基于数据挖掘聚类技术的信用评分评级

左子叶 朱扬勇
(复旦大学计算机与信息技术系 上海 200433)

摘 要 本文提出了一个基于数据挖掘聚类技术的信用评分评级方法。该方法使用数据挖掘的聚类算法,对传统信用评分模型进行了改进,本文给出了方法的理论证明,并在一个信用卡分析系统 DMCA 中实现了该方法,进行了详细的数据测试。理论证明及实验结果都表明,聚类技术在传统信用评分模型的 DM/MTM,分界值,均方差,交叉验证等问题上取得了良好的效果。

关键词 信用评分 信用评级 数据挖掘 聚类

CREDIT SCORING AND RATING BASED ON CLUSTERING TECHNOLOGY OF DATA MINING

Zuo Ziye Zhu Yangyong
(*Department of Computer and Information Technology, Fudan University, Shanghai 200433*)

Abstract This paper proposes a credit rating and scoring method based on clustering technology of data mining. The method promote traditional credit scoring model by the clustering algorithm of data mining. We have given theoretical proof of the method, implemented it in a credit card analysis system DCMA, and provide detailed data test. Our theoretical proof and experiments demonstrate that clustering technology does well in the problems of DM/MTM, Benchmarking, Average Square Sum, Cross-validation in traditional credit scoring models.

Keywords Credit scoring Credit rating Data mining Clustering

1 引 言

“信用是在某一段限定的时间内可获得一笔钱的预期,信用风险就是这个预期将来未能实现的可能性”[1]。目前,经济主体之间的信用活动亦日趋频繁,我国急需建立可靠的、全面的信用风险评估系统。这也为信用风险评估技术提出了新的挑战:易于使用,灵活性好,准确率高。

信用评分技术是信用风险评估技术的一种。本文提出了一个基于数据挖掘聚类技术的信用评分评级方法,使用数据挖掘的聚类算法,对传统信用评分进行了改进,本文给出了方法的理论证明,并在一个信用卡分析系统 DMCA 中实现了该方法。

2 信用评分方法及其问题暴露

2.1 相关工作

目前已有的信用评分模型分为单变量模型和多变量模型,代表性的包括:

基于 ROA(Return On Assets)的单变量模型,因为是单变量,它的算法复杂度是目前最低的;原始 Z 模型[2],它是建立在单变量度量指标的比率水平及绝对水平基础上的多变量信用评分模型。评分结果为关于若干个变量的线性函数,特点是:具有一个广泛使用的分界值。Z 模型后来被改进多次,主要体现在变量的选择和系数上,1998 年提出了简化 Z 模型;1998 年,[3] 提出了基于金融数据的危险模型;和 Z 模型一样,危险模型可以在选择的变量上根据实际有所变化;基于拖欠距离的 Merton 模型[4],1999 年 KMV 公司对 Merton 模型作了改进,使得 Merton 模型日益流行。

这些信用评分模型代表了目前使用的大部分评分模型。它们的基本工作原理类似:根据输入和采集的各项指标,经过模型运算,得到评分分数,然后根据一定的边界值,对分数进行分级,最后输出信用分数和级别。

信用评分模型可应用于各个行业,包括:公司,银行业,保险业的信用风险评分。

2.2 几个问题

传统的信用评分方法存在着几个共同的重要问题,这些问题与模型的性能和评价密切相关。

2.2.1 MTM vs DM

DM(Default mode)模式只定义了两种状态(defanlt vs non-default)。举例说明,只有当客户贷款超过期限时,银行才认为产生了贷款拖欠(default),如果没有拖欠情况,则认为客户贷款是 non-default。实质上就是 0/1 指示器,使用它来判断贷款是否已经超过期限。回到我们的原始定义,边界值就是一种指示器,它指明了客户属于哪一个状态。

与 DM 模式相反,MTM(market to market)根据客户行为的不同定义了多个状态。MTM 方法下,银行依据资产拖欠的恶化程度来逐级降低信用级别。MTM 方法注意了资产的信用价值变化和未来的经济状况变化。

