实验报告

MF 1933128 周慧聪

分析wmc, dit, noc, cbo, rfc和 lcom的缺陷预测能力

在**calculate.R**文件中，计算描述性统计参数最小值、25%处值、中位值、75%处值、 最大值、平均值、偏度(skewness)和峰度(kurtosis)。

计算代码

```
max <- max(data[,i])
min <- min(data[,i])
mean <- mean(data[,i])
median <- median(data[,i])
QL <- quantile(data[,i], probs = 0.25)
QU <- quantile(data[,i], probs = 0.75)
skew <- skewness(data[,i])
kurt <- kurtosis(data[,i]) - 3
```

输出结果

	name	min	QL	median	QU	max	mean	skewness	kurtosis
1	wmc	0	3	6	12.5	123	11.44952	3.478202	15.08622
2	dit	1	1	2	4	8	2.565698	0.656867	-0.29915
3	noc	0	0	0	0	29	0.608575	7.332323	63.16812
4	cbo	0	4	8	18	171	14.49793	3.472056	16.41106
5	rfc	0	8	19	41	355	30.16183	3.014723	14.6192
6	lcom	0	0	3	22.5	6589	130.0816	7.684377	67.13383

与bug数据的相关系数: **Spearman和Pearson相关系xiangguanxish 以及统计显著性。

Spearman相关系数

在统计学中，斯皮尔曼等级相关系数用来估计两个变量X、Y之间的相关性，其中变量间的相关性可以使用单调函数来描述。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数时（即两个变量的变化趋势相同），两个变量之间的ρ可以达到+1或-1 ：

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

Pearson相关系数

皮尔逊相关也称为积差相关（或积矩相关）是英国统计学家皮尔逊于20世纪提出的一种计算直线相关的方法。假设有两个变量X、Y，那么两变量间的皮尔逊相关系数可通过以下公式计算(其中E是数学期望，cov表示协方差)：

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

t 统计检验

$$t = r \times \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

n为样本容量。

计算代码

```
spear <-cor(data[,i],bug,method="spearman")
spearT <- spear*(sqrt(n-2))/sqrt(1-spear^2)
pear <- cor(data[,i],bug,method="pearson")
pearT <- pear*(sqrt(n-2))/sqrt(1-pear^2)
new.data <- data.frame(metrics[i],spear,spearT,pear,pearT)
final.data <- rbind(final.data, new.data)
```

输出结果

	name	Spearman	Spearman.T	Pearson	Pearson.T
1	wmc	0.314245	0.662027	0.378792	0.818584
2	dit	-0.02612	-0.05226	-0.00186	-0.00372
3	noc	0.090944	0.182645	0.054916	0.109998
4	cbo	0.217624	0.445936	0.223544	0.458696
5	rfc	0.356342	0.762754	0.459294	1.034114
6	lcom	0.259252	0.536859	0.307576	0.646491

其中，T为显著性差异。数据中样本容量大于700，不适宜用t检验。

两个表格数据保存在**description.csv**和**coefficient.csv**中。

10种机器学习方法建立多变量的缺陷预测模型

10种机器学习方法

```
learner_names = c("classif.randomForest",#随机森林
                  "classif.mlp",#多层感知机
                  "classif.naiveBayes",#朴素贝叶斯
                  "classif.nnet",#神经网络
                  "classif.svm",#支持向量机
                  "classif.multinom",#多元回归
                  "classif.probit",#单位几率回归
                  "classif.lda",#线性判别分析
                  "classif.ksvm",#支持向量机
                  "classif.mlp"#多层感知机
  )
```

训练模型

通过10折交叉运算来训练和测试模型。

```
  classif.task = makeClassifTask(data = train_data, target = "bugs")
  classif.lrn = makeLearner(learner, predict.type = "prob")
  mod = mlr::train(classif.lrn,classif.task,subset=train.set)
  task.pred = predict(mod, task = classif.task, subset = test.set)
```

