東南大學
SOUTHEAST UNIVERSITY

人工智能实验报告

学生姓名： 柳沿河

学生学号： 71117230

东南大学计算机科学与工程学院、软件学院

School of Computer Science & Engineering College of Software Engineering

Southeast University

二 0 一 九 年 十 二 月

一、 实验目的

- 掌握常见的机器学习任务：分类任务、或回归任务
- 掌握常用的机器学习算法，如线性模型、决策树、神经网络、支持向量机等，不仅包括算法的思想和具体过程，还包括算法的适用性、优缺点、参数（作用、敏感性、参数调节方法等）
- 掌握数据预处理的常用方法，如离散值、连续值的转换方法，缺失值的处理等等
- 深入了解一种机器学习算法平台，如 Weka 等；
- 能够调用算法，掌握参数选择方法
- 掌握机器学习实验方法的基础，包括实验方法（如留出法、交叉验证法等）、统计检验方法（如 t 检验、Friedman 检验等）
- 掌握算法分析和比较的方法和过程、包括结果的记录与展示，实验分析等

二、 实验内容

- 选取**分类**任务的 10 个数据集，并说明性能度量指标
- 对至少三个算法进行适当的参数选择（至少要有一个算法需进行参数选择。对使用默认参数的算法，需指明默认参数的取值，以及说明合理性，必要时，需进行参数敏感性分析实验）
- 使用某种实验方法进行实验
- 使用 t-test+sign-test 或 Friedman test 对算法之间进行对比
- 总结实验结果(对实验结果进行分析并得出相应的结论，比如哪个算法最优，各算法的优缺点，算法对参数的敏感性等)

三、 性能度量

名称	含义	趋势对结果的影响
Time taken to build model	构建模型所需的时间	越小越好
Correctly Classified Instances	正确分类的实例的百分比	越大越好
Incorrectly Classified Instances	错误分类的实例的百分比	越小越好
Kappa statistic	Kappa 统计量，[-1,1]范围的小数，用于评判分类器的分类的分类结果与随机分类的差异度	越接近1越好
Mean absolute error	平均绝对误差	越小越好
Root mean squared error	均方根误差	越小越好
Relative absolute error	相对绝对误差	越小越好
Root relative squared error	相对均方根误差	越小越好

四、 数据集信息

10 个 UCI 数据集的概要信息如下:

序号	数据集	样本数	属性数	类别数	各个类别样本数量	
					标签	数量
1	anneal	898	38	6	1	8
					2	99
					3	684
					4	0
					5	67
					U	40
2	balance-scale	625	4	3	L	288
					B	40
					R	288
3	credit-rating	690	15	2	+	307
					-	383
4	pima-diabetes	768	8	2	tested-negative	500
					tested-positive	268
5	Glass	214	9	7	bulid wind float	70
					bulid wind non-float	76
					vehic win float	17
					vehic win non-float	0
					containers	13
					tableware	9
					headlamps	29
6	hepatitis	155	19	2	DIE	32
					LIVE	123
7	iris	150	4	3	Iris-setosa	50
					Iris-versicolor	50
					Iris-virginica	50
8	kr-vs-kp	3196	36	2	won	1669
					nowin	1527
9	lymphography	148	18	4	normal	2
					metastases	67
					malign_lymph	46
					fibrosis	33
10	mushroom	8124	22	2	e	452
					p	4

五、 数据预处理

对数据集使用 weka.filters.supervised.ClassBalancer 进行预处理，该预处理器可以调整数据中的实例，使得各个类在数据集中的权重相同。

