

数据挖掘的流程改进和模型应用

高武奇^{1,2}, 康凤举¹, 钟联炯²

(1 西北工业大学 航海学院, 陕西 西安 710072; 2 西安工业大学 计算机科学与工程学院, 陕西 西安 710032)

摘要: 针对跨行业数据挖掘标准流程(CRISP-DM)的不足, 提出在结果部署和数据理解之间增加连线, 从而形成了数据挖掘过程的回环. 研究了模型部署的方法, 用数据库和批处理文件等技术解决了模型应用于信息系统中的平台限制. 以陕西省高校财务风险预警为例, 在 Clementine 上运用支持向量机作为挖掘算法, 按照改进的流程进行了实证. 改进的数据挖掘流程和实现方法对于数据挖掘模型的广泛使用具有推动作用.

关键词: 跨行业数据挖掘标准流程; Clementine; 财务风险; 模型应用

中图分类号: TP182

文献标识码: A

文章编号: 1000-7180(2011)07-0009-04

Process Optimization and Model Application of Data Mining

GAO Wu-qi^{1,2}, KANG Feng-ju¹, ZHONG Lian-jiong²

(1 College of Marine Engineering, Northwestern Polytechnical University, Xi'an 710072, China;

2 School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710072, China)

Abstract: This paper addresses the shortage of CRISP-DM (cross-industry standard process for data mining), propose add connection between evaluation and deployment to form the circle of data mining process. Based on the research of development method, the technologies of database and batch file is used to overcome the platform limitation of model in information system. According to the optimized process of data mining, early warning of financial risks of universities in Shaanxi province is implemented with support vector machines in Clementine. Improved data mining process and realization of the data mining model are useful to widen the data mining model.

Key words: CRISP-DM; Clementine; financial risk; model application

1 引言

文献[1]介绍了数据挖掘技术的方法——决策树方法, 并比较了逻辑回归、神经网络和决策树在我国上市公司财务困境预测问题上的优劣, 认为决策树在预测准确率、波动性及可解释性上具有综合优势.

文献[2]用 SVM 方法分析了 1998 年~2003 年台湾 22 家银行的信用评级变化, 并与用 BP 神经网络得出的结果作了比较, 结果 SVM 方法最优, 准确率达到 84.62%. 文献[3]基于改进 SVM 主动学习算法进行入侵检测, 取得了良好效果.

学者们对数据挖掘的研究多集中在数据挖掘方法的选择和具体技术优劣的比较上. 文中认为, 数据

挖掘项目的实现需要系统化的部署, 而不只是对方法本身的研究. 跨行业数据挖掘过程标准(CRISP-DM)^[4]是当今数据挖掘领域最有影响的通用标准, 其不只是对数据的组织或呈现, 也不仅是数据分析和统计建模, 而是一个从理解业务需求、寻求解决方案到接受实践检验的完整过程. 文中对标准的数据挖掘流程进行修正, 提出了一种改进的数据挖掘流程, 并以 Clementine 为平台, 按照改进的跨行业数据挖掘过程标准(CRISP-DM)来对陕西省高校的财务数据进行数据流构建, 对模型部署这个过程进行深入的研究.

2 数据挖掘流程

CRISP-DM 是一种用于指导数据挖掘工作的

方法. 该模型将一个知识发现(Knowledge Discovery in Database, KDD)工程分为 6 个不同的, 但顺序并非完全不变的阶段.

(1) 商业理解. 数据挖掘的第一步是定义目标, 首先要明确你要去度量或预测什么. 数据挖掘是面向商业应用的, 在数据挖掘的过程中, 都应该定义一个或者多个商业目标. 根据特定的目标选择或准备数据, 建立不同的模型, 同时可以评价数据挖掘的效果.

(2) 数据理解. 数据的理解以及收集, 对可用的数据进行评估.

(3) 数据准备. 对可用的原始数据进行一系列的组织和清洗, 使之达到建模需要.

(4) 建立模型. 应用数据挖掘工具建立模型.

(5) 模型评估. 重点具体考虑得出的结果是否符合第一步的商业目的.

(6) 结果部署. 部署就是使用新的深入见解在组织内部进行改善的过程. 此外, 部署还意味着可以使用从数据挖掘中获得的深入见解改善组织. 这些结果可能不会正式集成到信息系统中, 但它们对于计划和制定营销决策非常有用.

