状，在数据挖掘应用课程中提出开发与农牧类数据相关的教学案例，以层次聚类为例，对案例的设计进行了详细说明，为课程建设和课程改革提供思路。

[关键词]数据挖掘；教学案例；教学改革

[中图分类号] TP3-05 [文献标识码]A [文章编号]1893-9338(2026)-0019-82 [收稿日期]2025-12-30

一、引言

“四新”建设是高等教育应对未来挑战的战略先手旗，用好学科交叉融合对学科建设和人才培养有紧密联系[1]。随着教育部积极推进新工科建设、新农科建设，这对涉农高校的理工专业人才培养提出了新要求[2]。在新工科背景下，数据挖掘课程应以解决实际问题为导向，摒弃传统的以单一理论和算法为主的设计，数据挖掘应用涉及不同学科领域，有必要探索该交叉融合背景下的课程建设。如何结合理论与实践，激发学生的学习兴趣，将抽象的理论应用于实际问题中，是教育者必须深入思考的问题[3]。

二、课程特点及现状

数据挖掘应用是一门多学科交叉的综合性课程，课程主要讲授数据挖掘技术的理论知识和如何利用数据挖掘技术进行数据分析这两个教学目标。课程中理论知识涉及的内容广泛，包括很多经典算法，如：逻辑回归算法、关联规则、决策树、朴素贝叶斯、支持向量机、聚类算法等。以往采用的教学模式，首先通过理论课时讲授算法原理，然后通过传统的案例做一些验证性的实践，例如鸢尾花分类、波士顿房价预测、啤酒与尿布关联分析等，这些案例能够帮助学生领悟算法的原理。对于学生，仅仅局限于对算法原理的学习和理解是不够的，在未来的工作中，最重要的还是要具备利用所学的知识解决问题的能力，才能更好的适应工作岗位的需要[4,5]。

因此开发与农牧类数据相关的项目教学案例，将抽象的理论运用到真实场景中，让学生真切的感受自己所学的知识是如何解决实际问题的，激发他们的学习兴趣，提高他们的学习积极性。

三、教学案例设计原则

以学以致用为载体，导入并驱动理论知识教学。针对新工科人才培养的需求，结合学生的知识背景，采用案例式教学的原则进行案例设计，考虑到学校背景，作为地方农牧类高校，可以围绕农业和畜牧业本校传统

优势学科，探索学术科研项目中运用数据挖掘技术的典型问题或场景，建立与本课程教学内容相匹配的案例，将其他学科中的典型案例与数据挖掘技术中的经典算法进行融合重构，使枯燥的理论知识变得贴近现实，让学生意识到学有所用，以及数据挖掘应用的重要性。

四、教学案例

数据挖掘的过程是数据采集、数据预处理、特征工程、挖掘建模及结果分析。本文重点讲解挖掘的过程及分析。数据挖掘算法是本课程的重点，要求学生掌握常见算法的基本思想，并能使用python程序语言实现，常见挖掘算法包括分类、聚类和关联规则等。下面以聚类为例说明教学案例的设计思路。

（一）案例描述

某研究者收集了24种菌株，其中，17~22号为已知的标准菌株，分别取自牛、羊、犬、猪、鼠、绵羊，其他为未知菌株。现测得各菌株的16种脂肪酸百分含量，数据见表，试对此数据进行聚类分析，以便了解哪些未知菌株与已知的标准菌株在全部指标上最为接近[6,7]。24个菌株的16种脂肪酸百分含量数据集见表1。

表1 24个菌株的16种脂肪酸百分含量

菌株号	X_1	X_2	…	X_{15}	X_{16}
1	0.772 8	18.870 1	…	0.000 0	0.000 0
2	0.864 2	19.926 3	…	0.000 0	0.000 0
…	…	…	…	…	…
23	1.013 3	17.258 5	…	1.065 8	0.636 6
24	0.334 6	7.042 8	…	1.123 9	0.738 2

（X_1~X_{16}分别代表16种脂肪酸。）

（二）案例预备知识

1.聚类的概念

聚类是将样本集划分为若干互不相交的子集，即样本簇，同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。聚类结果的簇内相似度高且簇间相似度低，通常相似度是根据描述对象的属性值评估的[8]。