六、 实验方法

本实验采用 10 次 10 折交叉验证方法。

七、 参数设置

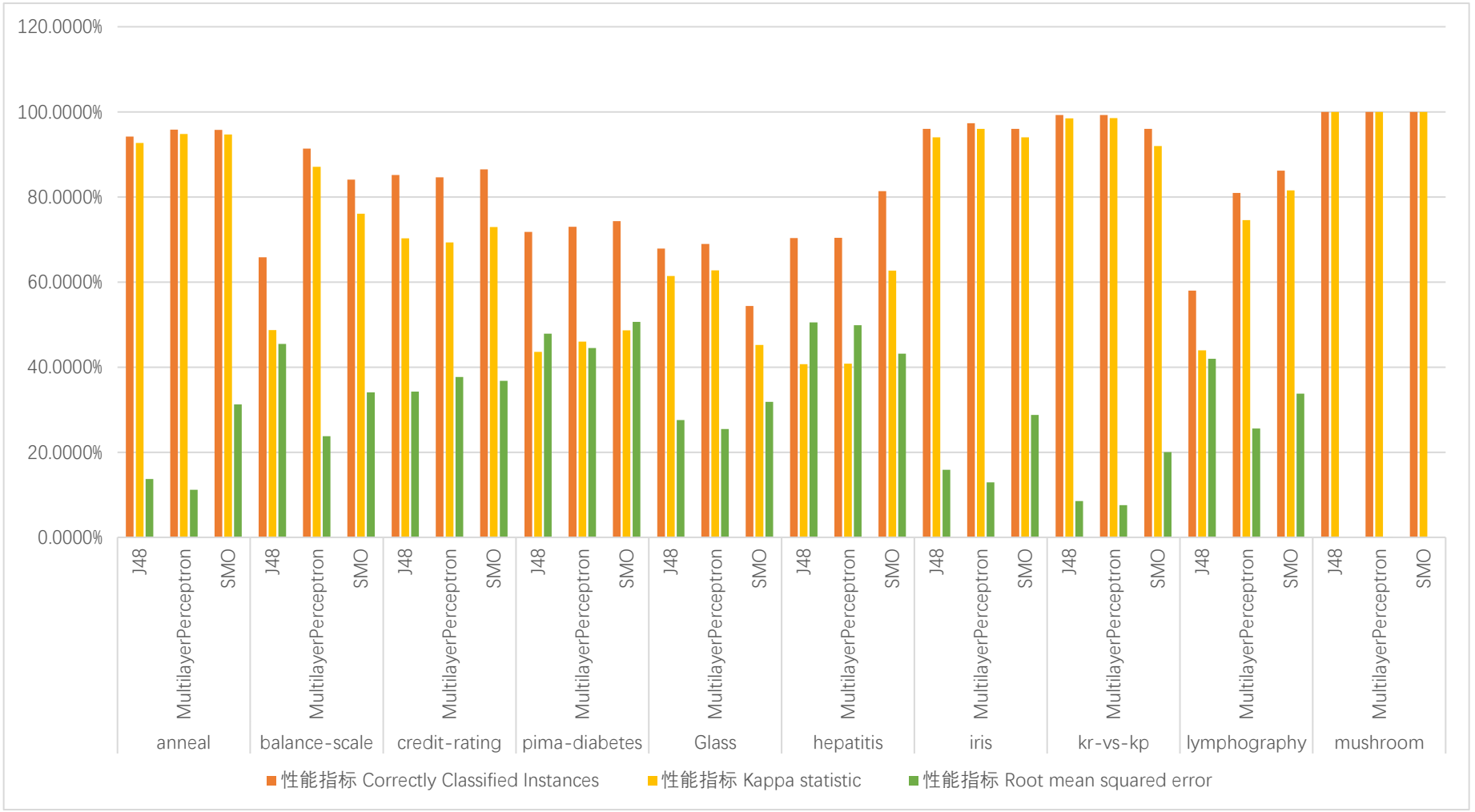
各个算法的参数设置如下:

算法	算法类型	参数	取值	说明
J48	决策树	置信因子C	0.2	该参数越小，剪枝越多（默认为0.25）
		最小对象数M	2	默认
MultilayerPerceptron	神经网络	学习率L	0.3	默认
		冲量M	0.2	默认
		训练周期数N	500	默认
		验证集百分比V	0	默认
		种子S	0	默认
		验证阈值E	20	默认
		隐藏层H	a	默认
SMO	支持向量机	复杂度C	1	默认
		收敛容忍参数L	0.001	默认且不能修改
		舍去误差P	1.00E-12	默认且不能修改