CRISP-DM 数据挖掘过程将结果部署, 形成数据挖掘报告作为一次挖掘任务的结束, 文中认为: 用于建立模型的数据一般来源于信息系统, 模型部署以后应该作用于信息系统, 从而完成对新数据的分析. 然后对利用模型产生的结果进行监视, 对于实证发现模型失效的数据, 应作为对该模型修正建立新模型的重要学习数据, 因为修正前的模型是不符合此数据, 从而实现对模型的修正, 完成真正意义的挖掘循环. 对于一个商业理解, 结果部署不是结束, 应该将部署后形成的数据, 根据数据理解过程的规范, 作为新的数据进入数据准备阶段, 完成挖掘模型的不断修正. 因此修正后的数据挖掘流程应该增加从结果部署到数据理解的连线. 改进的数据挖掘流程如图 1 所示.

3 模型部署应用的方法和原理过程

以往数据挖掘按照 CRISP-DM 规范进行, 模型部署大多以写出挖掘结果报告就视为结束, 因此对模型的应用很少有文章涉及. 文中以 Clementine13 为例, 给出可行的两种方法, 其他数据挖掘工具也可采用类似的方法实现.

(1) 挖掘模型共享的方法——PMML(也称预测模型标记语言)

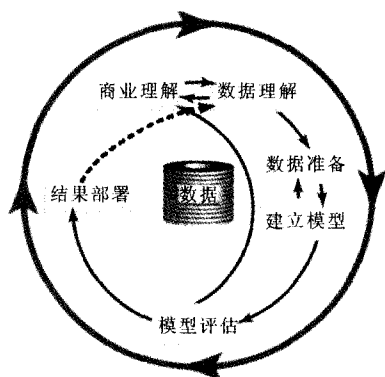


图 1 改进的数据挖掘流程

PMML 是一种 XML 格式, 用于描述数据挖掘和统计模型, 包括模型的输入、用于为数据挖掘准备数据的变换, 以及定义模型自身的参数. PASW Modeler 可导入和导出 PMML, 使得其能够与其他支持此格式的应用程序(如 PASW Statistics)共享模型^[5].

利用 PMML 作为中介, 可以将 Clementine 挖掘好的模型导出成 PMML 包, 然后导入 SQL Server2005 中, 利用 SQL Server2005 的模型部署方法将该模型引入信息系统中.

(2) 数据挖掘软件专用工具——Runtime 引擎

PASW Modeler Solution Publisher 是一个功能强大的工具, 用于将数据挖掘结果集成到业务过程中, 以解决实际问题. 使用 PASW Modeler Solution Publisher, 可以创建流的“打包”版本, 该版本的流可由外部 Runtime 引擎执行或者嵌入到外部应用程序中. 因此, 可以在生产环境中部署数据建模流以支持每天的业务过程, 以及使用挖掘数据所获得的知识增强组织决策者的能力.

(3) 两种方法的比较

与简单地导出模型(作为 PMML)相比, 使用 PASW Modeler Solution Publisher 提供了更强的性能, 因为能够通过它发布和部署完整的 PASW Modeler 流. 这就意味着, 在根据模型创建预测之前您可以执行数据准备, 也可以执行记录和字段操作(如聚合数据、选择记录或导出新字段). 然后可以进一步处理模型结果, 然后再保存这些数据, 所有这些通过执行发布的流即可完成.

(4) 模型部署原理过程

使用 PASW Modeler Solution Publisher 部署解决方案包括两个阶段:

① 发布. 当执行整个数据挖掘过程时, 最终将获得一个模型, 该模型会为业务问题提供很好的解

决方案.此时,将要采用该模型并将其应用到业务过程.发布流时,会向磁盘写入该流的详细描述(作为一个图像文件和一个参数文件).

② 执行.发布流之后,可以通过执行这个发布的流,重新创建在该流中实现的过程.此步骤是通过使用独立的 PASW Modeler Runtime (modelerrun.exe) 或通过开发一个使用 PASW Modeler Runtime Library 来执行流的应用程序完成的.要执行 PASW Modeler 外部的流(使用 Runtime 或自定义应用程序),则必须首先安装 PASW Modeler Solution Publisher Runtime.

