大数据架构下企业内部信用评级的实证研究

牟 刚, 袁先智*
(同济大学数学系, 上海 200092)

摘要: 研究了大数据背景下企业高级量化分析的方案, 并针对某企业内部的信用评级案例进行了实证研究. 通过考虑企业业务, 信息技术和数学模型三方面的整合, 提出了全新的企业高级量化分析平台, 把大数据和大计算有机的结合起来. 在这个平台上结合实际数据, 使用逻辑回归模型进行了信用评级. 新的评级模型很好的区分了不同客户的资信好坏. 研究发现, 模型选择和元模型库的建立对于企业高级量化分析工作至关重要.

关键词: 信用评级; Logistic 回归; 大数据; 金融模型
中图分类号: TP273 **文献标识码**: A **文章编号**: 1000−5781(2016)06−0808−08
doi: 10.13383/j.cnki.jse.2016.06.009

Empirical study for enterprise internal credit rating under big data framework

Mu Gang, Yuan Xianzhi*
(Department of Mathematics, Tongji University, Shanghai 200092, China)

Abstract: This paper focuses on enterprise advanced analytics solution under big data context. An empirical study for enterprise internal credit rating business is also in place. By considering factors of business process, information technology and mathematical model, a new enterprise advanced analytics platform is given, which integrates big data and big computation. On this platform, logistics regression model is used for credit rating. New rating model gives more details of credit of customer. In this process, it is found that model selection and meta-model library is vital important for enterprise advanced analytics implementation.

Key words: credit rating; logistic regression, big data, financial models

1 引 言

从 2010 年以来, “大数据” 成为全球非常流行的一个关键词, 原因在于大数据框架是革命性工具可以满足层出不穷的业界需求. McAfee 等[1]关于大数据对管理带来了革命性挑战的讨论. 另外, 魏伟[2] 也讨论了基于大数据背景下银行业面对互联网金融挑战的应对策略思考, 最近李平等[3]也对目前互联网金融的发展与研究进行了综述讨论. 本文的目的基于作者过去几年从学术理论研究和业界实践经验两方面对大数据的理解, 结合企业现状提出了全新的大数据架构, 并结合具体的企业信用业评级案例, 使用 Logistic 回归进行实际应用. 整个过程整合了企业的业务, 信息技术和数学模型三个方面的能力, 探索了一条从实际问题出发, 并结合大数据, 应用数学模型的解决管理问题的新思路. 试图把成熟的金融模型应用到实际的企业数据中, 从而结合模型和数据的双重价值. 研究发现 , 实践中复杂的模型在应用中可能出现难以解释和描述实际问

收稿日期: 2016−01−06; 修订日期: 2016−05−30.

* 通信作者

题的情况. 因此建议从简单模型入手, 逐渐增加需要考虑的因子, 建立更复杂的模型, 从而在实践中更具可行性.

1.1 大数据的定义和应用

大数据是一个特别流行的术语, 它通常指的是任意一个很大很复杂的数据集, 规模大到不能用传统软件在合理时间内进行抽取, 转化, 分析, 存储, 处理和可视化等操作. 通常大数据具有“4v”的特点, 分别是数据量大(volume),数据增长的速度高(velocity), 数据来源多样(variety)和数据的真实性(veracity). 目前对于大数据的基本理解是 TB 或者 PB 级别的数据就是大数据. 目前主流的大数据实现平台是基于 Google 2004 年发表的 MapReduce 算法. 目前 Apache 的 Hadoop 是这一算法的开源实现. Spark 是 MapReduce 算法的另外一种实现, 相对于 Hadoop 它使用了大量的内存计算, 同时提供了流数据处理, 图算法和机器学习的包. 需要注意到的是, 大数据当前更多的是应用于互联网搜索和社交网络, 而企业中的业务数据, 更多的是结构化的数据. 如何建立有效的大数据架构, 借鉴现有的数学模型, 应用于企业管理, 是当前面临的挑战.

