

数据挖掘技术在个人信用评估模型中的应用

葛继科¹, 赵永进¹, 王振华¹, 余建桥²

(1. 河南师范大学 计算机与信息技术学院, 河南 新乡 453007;

2. 西南大学 计算机与信息科学学院, 重庆 450052)

摘要:为了能够及时、恰当地进行个人信用评估分析, 加快信用卡发卡机构的决策速度, 介绍了数据挖掘技术在信用卡公司对用户评估中的应用, 对比分析了数理统计模型、分类-聚类个人信用评估模型等几种个人信用评估模型建模方法的优缺点。建立了一种决策树-神经网络个人信用评估模型, 针对该模型提出了一种近邻聚类算法。该算法不需要事先给定聚类的类别数, 可以进行无监督学习。通过对比分析可知, 该算法在个人信用评估应用中可以得到较理想的结果。

关键词:信用评估; 分类; 聚类; 决策树

中图分类号: TP391; F830.49

文献标识码: A

文章编号: 1673-629X(2006)12-0172-03

Application of Data Mining Technique to Personal Credit Evaluating Model

GE Ji-ke¹, ZHAO Yong-jin¹, WANG Zhen-hua¹, YU Jian-qiao²

(1. College of Computer and Information Technology, Henan Normal University, Xinxiang 453007, China;

2. College of Computer and Information Science, Southwest University, Chongqing 450052, China)

Abstract: For the purpose of process the personal credit evaluating timely and correctly, increase the decision rate, this paper describes the requirement of the credit card company for data mining and neural network technology which apply for personal credit evaluating. Contrast-ed and analyzed some of personal credit evaluating model, e. g. statistical model, classification - clustering model, and so on. Demonstrated those excellence and disadvantage. Constructed a decision tree - neural network personal credit evaluating model. At last, give a vicinage - extended clustering algorithm, the algorithm needn't give number of clustering, and can put up unsupervised learning. The algorithm is more fit for personal credit evaluating than other methods.

Key words: credit evaluating; classification; clustering; decision tree

0 引言

近几年,随着信用卡的出现和发展,银行及其他信用卡的发卡机构认识到了信用评估的作用及重要性。如何提高服务质量,改进服务方法,使公司的决策更为准确及时,是信用卡公司追求的一个目标。由于每天申请信用卡的人数众多,无论从经济的角度还是从人力的角度,发卡机构都不可能完全依赖人工对申请进行审批,必须有一套比人工主观判断具有更好预测能力的自动信用评估系统。随着市场竞争的加剧以及计算机技术的发展,一些非参数统计方法以及人工智能模型逐渐被引入到个人信用评估模型中,如神经网络、专家系统、基因算法等均被应用到信用评估卡的开发之中。这些方法的引入在一定程度上克服了传统分析方法的综合分析能力差、缺乏整体概括能力

的缺点,弥补了评价结果的一些不足^[1]。

神经网络(Neural Network, NN)是一种对数据分布无任何要求的非线性技术,它能有效解决非正态分布、非线性的信用评估问题,但它存在解释性差、训练样本集大和训练效率低等缺点^[2,3]。

数据挖掘(Data Mining, DM)是从存放在数据库、数据仓库或其他信息库中的大量数据中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程^[4,5]。随着研究的不断深入,出现了许多用于挖掘不同类型数据的算法和技术,常用的数据挖掘方法有:描述、分类、聚类、关联规则、孤立点检测等。利用基于聚类的分类信用评估方法,有效地克服了神经网络技术在信用评估中存在的某些问题。

1 常用信用评估方法

信用评估本质上是模式识别中的一类分类问题,将企业或个体消费者划分为能够按期还本付息(即“好”客户)和违约(即“坏”客户)两类^[6]。具体做法是根据历史上每

收稿日期:2006-03-16

基金项目:河南省自然科学基金(0511011500)

作者简介:葛继科(1977-),男,河南濮阳人,硕士,研究方向为数据挖掘、人工智能。

个类别(如期还本付息、违约)的若干样本,从已知的数据中找出违约及不违约者的特征,从而总结出分类的规则,建立数学模型,用于测量借款人的违约风险(或违约概率),从而为消费信贷决策提供依据。

国外已经有人做了大量的工作,提出了各种评估模型,如 FICO 评估模型、神经网络模型、贝叶斯分析模型等,采用了各种数学的、统计学的、信息学的方法,取得了一定的效果。尤其是 FICO 评估模型,现已成为西方发达国家信用评估方面事实上的标准^[7]。

随着信贷业务的需要,国内越来越多的金融机构也开始以业务对象的个人信用记录作为决策参考。

1.1 标准数理统计模型

基于标准数理统计理论的信用评估模型是对大量的个人消费贷款的历史信用数据进行科学的归纳、总结、计算而得到的量化分析公式。在美国,不同的行业有不同的信用评估模型,用来帮助专业人士进行信用风险管理,如表 1 所示。