模型性能

评价上述缺陷预测模型的性能，包括**分类性能(评价指标为AUC)**和**排序性能(评价 指标为CE)**

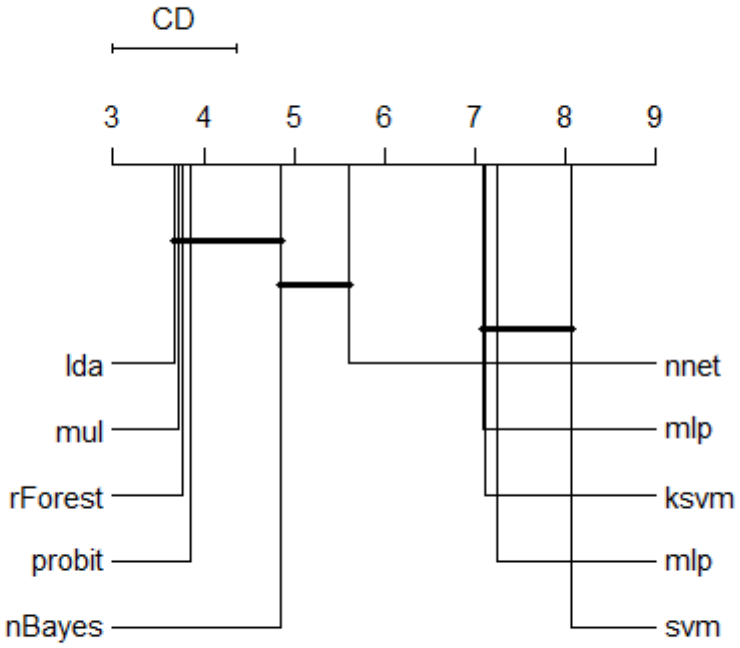
AUC可以通过performance函数获得，通过AUC可以计算得到**CE**:

$$\frac{CE_{\pi}(model) = Area_{\pi}(model) - Area_{\pi}(Random)}{Area_{\pi}(optimal) - Area_{\pi}(Random)}$$

其中，Random的AUC值为0.5，optimal的AUC值为1。带入计算可得：

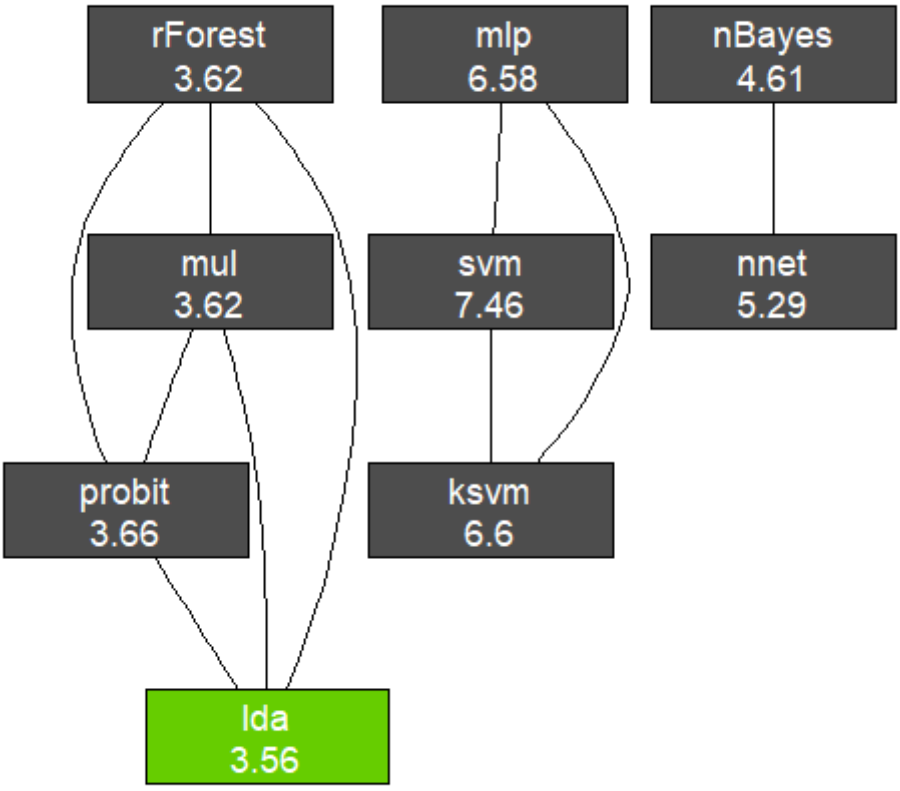
```
res = performance(task.pred, measures = list(mmce,auc ))
        model_auc = unname(res['auc'])
        model_ce = (m_auc - 0.5) / 0.5
```