1.2 大数据在企业中应用的解读

大数据在企业应用中需要关注四个方面: 1)如何应对结构化和非结构化的海量数据抽取, 转换和装载(ETL)并生成相应的汇总/报表, 2)如何对海量数据进行高效的科学计算, 3)如何把数据密集和计算密集的应用有机的结合起来建立使用的模型以及 4)如何将分析的结果通过可视化的工具展现. 目前在企业中, 对于 1)和 4)有很多应用和实践, 而对于中间两点, 缺乏相应的技术和模型. 高校/科研院所在 2)和 3)有很多前沿研究, 但是对于其他两点, 缺乏相关的工程化经验和数据. 从这个角度来说, 企业和高校/科研院所具有很高的互补性.

银行有很多成熟的风险管理和定价模型, 这些模型在应用的时候往往碰到数据不够, 参数难以估计的问题. 即使有数据, 也往往来自于企业提供的报表, 评级公司的评级, 各种行业报告数据等. 数据的准确性和可靠性值得商榷. 这也是直接导致模型的结果和实际的结果偏差比较大的一个主要原因. 在大数据的背景下, 使用已经收集的真实历史数据来动态实时估计模型的参数进行计算, 是解决现有金融数据/模型失真的一个关键. 既使用企业的真实数据, 套用成熟的金融模型, 解答业务提出的问题.

一般而言, 数据的数据量在 Terabyte 和 Petabyte 之间可以称为大数据, 一个典型的, 10 万人左右的全球五百强企业的数据量在 16 PB 级别, 其中管理/运营所产生的数据量在 4 PB 左右.

数据的多样性是指不同类型的数据, 包括了结构化, 半结构化和非结构化数据. 典型的数据类型包括 ERP 系统的历史数据和用于智能化工业生产的传感器数据等, 除此之外, 文本, 微博, 音频, 视频及日志等也作为数据的来源.

数据的速度是另外一个话题, 实时性会完全改变业务的模式和决策的结果. 传统的企业数据仓库, 往往由于性能限制, 只能给出隔天的分析结果. 实时的分析结果, 可以帮助决策者做出即时决定, 扁平化组织, 将组织现状透明化, 这些都对业务产生新的价值.

数据的真实性对于模型的准确性至关重要, 甄别数据真实性, 最关键的判别标准之一是数据是客观收集还是被动填写. 客观收集意味着在真实交易过程中得到的数据, 或者在被收集人不知道情况下采集的数据. 在这一过程中, 数据不确定性是可以接受的, 通过数据质量验证过程可以保证数据的准确性, 在需要的情况下可以进行数据的校准. 人为填写的数据受心理, 环境, 时间, 地点等多重因素的共同干扰, 往往不能真实的反应实际的业务情况.

落实大数据概念所需关注重点是通过适当的架构, 配合企业现有 IT 应用, 建立相应的数学模型库.

2 新的企业大数据架构

基于如上的大数据概念, 结合当前某企业的实际情况, 提出新的企业大数据架构. 这一架构分为数据层, 模型实现层和展示层三个部分. 数据层主要是通过数据的结构和收集, 体现业务人员对于业务的理解; 模型

实现层通过统计分析平台和 CPU/GPU 混合高性能计算平台来实现数学模型库, 体现了分析人员数学模型的构建能力; 而展现层, 则通过的 J2EE Web 服务将内嵌在工业标准 HTML5 的可视化图表在笔记本/平板电脑/手机等设备上展示出来, 体现了 IT 人员的实施能力. 而其中的核心就是通过最新的 SAP HANA 的内存数据库将业务, 数学模型和 IT 紧密结合在一起.