表 1 美国不同行业常用信用评估模型

行业	常用信用评估模型
消费者信贷	FICO 模型, Logit 模型
制造业	Z-Score
工业	Zeta score
普通企业	Risk calc, Z-Score
新兴市场企业	EM Score

信用评估模型的关键是科学合理地选出信用变量,并产生一个公式。信用评估模型的统计方法有:线性概率模型、Logit 模型、Probit 模型,以及判别(Discriminant)分析方法。

1.2 数据挖掘方法

数据挖掘是为了发现事先未知的规则和联系而对大量数据进行选择、探索和建模的过程,其任务可以分为两类:描述和预测。用于个人信用评估的常用方法包括分类、聚类、关联规则分析、预测、孤立点检测等^[8]。

(1)分类(Classification):按分析对象的属性、特征建立不同的组类来描述事物。它基于对类标记已知的数据对象的分析,导出描述并区分数据类或概念的模型(或函数),用以预测类标记未知的对象类,导出模式可以用分类规则、判定树、数学公式或神经网络等形式表示。

(2)聚类(Clustering):根据“物以类聚”的原理,将本身没有类别的样本聚集成不同的组,这样的一组数据对象叫做簇,并且对每一个这样的簇进行描述的过程。其目的是使得属于同一个簇的对象应该彼此相似,而不同簇的对象应该足够不相似。

2 分类-聚类个人信用评估模型

就个人信用评估建模问题而言,待建模数据库假设为一个信用数据库,它是一个由属性、元组组成的二维表,称之为信用决策表。属性分为条件属性和决策属性,各条件

属性的取值可以是某段区间的连续值,也可以是多个离散值,决策属性取值为百分制。初始样本集根据决策属性的分数作区段划分,得到多个大类别,为了进一步精确化,再对每一个大类别进行聚类分析,得到多个子聚类,对每个子聚类建立一个能拟合包含在其中的训练样本的子模型。

设训练样本集为 S , S 中共有 N 个样品,可以把它看作一个数据库, S 的每个样品是一个元组(即 \langle 属性, 值 \rangle 对),根据决策属性的取值对训练样本集进行类别划分,划分后可记为: S_1, S_2, \dots, S_k , 共 k 类样本子集。对第 i 类的样本子集进行聚类分析,可得到 N_i 个子聚类 C_i (第 i 类的样本子集 S_i 的第 j 个子聚类)。这里 $i = 1, 2, \dots, k, j = 1, 2, \dots, N_i, N = \sum_{i=1}^k N_i$ 。对每个子聚类 C_i 可建立一个子模型来拟合描述这个子聚类中的所有样本。如图 1 所示。

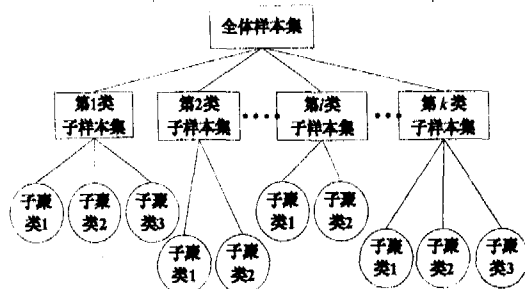


图 1 基于决策属性的分类-聚类模型图

按照上述方法,可以得到一个分类-聚类树,对于基层的子聚类,当某些子聚类满足一定的条件时就可以合并(融合)。设 $A = \{C_1, C_2, \dots, C_m\}$, $B = \{C_{m+1}, C_{m+2}, \dots, C_n\}$, 分别为某空间中的由多个子聚类组成的集合。集合 A 中的 $C_i (i = 1, 2, \dots, m)$ 与集合 B 中的 $C_j (j = m + 1, m + 2, \dots, n)$ 能否合并,可由以下判别方法决定:

- ① 若子聚类 C_i 的中心在 C_j 的边界所构成的区域内,且 C_i 与 C_j 有部分或全部空间重叠,则 C_i 可与 C_j 合并;
- ② 若子聚类 C_i 的中心在 C_j 的边界所构成的区域外,但 C_i 与 C_j 有部分空间重叠,此时需根据空间的比例及实际情况判断 C_i 与 C_j 是否可合并;
- ③ 若子聚类 C_i 与 C_j 完全不重叠,则 C_i 与 C_j 不能合并。

针对每个子聚类,具体的建模方法可以使用基于粗糙集的神经网络建模方法^[9]、神经网络二分类法、径向基函数 RBF 学习算法、范例类比模型法和模糊 C-均值聚类算法(FCM)^[10]等。

也可以采用 RBF 中的子聚类区域高斯函数描述法来确定其所辖范围,这相当于一个对待测样本判决其所属区域的开关;然后用 BP 神经网络模型来做结果评判(对于个人信用评估问题,其结果采用打分法)。

