德国信用数据集的描述。

1.标题：德国信用数据

2.源信息

汉斯霍夫曼教授

统计研究所和“经济汉堡系计量经济学大学”

冯·梅勒园5

2000汉堡13

3.实例数：1000

提供了两个数据集。原始数据集，由霍夫曼教授以提供的形式，包含分类/符号属性和是在文件“german.data”。

对于需要数值属性的算法，斯特拉斯克莱德大学产生文件“german.data-numeric”。此文件已编辑

并添加了几个指示符变量以使其适合不能处理分类变量的算法。一些排序分类（例如属性17）的属性被编码为整数。这是StatLog使用的形式。

6. german文件包括20个变量: 20 (7 numerical, 13 categorical)

german.numer文件包括24个变量: 24 (24 numerical)

7.德国人的属性描述

属性1：（定性）现有支票帐户的状态

   A11：... <0 DM

A12：0 <= ... <200DM

A13：...> = 200 DM/工资分配至少1年

   A14：不检查帐户

属性2：（数值）月份持续时间

A2

属性3：（定性）信用记录

A30：没有学分/所有积分支付回来

A31：所有在这家银行的信用卡已收回

A32：到现在支付的现有贷项

   A33：延迟过去的回报

A34：关键帐户/存在的其他信用（不在此银行）

属性4：（定性）目的

A40：汽车（新）

A41：汽车（使用）

A42：家具/设备

A43：广播/电视

A44：家用电器

A45：修理

A46：教育

A47：（vacation - 不存在？）

A48：再培训

A49：业务

A410：其他

属性5：（数值）信用额

A5

Attibute 6：（定性）储蓄账户/债券

A61：... <100 DM

A62：100 <= ... <500DM

A63：500 <= ... <1000DM

A64：..> = 1000 DM

   A65：未知/不存储储蓄帐户

属性7：（定性）目前就业

A71：失业

A72：... <1年

A73：1 <= ... <4年

A74：4 <= ... <7年

A75：..> = = 7年

属性8：（数值）分期付款率按可支配收入的百分比计算

A8

属性9：（定性）个人地位和性别

A91：男：离婚/分居

A92：女性：离婚/分居/已婚

   A93：男：单身

A94：男性：已婚/丧偶

A95：女：单身

属性10：（定性）其他债务人/担保人

A101：无

A102：共同申请人

A103：担保人

属性11：（数值）现在居住地

A11

属性12：（定性）属性

A121：房地产

A122：如果不是A121：建筑社会储蓄协议/人寿保险

   A123：如果不是A121 / A122：汽车或其他，不属于属性6

A124：未知/无属性

属性13：（数值）年龄

A13

属性14：（定性）其他分期付款计划

A141：银行

A142：商店

A143：无

属性15：（定性）住房

A151：租金

A152：自己

A153：免费

属性16：（数值）此银行现有的信用额

A16

属性17：（定性）工作

A171：失业/非技术 - 非居民

A172：不熟练

A173：熟练的员工/官员

A174：管理/个体经营/高度合格的员工/官员

属性18：（数值）负责提供维修的人数

A18

属性19：（定性）电话

A191：无

A192：是，以客户名称注册

属性20：（定性）外国工人

A201：是的

A202：不

8.成本矩阵

此数据集所需要使用成本矩阵（参见下文）

1 2

----------------------------

1 0 1

-----------------------

2 5 0

（1 =好，2 =差）

其中，行表示实际分类和列表示预测分类。

将坏客户分类为好客户的成本较大为5个单位，而将好客户分类为坏客户的成本为1个单位。

该数据集包含了1000个样本，每个样本包括了21个变量（属性），其中包括1个违约状态变量“credit\_risk”，剩余20个变量包括了所有的定量和定性指标，分别如表1所示。 