将现有的传统 ERP 数据实时的通过 Sybase Replication Server 复制到 HANA 内存数据库中. 这个复制不单是抽取某一个 ERP 系统的数据而是将企业全球各个地区的数据实时复制, 从而达到企业运营管理透明的目的. 在这一过程中, 最关键是主数据的管理. 在企业实践中, 往往从最小的公共子集开始拓展, 定义元数据, 通过数据总线把元数据同步到各个 ERP 系统的实例中. 通过这个步骤, 增加企业运营的透明度, 同时在企业中提倡数据民主的概念, 即数据是企业的共有财产, 而非企业某个地区或部门的私有财产.

将现有的企业中其它系统的数据, 业务人员使用的类似于 Excel 表/Access 数据库等本地数据库中的数据, 通过 HANA 的 Data Service 导入 HANA 数据库. 将存储在 Hadoop/Spark 平台上的超大数据集, 使用 MapReduce 等方法, 提取数字特征, 将统计/汇总结果作为结构化数据导入 HANA 数据库. Hadoop 和 Spark 平台在这一过程中也担任了把计算带到数据中的角色.

数据层通过对数据结构的理解, 获取模型所需的各种原始数据. 值得一提的是, 载入的除了企业私有的各种类型的数据之外, 也包括公开的信息, 如 Internet 上通过爬虫搜索到的各种信息和公共的信息, 如政府提供的气象部门, 工商部门和疾控中心等所拥有的可以作为公共信息发布的数据. 这些数据往往都存放在 Hadoop 平台中, 通过 Data Service 传输到 HANA 数据库.

通过数据理解了业务流程以后, 使用数学模型, 给出业务洞察. 目前统计分析和统计学习是企业中使用最多的数学工具, 除此之外, 针对生产制造和财务管理的最优化问题, 针对物流网络的图论的问题也在企业实践中经常碰到. 特别想指出, 企业管理有很多倒向随机微分方程的一类问题, 例如, 把每个月的销售额看成股票价格, 假设其遵循几何布朗运动, 把销售的业绩目标作为行权价, 则员工奖金的计算和亚式期权的计算是一样的. 可以使用亚式期权的定价模型来计算相关的管理问题. 相应的可以考虑其它金融风险和定价模型, 对于管理问题使用企业内外部数据进行解答. 数学模型库通过 SVN 等工具进行版本控制, 使用知识管理工具进行发布, 帮助企业建立数学能力. 这个模型库包括了针对企业管理人员的基本概念, 针对分析人员的模型应用和针对建模人员的模型推导过程.

内存数据库是整个架构的核心和结合点. 内存数据库(IMDB) 是一种将数据放在内存中直接操作的数据库, 相对于存放在外部存储器中的数据库, 内存的数据读写速度要高出几个数量级, 将数据保存在内存中相比从磁盘上访问能够极大地提高应用的性能. 它最大的特点是传统关系型数据库(RDBMS)里表的存储方式从行存储变为列存储. 列存储往往应用表里一列的数据冗余度来大幅压缩数据, 求和, 平均等聚集的计算在列上操作效率非常高. 可以说内存数据库天生就是为统计计算而准备的. SAP HANA (high-performance analytic appliance) 是一个软硬件结合体, 是内存数据库的一个实现. 它提供高性能的数据查询功能, 用户可以直接对大量实时业务数据进行查询和分析, 而不需要对业务数据进行建模, 聚合等. 用户拿到的是一个装有预配置软件的设备. 它基于内存计算技术的高性能实时数据计算平台, 是全球一个发布商用的基于内存计算的产品. 通过构建一个 1 TB 的 HANA 平台, 作为大数据的分析平台, 汇总数据层的各个数据源. HANA 平台中通过 R 存储过程来实现模型.