2.层次聚类

层次聚类算法可分为自底向上的凝聚法和自顶向下的分裂法。本教学案例使用凝聚法，它的思想是先将每个样本各自作为一个簇，计算所有样本两两之间的距离，将距离最近的样本合并成一个簇，然后重新计算簇与簇之间的距离，每次都将距离最近的簇合并，如此迭代，直到所有的样本都聚成一个簇。

3.距离度量方法

最小连接法，两个簇之间的距离定义为

两簇中元素之间距离最小者。最大连接法，两个簇之间的距离定义为两簇中元素之间距离最远者。平均连接法，两簇之间的距离定义为两簇中所有点对之间距离的平均值。Ward法，两个簇的邻近度定义为两个簇合并时导致的平方误差的增量。

4.聚类模型评价指标

轮廓系数评价法、Calinski-Harabasz指数评价法。

（三）实验环境

Python、Pycharm、Pandas、Matplotlib、Sklearn。

（四）案例教学目标

理解聚类的概念。

理解层次聚类算法基本思想。

使用pyhton语言实现层次聚类模型构建。

掌握聚类模型的评价方法。

（五）案例分析过程

根据研究对象的不同，聚类分析分为样本聚类分析和指标聚类分析。前者使用距离法衡量样本个体的属性相似程度，后者使用相似系数衡量指标变量的属性相似程度。根据案例描述，聚类分析的目标是找出未知菌株在全部16种脂肪酸含量上与已知标准菌株最为接近，分析对象是每一个菌株样本，因此该案例属于样本聚类分析，使用距离法度量菌株相似性，距离越近，表明两个菌株的脂肪酸越相似。案例分析过程如图1所示。

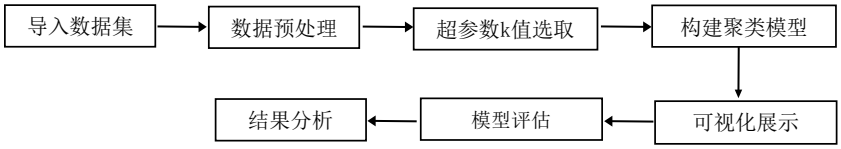

图 1　案例分析过程

1.超参数k值选取

在层次聚类中，可以通过轮廓系数判断哪个n_clusters值更适合当前数据，轮廓系数取值越高，代表当前的聚类效果越好。如图2所示，对于当前数据集超参数k值取7时效果最佳。

