Advances in Computer Science Research (ACSR), volume 61
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

Personal Credit Scoring Based on Decision Tree C5.0 Algorithm

Shang Gao* and Changbao Wang

School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China

gao_shang@just.eud.cn

*The corresponding author

Abstract. There are some problems still exist in traditional individual credit assessment system. To solve the problems, a decision tree individual credit assessment model is proposed. Using SPSS Clementine data mining tool, the personal credit data is clustering analysis by decision tree C5.0 method. It is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good. It is discussed as the different proportion of loss.

Keywords: Personal credit scoring; Decision tree; C5.0 algorithm

基于决策树 C5.0 算法的个人信用评估

高尚，王长宝

（江苏科技大学计算机科学与工程学院，江苏 镇江 212003）

摘要：针对当前传统个人信用评价中的种种问题,提出了利用决策树进行个人信用评估的方法。该方法利用 SPSS Clementine 数据挖掘工具，使用 C5.0 方法进行聚类分析。个人信用评估时，接受“坏”客户损失比拒绝“好”客户的损失的要高，对不同的比例进行了分析。

关键词：个人信用评估；决策树；C5.0 算法

中图分类号：**TP391**　　**文献标志码**：**A**

引言

随着中国经济的快速发展，信用消费已逐步浮出水面，住房按揭、汽车贷款、教育贷款、信用卡等各种个人消费贷款的规模在迅速扩大。在消费信贷热不断升温的形势下，各商业银行均把发展消费贷款作为未来发展战略的重要组成部分。但是目前国内商业银行对消费贷款的风险管理水平较低，管理手段与方法均较落后，其中缺乏一套有效的个人信用评估方法是阻碍了个人消费信贷业务进一步开展的主要因素之一[1-2]。信用评估可以较精确地估计消费信贷的风险，给贷款人提供了一个可靠的技术手段，避免不良贷款，控制债务拖欠和清偿。个人信用评估可以使贷款人更加精确地界定可以接受的消费信贷的风险，扩大消费信贷的发放。

在我国，个人信用评估体系尚不健全，个人信用长期没有评估，良好的信誉没有得到合理的优惠，欠债不还也没有受到相应的惩罚，使我国的个人信用较为脆弱，个人资信程度降低。由此导致以下结果：（1）缺乏完善的消费者个人信用评估体系必将导致贷款审批时间延长、手续复杂。（2）缺乏完善的消费者个人信用评估体系容易导致非个人因素的信贷风险。目前国内商业银行对消费贷款的风险管理水平较低，管理手段与方法均较落后，在消费信贷的发放过程中，仍然采用传统的比率分析方法，根据“5C”原则来评价消费信贷申请者的信用状况及还款能力，主观成分较多；在个人信用评估方法上仍然没有形成稳健可靠的模型，没有一套科学合理的个人信用评价指标体系。刚刚启动的个人征信数据库虽然解决了个人信用记录

有限和样本数据缺失的问题，但如何充分有效的利用这些数据，提高信贷审批的效率，避免不良贷款的出现，在目前仍然是一个亟待解决的问题。(3) 缺乏完善的消费者个人信用评估体系资料影响了消费者申办消费信贷的积极性。进行科学的个人信用评估是贷款决策最重要的环节，除了建立个人信用档案系统并能够提供全社会交流外，就是建立科学客观的数学模型[3-4]。目前，国外商业银行信用评估中应用最为广泛的是多元统计分析方法，其基本思路是，根据已经掌握的历史上每个类别(违约类、不违约类或正常类)的若干样本点，从中总结出分类的规律，建立判别公式，用于对新样本点的分类。近年来，随着技术的突破性进展，许多学者将其应用于信用评估中[5-9]。文[7]提出了一种将分类树和支持向量机结合起来处理个人信用评估的新方法. 针对信用评估指标维数较高的问题，文[8]运用主成分分析与支持向量机理论建立了一个新的个人信用评估预测模型。本文将采用决策树运用到个人信用评估中。

1 决策树方法

决策树(Decision Tree)是一个可以自动对数据进行分类的树形结构，是树形结构的知识表示，其每一个内部节点(非叶子节点)都代表了一个属性上的测试，也就是一个分裂属性。决策树也可解释为一种特殊形式的规则集，其特征是规则的层次组织关系。

决策树算法有很多，如 ID3 算法、C4.5 算法、CART 算法、CHAID 算法、PUBLIC 算法、SLIQ 算法以及 SPRLNT 算法。C4.5 算法是在 ID3 算法基础上改进的决策树生成算法，并且凭借其独特的特点和突出的优势在各行各业的数据挖掘中得到了成功的应用。