算法的并行计算是通过 CPU/GPU 混合的高性能计算平台实现的. 由于散热, 功耗以及材料的物理特性极限, 通过不断增加晶体管密度来提高单个处理器的运算能力可能性越来越低. 多核技术, 多CPU 的并行处理技术和异构多核集成技术(CPU 与 GPU 的组合)已经成为提高计算性能的主流途径. 在企业中, 往往重视数据密集型的问题, 而随着大数据不断深入, 出现了越来越多的计算密集型的问题. 而基于 Nvidia CUDA 编程模式的 CPU/GPU 混合架构的 Tesla K40(很快将推出 K80)平台是目前最先进的平台之一. 它有 2 880 个处理器, 单精度计算高达 4.29 TFLOPS. 使用 R 语言回归, 估计出来的模型参数, 通过 C 语言的扩展编程 CUDA 进行实现模型, 利用并行化算法, 通过 MPI 进行通讯, 达到集群计算的目的. 这个架构对于企

业管理中最常用方法之一的蒙特卡罗法的加速, 具有非常好的效果.

计算结果通过 Java JBoss 应用服务器, 使用 Spring MVC 框架, 采用 Web 方式, 通过 HTML5, 展现数据分析结果. Spring MVC 使用基本的 JavaBean 来完成以前只可能由 EJB 完成的事情, 可以提供简单性, 可测试和松耦合的 Java 解决方案.

使用开源的 D3js 框架来进行数据可视化, 帮助用户理解分析的结果. D3js 是一个基于数据操作文档JavaScript 库. D3 通过使用 HTML, SVG 和 CSS 提供动态的可视化效果. D3 允许绑定任何数据到 DOM 对象模型, 然后应用数据驱动转换到文档. 例如, 可以用 D3 从数组生成 HTML 表格, 或者使用相同数据平滑和动态创建一个 SVG 图表.

客户使用多种设备, 通过公开或者公司内部网络, 访问分析结果. 结合业务的具体情况作出更准确的业务决策. 在这一企业管理大数据应用过程中, 整个大数据系统并非代替用户进行决策, 而是为企业决策提供更多实时的信息和知识. 这里的讨论架构和应用于工业 4.0 的大数据架构有很大区别.

3　基于企业大数据架构的实证研究

1) 结构模型

结构模型认为公司违约的发生是公司资产价值降低导致的结果. 该理论将公司的股权视为以公司资产价值为标的的欧式看涨期权. 如果股票市场是有效的, 那么在知道公司股价和股价波动率, 以及公司债务结构的情况下, 可以估计出该公司的违约概率. 1997 年, KMV 公司(现被穆迪收购)对基于 Black-Scholes 公式的 Merton 模型进行了重要的改进, 推出 KMV 模型[4]. KMV 模型将企业负债看作一份欧式看涨期权, 利用 Black-Scholes 期权定价公式, 根据企业股价 E, 股价波动率 σ_E, 债券到期时间 T, 无风险借贷利率 r, 负债 D, 来估计出企业的资产价值 V 和资产波动率 σ_V. KMV 模型中的两个未知变量 V 和 σ_V, 可以从以下联立方程组中求得, 即

$$\begin{cases} E = V\Phi(d_1) - D\mathrm{e}^{-rT}\Phi(d_2) \\ \sigma_E = V\Phi(d_1)\sigma_V/E, \end{cases} \tag{1}$$

其中 $\Phi(\cdot)$ 为标准正态分布函数, $d_1 = (\ln(V/D) + (r + \sigma_V^2/2)T)/(\sigma_V\sqrt{T}), d_2 = d_1 - \sigma_V\sqrt{T}$.

根据公司的违约点 DP(一般采用短期负债加长期负债的一半), 计算借款人的违约距离

$$\mathrm{DD} = (V - \mathrm{DP})/(V\sigma_V). \tag{2}$$

最后, 根据历史违约数据得到违约距离所对应的违约概率.

张大斌等[5]提出一种中国上市信用风险测度的不确定性 DE-KMV 模型, 对 KMV 模型针对中国国内的情况进行了优化, 使用用差分进化算法(DE) 来优化了违约点系数, 该模型通过分位数回归分析, 其系数在置信区间内显著性更好. 因此, 相对于常用的 KMV 模型, 该模型更据灵活性, 能提高上市公司信用风险测度的准确性.

