数据挖掘实验指导书

2020.8修订

数据挖掘实验是深入数据挖掘课程内容的必要环节，是将理论模型应用于实际问题的必要途径。

数据挖掘按照任务可划分为分类，关联和聚类三大类任务，从下列任务中选择一个任务来实现，平台环境不限。每个任务又分为验证型实验和设计性实验两个层次。验证型实验旨在通过自己编程实现“数据来源—建模—模型分析”的完整过程，使学生能更深入的了解数据，模型并能对挖掘结果进行多维度的分析。设计性实验则旨在针对实际应用中可能出现的数据条件不充分，挖掘结果不理想等问题，从数据处理、模型优化等角度进行改进，从而提高任务的效果，旨在训练学生对新问题的解决能力以及创新思维的训练。学生可根据自己的实际情况选作验证型或设计性实验。

目录：

1. 分类任务
2. 验证型实验
3. 设计型实验
4. 关联分析
5. 验证型实验
6. 设计型实验
7. 聚类分析
8. 验证型实验
9. 设计型实验

## 一、分类任务

### 1．验证型实验

1.1目的

1）理解分类任务

2）考察学生对数据预处理步骤的理解，强化预处理的重要性

3）基模型可以调用已有的包，训练学生熟悉数据挖掘的基本框架

4）学会多维度的对模型进行评估以及模型中参数的讨论

1.2数据集

1）UCI的soybean数据集， 数据中包含缺失值

2）UCI蛋白质数据集，特征处理

3）其他数据集

数据说明：给一个train和一个test

1.3 方法要求

1）要有针对数据特点的预处理步骤

2）原则上不限制模型，决策树，NB，NN，SVM，random forest均可，且不限于上述方法

1.4 结果要求

1）实现一个基本分类模型，并计算其评估指标如准确率，召回率等

2）对模型中的关键参数，（如决策树中停止分裂条件，NN中层数等参数）进行不同范围的取值，讨论参数的最佳取值范围。

3）对进行数据预处理和未进行预处理的分类结果进行对比，以验证预处理步骤的有效性

4）若对同一数据采用两种或多种模型进行了分类，对多种模型结果进行对比，以评估模型对该数据集上分类任务的适用性。

1.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）填写了测试数据集的预测标签的文档文本或excel

### 2．设计型实验

2.1目的

1）在理解设计性实验的基础上，针对实际应用中存在的问题对模型进行优化改进，具体可参考以下内容：

针对模型的稳定性差或分类效果差，尝试采用多模型的组合方法如 bagging， boosting方法进行优化

针对数据中大量缺失标签的问题，尝试采用半监督等方法进行建模

针对数据中01类比重失衡的不平衡数据，如入侵检测，信用卡欺诈，垃圾邮件检测等实际问题，进行有效建模

针对多标签数据进行多标签分类

针对数据中特征过少或过于稀疏的问题，进行分类

2）考察学生针对实际应用问题的分析和解决能力

2.2数据集

1）提供部分数据集，如短文本，垃圾邮件检测等

2）学生也可自行下载其他数据集

数据说明：给一个train和一个test

2.3 方法要求

1）要有针对数据特点和任务特定要求的具体方法，方法可自行改进设计，也可参考相关文献中已有的方法思路，自行进行方法的编码实现

2）原则上不限制模型方法

2.4 结果要求

1）实现一个基本分类模型，并计算其评估指标如准确率，召回率等

2）对方法的有效性进行对比、评估和分析

2.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）填写了测试数据集的预测标签的文档文本或excel

设计型方法的具体参考资料：

1）数据中特征过少：

数据集1：短文本分类数据集Spin

问题描述：特征少导致分类效果不理想。

参考方案1：借助外部资源如同义词库，查找同义词来扩充原文数据

例：the food is awful 通过词库查找bad 是awful的近义词

则原文数据替换为 the food is awful bad。

参考方案2：借助聚类，相关性评估等指标来在train数据中找近义词，并将该近义词填充到原文数据中。

要求：1）对比未扩充与扩充数据的分类效果；

2）其他优化方法可自行增加

2）数据中标签大量缺失的半监督分类

数据集1：垃圾邮件过滤spam

问题描述：实际应用中，标签的获取是十分昂贵的。往往能获取大量数据，但只能获取少量的标签，因此需要开展标签缺失条件下的分类。半监督分类方法是解决该问题的一个有效途径。半监督方法简单的可描述为train中仅有少量的标签数据，需要完成对test的预测