ID3 算法最初的定义是假设属性值是离散值，但在实际环境中，有很多属性是连续的，不能够用一个确定的标准来对其进行划分。C4.5 使用下面的一系列处理过程来对连续的属性划分成离散的属性，进而达到能够建立决策树的目的。C4.5 算法的步骤如下：

Step1　根据训练数据集 D 中各个属性的值对该训练数据集进行排序；

Step2 利用其中各属性的值对该训练数据集动态地进行划分；

Step3 在划分后的得到的不同的结果集中确定一个阈值，该阈值将训练数据集数据划分为两个部分；

Step4 针对这两边各部分的值分别计算它们的增益或增益比率，以保证选择的划分使得增益最大。

增益比率的公式如下所示：

$$GainRatio(A)=\frac{Gain(A)}{SplitI(A)} \tag{1}$$

C5.0 与 C4.5 不同之处在于 C5.0 可以处理多种数据类型，包括了日期(date)、时间(times)、时间戳(timestamps)、序列性的离散型数据(ordered discrete attributes)等等。除了处理数据部分丢失的问题，C5.0 还可以将部分属性标记为不适合，以使得分析时仍能保持资料的完整性。

本文采用 Clementine 数据挖掘工具，使用 C5.0 模型进行聚类分析。Clementine 是 SPSS 公司开发的数据挖掘软件，它将聚类、决策树、神经网络、关联规则等多种数据挖掘技术集成在直观的可视化图形界面中[9]。Clementine 结合商业技术可以快速建立数据模型，进而应用到商业活动中，帮助人们改进决策过程。在 Clementine 中建立的挖掘模型，如图 1 所示，具体构建方法参见文献[10]。

本文选用德国一银行信贷评估实例数据进行实证分析。其网址为：http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29。该网站提供了 2 套数

据，这里选则是把各指标都数字化的数据集 german.data-numeric，该数据集由 1000 个样本构成，每个样本含有 24 维数据和每个样本的类别，变量要么取“1”(不违约，“好”客户)，要么取“2”(违约，“坏”客户)。测试时，采用的方法是将该样本按照 2：1 的比例分为训练样本(667 个，从第 1 到第 667 个样本)和测试样本(333 个，从第 668 个到 1000 个样本)。

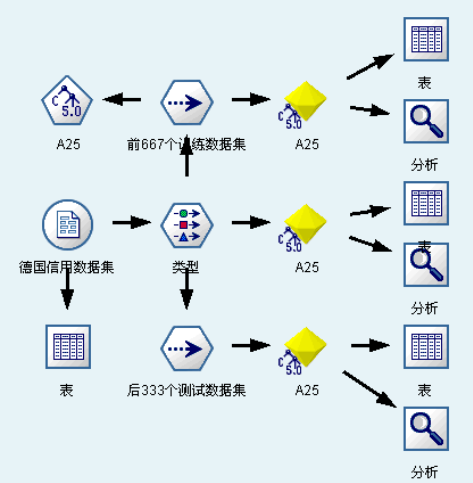

图 1 数据流

图 2 误分类损失矩阵

对于银行来说，拒绝“好”客户和接受“坏”客户这两种错误造成的损失并不相等。接受“坏”客户，银行可能遭受较大的违约风险；而拒绝“好”客户，可能的损失则是贷款利息。也就是说，接受“坏”客户比拒绝“好”客户的成本高。该数据集提供的建议：接受“坏”客户损失与拒绝“好”客户的损失比例是 5：1（图 2）。

图 3 的左侧以层的形式显示决策树，右侧显示的每个变量的相对重要性。决策树也可以用规则集来表示，如图 4 所示。图 5 给出了详细的分析结果。对 333 个测试样本进行测试，正确率为 59.16%。

2 误分类损失分析

很显然，拒绝“好”客户和接受“坏”客户这两种错误造成的损失并不相等。由于接受“坏”客户损失与拒绝“好”客户的损失的比例不同，结果也不一样，表 1 给出了不同比例时的分类正确率比较。比例相差越大，可能会把“好”的客户误判为“坏”客户的可能性也加大，其正确率降低，表 1 的数据也反映了该情况。

表 1 不同误分类损失比例的结果

接受“坏”客户损失与拒绝“好”客户的损失比例	训练集（667 个样本）的正确率	测试集（333 个样本）正确率	全部数据（1000 个样本）正确率
6:1	62.67%	54.65%	60%
5:1	64.32%	59.16%	62.6%
4:1	70.16%	63.66%	68%
3:1	70.31%	63.06%	67.9%
2:1	77.66%	72.67%	76%