4 模型应用的几个关键技术

(1) 用于挖掘的样本数据和模型部署阶段的输入的数据均用数据库作为数据源,这样可以通过 ODBC 连接不同的数据库,从而解决模型受数据库平台的限制.通过模型预测的结果也用数据库保存,从而可以在信息系统中以 WEB 形式查看.图 2 为 Clementine 挖掘流程,上半部分为挖掘过程,下半部分为模型的应用.为使模型脱离 Clementine 应用,可以采用导出 PMML 文件或用 Runtime 运行,根据第 3 部分的分析,这里采用效果较优的 Runtime 引擎部署模型.

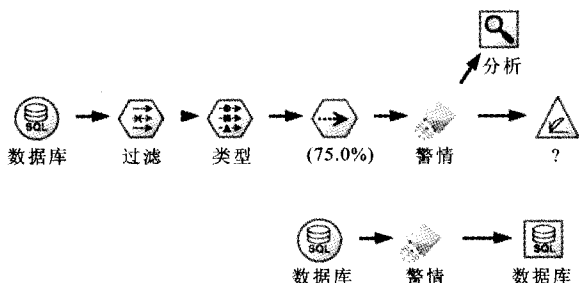


图 2 Clementine 数据挖掘流程

(2) 采用 Runtime 运行有两种方式,一种是使用 PASW Modeler Solution Publisher Runtime 程序执行使用,格式为: modelerrun -nobanner -p <parameter-file> -o <options> <image-file>. 这种方式执行时脱离不了命令行方式. 另外 CLEMRTL API 可以在应用程序中控制已发布流的执行, API 含有 C 连接,从而具有最大兼容性,但是库依赖于 C++. 在某些平台上,这可能意味着仅能使用具有可识别 C++ 连接器的 API.

(3) 对于(2)中两种方式的不足,文中采用批处理文件的方式给予很好的解决,既脱离了命令行方式,也不受 C++ 平台的限制. 其原理就是用批处理文件执行命令行内容,而所有语言都可以调用批处

理文件,从而避免了直接执行命令行方式和只能用 C++ 嵌入程序中的不足.

5 改进的数据挖掘过程实证

鉴于本次实证主要验证在结果部署和数据理解之间增加连线后的数据挖掘流程,因此对其他过程作了简化,主要由数据理解和准备→建立模型→模型评估→结果部署→数据理解和准备的回环组成.

(1) 数据理解和准备

① 确定业务对象. 数据挖掘是要从大量数据中发现有用的知识或规则,虽然挖掘的结果不能预先知道,但是要探索的问题应是可预见的. 通过前面的分析,本次数据挖掘的对象是陕西省普通高校的财务指标数据. 目的是要通过对这些指标的分析发现建立的模型能否预测高校的财务危机^[6]. 文中研究中设高校财务危机的概率为预测变量.

② 确定指标和样本. 文中选择预警指标体系由偿债能力、投资能力、运行绩效、发展潜力和非财务因素五个模块构成. 偿债能力和运行绩效是高校财务评价的两个基本部分,投资能力和发展潜力也体现了高校的财务管理水平,此外,非财务因素也尤为重要. 在具体指标的选择方面,考虑到各指标间应既相互补充,又不重复,尽可能全面综合地反映高效的财务状况,故每个预警模块各取几个具有代表性的指标.

在数据挖掘过程中样本的选取非常关键,样本的选择正确与否决定了研究结果是否有效. 在财务危机预测方面,我国研究者在选取配对样本时一般是借鉴国外的做法,即根据学校的规模选取配对样本. 目前利用 2005、2006 两年 36 所陕西省省属高校的财务指标的数据中取出指标数据.

(2) 高校财务风险的支持向量机模型

支持向量机用于对高校财务风险评估的原理是:把所选择的高校做为样本集,将代表分类结果的量作为类别属性输出. 由于所选样本集是非线性的,因此,采用非线性 SVM 分类算法. SVM 构建设置如表 1.

表 1 支持向量机构建设置

参数	构建设置
使用分区数据	假
模式	简单
停止条件	默认: 1.0E-3
核函数类型	RBF
RBF gamma	0.1

(3) 模型分析

支持向量机用于对高校财务风险评估的原理是:把所选择的高校做为样本集,将代表分类结果的量作为类别属性输出.由于所选样本集是非线性的,因此,采用非线性 SVM 分类算法.非线性 SVM 分类算法通过内积核函数的运算,建立一个最优分类超平面,根据样本离分类超平面的距离,来判断误判率.

