北京邮电大学软件学院

2019-2020 学年第一学期实验报告

课程名称： 数据挖掘

项目名称： 实验二——分类

项目完成人：

姓名： 罗舒婷 **学号：** 2017211951

指导教师： 牛琨

日　　期： 2019 年 10 月 28 日

一、 实验模型

我选择了对数据除去产品 ID 和型号两种属性，剩余属性数量为 34，对有缺失值的数据做过滤删除，对所有属性进行离散化处理。分类器使用决策树 J48，生成 model 的时候对分支大小 minNumObj 大小设置为 3.训练集和测试集的百分比为 8 比 2，即 Percentage split=50%，观察选择(Nom)文字输入系统.

二、 解释模型和选择原因

此为我选择的模型选择文字输入系统，在我尝试其余类时，明显可见键盘模型、RAM、ROM 识别准确率为 90%-95%之间时，除了键盘类型的其他属性 Kappa 系数落于低于 0.5 的区间。

Kappa 系数是一种衡量分类精度的指标。

所以我在选择模型时，识别准确率和 Kappa 系数是我的衡量指标。权重比例为 50%，50%。即二者之和。

在对分支大小、训练集和测试集百分比、选择类、数据是否离散化控制变量时，我选择的模型是识别准确率和 Kappa 系数两个数值都达到了比其他变量选择更高的数值。

以下数据具有代表性，其余数值均落于区间内，因此不于此放入表格展示。

控制变量共做尝试 50 余次。

类	minNumObj	准确率	Kappa 系数	percentage split	K+C
文字输入系统	3	90.50%	0.8346	50	1.739565
文字输入系统	20	90.21%	0.8302	50	1.732328
文字输入系统	3	89.71%	0.8249	80	1.722
文字输入系统	20	89.71%	0.82	80	1.7171
键盘类型	3	91.13%	0.8016	80	1.712948
键盘类型	20	90.78%	0.7942	80	1.702001
键盘类型	3	89.36%	0.7592	50	1.652817

键盘类型	20	89.65%	0.7563	50	1.652754
ROM	20	98.28%	0.4396	80	1.42237
RAM	3	93.90%	0.4509	50	1.389907
颜色数	3	74.47%	0.3293	50	1.073981
颜色数	20	73.76%	0.3093	80	1.046889
颜色数	3	72.34%	0.3085	80	1.031904
ROM	20	98.16%	0	50	0.98156
ROM	3	97.52%	0	80	0.975177
颜色数	20	68.79%	0.0687	50	0.75664