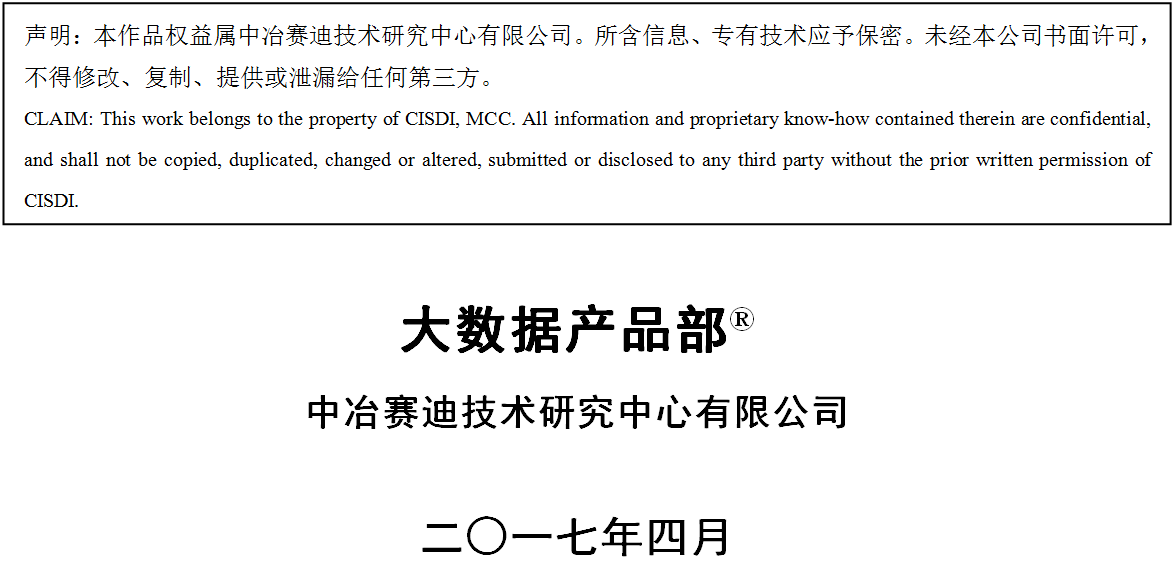
|  |
| --- |
| **CISDI中冶赛迪** |
| 测试题目——机器学习 |
| 主 研 人：Franz、Dargon  参 研 人：Jacky |
| 审 核 人：Mark |
|  |



目 录

[1.概述 3](#_Toc484523603)

[2.测试题目 3](#_Toc484523604)

[2.1问题1—算法设计 3](#_Toc484523605)

[2.2问题2—R与C混编 6](#_Toc484523606)

[3.交付物要求 6](#_Toc484523607)

# 1.概述

本文档包含了机器学习方向实习生的测试题目，请各位同学严格按照以下要求来完成测试。

# 2.测试题目

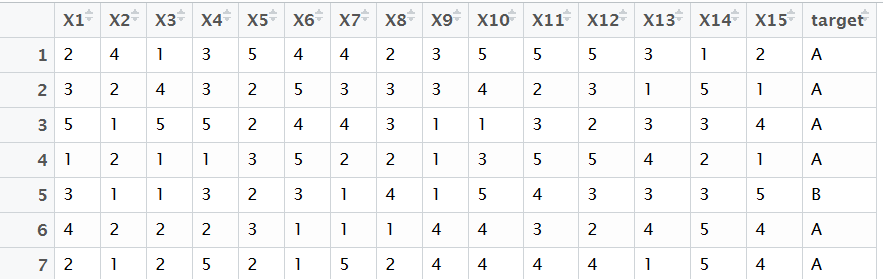
## 2.1问题1—算法设计

1）数据说明

测试题使用数据见res.RData。

* df

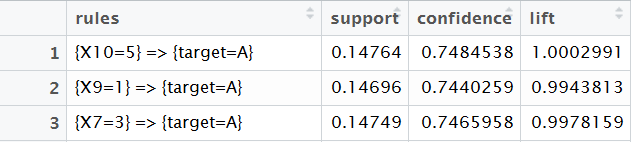
数据集（见图2.1-1示例）。用于生成规则和计算规则的各项指标。由10万条数据构成，每一条数据有16个属性（X1~X15和target），其中target仅可能为A或B。