2) 生存分析模型

生存分析模型不仅仅是被评级对象的违约率进行判断, 而且对违约率的期限结构进行研究. Tyler[6]使用了带有与时间相关变量的离散型模型. 该模型等同于一个多阶段的 Logit 模型, 但 Logit 模型的标准误差(standard error) 需要进行调整. 其形式如下

$$S(t, \boldsymbol{\xi}; \boldsymbol{\theta}) = 1 - \sum_{\tau<t} f(\tau, \boldsymbol{\xi}; \boldsymbol{\theta}), \tag{3}$$

$$\phi(t, \boldsymbol{\xi}; \boldsymbol{\theta}) = f(t, \boldsymbol{\xi}; \boldsymbol{\theta})/S(t, \boldsymbol{\xi}; \boldsymbol{\theta}), \tag{4}$$

其中 t 是可能发生违约事件的时间, $f(t, \boldsymbol{\xi}; \boldsymbol{\theta})$是违约的概率质量函数, $\boldsymbol{\theta}$ 代表了 f 的参数向量, $\boldsymbol{\xi}$ 代表了解

释违约原因的向量. S 是生存函数, 而 ϕ 是风险函数.

Duffie 等(以下简称 DWS)[7]推进了对解释变量的时间序列动态机制(time-seires dynamics) 的研究, 并可预测多时间点的违约率(每季度或每年). 该模型定义了代表公司特殊因素以及整体宏观因素的马尔可夫向量 $\boldsymbol{X}_t$. 违约强度为 $\lambda_t = \Lambda(\boldsymbol{X}_t)$, λ_t 表示平均每年违约的次数, 其它退出情况(如并购)强度为 $\alpha_t = A(\boldsymbol{X}_t)$, 那么总退出强度即为 $\lambda_t + \alpha_t$. 目前在 t 年存活的公司, 在 $t+s$ 年首次违约但没有发生其它退出情况的条件概率为

$$q(\boldsymbol{X}_t, s) = \mathrm{E}\left[\int_t^{t+s} \mathrm{e}^{-\int_t^z (\lambda(u)+\alpha(u))\mathrm{d}u}\lambda(z)\mathrm{d}z|\boldsymbol{X}_t\right]. \tag{5}$$

Duan 等[8]对 DWS 模型进行了改进, 使用了远期强度(forward intensity)技术和伪似然函数(pseudo-likelihood function)估计技术, 并增加了违约距离的趋势项和公司财务指标等解释变量.

3) 模型选择标准与分析

信用风险评估专家 Galindo 等[9]提出了好的信用评估模型的质量要求是: 1) 精确度: 评级结果的误差率较低; 2) 变量较少: 不包含太多的解释变量; 3) 可行性: 采用可获取的数据资源; 4) 透明性和解释性: 能高水平地反映数据之间的关系和趋势, 模型结果易读. 针对企业内部的信用评级问题, 根据上述原则来进行分析和筛选.

结构模型在国内的应用存在如下问题: 1) 国内股票市场不够成熟, 市场有效性低, 这会影响模型的效力. 2) 对于未上市企业模型并不适用, 这限制了模型的应用范围. 3) 在没有 KMV 的违约数据库的情况下, 使用 Merton 模型计算违约率和实际情况出入比较大.

生存分析模型中对于参数的估算非常困难, 在企业的实际操作中很难实现.

综上所述, 二元选择模型是比较现实和可行的选择. 通过实证研究, 发现其精确度高, 变量要求较少, 可行性, 透明性和可解释性都比其它模型要好.

