**MIS演示软件需求说明**

**一、需求说明**

**1、本演示软件的主要内容和教学目的**

本演示软件采用朴素贝叶斯算法，对银行信用卡客户数据进行分类，通过分类估计银行信用卡客户的信用额度（信用等级）。其中信用等级数值越大，表明该客户应该授予较大的信用额度。系统中的方法可应用于其他已知客户群体的某些信息，推断客户群体某一未知信息的场合。

**2、知识点举例**

根据教学大纲确定软件演示所需的知识点，设计演示软件的目的是理解该章所有的知识点，表1是知识点一个例子：

**表1 知识点**

|  |  |
| --- | --- |
| **知识点1** | 朴素贝叶斯算法 |
| **知识点2** | 面向对象软件分析和设计方法 |

**3、系统功能规划**

**3.1用例图示例**

（1）确定系统用户；

（2）确定每个用户需要的功能；

（3）各个功能模块之间的调用关系。

用例图描述系统的静态属性，即系统组成和功能属性。

用例图如下：

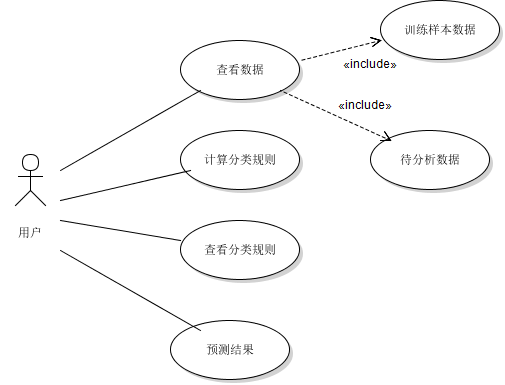


图1 UML用例图例子

**3.2 序列图示例** 描述系统各个子模块的协作关系，通过相互协作完成用例图中描述的每个子功能。

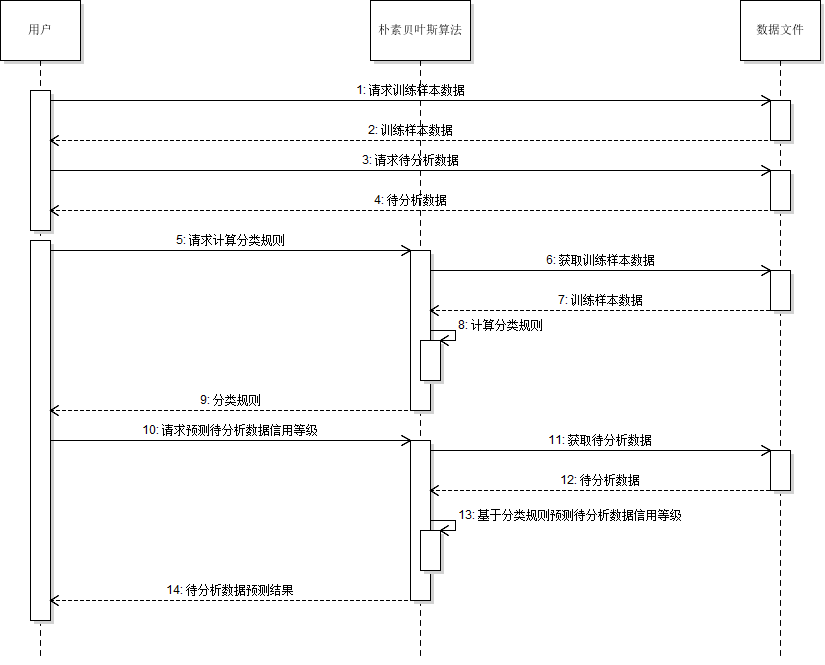


图2 UML序列图例子

**3.3 状态图示例**

状态图表示一个系统在接收外来事件后处理机状态迁移过程，表示系统的动态属性。首先分析系统处于哪些状态，外部有哪些时间，在某一状态下针对某一外部事件到来如何处理以及系统迁移到何种状态。

