数据挖掘技术
分类:替代技术
第四章的课堂笔记
基于规则
数据挖掘导论，第二版
经过
谭、斯坦贝克、卡帕特内、库马尔

02/14/2018 数据挖掘导论，第 2 版 2
基于规则的分类器
●使用以下集合对记录进行分类
“如果…那么…”
●规则:(条件)→ y
在哪里
条件是属性的连词
u y 是班级标签
LHS:规则前提或条件
RHS:规则结果
分类规则示例:
u(血型=温暖)∧(产卵=是)→鸟类
u(应税收入< 50K) ∧(退款=是)→逃避=否

02/14/2018 数据挖掘导论，第 2 版 3
基于规则的分类器(示例)
R1:(生=否)∧(会飞=会)→鸟
R2:(生=否)∧(水中生活=是)→鱼
R3:(分娩=是)∧(血型=温暖)→哺乳动物
R4:(生=否)∧(会飞=否)→爬行动物
R5:(生活在水中=有时)→两栖动物
名字血型分娩能在水上飞行生活吗
人类温暖是不是没有哺乳动物
巨蟒冷不不不爬行动物
鲑鱼冷不不是的鱼
鲸鱼是温暖的是不是哺乳动物
青蛙冷不不有时两栖动物
科莫多冷不不不爬行动物
蝙蝠温暖是的是的没有哺乳动物
鸽子温暖不是的不鸟
猫是温暖的是不不哺乳动物
豹鲨冷是的不是的鱼
龟冷不不，有时是爬行动物
企鹅不暖和，有时是鸟
豪猪温暖是不是没有哺乳动物
鳗鱼冷不不是的鱼
蟒蛳冷不不，有时是两栖动物
吉拉怪物冷不不不爬行动物

鸭嘴兽温暖不不不哺乳动物
猫头鹰温暖不是的不鸟
海豚温暖是的不是的哺乳动物
鹰温暖不是的不鸟

02/14/2018 数据挖掘导论，第 2 版 4
基于规则的分类器的应用
●规则 r 覆盖实例 x，如果
实例满足规则的条件
R1:(生=否)∧(会飞=会)→鸟
R2:(生=否)∧(水中生活=是)→鱼
R3:(分娩=是)∧(血型=温暖)→哺乳动物
R4:(生=否)∧(会飞=否)→爬行动物
R5:(生活在水中=有时)→两栖动物
R1 的规则涵盖鹰= >鸟
R3 规则适用于灰熊= >哺乳动物
名字血型分娩能在水上飞行生活吗
霍克温不是的不?
灰熊温暖是不是不?

02/14/2018 数据挖掘导论，第 2 版 5
规则覆盖范围和准确性
●规则的覆盖范围:
记录分数
满足
规则的前身
●规则的准确性:
记录分数
满足
在那之前
也满足
a 的结果
规则
Tid 退款婚姻
状态
应纳税的
收入等级
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不

没有单个 90K 是 10
(状态=单身)→否
覆盖率= 40%，准确度= 50%

02/14/2018 数据挖掘导论，第 2 版 6
基于规则的分类器是如何工作的?
R1:(生=否)∧(会飞=会)→鸟
R2:(生=否)∧(水中生活=是)→鱼
R3:(分娩=是)∧(血型=温暖)→哺乳动物
R4:(生=否)∧(会飞=否)→爬行动物
R5:(生活在水中=有时)→两栖动物
狐猴触发规则 R3，所以它被归类为哺乳动物
海龟触发了 R4 和 R5
一条狗鱼鲨鱼不会触发任何规则
名字血型分娩能在水上飞行生活吗
狐猴温暖是不是不?
甲鱼冷不不有时候?
狗鱼鲨鱼冷是不是是的?