3.1 Logistic 回归模型

Logistic 回归和多重线性回归都属于广义线性回归模型. 如果因变量是连续的, 没有范围限制就是多重线性回归; 如果因变量是$\{0,1\}$取值的二项分布, 则为 Logistic 回归. 1920 年 Raymond 等[10]在研究果蝇的繁殖中发现和使用该函数, 并在人口估计和预测中推广使用. Logistic 函数的形式为

$$f(\zeta) = \mathrm{e}^{\zeta}/(1+\mathrm{e}^{\zeta}). \tag{6}$$

易知其值域为$[0,1]$. 用 $p_i = \Pr(Y = y_i|x_{i1}, x_{i2}, \ldots, x_{ik})$ 作为因变量得到 Logistic 回归模型

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}, \tag{7}$$

其中 Y 是服从两点分布的随机变量, $\Pr(Y = 1|x_{i1}, x_{i2}, \cdots, x_{ik}) = p_i$, $\Pr(Y = 0|x_{i1}, x_{i2}, \ldots, x_{ik}) = 1 - p_i$, 从而得到

$$\ln(p_i/(1-p_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{8}$$

为了将 Logistic 回归模型转换为线性模型, 定义 logit（logistic probability unit）变换为

$$\mathrm{logit}(p_i) = \ln(p_i/(1-p_i)), \tag{9}$$

从而有

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{10}$$

令得到实际观测值 y_i 的概率为

$$\Pr(Y = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}, \tag{11}$$

似然函数为

$$L=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}, \tag{12}$$

对上式两边取对数

$$\ln L=\sum_{i=1}^{n}\left(y_i\ln p_i+(1-y_i)\ln(1-p_i)\right)=\sum_{i=1}^{n}\left(y_i\ln\frac{p_i}{1-p_i}+\ln(1-p_i)\right), \tag{13}$$

代入 p_i 得到

$$\ln L=\sum_{i=1}^{n}\left(y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik})-\ln(1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}))\right), \tag{14}$$

当

$$\frac{\partial\ln L}{\partial\beta_i}=0,\ i=0,1,\ldots,k, \tag{15}$$

$\ln L$ 取得最大值. 采用 Newton-Raphson 迭代可以得到参数 β_i 的估计值 b_i.

3.2　实证研究

某五百强企业, 其财务授信管理部门采用专家判别的方法对企业进行评级, 以便进行授信额度管理. 评级分为 A, B1, B2, C 四个级别. 对其 50 家客户的评级如图 1.

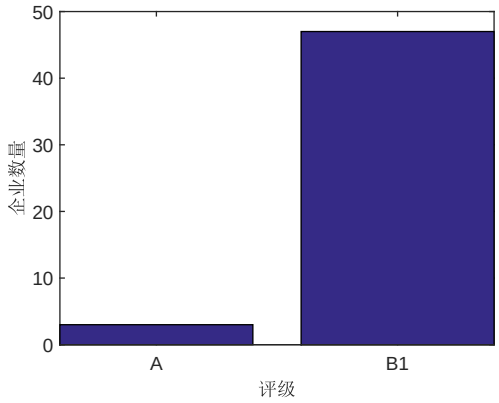

图 1　企业 2013 年专家判别法评级结果

Fig. 1　Corporate ratings in 2013 by using experts criterion method

在实际操作中, 发现评级的结果很难区别好的客户和坏的客户. 如何进行有效的评级成为该公司的难题. 把公司看作银行, 把公司的客户看作银行的客户, 把营收账款看作利率为 0 的短期债券. 借鉴银行对企业的信用风险指标进行主成分分析, 可以选取合适的指标来对客户进行评级.

银行对企业信用风险指标的选择, 主要考虑了可操作性, 系统科学性, 定量指标与定性指标相结合, 风险度量方法与度量目的相结合和企业整体信用与局部信用相协调. 银行在选择度量指标时注重企业的发展, 创新和成长, 考虑宏观经济的影响, 在极端情形下风险仍然可控. 为了全面综合评价企业主要领导者及企业管理者素质, 市场竞争力, 银行信用状态, 偿债能力, 盈利能力, 规模及经营能力, 发展能力以及担保和抵押情况. 银行使用了六个维度共 39 个指标, 其中包含 22 个定性指标和 17 个定量指标作为风险度量指标池.