利用CD图比较这10种模型在统计上的差别



根据图上可以得出，lda（线性判别分析）和naiveBayes（朴素贝叶斯）和神经网络没有比较大的区别，mlp（多层感知机）和svm(支持向量机)没有显著的区别。

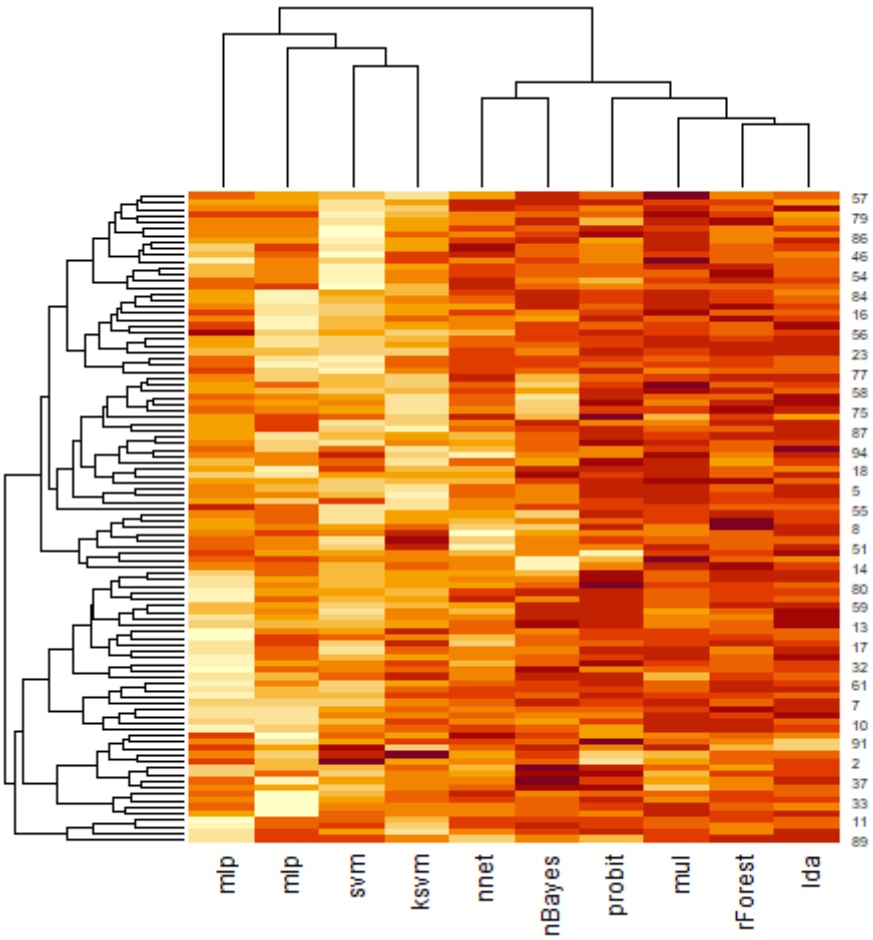
利用Algorithm图比较这10种模型在统计上的差别



从algorithm图中，可以看出，lda,mul,rForest,problit之间没有显著的差异。mlp，svm，ksvm之间没有显著差异。nBayes，nnet之间没有显著差异。没用显著差异的模型之间会用一条线连接。

相较于CD图，algorithm图中的比较信息更加详细。

利用heatmap展示10个模型在100个测试集上的结果



总结

在使用R的时候，由于不熟悉出现了很多问题。R的mlr包对于机器学习的方法集成度很高，非常便于使用，不过具体的参数以及他们的含义需要经过深入的学习和研究。