3 决策树-神经网络个人信用评估模型

当完全采用决策树方法时,由于它使用信息熵或其它

2 系统实现

Oracle 9iFS 是 Oracle 公司开发的基于 Oracle 数据库的新型文件系统。Oracle 9iFS 将企业所有数据统一成单个、统一的信息库,企业的所有数据都存放在 Oracle 数据库中。Oracle 9iFS 被设计成数据库层、中间层和客户端三层架构来提供较好的性能、可扩展性和可靠性,它已经为文件管理提供了一系列的 Java API 开发包,所以笔者设计的面向内容管理的浏览器是基于 9iFS 开发的。

面向内容管理的浏览器的用户界面被设计成类似于传统文件系统的风格,用户可以在客户端进行简单的逻辑操作。

2.1 右键菜单

为了用户操作的方便,将对文件的操作加入到右键菜单中,每项菜单命令对应一个文件操作。用户可以通过右键菜单,调用系统的各项功能。如图 4 所示。

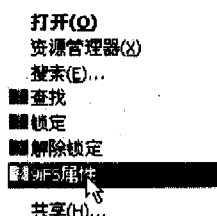


图 4 右键菜单

2.2 高级搜索页面

面向内容管理的浏览器除了提供传统文件系统提供的搜索功能外,还提供了全文搜索、正则搜索、类别搜索和关联搜索等高级搜索。不同类型的搜索分布在不同的 Tab 页中,用户可以根据提供的搜索功能进行相应的查询。在一次搜索中,用户还可以进行组合条件搜索。

高级搜索如图 5 所示。

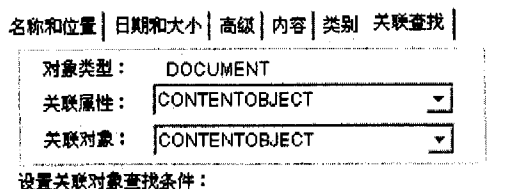


图 5 高级搜索

全文搜索时,在查找操作界面的“内容”Tab 页中输入全文搜索的关键字,然后执行搜索,就可以将内容包含关键字的所有文件搜索出来。

在“名称和位置”Tab 页中为高级用户提供了正则搜索功能,用户可以对字符串类型的属性构建复杂的正则查询条件,准确地搜索到目标文件。

在面向内容管理的浏览器中,可以通过类别属性对文件进行分类,这样存放在不同位置的文件可以通过类别将同类文件搜索出来。

关联搜索能够为用户进一步提供搜索功能,如某用户需要根据邮件发送者的一些个人信息(如:爱好、生日)来搜索满足条件的邮件。

3 小结

本系统是针对用户对内容管理的迫切要求和当前内容管理系统的不完善而设计的,为用户提供了强大的文件管理功能,提高了文件访问的安全性和效率,实现了跨平台性,并将所有的功能操作集成到了右键菜单中。面向内容管理的浏览器实现了基本的内容管理功能,体现了内容管理的优越性,具有良好的研究价值和实用意义。

参考文献:

- [1] Hyman F, Garg S, Harrison S. Oracle® Internet File System Installation Guide Release 9.0.1.1.0 for Microsoft Windows NT/2000[EB/OL]. 2001-09. <http://www.oracle.com>.
- [2] Stokes A, Dawson D. Oracle® Internet File System Developer Reference Release 9.0.1.1.0 Oracle 9iFS and Oracle9i database documentation sets[EB/OL]. 2001-09. <http://www.oracle.com>.
- [3] Rizzo T. 新型 Windows 文件系统简介[EB/OL]. 2004-08-25. <http://www.microsoft.com/china/MSDN/library/windev/longhorn/winfs03112004.msp>.
- [4] 车敦仁,周立柱,王令赤. 面向对象数据库系统的体系结构[J]. 软件学报,1995(10):599-606.
- [5] 周 军. 应用版本控制软件管理软件开发[J]. 计算机系统应用,2000(10):50-52.
- [6] 潘 定,沈钧毅. 数据仓库中实时元数据管理的研究[J]. 计算机工程,2000(5):29-31.
- [7] 王 强,刘东波,王建新. 数据仓库元数据标准研究[J]. 计算机工程,2002(12):123-125.
- [8] 朱 斐. 一种结构化文件的访问控制模型的设计和实现[J]. 微机发展,2005,15(4):132-134.
- [9] 诸 晔. 用 ACL 实现系统的安全访问控制[J]. 计算机应用与软件,2005(3):111-114.

(上接第 174 页)

- [8] Han Jiawei. 数据挖掘概念与技术[M]. 范 明,孟小峰等译. 机械工业出版社,2001:14-17.
- [9] 何 明,李 博. 粗糙集理论框架下的神经网络建模研究及应用[J]. 控制与决策,2005(7):65-68.
- [10] 尹 松,周永权,李陶深. 基于稀疏差异度的聚类方法在信息分类中的应用[J]. 计算机技术与发展,2006,16(1):117-119.

- [11] 李宝林,王秀峰. CBRDI:一种基于范例推理的数据集成方法[J]. 计算机工程与应用,2003(16):52-55.
- [12] Beckmann N, Kriegel H P, Schneider R, et al. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles[C]//Proc. ACM SIGMOD Int. Conf. on Management of Data. Atlantic City, NJ:ACMPress,1990:322-331.