02/14/2018 数据挖掘导论，第 2 版 7
规则集的特征:策略 1
●相互排斥的规则
分类器包含互斥规则，如果
这些规则相互独立
每条记录最多由一条规则覆盖
●详尽的规则
如果分类器
解释了所有可能的组合
属性值
每条记录至少包含一条规则

02/14/2018 数据挖掘导论，第 2 版 8
规则集的特征:策略 2
●规则并不相互排斥
一条记录可能会触发多个规则
解决方案?
有序规则集
使用投票方案无序规则集
●规则并不详尽
记录不能触发任何规则
解决方案?
u 使用默认类别

02/14/2018 数据挖掘导论，第 2 版 9
有序规则集

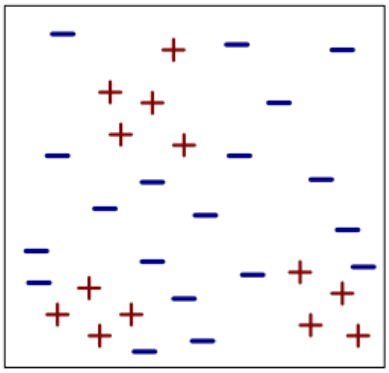
●规则根据其优先级排序
有序规则集被称为决策列表
●当测试记录呈现给分类器时
它被分配到它拥有的最高等级规则的类别标签
触发的
如果没有触发任何规则，则将其分配给默认类
R1:(生=否)∧(会飞=会)→鸟
R2:(生=否)∧(水中生活=是)→鱼
R3:(分娩=是)∧(血型=温暖)→哺乳动物
R4:(生=否)∧(会飞=否)→爬行动物
R5:(生活在水中=有时)→两栖动物
名字血型分娩能在水上飞行生活吗
甲鱼冷不不有时候?

02/14/2018 数据挖掘导论，第 2 版 10
规则排序方案
●基于规则的排序
单个规则根据其质量进行排序
●基于类别的排序
属于同一类的规则一起出现
基于规则的排序
(退款=是)= >否
(退款=否，婚姻状况= {单身，离婚}，
应税收入< 80K) == >否
(退款=否，婚姻状况= {单身，离婚}，
应税收入> 80K) =是
(退款=否，婚姻状况= {已婚})= = = >否
基于类别的排序
(退款=是)= >否
(退款=否，婚姻状况= {单身，离婚}，
应税收入< 80K) == >否
(退款=否，婚姻状况= {已婚})= = = >否
(退款=否，婚姻状况= {单身，离婚}，
应税收入> 80K) =是

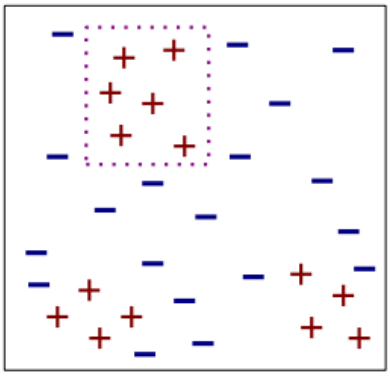
02/14/2018 数据挖掘导论，第 2 版 11
建筑分类规则
●直接方法:
u 直接从数据中提取规则
u 示例:开膛手、CN2、霍尔特 1R
●间接方法:
u 从其他分类模型中提取规则(例如
决策树、神经网络等)。
u 示例:C4.5 规则

02/14/2018 数据挖掘导论，第 2 版 12
直接方法:顺序覆盖
1.从空的规则开始
2.使用 LearnOneRule 函数增长规则
3.删除规则涵盖的培训记录
4.重复步骤(2)和(3)，直到停止标准
遇到了