对上述指标进行筛选和主成分分析. 企业素质方面, 该 500 强公司的客户均为国内大中型企业, 管理者均为职业经理人, 生产情况, 企业员工基础能力等没有显著差别. 宏观经济, 市场评价等维度的指标, 由于企业客户的行业同质性, 也可以不予以考虑. 对于其它指标进行主成分分析, 并增加针对应收账款考虑周转天数和资本周转等指标. 发现对于该企业评价客户信用等级, 最重要的指标包括了 8 个, 如表 1 所示.

由向该企业的商务部门向 50 家客户按年度收集以上数据, 并要求客户的财务报表经过外部审计. 从而得到 2013 年该 50 家企业的相关的 8 个指标共计 50 条数据.

表 1 信用等级评价涉及的指标

指标变量	中文描述	英文描述
X_1	净利率	Net Profit Margin
X_2	应收账款周转天数	Receivable Turnover
X_3	速动比率	Quick Ratio
X_4	库存周转率	Inventory Turnover
X_5	投资回报率	Return on Investment
X_6	净资产负债率	Net Worth on Liabilities
X_7	营运资本周转天数	Working Capital Days of Sales
X_8	营业利润率	Operation Profit Rate

将该企业 SAP 数据库中 2013 年和这 50 家企业相关的历史交易数据导出. 查看这些企业应收账款的实际付款情况. 参考银行的信用卡宽限期管理的办法, 定义 5 d 为应收账款的宽限期, 即超期付款时间在 5 d 以内不认为延期. 统计客户 2013 年付款延期的笔数和客户总的交易笔数, 得到每个客户在 2013 年延期付款的比例. 分两种情况定义客户违约

1) 如果客户延期比例大于零, 则定义为违约 $Y=1$, 延期比例等于零为不违约 $Y=0$;

2) 如果客户延期比例大约 5%, 则定义为违约 $Y=1$, 否则定义为不违约 $Y=0$.

根据这两种情况使用统计软件R对上述的数据按照 Logistic 回归进行计算.

对于情况 1), 有

$$\begin{aligned}\text{logit}(p)=&1.420\,6-46.097\,1X_1-0.034\,0X_2+3.403\,0X_3+0.005\,9X_4-2.267\,7X_5-\\&8.764\,7X_6+0.060\,1X_7+25.249\,0X_8,\end{aligned}\tag{16}$$

对于情况 2), 有

$$\begin{aligned}\text{logit}(p)=&-4.559\,0-14.524\,6X_1-0.003\,8X_2+5.200\,5X_3-0.006\,7x_4-0.230\,8X_5-\\&6.715\,7X_6+0.000\,5X_7-5.890\,9X_8,\end{aligned}\tag{17}$$

其中 p 违约概率.

企业的实际情况是客户付款延期有很多种原因, 比如, 商务操作流程, 银行付款流程等. 本文采用延期比例大约 5% 作为违约的边界条件. 马若微等[11]讨论了通过选择适当的切割点使总期望判断损失最小. 本文采用同样的思路, 使用第二种情况作为评级的参数.

使用公式

$$p_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik})}{1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik})},\tag{18}$$

将客户的指标 $X_1\sim X_8$ 代入得到每个客户的违约概率.

参考 CSFP 信用风险附加模型, 在任何时期该模型和本文中的案例一样, 只考虑违约和不违约这两种状态, 计量相应的损失. 在 CSFP 信用风险附加计量模型中, 违约概率不是离散的, 而是被模型化为具有概率分布的连续变量. 对于企业交易, 每一个客户的违约都是小概率违约事件, 并且每个客户的违约概率都独立于其它客户, 这样, 客户的违约概率的分布接近泊松分布. 相应的其对应的分布进行划分, 从而得到评级.