从表 2 中可以看出,支持向量机对训练集的误判数为 1,误判率为 1.92%;对测试集误判数为 1,误判率为 5%,总样本的误判率为 2.78%;52 个训练集的分类正确率为 98.08%,20 个测试样本的分类正确率为 95%,总的 72 个样本的正确率为 97.22%.

表 2 支持向量机模型分类正确率表

分类名称	分类准确率/%		
	训练集(52)	测试集(20)	总样本(72)
误判数	1	1	2
分类正确率/%	98.08	95.00	97.22

(4) 模型部署与应用

① 模型的发布. 点选数据数据库的发布流即可生成 publish. pim(提供了 Runtime 执行发布的流时所需的所有信息), publish. par(包含有关数据来源、输出文件和执行选项的可配置信息), publish. xml(介绍图像的输入和输出以及它们的数据模型)三个文件.

② 建立批处理文件:C:\shujuwajue. bat
内容如下为 d:

cd d:\shujuwajue // d:\shujuwajue 为发布流文件和参数文件所在位置

modelerrun -nobanner -p publish. par publish. pim

③ 用 jsp 执行该批处理文件实现与信息系统的集成. 有关 jsp 运行批处理文件的代码如下:

```
<%Runtime.getRuntime().exec("c:/shujuwajue.bat");%>
```

ASP 和 VC 等软件也可采用相似的方法实现.

④ 模型应用截图如图 3 所示.

(5) 模型部署应用后对模型的两次修正

通过对部署以后模型的应用,将模型预警不准确的数据再次加入进入数据理解和准备阶段,建立模型,用相同的测试样本进行测试. 警情收益图如图 4 所示.

从三种模型分别对巨警、重警、中警、轻警和无

指标	值	指标	值
流动比率(X1)	67.97	学校偿债率(X2)	49.09
学校资金储备能力(X3)	68.74	学校资产负债率(X4)	89.11
后勤及教学资产负债率(X5)	75.11	基建投资借款支付率(X6)	71.63
学校收支平衡比(X7)	62.22	学校债务率(X8)	75.88
学校收入款的期限结构(X9)	60.00	学校年末应付账款占年末流动资产总额的比重(X10)	97.73
非自有资金动用比率(X11)	62.53	中国大学排行榜(X12)	73.54
中国大学评价(X13)	64.41		

应用模型判断警情

图 3 模型应用截图

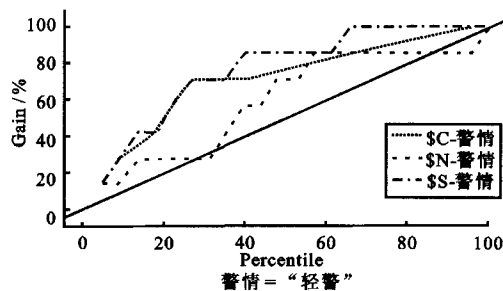


图 4 轻警收益图

警的收益图可以看出,第三次建立的模型基本陡峭地上升到 100% 然后渐渐变得平缓,预测模型的效果较为理想.说明了将模型预警不准确的数据再次加入进入数据理解和准备阶段,完成对模型的修正很有必要.

6 结束语

CRISP-DM 是一个通过工业界验证的指导数据挖掘工作的方法.该模型将一个 KDD 工程分为商业理解、数据理解、数据准备、建立模型、模型评估到结果部署 6 个阶段.文中提出在数据挖掘标准过程中增加模型部署到数据理解之间的连线,从而形成了数据挖掘过程的回环.模型应用将挖掘出的模型应用于信息系统,使挖掘模型回到新的业务流程中.模型应用以后,根据现实情况对模型的评估是对现有模型的验证和优化过程,对于已验证模型预测不正确的数据,应该及时将其加入模型训练样本,重新学习,从而实现对现有模型的优化.

以陕西省高校财务风险预警为例,在 Clementine 上运用支持向量机作为挖掘算法,按照改进的流程进行了实证,按照回环的流程对数据进行挖掘,从收益图来看,修正以后的模型预测准确率更好.改进的数据挖掘流程和实现方法对于数据挖掘模型的广泛使用具有推动作用.