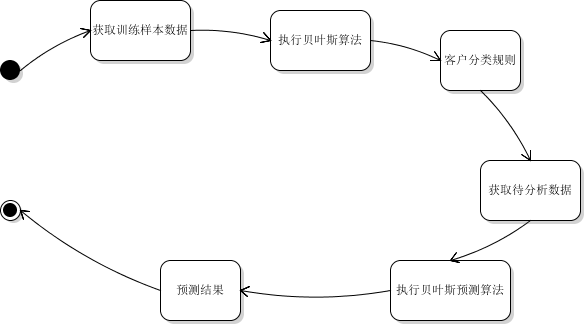


图3 UML状态图例子

**4.数据及算法说明**

（1）数据说明

本系统的训练样本数据格式如表1所示，每条数据12列，共计460条数据，数据详见附件excel文件，文件名“训练样本数据.xlsx”。

表1 训练样本数据格式

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **开卡时间差** | **信用等级改变情况** | **信用等级** | **年收入** | **年龄** | **性别** | **婚姻** | **住房状况** | **教育程度** | **上次账单余额** | **工作年限** |
| 359 | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 3 |
| 360 | 3 | 0 | 1 | 1 | 2 | 2 | 0 | 0 | 0 | 1 | 1 |
| 361 | 3 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 1 |
| 362 | 3 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 1 |
| 363 | 1 | 0 | 1 | 3 | 2 | 1 | 1 | 2 | 5 | 3 | 2 |
| 364 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 4 | 4 | 1 | 3 |

本系统的待分析数据格式如表2所示，数据格式与训练数据一致，共计10条数据，数据详见附件excel文件，文件名“待分析数据.xlsx”。其中，“信用等级”为空，是朴素贝叶斯算法的预测列。

表2 待分析数据格式

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **开卡时间差** | **信用等级改变情况** | **年收入** | **年龄** | **性别** | **婚姻** | **住房状况** | **教育程度** | **上次账单余额** | **信用等级** | **工作年限** |
| 1 | 1 | 0 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |  | 2 |
| 2 | 0 | 1 | 3 | 1 | 1 | 1 | 4 | 4 | 1 |  | 2 |
| 3 | 0 | 0 | 3 | 1 | 2 | 2 | 3 | 3 | 1 |  | 3 |
| 4 | 1 | 0 | 4 | 1 | 2 | 1 | 3 | 4 | 1 |  | 2 |
| 5 | 0 | 1 | 3 | 1 | 2 | 1 | 4 | 4 | 3 |  | 2 |
| 6 | 1 | 0 | 1 | 1 | 1 | 1 | 3 | 4 | 1 |  | 1 |
| 7 | 0 | 0 | 2 | 1 | 2 | 2 | 4 | 3 | 3 |  | 3 |
| 8 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 4 | 3 |  | 2 |
| 9 | 1 | 1 | 3 | 1 | 2 | 2 | 1 | 5 | 3 |  | 3 |
| 10 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 1 |  | 1 |

上表中数据是经过编码离散化后的数据，数据的实际含义如下所示：

性别：1代表男；2代表女。

婚姻：0数据缺失；1未婚；2已婚。

住房状况： 0未填；1 与父母同住；2 租用；3 自购无贷款；4 自购有贷款；5 其他。

教育程度： 0 未填；1 初中以下；2 高中及中专；3 大专；4 本科；5 硕士；6 博士。

信用度改变情况: 0未改变；1 改变。

开卡时间差：0年；1年；2年；3年；3年以上。（N为负值—异常）

信用额度（信用等级）根据银行实际管理需要确定为若干等级，等级数值越大，表明客户应该授予越高的信用额度。

1. 朴素贝叶斯算法设计

## 1）算法理论

贝叶斯分类算法的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

设*X*是类标号未知的数据样本。设*H*为某种假设，如数据样本*X*属于某特定的类*C*。对于分类问题，我们希望确定*P*(*H*|*X*)——给定观测数据样本*X*，假设*H*成立的概率。换言之，给定*Ｘ*的属性描述，找出样本*Ｘ*属于类*Ｃ*的概率。

*P*(*H*|*X*)是后验概率，或条件*X*下，*H*的后验概率。例如，假定数据样本世界限于分别由属性“年龄”和“收入”来描述的顾客，而*Ｘ*是一位35岁的顾客，其收入为40000美元。假定*Ｈ*表述我们的顾客将购买计算机。则P(*H*|*X*)反映当我们知道顾客的年龄和收入时，顾客*Ｘ*将购买计算机的概率。