根据以上的数值结果, 使用公司内部评级标尺(违约概率 $\leqslant 10\%$ 为 A 级客户, $10\%<$违约概率$\leqslant 20\%$ 为 B1 级客户, $20\%<$ 违约概率 $\leqslant 30\%$ 为 B2 级客户, 违约概率 $>30\%$ 为 C 级客户对以上客户进行评级, 得到如下图 2 的图示结果.

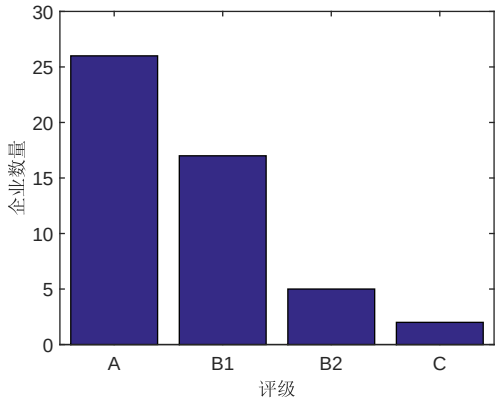

图 2　企业 2014 年评级的结果

Fig. 2　Corporate ratings in 2013

这个结果和实际情况吻合得很好, 表明大多数的客户都是资信比较好的, 极少客户资信比较差. 在实际操作中, 大多数的客户都是 A 以上的评级, 该评级相当于对于标普 A 以上评级进行了进一步的细化, 从而适用到企业的日常管理中. 新的评级模型很好的区分了不同客户的资信好坏. 该模型也通过四大审计公司之一的验证, 应用到实际的信用限额管理和账期管理中, 也得到了比较好的效果.

4　结束语

本文简要回顾了大数据的概念, 并加以进一步的阐述. 提出了基于企业历史数据的新的大数据的架构, 并通过实际的企业信用评级的案例, 使用新的架构进行了实证研究. 把企业管理实践, 信息技术和数学模型紧密结合在一起. 该实施路径也是一个全新的尝试, 在企业管理中受到了全球管理层的一致好评.

在具体实证中, 回顾了信用评级的一系列方法, 结合某企业实际情况, 选取了二元选择模型. 通过收集 SAP 运营的数据和客户的财务数据, 使用 Logistic 回归, 建立了可用于企业日常实际运营的评级过程, 并应用于企业的实际运营中, 取得了很好的效果.

研究发现, 过分复杂的模型由于模型的前提假设过于严格, 数据采集困难和本身的偏差等问题, 很难应用于实际企业实践中去. 过于简单的线性模型由于实际情况均为非线性的, 很难用简单的线性模型来模拟和预测. 如何选择复杂程度合适、易于理解的模型, 是模型选取关键.

在数据方面, 利用企业运营的数据, 可以更准确的反映实际的情况, 并建立动态的模型, 这也为大数据背景下如何准确的收集和提取信息/知识提出了新的思路.

参考文献:

[1] Andrew M, Brynjolfsson E. Big data: The management revolution. Harvard Business Review, 2012, 90(10): 61–67.

[2] 魏　伟. 银行业面对互联网金融挑战的应对策略: 基于大数据背景下的思考. 上海金融学院学报, 2014 (4): 45–51.
Wei W. Strategy for banking industry to facing challenges from internet finance: Thoughts under big data context. Journal of Shanghai Finance University, 2014 (4): 45–51. (in Chinese)

[3] 李　平, 陈　林, 李　强, 等. 互联网金融的发展与研究综述. 计算机工程与应用, 2015 (2): 245–253.
Li P, Chen L, Li Q, et al. Overview of internet finance development and research. Computer Engineering and Applications, 2015 (2): 245–253. (in Chinese)

[4] Yeh C C, Lin F y, Hsu C Y. A hybrid KMV model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems, 2012(33): 166–172.

(下转第 849 页)