收稿日期:2003-02-13。国家 863 高技术基金项目(2001AA13181);上海市科学技术发展基金项目(01511501O)。左子叶,硕士生,主研领域:数据挖掘。

MTM 和 DM 方法都试图衡量客户的信用级别, DM 方法易于使用, 只需要分为两个级别, 比较简单, 如果需要不止两元的划分, 则不再适用; MTM 方法提供多级划分, 相对 DM 方法更有灵活性, 适用于各种变化, 如流动性资产的贷款级别评估等, 但是需要人为指定多个分界值, 不能自动分级, 假设需要 K 个级别, 则需要用户指定 $k-1$ 个分界值。

2.2.2 交叉验证

由于模型是数据驱动的, 模型的参数对选择的数据训练集敏感。为了获取健壮的参数, 需要使用样本数据的不同子集对一组变量进行反复测试, 称为交叉验证(注意样本数据不能包括验证数据 hold-out)。由于交叉验证的数据敏感性, 我们必须保证, 我们的采样方法是可行的, 即, 需严格证明采样样本有能力代表整个数据集。

2.2.3 边界值

数据和长期影响数据(可能跨越多个信用周期)的缺乏, 造成了寻找边界值的困难。而边界值的选择直接关系到模型的评分能力和准确性, 对信用风险管理有着重要的影响。

2.2.4 均方差

信用评分模型得到评分结果后, 根据边界值, 划分为多个级别, 一个判断级别划分的质量的指标是级别内的均方差[5], 一个级别内的均方差越小, 我们认为这个级别内的数据对象越接近, 它们之间的分数相近性比较好, 对应的公司/个人信用情况比较相似, 分级质量也相对越好; 级别内的分数越相近, 对应的公司/个人信用情况越相似, 分级质量也相对越好。

3 基于聚类的信用评分系统

3.1 DMCA 系统概述

DMCA(Data Mining for Credit card Analysis)系统以银行信用卡部门的客户及其交易数据为数据源, 通过数据清洗、转换、汇总、抽取等技术手段, 构建信用卡部门数据集市(CADataMart: Credit card Analysis Data Mart), 将业务数据与分析数据隔离。在数据集市 CADataMart 的基础上, 利用数据挖掘技术, 以判断客户信用级别, 了解客户消费模式, 预测客户行为, 辅助管理和决策。

DMCA 的体系结构如图 1 所示, 分为三个层次:

第一层次是银行信用卡部门的客户及其交易数据源, 提供原始数据;

第二层次是 DMCA 应用服务器, 该层首先将第一层的原始数据经过清洗、转换、抽取、汇总, 装载到以客户数据为中心的信用卡部门数据集市 CADataMart, 然后以数据集市 CADataMart 为基础, 以数据挖掘技术为核心, 存放分析结果;

第三层次是 DMCA 客户端软件。

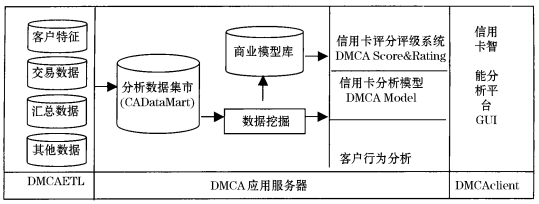

图 1 DMCA 解决方案体系结构图

DMCA 包括四个功能子系统: DMCA ETL、DMCA Model、DMCA Client、DMCA Score & Rating。

DMCA ETL: 信用卡分析的数据抽取工具, 负责将信用卡部门的业务数据转换成 DMCA 系统所需要的数据格式。它位于第一层次。

DMCA Model: 在信用卡分析数据集市 CADataMart 基础上, 创建信用卡分析模型的系统, 同时负责对这些创建的模型进行管理。是第二层次的核心模块。