八、 实验结果

数据集	算法	性能指标							
		Time taken to build model(second)	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
anneal	J48	0.02	94.1784%	5.8216%	0.9272	0.0208	0.1371	7.7991%	37.5114%
	MultilayerPerceptron	20.17	95.8246%	4.1754%	0.9478	0.0156	0.1121	5.8467%	30.6504%
	SMO	0.21	95.7456%	4.2544%	0.9468	0.2237	0.3125	83.7700%	85.4801%
balance-scale	J48	0.01	65.8187%	34.1813%	0.4873	0.2507	0.4544	56.4008%	96.3977%
	MultilayerPerceptron	0.23	91.3761%	8.6239%	0.8706	0.0988	0.2376	22.2222%	50.4088%
	SMO	0.04	84.0561%	15.9439%	0.7608	0.2641	0.3405	59.4155%	72.2252%
credit-rating	J48	0.02	85.1328%	14.8672%	0.7027	0.1999	0.3423	39.9725%	68.4886%
	MultilayerPerceptron	3.71	84.6455%	15.3545%	0.6929	0.1710	0.3771	34.2069%	75.4283%
	SMO	0.46	86.4617%	13.5383%	0.7292	0.1354	0.3679	27.0761%	73.5874%
pima-diabetes	J48	0.01	71.8075%	28.1925%	0.4361	0.3433	0.4789	68.6640%	95.7865%
	MultilayerPerceptron	0.38	72.9940%	27.0060%	0.4599	0.3157	0.4450	63.1457%	88.9998%
	SMO	0.01	74.3343%	25.6657%	0.4867	0.2567	0.5066	51.3310%	101.3219%
Glass	J48	0	67.8473%	32.1527%	0.6142	0.1043	0.2755	43.7243%	79.7550%
	MultilayerPerceptron	0.27	68.9726%	31.0274%	0.6277	0.1002	0.2546	41.9856%	73.7141%
	SMO	0.03	54.3537%	45.6463%	0.4522	0.2147	0.3182	89.9719%	92.1112%
hepatitis	J48	0	70.3379%	29.6621%	0.4068	0.3204	0.5052	64.0408%	100.9734%
	MultilayerPerceptron	0.24	70.4014%	29.5986%	0.4080	0.2961	0.4988	59.1890%	99.7096%
	SMO	0	81.3389%	18.6611%	0.6268	0.1866	0.4320	37.3015%	86.3462%
iris	J48	0	96.0000%	4.0000%	0.9400	0.0350	0.1586	7.8705%	33.6353%
	MultilayerPerceptron	0.06	97.3333%	2.6667%	0.9600	0.0327	0.1291	7.3555%	27.3796%
	SMO	0.01	96.0000%	4.0000%	0.9400	0.2311	0.2880	52.0000%	61.1010%
kr-vs-kp	J48	0.02	99.2504%	0.7496%	0.9850	0.0114	0.0853	2.2717%	17.0543%
	MultilayerPerceptron	14.6	99.2803%	0.7197%	0.9856	0.0090	0.0755	1.7938%	15.0909%
	SMO	0.47	95.9850%	4.0150%	0.9197	0.0404	0.2004	8.0301%	40.0750%
lymphography	J48	0	57.9728%	42.0272%	0.4396	0.2296	0.4199	58.6303%	92.6326%
	MultilayerPerceptron	0.76	80.9325%	19.0675%	0.7458	0.1166	0.2561	29.7787%	56.5026%
	SMO	0.03	86.1554%	13.8446%	0.8154	0.2667	0.3375	68.1253%	74.4637%
mushroom	J48	0.02	100.0000%	0.0000%	1.0000	0.0000	0.0000	0.0000%	0.0000%
	MultilayerPerceptron	337.41	100.0000%	0.0000%	1.0000	0.0002	0.0005	0.0316%	0.1026%
	SMO	0.76	100.0000%	0.0000%	1.0000	0.0000	0.0000	0.0000%	0.0000%

每个数据集上各个算法的性能（红色表示对应算法在对应数据集上正确率最低，绿色表示最高）



每个数据集上的各个算法的性能度量的柱状图对比

由上述数据大致分析可得：神经网络在正确率方面获得最好性能的次数最多，但是对于较大的数据集，建模时间较长；支持向量机在正确率方面的表现没有神经网络好，但是强于决策树；决策树很少能取得最好的表现，但建模速度往往最快。

