**《数据挖掘》期末大作业要求**

项目目标

本学期，《数据挖掘》课程中共学习了多种不同的数据挖掘方法：划分聚类、层次聚类、密度聚类、朴素贝叶斯分类、基于规则的分类以及最近邻分类器等。对于任何数据挖掘方法，一个方法不可能适用于所有的数据，同样的数据可以使用多种方法，但是每个方法的性能是有差别的。因此，确定针对特定问题采用哪种学习算法的唯一方法是**尝试多种算法**，看看哪种算法效果最好。随着这些算法的经验，你将开始看到某些方法在某些问题上比其他方法更有效，所以，你可以更有经验地选择适合当前问题的算法。

为此，本次大作业的主要任务是将本学期学习的数据挖掘中的3种以上方法，应用于更大规模的数据挖掘任务。具体步骤如下：

（1）需要自行查找选择一个具有超过1000个标记的数据集，并使用3种以上方法对其进行挖掘。

（2）对数据集进行分类后，需要比较所选分类器的结果。必须包括以下分析（需要体现在大作业报告中）：

* 训练和测试的计算时间
* 训练和测试的计算复杂度（近似值）
* 参数选择的交叉验证
* ROC曲线
* 混淆矩阵

（3）开发**基于Web的数据挖掘系统**。必须包括以下模块：

* 数据存储和调用：利用数据库对原始数据集进行存取。
* 数据可视化：对原始数据集进行可视化展示，可利用现有Web前端可视化框架（比如Echarts等）。此步骤必须从数据库中实时读取数据，不允许利用文本进行中转。
* 输入：可以选择不同的训练好的模型（从前两步对训练好的模型进行导出）、可以输入新的数据进行实时分析。
* 输出：对挖掘结果进行展示。

（4）大作业报告撰写，至少包括以下章节：

* 背景
* 相关工作
* 方法
* 实验
* 可视化系统实现
* 总结
* 参考文献

项目方法

可以自由选择你喜欢的任何数据集，但要遵守下面提到的数据类型的约束。数据集必须包含至少1000个两个类以上的样本示例。数据集来源不局限于以下数据网站：[http://archive.ics.uci.edu/datasets、https://www.kaggle.com/datasets、https://huggingface.co/datasets](http://archive.ics.uci.edu/ml、https://www.kaggle.com/datasets、https://huggingface.co/datasets)。

如上所述，本项目必须使用从已经学习的数据挖掘方法中选择3种以上不同的方法来对数据进行数据挖掘。训练分类器必须至少使用750个数据点来训练分类器（剩余的250个或更多数据点将用于测试）。在执行训练时，必须执行三个分析：

计算时间：可以使用Python时间函数将此测量内置到你的代码中，或者你可以简单地用手表记录时间（到最接近的秒就足够准确了）。

训练参数：指算法中的超参数的选择，这些超参数必须调整以提供合理的结果。使用交叉验证来确定超参数。要执行交叉验证，只需将训练集分成几部分，在一个部分进行测试，在另一个部分上进行训练。通常，我们可以执行所谓的d折交叉验证。这个想法是我们将训练数据集划分为d个集合，在d-1集合的每个可能组合上进行训练，并在剩余的集合上进行验证。例如，使用750个数据点进行训练，我们可以通过将设置的训练数据随机分为3组，每组250个数据元素来执行3折交叉验证。然后，我们将执行3个实验，每个实验使用不同的250个元素进行验证（因此500个元素用于训练）。对于要确定的参数的每个值，将重复3次实验，并且该参数值的分类误差将在3个实验中取平均值。

记录分类结果：必须至少保留 25%的数据集用于测试。请注意，此数据不得在任何算法的训练期间使用，包括在确定参数时执行交叉验证。算法的测试结果必须以两种方式显示。首先，必须计算混淆矩阵，混淆矩阵形式如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 预测 | |
| 类A | 类B |
| 实际 | 类A |  |  |
| 类B |  |  |

必须显示结果的第二种方法是使用 ROC（受试者工作特性）曲线。

要求

1. 不允许抄袭（重复率大于15%），包括网络及班内，一旦发现0分处理；数据集选用相同组将被重点审查。
2. 分小组完成期末大作业，3-4人一组。
3. 必须实现至少3种数据挖掘算法，超过3种有加分。
4. 按小组形式提交大作业，将每个小组的所有电子版上交材料（主要包括电子版报告、源码、数据）放在以“组长学号\_姓名”命名的文件夹中。压缩后上传至学习委员处文件夹。
5. 提交纸质大作业报告。大作业封面模版和评分表见附件，
6. 期末大作业提交时间（提交项目+报告）：做课堂报告时提交。地点：力行楼2405（从17周开始，随堂汇报，每组汇报时长12分钟，3分钟提问）
7. 期末课程报告结构：见期末报告模板