相反，*P*(*H*)是先验概率，或*H*的先验概率。在上面的例子中，它表示任意给定的顾客将购买计算机的概率，而不管他的年龄、收入和其他信息。后验概率P(*H*|*X*)比先验概率*P*(*H*)基于更多的信息（如顾客的信息）。*P*(*H*)独立于*X*。

类似地，*P*(*X*|*H*) 是条件*H*下，*X*的后验概率。即，它是已知顾客*X*购买计算机，该顾客是35岁并且收入为40000美元的概率。

*P*(*X*)是*X*的先验概率。使用我们的例子，它是我们的顾客集合中的一个人年龄为35岁并且收入为40000美元的概率。

“如何估计这些概率？”正如我们下面将看到的，*P*(*X*)，P(*H*)和*P*(*X*|*H*)可以由给定的数据估计。贝叶斯定理提供了一种由*P*(*X*)，*P*(*H*)和*P*(*X*|*H*)计算后验概率*P*(*H*|*X*)的方法。

贝叶斯定理用公式表示如下：



我们在朴素贝叶斯分类法中将使用这一公式。

朴素贝叶斯分类的工作过程如下：

（1）. 设*D*为训练样本和相关联的类标号集合。每个数据样本用一个*n*维特征向量*X* ={*x*1, *x*2,..., *xn*}表示，描述由属性*A*1, *A*2 ,..., *An*对样本的*n*个度量。

（2）. 假定有*m*个类*C*1, *C*2, ..., *Cm* 。给定一个未知的数据样本*X*（即，没有类标号），分类法将预测*X*属于具有最高后验概率（条件*X*下）的类。即，朴素贝叶斯分类将未知的样本分配给类*Ci*，当且仅当：

*P*(*Ci*|*X*) > *P*(*Cj*|*X*) 1≤*j*≤*m, j*≠*i*

这样，最大化*P*(*Ci* | *X* ) 。其*P*(*Ci*|*X*)最大的类*Ci*称为最大后验假定。

（3）．根据贝叶斯定理：



由于*P*(*X*)对于所有类为常数，只需要*P*(*X* |*Ci*)*P*(*Ci* )最大即可。如果类的先验概率未知，则通常假定这些类是等概率的，即，*P*(*C*1)= *P*(*C*2)= ... = *P*(*Cm*)。并据此对只*P*(*Ci*|*X*)最大化。否则，最大化*P*(*X*|*Ci*)*P*(*Ci*)。注意，类的先验概率可以用*P*(*Ci*)=|C*i,D*|/|D|估计，其中，|C*i,D*|是*D*中*Ci*类的训练样本数。

【注】本部分内容摘自Jiawei Han编著的《数据挖掘概念与技术》，机械工业出版社，2007年7月。

（4）．给定具有许多属性的数据集，计算*P*(*X*|*Ci*)的开销可能非常大。为降低计算*P*(*X*|*Ci*)的开销，可以做类条件独立的朴素假定。给定元组的类标号，假定属性值条件地相互独立。即，在属性间，不存在依赖关系。这样，



可以容易地由训练元组估计概率*P*(*x*1|*Ci*), *P*(*x*2|*Ci*), ..., *P*(*xn*|*Ci*)。

（5）．为预测*X*的类标号，对每个类*Ci*，计算*P*(*X*|*Ci*)*P*(*Ci*)。分类法预测元组X的类标号为*Ci*，当且仅当：

*P*(*X*|*Ci*)*P*(*Ci*) > *P*(*X*|*Cj*)*P*(*Cj*) 对1≤*j*≤*m*，*j*≠*i*

换言之，预测类标号是使*P*(*X*|*Ci*)*P*(*Ci*)最大的类*Ci*。

## 2）算法设计

由3.1可知，预测待分析数据的“信用等级”，需要利用训练样本数据，计算分类规则：

P(“开卡时间差，信用等级改变情况，年收入，年龄，性别，婚姻，住房状况，教育程度，上次账单余额，工作年限”|信用等级 ）P（“信用等级”）=

P(“开卡时间差”|信用等级 ）×

P(“信用等级改变情况”|信用等级 ）×

P(“年收入”|信用等级 ）×

P(“年龄”|信用等级 ）×

P(“性别”|信用等级 ）×

P(“婚姻”|信用等级 ）×

P(“住房状况”|信用等级 ）×

P(“教育程度”|信用等级 ）×

P(“上次账单余额”|信用等级 ）×

P(“工作年限”|信用等级 ）P（“信用等级”）

根据训练样本数据中各列的数据范围，以“信用等级=1”为例，设计分类规则如表3所示。

表3 分类规则“信用等级=1”