参考方案1：结合聚类和分类两个方法进行处理

以少量标签信息为中心点，对train中的未标签信息进行聚类，给未标签数据进行预标签；

基于train进行分类器构建，并对test预测

参考方案2：

以少量标签数据作为train，训练多个分类器

利用多个分类器对未标签数据进行预标签，选出其中最可信的标签作为其最终标签

基于train进行分类器构建，并对test预测

3）数据集是非一次性读入的

数据集1：Tweet

数据集2：KDDcup

任务描述：在在线的网络入侵检测/在线的信用卡欺诈检测/在线的垃圾邮件检测等应用环境中，数据是分批到来，不能一次性读入，数据中的分类可能存在变化

评估标准：以第i个窗口构建的分类器对第i+1个窗口的数据进行测试，计算整个数据流动过程中（即第2个到最后一个窗口）的所有准确率的平均值

参考方案：

给定一个窗口尺寸w，每次读入w个样本，对这w个样本构建分类器

检测是否发生漂移：当前窗口与上次窗口的数据差异，可通过分类精度评估

对分类模型进行更新：更新策略

4）不平衡数据分类

数据集：信贷类数据Lonestate

任务描述：针对不平衡数据完成分类任务。

## 二、关联分析

### 1．验证型实验

1.1目的

1）理解关联规则挖掘的基本任务

2）实现基本的关联规则算法，鼓励自主实现，也可调用已有的包

4）对关联规则结果和算法的性能进行评估和分析

1.2数据集

1）Groceries数据集，记录了某个杂货店一个月的真实交易记录。具体是9835行，169列，即9835条消费记录、169个不同商品[//download.csdn.net/download/beauty0522/10616314?utm\_source=bbsseo](https://download.csdn.net/download/beauty0522/10616314?utm_source=bbsseo" \t "_blank)

2）T10I4D100K（IBM数据生成器，网站访问数据）

1.3 方法要求

1）Apriori和FP-growth任选一个算法实现关联规则

2）也可选择改进版的关联规则方法

1.4 结果要求

1）给出频繁项集和关联规则的结果，并计算其支持度，置信度和其他指标

2）图表分析算法的时间性能和空间性能跟数据集大小，支持度阈值等因素的定性定量关系。例如，讨论support在[0.1-0.4]之间算法的运行时间变化，频繁项集数目变化等。

1.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）关联规则结果文档

### 2．设计型实验

2.1目的

1）在理解设计性实验的基础上，针对实际应用中存在的问题对模型进行优化改进，具体可参考以下内容：

针对模型的时空性能进行算法改进

设计引入多层或多维支持度阈值的算法实现

根据用户的历史购物数据进行推荐或个性化推荐。

2）考察学生针对实际应用问题和新问题的分析和解决能力

2.2数据集

1）参考验证型实验的数据集

2）学生也可自行下载其他数据集

2.3 方法要求

1）要有针对数据特点和任务特定要求的具体方法，方法可自行改进设计，也可参考相关的文献中已有的方法思路，自行进行方法的编码实现

2）原则上不限制模型方法

2.4 结果要求

1）实现并改进关联规则方法，并计算关联规则的各项评估指标

2）对方法的有效性进行对比、评估和分析

2.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）关联规则结果

设计型方法的具体参考内容：

1. 针对模型的时空性能进行算法改进

参考课件中的改进方法，从底层数据存储，数据库投影，划分等角度对关联规则方法及逆行改进，提高其时空性能。

1. 设计引入多层或多维支持度阈值的算法实现

实际应用中单一的支持度阈值是不适用的，考察多阈值或变阈值的关联规则分析方法。例如针对快速消耗类项目（牙膏，牛奶等）和慢消耗类项目（电视等）设定不同的阈值。

1. 根据用户的历史购物数据进行推荐或个性化推荐。

在关联分析的基础上，根据给定的用户数据进行有效的个性化推荐。

## 三、聚类分析

### 1．验证型实验

1.1目的

1）理解聚类任务

2）考察学生对数据预处理步骤的理解，理解数据归一等预处理步骤对聚类结果的影响

3）算法可自行设计也可调用已有的包，训练学生熟悉聚类的基本算法思想

4）学会多维度的对模型进行评估以及模型中参数的讨论

1.2数据集

1）UCI的中的聚类数据集或分类数据集（忽略标签）

2）其他数据集

1.3 方法要求

1）要有针对数据特点的预处理步骤

2）原则上不限制方法和模型，KNN，二分KNN，层次聚类，随机游走等，且不限于上述方法

1.4 结果要求

1）实现一个基本聚类模型，并计算其评估指标如准确率，召回率，内聚性，分离性等

2）对模型中的关键参数，（如k）进行不同范围的取值，讨论参数的最佳取值范围。

3）对进行数据预处理和未进行预处理的分类结果进行对比，以验证预处理步骤的有效性

4）若对同一数据采用两种或多种模型进行了聚类，对多种模型结果进行对比，以评估模型对该数据集上分类任务的适用性。

5）对同一算法在多个数据集上的结果进行对比分析，分析算法的特点和适用场景。

1.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）聚类结果文档

### 2．设计性实验

2.1目的

1）在理解设计性实验的基础上，针对实际应用中存在的问题对模型进行优化改进，具体可参考以下内容：

针对聚类中的噪音数据进行方法优化

2.2数据集

1）客户群体划分数据集（待整理）

2）学生也可自行下载其他数据集

2.3 方法要求

1）要有针对数据特点和任务特定要求的具体方法，方法可自行改进设计，也可相关的文献参考已有的方法思路，自行进行方法的编码实现

2）原则上不限制模型方法

2.4 结果要求

1）实现一个基本聚类模型，并计算其评估指标如准确率，召回率和内聚性等无监督指标

2）对方法的有效性进行对比、评估和分析

2.5 提交材料清单

1）代码源文件

2）报告文档，文档中需说明数据，方法，运行结果，对比效果图等

3）聚类结果文本

数据集参考：

UCI：

Kaggle：https://www.kaggle.com/competitions

KDnuggets：<https://www.kdnuggets.com/datasets/index.html>

其他汇总：

<https://blog.csdn.net/weixin_44784939/article/details/96453965>

链接: https://pan.baidu.com/s/18Wv1ypl9V99YaM7F8hLHIg 提取码: d4m4