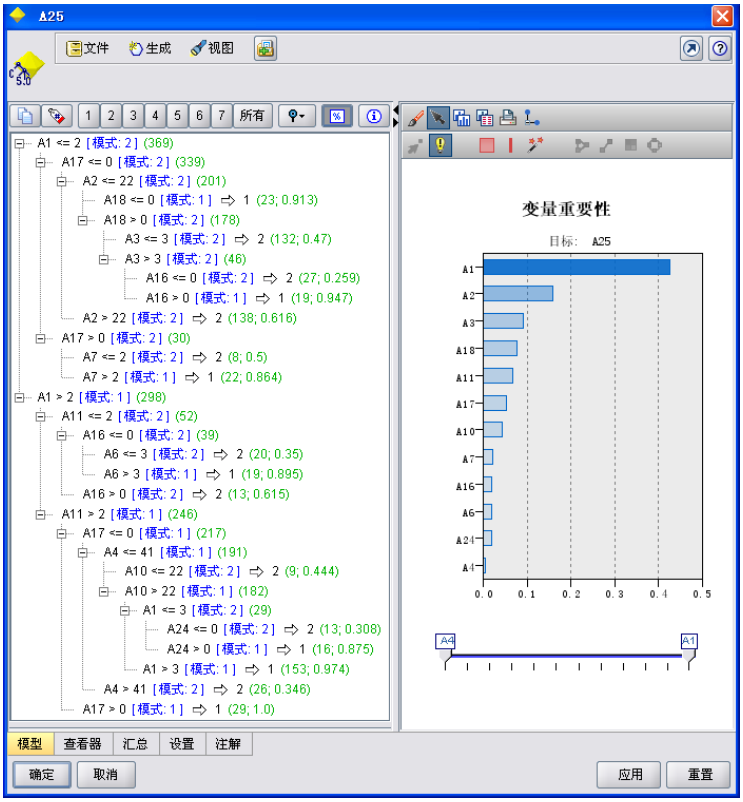

图 3 以层的形式显示决策树

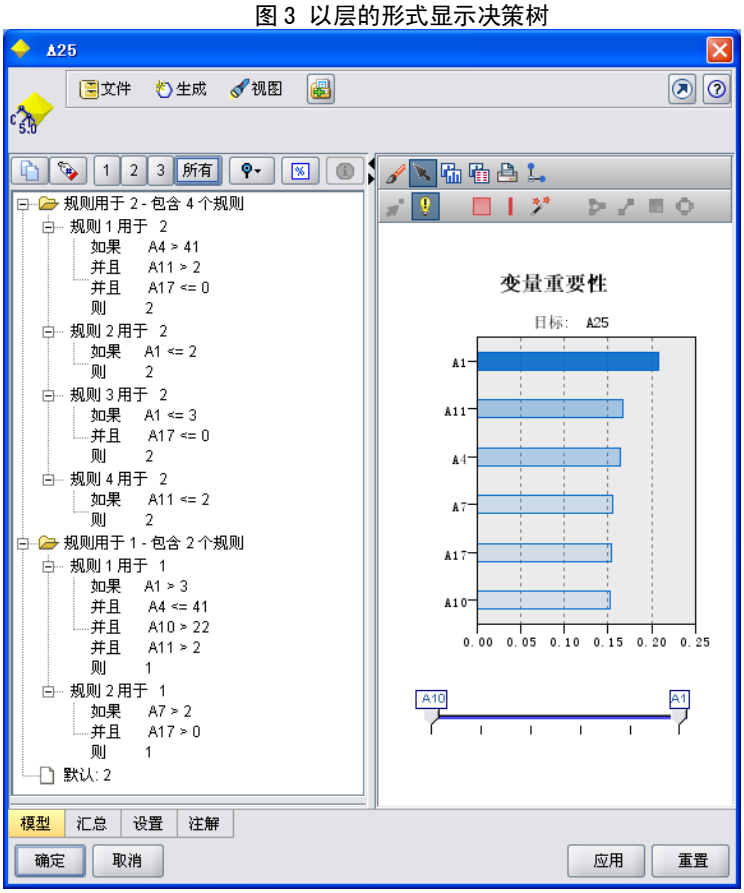

图 4 规则集

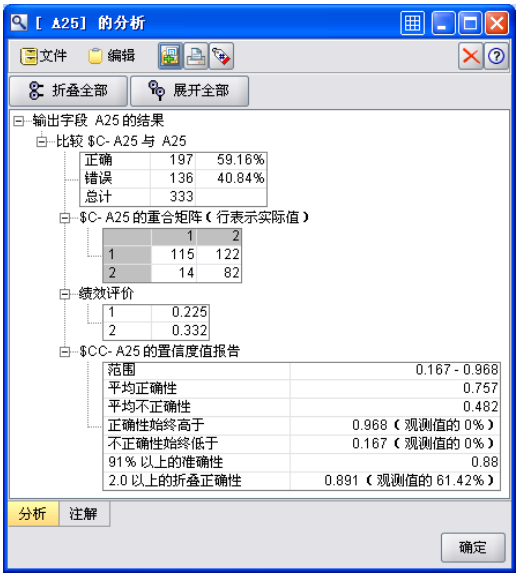

图 5 分析结果

3 结束语

随着消费信贷规模的扩大，如何提高信用评估准确率已经成为信贷行业的一个重要问题，因为信用评估准确率直接影响信贷机构的收益。决策树是归纳学习和数据挖掘的重要方法，通常用来形成分类器和预测模型。本文利用决策树方法对个人信用进行了评估。由于 C5.0 决策树既可以非常直观的解释，也可以根据规则进行解释，适合对某一属性作深入的分析。本文利用 SPSS Clementine 数据挖掘工具，使用 C5.0 方法进行聚类分析。试验结果表明，该算法能够将数据分类和预测。

4 致谢

论文得到人工智能四川省重点实验室开放基金(2016RYJ03)支持。

Acknowledgement

This work was supported by Artificial Intelligence of Key Laboratory of Sichuan Province (2016RYJ03).

参考文献：

[1] 石庆众，靳云汇.个人信用评分的主要模型与方法综述[J].统计研究，2003，8(20):36-39.

[2] R. A. Fisher. The Use of Multiple Measurement in Taxonomic Problem[R]. Annuals of Eugenie，1936，7:179-188.

[3] D Durand. Risk Elements in consumer Installment financing[M]. National Bureua of Economic Research，New York，1941.

[4] J. Myers，W. Forgy. The Development of Numerieal Credit Evaluation System[J]. Jounral of the American statistical Association，1963，303(58):779-806.

[5] 于兆吉，胡祥培，毛强. 电子商务环境下信用评级的一种新方法[J].控制与决策，2009，24(11): 1668-1672.

[6] 姜明辉，谢行恒，王树林，温潇.个人信用评估的 Logistic-RBF 组合模型[J]. 哈尔滨工业大学学报，2007，39(7):1128-1130.

[7] 高莉.基于分类树和支持向量机的个人信用评估方法[J].内江师范学院学报.2009,24(8):58-61.

[8] 肖智,李文娟.基于主成分分析和支持向量机的个人信用评估[J].技术经济. 2010, 29(3):69-72.

[9] 于兆吉，郭亚军. 基于信用级别的银行信贷优化模型[J].控制与决策， 2006，21(12):1429-1431.

[10] 熊平.数据挖掘算法与 Clementine 实践[M].清华大学出版社,2011:15-53.

References

