# 数据挖掘Lab3实验报告

161220096 欧阳鸿荣

## 实验要求

在提供的数据集上，基于10折的交叉验证，使用J48(C4.8)，朴素贝叶斯，SVM，神经网络，kNN算法以及它们使用装袋(Bagging)的集成学习的版本对数据集训练。基于如accuracy，AUC等指标比较各种方法的表现。讨论各种方法的表现并且说明如何优化基于Bagging的kNN算法的性能。

## 实验配置

考虑到数据集格式(arff)是Weka的数据格式，因此本次实验采用开源的机器学习工具Weka进行数据挖掘。使用的版本是Weka 3.8版本。

对于实验要求的算法，J4.8 (C4.5)、Naïve Bayes、神经网络和kNN在Weka中都有提供。其中，神经网络采用的是Weka中的Multilayer Perceptron多层感知机，kNN算法采用Weka中的IBk(K-nearest neighbours classifier)分类器。而对于Weka默认安装里缺少的SVM，使用的是weka库中的libSVM进行实验。

在实验中，在对比多个方法的任务中，统一采用原始数据，10-fold cross validation和weka中的默认配置进行分类器的训练。在对kNN的优化中，才考虑对数据预处理和修改默认配置等因素。同时，在实验过程中发现libSVM的表现很不稳定，因此对其进行数据预处理实验进行对比，从此发现了对数据标准化等预处理手段的重要性。

## 实验原理

## 4. 实验过程与实验结果

## 5. 实验感悟

## 6. 参考链接