Example of Sequential Covering

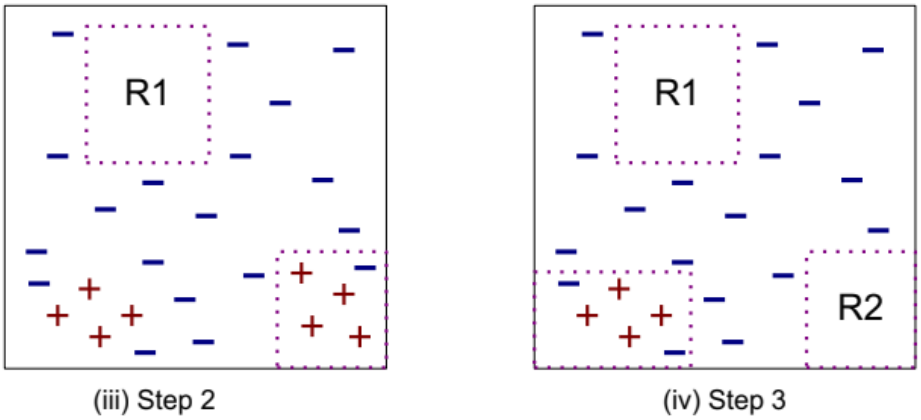


(i) Original Data



(ii) Step 1

Example of Sequential Covering…



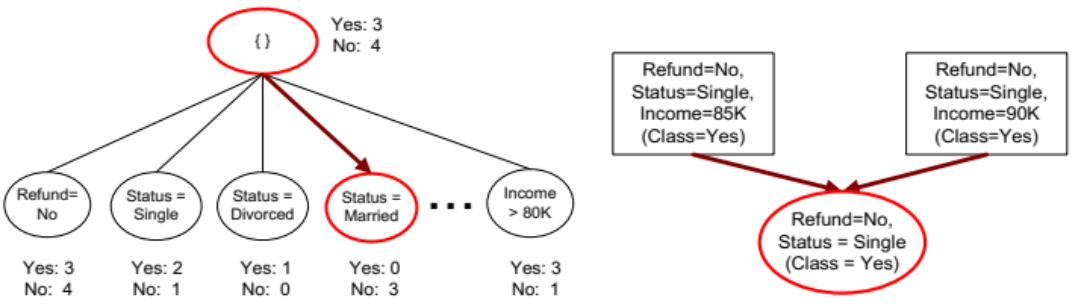
(iii) Step 2 (iv) Step 3

02/14/2018 数据挖掘导论，第 2 版 15
实例消除
●为什么我们需要
消除实例?
否则，下一个规则是
与以前的规则相同
●我们为什么要移除
积极的例子?
确保下一个规则是
不同的
●我们为什么要移除
负面实例?
防止低估
规则的准确性
比较规则 R2 和 R3
在图表中
class = +
类=
+
+ +
+ +
+
+ +
+ +
+
+

+
+
+
+
+ +
+
+
-
-
- -
-
-
-
-
-
-
-
-
-
-
+
+
+ +
+
+
+
R1
R3 R2
+
+

Rule Growing

- Two common strategies



(a) General-to-specific

(b) Specific-to-general

02/14/2018 数据挖掘导论，第 2 版 17
规则评估
●铝箔的信息增益
R0: {} = >类(初始规则)
R1: {A} = >类(添加连词后的规则)
增益(R0，R1) = t [对数(p1/(p1+n1))对数(p0/(p0 + n0))]
其中 t:涵盖的阳性实例数
R0 和 R1
P0:R0 涵盖的阳性实例数
n0:R0 覆盖的负实例数
P1:R1 涵盖的积极实例数量
n1:R1 涵盖的负面实例数量
铝箔:一级归纳
学习者早期规则-
基于学习的算法

02/14/2018 数据挖掘导论，第 2 版 18
直接方法:开膛手
●对于 2 类问题，选择下列类别之一
积极类，另一个作为消极类
学习积极课堂的规则
负类将是默认类
●对于多类问题
根据增加的班级排列班级
流行率(属于
特定类别)
首先学习最小班级的规则集，然后对待其余班级
作为负面类
重复下一个最小的类作为正类

02/14/2018 数据挖掘导论，第 2 版 19
直接方法:开膛手
●发展规则:
从空规则开始
添加连词，只要它们能改善铝箔
信息增益
当规则不再包含负面例子时停止
使用增量缩减立即修剪规则
错误修剪
修剪措施:v = (pn)/(p+n)
u p:中的规则所涵盖的正面示例的数量
验证集
联合国:中的规则所涵盖的负面例子的数量
验证集
修剪方法:删除任何最终序列
最大化 v 的条件