九、 比较检验

- 检验方法：本实验采用 Friedman 检验与 Nemenyi 后续检验对三个算法进行比较检验
- 检验指标：分类正确率
- 比较步骤

1. 列出算法比较序值表

   算法比较序值表如下：

数据集	算法序值		
	J48	Multilayer	SMO
anneal	3	1	2
balance-scale	3	1	2
credit-rating	2	3	1
pima-diabetes	3	2	1
Glass	2	1	3
hepatitis	3	2	1
iris	2.5	1	2.5
kr-vs-kp	2	1	3
lymphography	3	2	1
mushroom	2	2	2
平均序值r	2.55	1.6	1.85

2. 使用 Friedman 检验来判断算法性能是否相同

   a) 三个算法平均序值不同，故性能不同

   b) 计算τ_{χ^2}：令N为数据集数量，k为算法数量，r_i为第i个算法的平均序值，则变量τ_{χ^2}的计算公式为：

$$\tau_{\chi^2}=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}r_i^{\,2}-\frac{k(k+1)^2}{4}\right)=\frac{12\times10}{3(3+1)}\times\left(2.55^2+1.6^2+1.85^2-\frac{3(3+1)^2}{4}\right)=4.85$$

该变量服从自由度为$k-1$的χ^2分布

c) 计算τ_F：

$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}=\frac{(10-1)\times4.85}{10\times(3-1)-4.85}=2.881$$

该变量服从自由度为$k-1$和$(k-1)(N-1)$的F分布

d) 判断“所有算法的性能相同”的假设是否成立

由F检验常用的临界值可得：在$\alpha=0.05,N=10,k=3$的临界值为3.555，而$2.881<3.555$,故可得：三种算法的性能没有显著区别

3. 使用 Nemenyi 后续检验区分算法

a) 计算平均序值差别的临界值域CD：

$$CD=q_\alpha\sqrt{\frac{k(k+1)}{6N}}$$

查表可得相应的$q_\alpha=2.344$，故可得$CD=2.344\times\sqrt{\frac{3(3+1)}{6\times10}}=1.048$

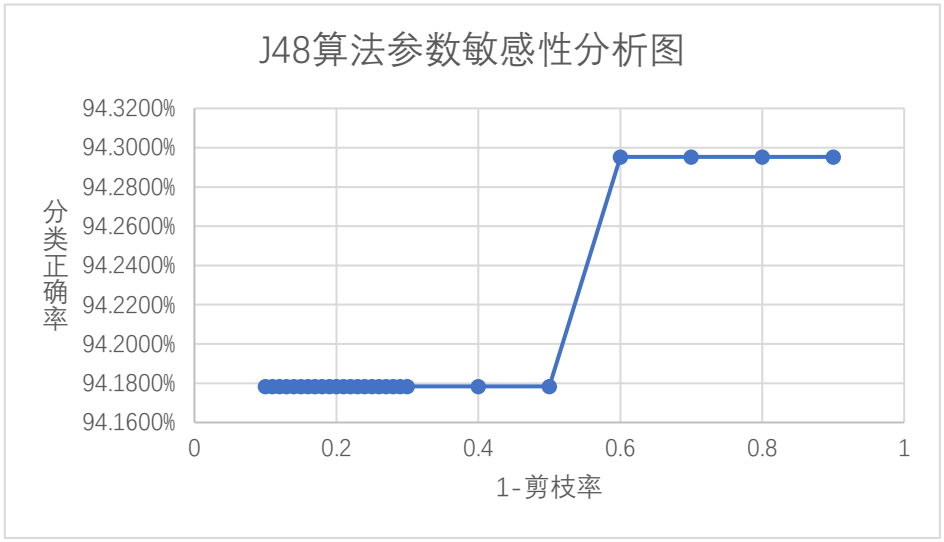
b) 计算各个算法两两之间的平均序值之差：

$$d_1=2.55-1.6=0.95;$$
$$d_2=2.55-1.85=0.7;$$
$$d_3=1.85-1.6=0.25;$$

可得d_1,d_2,d_3均小于CD，故进一步说明这三个算法之间性能差距不显著。

十、 参数敏感性分析

在上述实验中，对 J48 决策树算法的剪枝率参数进行了选择，得到算法性能与剪枝率变化之间的关系如下：



算法性能（分类正确率）与参数（剪枝率）之间的关系

可以看出在较大范围内参数取值的改变对算法性能影响不大，因此该算法对此参数不敏感。