DMCA Client 子系统: 信用卡分析客户端, 便于用户能够直观地了解信用卡分析模型的内容。它位于第三层次。

DMCA Score & Rating 子系统: 利用 DMCA Model 创建的信用卡分析模型和数据挖掘聚类技术, 预测客户行为, 给相应的信用卡客户行为打分与评级, 生成信用风险评估报告。信用报告可以在实际工作中直接使用, 作为投资、交易的重要参考。

3.2 数据挖掘的聚类算法在 DMCA 中的应用

数据挖掘能够自动地从大量的数据中抽取出潜在的、有价值的知识、模型或规则, 属于发现型技术。它能够从大量数据中发现并提取隐藏在数据中的有效、合理的信息, 帮助决策者寻找规律、发现要素、预测趋势。

数据挖掘的聚类算法将数据对象分组为多个类或簇, 把相似度高的对象集中在一个簇里。同一个簇里的对象彼此相似, 与不同簇里的对象相异, 它是一种无监督的学习过程。聚类算法根据数据自身的属性, 自动地把大数据集分为若干个簇间相异, 簇内相似的子集, 从而解决信用评分评级的边界值、均方差、MTM、交叉验证等问题。

聚类算法主要分为以下几种:

• 基于划分的方法: 代表性的算法有 K-means, K-mediods, CLARANS 等。主要思想是: 选择初始区域, 反复在簇之间移动数据点, 使得目标函数最优。K-means 方法把簇的中心作为簇的代表, K-mediods 则把簇的某一中心位置上的点作为代表点。这种方法的结果是局部最优。

• 基于层次的方法: 它通过反复合并最近的两个簇, 或者分裂比较远的簇, 从而形成几个层次上的簇。有时候, 也把基于划分和基于层次的方法归为一类。

• 基于密度的方法: 它认为, 对于给定类中的每个数据点, 在一个给定范围的区域中必须包含至少某个数目的点。它可以过滤孤立点和噪音。

• 基于网格的方法: 把对象空间量化为有限数目的单元, 形成了一个网格结构。所有的操作在网格结构上进行。

K-means 算法[6]的算法步骤过程为: 1) 随机选择 k 个数据对象(Z 记录)作为初始的中心点。2) 使用距离函数, 计算每个对象点与每个簇的平均值的距离, 将每个对象点分配到最近的簇里。3) 重新计算簇里的平均值。4) 重复第 2、3 步, 直到簇里不再发生变化。

DMCA Score & Rating 子系统的流程为:

1) 采用 K-means 算法使用的采样数据进行聚类;

2) 得出各个级别的分界值;

3) 根据得到的分界值对全部数据进行分级评定。

3.2.1 交叉验证与采样的科学性

对于信用评分的交叉验证性问题, 我们需要证明我们的采样可以代表全部数据集。

假设有样本 S, 大小为 s, 证明: S 中包含了属于每个簇的一定数量的点的概率很大。设 N 为数据集的大小, u 代表某个簇, $|u|$ 为该簇的大小, s 为采样的样本大小, δ 是任意小的正数, $0 \leqslant \delta \leqslant 1$。

根据 Chernoffbounds, 可得以下定理:

定理 3.1　对于一个簇 u, 如果样本大小 s 满足

$$s \geqslant fN + \frac{N}{|u|}\log\left(\frac{1}{\delta}\right) + \frac{N}{|u|}\sqrt{\left(\log\left(\frac{1}{\delta}\right)\right)^2 + 2f|u|\log\left(\frac{1}{\delta}\right)}$$

则样本包含属于簇 u 的少于 $f|u|$ 个点的概率小于 δ, $0 \leqslant \delta \leqslant 1$, $0 \leqslant f \leqslant 1$。