[1] Shi Qing-zong, Le Yun-hui: “Consumer credit scoring model:a survey” [J]. Statistical Research,Vol.8(2003), No.20,p.36-39(in Chinese)

[2] R. A. Fisher. The Use of Multiple Measurement in Taxonomic Problem[R]. Annuals of Eugenie, 1936, 7:179-188.

[3] D Durand. *Risk Elements in consumer Installment financing*[M]. National Bureua of Economic Research, New York, 1941.

[4] J. Myers, W. Forgy. The Development of Numerieal Credit Evaluation System[J]. Jounral of the American statistical Association, 1963, 303(58):779-806.

[5] Yu zhao-ji, Hu Xiang-pei, Mao Qiang: “Novel credit rating method under electronic commerce” [J]. Control and Decision, Vol.24 (2009), No.11,p.1668-1672. (in Chinese)

[6] Jiang Ming-hui, Xie Xing-heng, Wang Shulin, et al: “Personal credit scoring based on Logistic and RBF combined model” [J], Journal of Harbin Institute of Technology, Vol.39(2007), No.7,p. 1128-1130. (in Chinese)

[7] Gao Li: “Personal Credit Scoring And Support Based on Classification Tree Vector Machines”[J], Journal of Neijiang Teachers College, Vol.24(2009), No.8, p.58-61. (in Chinese)

[8] Xiao Zhi, Li Wen-juan: “Personal Credit Scoring Based on PCA and SVM” [J], Technology Economics. Vol.29(2010),No.3,p.69-72. (in Chinese)

[9] Yu Zhao-ji, Guo Ya-jun. Credit Level Based Optimization Model for Bank Loan [J]. Control and Decision. Vol. 21(2006),No.12,p.1429-1431. (in Chinese)

[10] Xiong Ping, *Data mining algorithms and Clementine practice*[M]. Tsinghua University press, 2011,p.15-53. (in Chinese)