|  |  |  |  |
| --- | --- | --- | --- |
| **计算公式** | **信用等级** | **列** | **值** |
| P(“开卡时间差”|“信用等级”） | 1 | 开卡时间差=0 |  |
| 1 | 开卡时间差=1 |  |
| 1 | 开卡时间差=2 |  |
| 1 | 开卡时间差=3 |  |
| 1 | 开卡时间差=4 |  |
| P(“信用等级改变情况”|“信用等级”） | 1 | 信用等级改变情况=0 |  |
| 1 | 信用等级改变情况=1 |  |
| P(“年收入”|“信用等级”） | 1 | 年收入=1 |  |
| 1 | 年收入=2 |  |
| 1 | 年收入=3 |  |
| 1 | 年收入=4 |  |
| P(“年龄”|“信用等级”） | 1 | 年龄=1 |  |
| 1 | 年龄=2 |  |
| 1 | 年龄=3 |  |
| 1 | 年龄=4 |  |
| P(“性别”|“信用等级”） | 1 | 性别=1 |  |
| 1 | 性别=2 |  |
| P(“婚姻”|“信用等级”） | 1 | 婚姻=0 |  |
| 1 | 婚姻=1 |  |
| 1 | 婚姻=2 |  |
| 1 | 婚姻=3 |  |
| P(“住房状况”|“信用等级”） | 1 | 住房状况=0 |  |
| 1 | 住房状况=1 |  |
| 1 | 住房状况=2 |  |
| 1 | 住房状况=3 |  |
| 1 | 住房状况=4 |  |
| 1 | 住房状况=5 |  |
| P(“教育程度”|“信用等级”） | 1 | 教育程度=0 |  |
| 1 | 教育程度=1 |  |
| 1 | 教育程度=2 |  |
| 1 | 教育程度=3 |  |
| 1 | 教育程度=4 |  |
| 1 | 教育程度=5 |  |
| 1 | 教育程度=6 |  |
| P(“上次账单余额”|“信用等级”） | 1 | 上次账单余额=1 |  |
| 1 | 上次账单余额=2 |  |
| 1 | 上次账单余额=3 |  |
| P(“工作年限”|“信用等级”） | 1 | 工作年限=1 |  |
| 1 | 工作年限=2 |  |
| 1 | 工作年限=3 |  |
| 1 | 工作年限=4 |  |

按照表3格式，同样可以计算“信用等级=2”、“信用等级=3”、“信用等级=4”的分类规则。

从训练样本数据计算得到分类规则后，就可以对待分析数据进行信用等级预测。预测待分析数据属于那个信用等级，即，计算

P(“开卡时间差，信用等级改变情况，年收入，年龄，性别，婚姻，住房状况，教育程度，上次账单余额，工作年限”|信用等级 ）P（“信用等级”）的最大值，以第一条待分析数据为例，计算其属于“信用等级=1”的可能性，计算方法为：

P(“开卡时间差=1”|信用等级=1 ）×

P(“信用等级改变情况=0”|信用等级=1 ）×

P(“年收入=2”|信用等级=1 ）×

P(“年龄=1”|信用等级=1）×

P(“性别=1”|信用等级=1 ）×

P(“婚姻=1”|信用等级=1 ）×

P(“住房状况=2”|信用等级=1）×

P(“教育程度=2”|信用等级=1）×

P(“上次账单余额=2”|信用等级=1）×

P(“工作年限=2”|信用等级=1）P（“信用等级=1”）

而上面计算方法中的值可以从表3中获得。

因此，设计朴素贝叶斯算法如图4所示。



图4 朴素贝叶斯算法流程图