证明: 见文献[7]。

根据定理 3.1 可知, 使用采样数据进行聚类, 采样数据无法代表全部数据特征的概率非常小, 即在很大的概率上, 我们的采样可以代表全部数据集。

3.2.2 均方差局部最小化

K-means 算法是基于划分的方法。DSS(Distance Square Sum)是判别 K-means 聚类结果的重要指标。K-means 算法通过它的算法步骤, 试图找出使 DSS 函数值最小的划分, 一般实际可达到局部最优, 即 K-means 算法的最终目的就是要使 DSS 函数值最小化。也就是说, K-means 算法的准则函数与信用评分评级的"均方差"准则相吻合。

定义 1　一个簇 p 的距离平方和(DSS)为:

$$E_p = \sum_{i=1}^{k}\sum_{p\in C_i} d^2_{pm_i}$$

其中 d_{pm_i} 是簇 p 的点 m_i 与簇 p 的平均值的距离(见定义 2)。整个聚类的 DSS 为:

$$E = \sum_p E_p$$

定义 2　数据对象 i 与 j 的相异度为:

$$d^2_{ij} = \frac{\sum_k w_{ijk} d^2_{ijk}}{\sum_k w_{ijk}}$$

其中, d^2_{ijk} 是第 k 个值的距离的平方, 它可以是 Euclidean 距离, 也可以是 Manhattan 距离, 或者其它的距离函数。w_{ijk} 是值 0~1 的权重, 它决定第 k 个值的重要性。

除了以上的理论证明, 我们从第 4 节的实验数据也可以看出, K-means 算法大大减小了簇内均方差。

3.2.3 自动化的 MTM 方法

DM 方法虽然只需要一个分界值, 但是灵活性不够, 限制了它的应用。MTM 方法则需要人为指定多个分界值, 假设需要 K 个分级, 则要预先指定 $K-1$ 个分界值, 由于分界值的数据敏感性, 往往会受到各种因素的影响, 只能是经验性的数据, 也就违背了信用评分系统的基本准则: 必须要经过严格的验证并且应该是客观的, 即经过一个统计的、机械的过程, 独立于分析人员的主观观念。聚类算法是无监督的学习过程, 根据数据的自然分布和用户输入的级别数目, 自动把数据分为指定数量的组, 并得出各组的特征值。

3.2.4 自发产生边界值

聚类算法是无监督的学习过程, 不需要人为指定分界值, 自动产生不相交的数值子集, 根据最大、最小值等数据, 很容易即可得出边界值。更详细的信息, 请参看第 4 节的实验与结果分析。

4　实验与结果分析

4.1 实验环境

PC 机, 处理器: Pentium III; 128MB 内存; 操作系统: Microsoft Windows 2000 Professional; 聚类算法实现: J++ Builder6.0; 数据生成器: C++ Builder5.0。

聚类算法采用改进后的 K-means, 信用评分模型使用原始 Z 模型, 原始 Z 模型包括五个变量, 评分公式如下:

$$z = 1.2x_1 + 1.4x_2 + 3.3x_3 + 0.6x_4 + 0.999x_5$$

其中, X_1: 营运资本/总资产; X_2: 留存收益/总资产; X_3: 税息前利润/总资产; X_4: 权益市场值/总债务的账面值; X_5: 销售收入/总资产。

4.2 实验数据

由于获取大量实际的金融数据比较困难, 系统开发了随机数据生成器, 对应于 Z 模型, 一共生成 5 个数值指标, 分别对应: X_1(−1~1); X_2(−0.9~0.6); X_3(−0.6~0.4); X_4(0~3); X_5(0~5)。数据分布服从正态分布或随机分布。

说明: 实际实现时, 由于值域较小, 质量测试的结果差异很小, 为了便于比较, 将每个指标放大了 100 倍, WC/TA 的值域由−1~1 变为了−100~100, 依此类推。

测试数据量为 1000 条数据记录。

4.3 实验结果与分析

4.3.1 原始数据记录(1000 条)

测试数据集 1 的原始数据分布和方差如表 1 所示:

表　1

	最大值	最小值	均方差	平均值
Z 评分	6.95981216	−2.28476905 8	3.816231	2.913523
营运资本/总资产 x1	98.8433456	−98.7334823	3181.1042993	−1.8962982
留存收益/总资产 x2	49.5641975	−89.1112976	1704.1508653	−21.6726765
税息前利润/总资产 x3	38.883022 30	−58.9135398	750.89792226	−12.2697531
权益市场值/总债务 x4	295.957824 7	4.3168430	6628.0390059	160.5763107
销售收入/总资产 x5	4.9612722	0.0158085	2.02938695	2.6838245

4.3.2 聚类结果与分析

1) 基本的 Z 模型: 分成 5 个簇, 指定边界值为: 0.91, 1.81, 2.99, 5.5。

分类后, 每个簇内的最大值、最小值, 平均值和方差, 如表 2 所示。

表　2

簇标号	最大值	最小值	平均值	均方差
1	0.905627	−2.28477	0.483236	1.189046
2	1.804847	0.918574	1.524653	1.142356
3	2.983543	1.856909	2.02746	0.921493
4	5.402146	3.010559	4.235642	1.124863
5	6.959812	5.510123	4.29363	1.164683

2) K−means 算法, 只需指定簇的数目, 数据自动分成 5 个簇: 聚类结果, 每个簇里的最大值、最小值, 平均值和方差, 如表 3 所示:

表　3

簇标号	最大值	最小值	平均值	均方差	聚类前的均方差
1	0 0	−2.28477	−1.73979	0.110384	1.189046
2	0.899399	−0.68669	0.176234	0.163596	1.142356
3	2.558041	0.974683	1.704716	0.235908	0.921493
4	4.471203	2.617802	3.487793	0.226554	1.124863
5	6.959812	4.6264	5.632267	0.409703	1.164683

表 3 中, 每个簇的 z 评分均方差最大为 0.409703, 最小为 0.110384, 比表 2 中传统评分固定标准得到的簇的均方差减小了一个数量级, 可见, 聚类算法把最接近的分数聚集在同一个簇里。

(下转第 101 页)

或S_A= 1。这个结果对于序列的分析是无意义的。随着序列长度的改变S_A变化不大。也即是说,改变序列的长度对使用 C 语言库调用特征检测入侵影响不大。当序列长度大于 6 时,入侵检测的结果是稳定的。在一般情况下选择序列长度等于 10。

4 结　论

语言库调用序列作为特征在应用级检测入侵的方法比用系统调用特征更有应用方向。从应用的观点看,语言库调用特征不是 OS 的特性,它允许使用丰富的语义。使用了不同的异常度量方法(不匹配的数量,局部范围的异常数,异常信号的标准化)检测应用的异常。从实验中可以看出语言库调用序列能够描述不同应用的特征。用库调用方法可以检测到来自内部和外部的入侵。

参 考 文 献

[1] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A Sense of Self for Unix Processes In Proceedings of 1996 IEEE Symposium on Computer Security and Privacy, 1996.

[2] S. Forrest, S. Hofmeyr, and A. Somayaji. Computer Immunology. In Communications of the ACM, Vol. 40, No. 10, pp. 88~96, 1997.

[3] R. Heady, G. Luger, A. Maccabe, and M. Servilla. The Architecture of a Network Level Intrusion Detection System. Technical Report CS90—20, Dept. of Computer Science, Univ. of New Mexico, August 1990.

[4] S. A. Hofmeyr, A. Somayaji, and S. Forrest. Intrusion Detection using Sequences of System Calls. In Journal of Computer Security, Vol. 6, pp. 151~180, 1998.

[5] Debian ltrace home page, http://packages.debian.org stable/utils/ltrace.html.

[6] Matthew Stillerman, Carla Marceau and Mareen Stillman. Intrusion Detection for Distributed Applications. In Communications of ACM, 1999.

[7] Rebecca Curley Bace, 入侵检测, 人民邮电出版社.

[8] Rechard Petersen 著,陶华敏、龚志翔、任宇飞、谢晓竹等译, Linux 技术大全, 机械工业出版社.