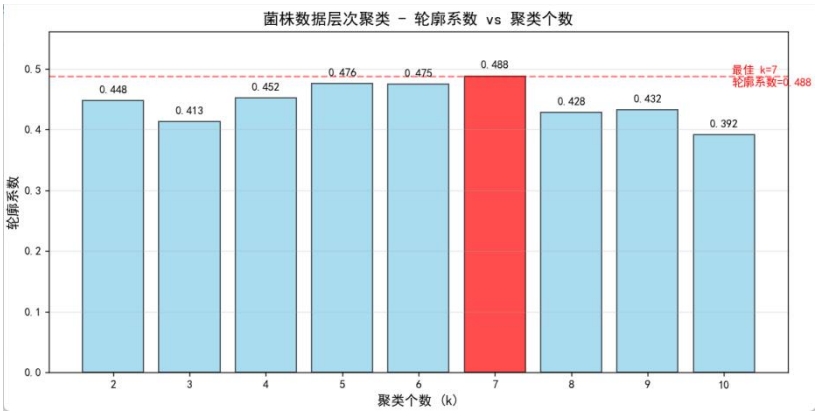

图 2　超参数k值选取

2.模型构建

首先，使用StandardScaler对原始数据

进行标准化处理，消除各脂肪酸指标量纲差异。然后调用AgglomerativeClustering类构建聚类模型。设置参数如下：

n_clusters=7：根据轮廓系数分析确定最优聚类数为7；

affinity='euclidean'：采用欧氏距离作为样本间相似性度量；

linkage='average'：使用平均连接法作为簇间距离的计算方式。

模型训练完成后，通过fit_predict方法获取每个样本的聚类标签，并将其添加到原始数据框中，聚类结果以Cluster列的形式保存在数据集中。

程序代码如下。

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import AgglomerativeClustering
df = pd.read_excel('../菌株数据.xlsx') #读取数据
df.set_index('菌株号', inplace=True)
scaler = StandardScaler() #数据标准化
X_scaled = scaler.fit_transform(df)
clustering = AgglomerativeClustering( #执行层次聚类分析
    n_clusters=7,                    # 设置K值为7，
    affinity='euclidean',            # 距离度量
    linkage='average'                # 使用平均连接方式
)
cluster_labels = clustering.fit_predict(X_scaled)

df['Cluster'] = cluster_labels
print(df.loc[:,'Cluster'])
```

3.可视化展示

层次聚类可以利用系统树图进行可视化展示，具体实现过程为，首先在标准化后的数据矩阵 X_scaled基础上，使用 linkage 函数计算样本间的层次聚类链接矩阵，连接方法设置为平均连接法。调用 dendrogram 函数绘制系统树图，横坐标标注各菌株编号，纵坐标表示样本或簇间的欧氏距离。系统树图能展示层次聚类的每一个步骤及不同合并簇的距离变化。

程序代码如下。

```
plt.figure(figsize=(12, 8))
Z = linkage(X_scaled, method=linkage_method)
dendrogram(Z, labels=labels, leaf_rotation=90)
plt.title(f'Cluster Analysis ({linkage_method} method)')
plt.xlabel('菌株号')
```