02/14/2018 数据挖掘导论，第 2 版 20

直接方法:开膛手

●建立规则集:

使用顺序覆盖算法

找到覆盖当前集合的最佳规则

正面例子

消除正面和负面的例子

被规则覆盖

每次向规则集中添加规则时，

计算新的描述长度

u 停止添加新规则，当新描述

长度比最小的描述长 d 位

迄今获得的长度

02/14/2018 数据挖掘导论，第 2 版 21

直接方法:开膛手

●优化规则集:

对于规则集中的每个规则

u 考虑 2 个备选规则:

替换规则(r*):从头开始增长新规则修订规则(r′):添加连词以扩展规则 r

u 将 r 的规则集与 r*的规则集进行比较

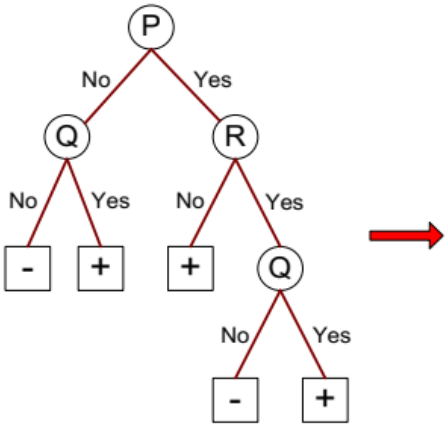
和 r′

u 选择最小化 MDL 原则的规则集

重复规则生成和规则优化

对于其余的正面例子

Indirect Methods



Rule Set

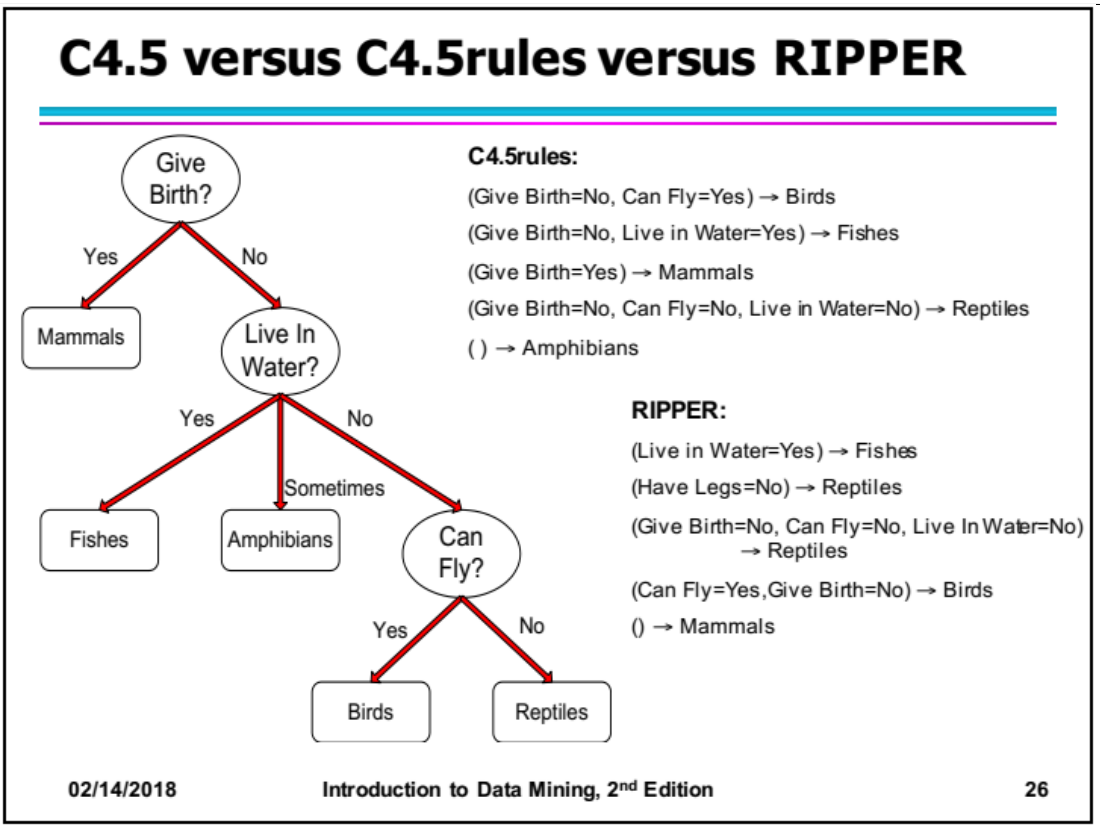
r1: (P=No,Q=No) ==> -
r2: (P=No,Q=Yes) ==> +
r3: (P=Yes,R=No) ==> +
r4: (P=Yes,R=Yes,Q=No) ==> -
r5: (P=Yes,R=Yes,Q=Yes) ==> +