(上接第 3 页)

从每个簇的最大值和最小值还可以看出,每个簇之间是不相交的子集。

由表 2 和表 3 可知,聚类算法把最接近的一些分数聚集在一起,并能够给出簇特征。解决了传统评分指定临界值、手动划分的方式划分数据。

表 4

簇标号	最大值	最小值	平均值	均方差	聚类前的均方差
2	6.959812	6.189983	6.55400238	0.0879	0.589046
0	5.863289	5.000843	5.405485789	0.0576	0.842356
6	4.750832	4.044838	4.320374807	0.0584	0.921493
8	3.95009	3.443657	3.70676589	0.0263	0.624863
5	3.301683	2.903294	3.156416297	0.012802	0.664683
3	2.71541	2.261889	2.531480577	0.0215	0.582365
9	2.056385	1.515071	1.743447911	0.0307	0.652356
4	1.340792	0.61586	1.005220461	0.0512	0.365458
7	0.319856	−0.68669	−3.57E−03	0.0832	0.862354
1	0	−2.28477	−1.73979377	0.110384	0.865345

3) 使用传统评分和聚类计算对测试数据集再次进行比较,聚类数为 10,聚类结果如表 4 所示。

从表 4 可以看到,这次的簇的均方差已经降到 0.01 的数量级,簇内的相似度更加明显,而聚类前的均方差都在 0.5 以上。这样,我们通过对 Z 评分进行聚类,可以自动把分数最相近的记录分到同一个簇里,确保了每个等级里的 Z 评分都是最接近的。

4) 传统评分评级方法和基于数据挖掘聚类方法得到的簇内均方差的平均值,如图 1 所示。

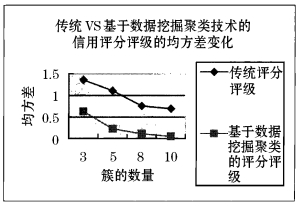

图 1

5 小　结

综上所述,数据挖掘聚类算法对传统信用评分模型的交叉验证,DM/MTM,边界值,均方差等问题上有良好的表现。聚类算法可自动生成级别,轻易变 DM(二元划分)方法为 MTM 方法,多元评级,便于用户根据需要随意调整等级数;自动得出边界值;大大减小簇内的均方差。本文在概率上对聚类算法采样验证的方法进行了证明。实验与分析结果表明,聚类算法得出的评级结果质量比较高。

以上的理论证明、论述和大量的实验结果都表明:聚类算法对传统的信用评分模型有着良好的改进作用。

参 考 文 献

[1] Anthony Saunders, Credit Risk Measurement: the next great financial challenge, 2nd edition, amozon.com, 1998 p. 278

[2] Altman, E. I., ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations, 1977.

[3] Shumway, T., Forecasting Bankruptcy More Accurately: A Simple Hazard Model, 1998.

[4] Kealhofer, s., Credit Risk and Risk Management, 1990.

[5] Anthony Saunders, Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms, 2nd edition, amozon.com, 1998.

[6] David Wishart, k-Means Clustering with Outlier Detection, GfKl 2001, the 25th Annual Conference of the German Classification Society, University of Munich, March 14~16, 2001.

[7] Sudipto Guha, CURE: A clustering algorithm for large databases, Proceedings of the 1998 ACM SIGMOD international conference on Management of data, 1999.

[8] Liadan O'Callaghan, Streaming-Data Algorithms For High-Quality Clustering, Proceedings of IEEE International Conference on Data Algorithms For High-Quality Clustering, Proceedings of IEEE International Conference on Data Engineering, March 2002.

[9] Haixun Wang, Clustering by pattern similarity in large data sets, Proceedings of the 2002 ACM SIGMOD international conference on Management of data, 2002.

[10] FAS—A Freshness-Sensitive Coordination Middleware for a Cluster of OLAP Components, 28th International Conference on Very Large Data Bases, VLDB 2002.