**图2.1-1 df示例**

* res

由df通过apriori生成的59575条规则（见图2.1-2示例）。Apriori算法思路参见链接：<http://blog.csdn.net/lizhengnanhua/article/details/9061755>



**图2.1-2 res示例**

其中rules为规则，support为支持度，confidence为置信度，lift为提升度。规则由两部分组成：前项 => 后项，分别由大括号括出，前项为属性特征值对（即X10为属性，5为该属性的值），可以包含多个属性特征值对，由逗号分隔；后项全部为target=A。

2）具体要求

Step1：将res中的所有的规则按照等价规则分类，相互等价的规则为一类。

要求：运行时间尽可能少，此步算法复杂度必须小于等于O(R\*K)。

Step2：在每一类中，筛选出interest值最大的规则，如interest最大的规则存在多个，则均保留，其余规则删除。

要求：运行时间尽可能少，此步算法复杂度必须小于等于O(R\*LogR)。

注：在计算算法复杂度时，R为规则条数、K为规则平均项数、N为样本个数。

3）定义说明

* 定义1：等价属性特征值对

对于两个属性特征值对Xi=j, Xm=n, 如果i+j=m+n, 那么认为这两个属性特征值对是等价的。

示例：X1=7，X4=4这两个特征值对是等价的

* 定义2：等价规则

如果两个规则中的属性特征值对两两等价（不管顺序），那么我们认为这两个规则是等价的。

示例：{X1=7，X4=8，X2=6}=>{target=A}

{ X9=3，X3=5，X4=4}=>{target=A}

{X3=5，X2=10}=>{target=A}

第一个规则可写为{8,12,8}，第二个为{12,8,8}，第三个为{8,12}，可以看出，第一个规则与第二个规则是等价的，第三个规则与前两个规则不等价。

* 定义3：支持度



* 定义4：置信度



* 定义5：兴趣度



4）输出样例

假定我们已计算出下面规则的interest值

|  |  |  |
| --- | --- | --- |
| 序号 | rules | interest |
| 1 | {X1=7，X4=8，X2=6}=>{target=A} | 0.3 |
| 2 | { X9=1，X3=5，X4=4}=>{target=A} | 0.2 |
| 3 | {X3=5，X2=3}=>{target=A} | 0.3 |
| 4 | {X3=2，X3=5}=>{target=A} | 0.2 |
| 5 | { X9=1，X3=5，X1=X7}=>{target=A} | 0.3 |
| 6 | {X3=1，X2=10}=>{target=A} | 0.5 |

通过等价分类，我们得到1、2、5为一类，3、4为一类，6为一类

则我们保留：1、5、3、6

最终输出保留的规则及其interest指标，如下：

rules interest

{X1=7，X4=8，X2=6}=>{target=A} 0.3

{ X9=1，X3=5，X1=X7}=>{target=A} 0.3

{X3=5，X2=3}=>{target=A} 0.3

{X3=1，X2=10}=>{target=A} 0.5

## 2.2问题2—R与C混编

1）具体要求：将上述算法改写为C或C++语言（windows环境），并在R中调用且输出。（提示，可以使用Rcpp包和Rtools编译环境）

# 3.交付物要求

1）仔细参考《0 交付文档（标准格式-机器学习）》，针对以上测试题目撰写一篇文档。文档中必须包含设计的算法思路、算法执行时间统计和算法渐进时间复杂度计算。

2）针对以上问题分别编写程序（需有必要的代码注释），需注意代码的规范性、可读性、模块化等。