02/14/2018 数据挖掘导论，第 2 版 23
间接方法:4.5 规则
●从未绘制的决策树中提取规则
●对于每个规则，r: A → y，
考虑另一条规则 r′:A′→y，其中 A′
是通过去掉一个连词得到的
在 A 中
比较 r 的悲观错误率
与所有 r 相对
如果其中一个替代规则的优先级较低，则进行修剪
悲观错误率
重复直到我们不能再提高
泛化误差

02/14/2018 数据挖掘导论，第 2 版 24
间接方法:4.5 规则
●不是对规则进行排序，而是对子集进行排序
规则(类排序)
每个子集都是规则的集合
相同的规则结果(类)
计算每个子集的描述长度
u 描述长度= L(误差)+ g L(模型)
u g 是一个参数，它考虑了
规则集中存在冗余属性

(默认值= 0.5)

02/14/2018 数据挖掘导论，第 2 版 25
例子
名生蛋能飞能在水中生存有腿类
人类是的不不不是的哺乳动物
蟒蛇不是的不不不爬行动物
鲑鱼不，是的，不，是的，不，鱼
鲸鱼是不是不是不是哺乳动物
青蛙不是的不有时是的两栖动物
科莫多没有是的没有是的爬行动物
蝙蝠是否是否是哺乳动物
鸽子不是的是的是的是的是的鸟
猫是不是不不是哺乳动物
豹鲨是否否是否鱼
乌龟不，是的，不，有时是爬行动物
企鹅不是的不有时是的鸟
豪猪是不是不不是哺乳动物
鳗鱼不是的不是的不鱼
蠑螈不是的不有时是的两栖动物
吉拉怪物不，是，不是，是，爬行动物
鸭嘴兽不是的不是的哺乳动物
猫头鹰不是的是的是的是的是的鸟
海豚是否否是否哺乳动物
鹰不是的是的是的是的是的鸟



C4.5 versus C4.5rules versus RIPPER

C4.5 and C4.5rules:

			PREDICTED CLASS			
		Amphibians	Fishes	Reptiles	Birds	Mammals
ACTUAL CLASS	Amphibians	2	0	0	0	0
	Fishes	0	2	0	0	1
	Reptiles	1	0	3	0	0
	Birds	1	0	0	3	0
	Mammals	0	0	1	0	6

RIPPER:

			PREDICTED CLASS			
		Amphibians	Fishes	Reptiles	Birds	Mammals
ACTUAL CLASS	Amphibians	0	0	0	0	2
	Fishes	0	3	0	0	0
	Reptiles	0	0	3	0	1
	Birds	0	0	1	2	1
	Mammals	0	2	1	0	4

02/14/2018 数据挖掘导论，第 2 版 28

基于规则的分类器的优势

●具有与决策树非常相似的特征

像决策树一样有表现力

易于解释

性能与决策树相当

可以处理冗余属性

●更适合处理不平衡的课程

●更难处理测试集中的缺失值