```
plt.ylabel('距离')
plt.rcParams['font.sans-serif'] = ['SimHei'] # 设置中文字体
plt.rcParams['axes.unicode_minus'] = False
plt.tight_layout()
plt.show()
```

系统树图结果如图3所示。

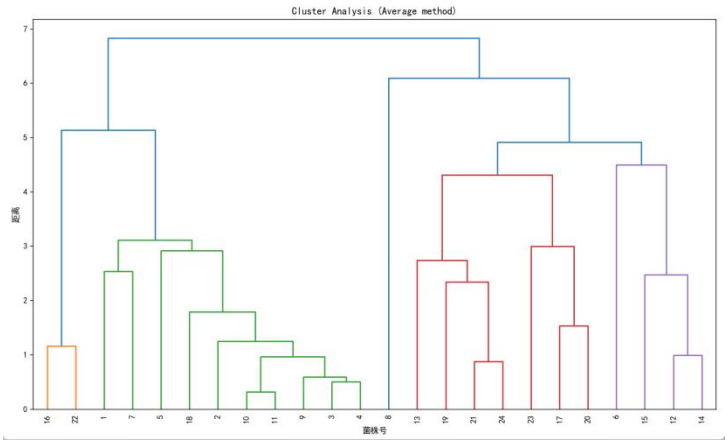

图 3　层次聚类系统树图

采用主成分分析方法对聚类结果进行进一步可视化展示。选取贡献率最高的两个主成分，将16维的脂肪酸含量数据降维至二维平面绘制散点图，如图4所示。

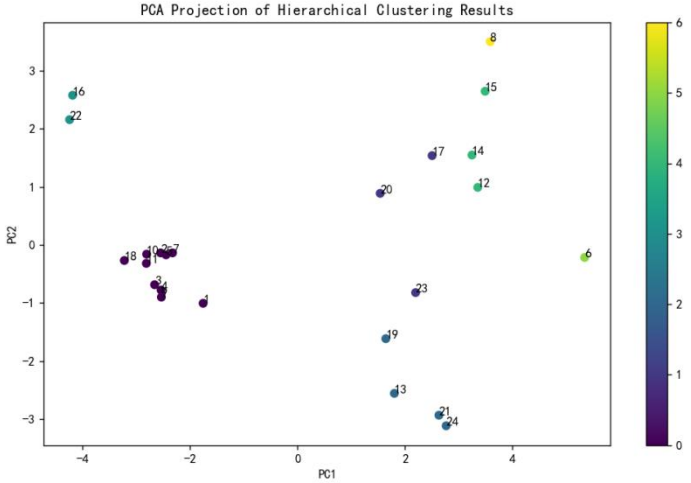

图 4　主成分分析二维散点图

（六）结果分析

因17~22号样品是已知菌株，故得知：24号与21号最接近，16号与22号最接近、23号与（17，20）号最接近、（2，3，4，5，9，10，11）号与18号最接近。通过对已知菌株和未知菌株的聚类分析，当未知菌株与某已知菌株分在同一类中时，可以判断该未知菌株与同类的已知菌株属于同一种类。

（七）模型评估

评估聚类模型采用CH指数（Calinski-Harabasz指数评价法），该指标通过计算簇间离散度与簇内离散度的比值来衡量聚类效果，其值越高，表明簇结构越紧密，且簇间分离度越好。不同连接方法的评估结果如表2所示。

表2 不同连接方法对比

连接方法	CH指数
最小连接法	20.44
最大连接法	20.45
平均连接法	20.58
Ward法	20.58

四种连接方法所得的CH指数非常接近，在本数据集上表明：不同连接方式得出的聚类效果实际相当。其中平均连接法与Ward法均为20.58，略高于最小与最大连接法，平均连接法与Ward法构建的聚类效果更佳。

（八）案例总结

以未知菌株和已知标准菌株为样本，以16 种脂肪酸含量为变量，根据距离系数作为统计量进行聚类分析。通过观察生成的聚类树状图，能识别出哪些未知菌株与已知的标准菌株被聚到了同一个类别下。通过此项目，让学生理解层次聚类分析中 k 值的含义以及学会如何找到最优的 k 值。通过衡量不同菌株个体之间在 16 种脂肪酸属性上的相似程度，让同学们理解聚类分析的实际意义。

五、结语

将数据挖掘技术融入农牧业案例教学，不仅能提升学生将专业知识应用于行业的实践能力，满足其职业发展需求，也能促使教学从书本走向实际，反哺教师对理论知识的深入理解。在教学中探索数据挖掘与农牧科学相融合，有助于培养高层次的复合型计算机人才，从而更好地服务社会发展。

基金项目：基于新工科背景下的学科交叉课程的研究应用——以《数据仓库与数据挖掘》为例，2023 年度内蒙古农业大学教育教学改革研究项目，RC2300000868；智慧农牧业科技创新团队，2023 年度，TDE202308。

作者简介：胡皎月（1992-），内蒙古包头市，女，讲师，硕士，研究方向为数据挖掘应用;侯佳（1981-），内蒙古呼和浩特市，女，副教授，硕士，研究方向农学。

参考文献：

[1]马陆亭.新工科、新医科、新农科、新文科——从教育理念到范式变革[J].中国高等育,2022,(12):9-11.

[2]何菁,周统建."双一流"背景下农林院校跨学科建设的战略选择——知识创新中重构的视角[J].

中国高校科技，2022(1):7-11.
[3]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.
[4]曾萍,韦杰,黄颖琦,等.医学信息分析技术与应用课程教学案例建设研究[J].软件,2020,41(11):19-22
[5]殷丽凤,刘震.基于产教融合的机器学习课程实践能力提升教学改革[J].计算机教育,2025,(04):156-161
[6]朱永平,和凤美.田间试验与统计分析实验教程[M].北京:中国农业出版社,2019:160-165.
[7]胡纯严,胡良平.无序样品聚类分析——基于马氏和中间距离法[J].四川精神卫生,2024,37(S1):60-65.
[8]周志华.机器学习[M].北京:清华大学出版社,2020:197-214.

Curriculum Development of Teaching Cases for Applications of Data Mining

Hu Jiaoyue,Hou Jia
Vocational and Technical College, Inner Mongolia Agricultural University, Baotou, 014109, China

Abstract: Addressing the need to cultivate talent in Emerging Engineering and Agricultural Disciplines, this paper analyzes the specific context of the Data Mining Applications course and proposes the creation of teaching cases based on agricultural and pastoral data.A detailed design of such a case, focusing on the hierarchical clustering algorithm, is provided to inform and inspire ongoing curriculum construction and innovation.

Key words: data mining; teaching cases; teaching reform