参考文献:

- [1] 吴俊杰. 财务困境预测:数据挖掘方法的比较与运用[J]. 清华大学学报:哲学社会科学版, 2006, 21(1): 45-53.

(下转第 16 页)

下具有的性质: 模块层数序列 $\{M_j\} = \{j+1\} = \{\log \ell^{N_j}\}$, ($N_j = \ell^{N_j} = N_0 \ell^{j+1}$, $j \in [0, k]$) 与 $\{M_j\}$ 相对应的模块操作位数序列 $\{N_j\} = \{\ell^{j+1}\} = \{(\ell-1)S_j + \ell\}$, 序列操作位数 $n = S_K = \frac{\ell(\ell^K - 1)}{\ell - 1}$, $S_j = \sum_{i=0}^{j-1} N_i = \frac{\ell(\ell^j - 1)}{\ell - 1}$, ($j \in [0, k]$, $K = k+1$).

4 结束语

定义 $\{M_j\} = \{M_0 + j\}$ ($j \in [0, k]$, $M_0 \geq 1$ 的整数) 为混合模块 TC^2CLA 的无等待时间序列; 定义 $\{M_j\} = \{M_0 + j\} = \{j+1\}$ ($j \in [0, k]$, $M_0 = 1$) 为混合模块 TC^2CLA 无等待时间完全序列.

混合模块 TC^2CLA 无等待时间序列延迟时间 $T_{pd}^{k,c^0}(M_0 \cdots M_k, k) = (M_0 + k + 1)t_{pd}$, $T_{pd}^{k,s(max)}(M_0 \cdots M_k, k) = (2M_0 + 2k + 1)t_{pd} = (2M_k + 1)t_{pd}$ 是最小或最大模块层数 (M_0 或 M_k) 与级数 k 的函数.

混合模块 TC^2CLA 无等待时间完全序列延迟时间 $T_{pd}^{k,c^0}(k) = (k + 2)t_{pd} = (M_k + 1)t_{pd}$, $T_{pd}^{k,s(max)}(k) = (2k + 3)t_{pd} = (2M_k + 1)t_{pd}$ 仅是最大模块层数 M_k 或级数 k 的函数.

推导出混合模块 TC^2CLA 无等待时间序列和完全序列在组位数 ℓ_m 相同和不同的条件下的 15 条性质, 为混合模块 TC^2CLA 无等待时间序列和完全序列在不增加 TC^2CLA 延迟时间条件下的操作位数的扩展和优化设计提供了基础、方法和途径.

参考文献:

[1] Huong Yenmou, James B Kuo. A high-speed condi-

tional carry select (CCS) adder circuit with a Successively incremented carry number block (SICNB) structure for low-voltage VLSI implementation [J]. IEEE Transactions on Circuits and systems, 2000, 47 (10): 1074-1079.

[2] Bui Hungtien, Wang Yuke, Jiang Yingtao. Design and analysis of low-power 10-transistor full adders using novel XOR-XNOR gates [J]. IEEE Transactions on Circuits and systems, 2002, 49(1): 25-30.

[3] 贾嵩, 刘飞, 刘凌, 等. 对数跳跃加法器的算法及结构设计[J]. 电子学报, 2003, 31(8): 1186-1189.

[4] 王礼平, 王观凤. 顶层进位级联 CLA 的算法与设计规则 [J]. 华中科技大学学报: 自然科学版, 2004, 32(7): 88-91.

[5] 王元媛, 王礼平. TC^2CLA 的混合模块延迟公式及优化序列 [J]. 微电子学与计算机, 2008, 25(11): 64-67.

[6] Wang C, Lee PoMing, Lee Rongchin, et al. A 1.25GHz 32bit tree-structured carry lookahead adder [C]// IEEE Intel Synop On Circuits and Systems. Sydney, Australia: IEEE, 2001, 4: 80-83.

[7] Wang Yuke, Pai C, Song Xiaoyu. The design of hybrid carry-lookahead/carry-select adders [J]. IEEE Transactions on Circuits and systems, 2002, 49(1): 16-24.

[8] 王元媛, 王礼平. 混合模块无等待时间序列超前进位加法器设计 [J]. 微电子学与计算机, 2005, 25(12): 12-15.

作者简介:

王元媛 女, (1982-), 博士, 讲师. 研究方向为计算数学、计算技术及应用.

黄娟 女, (1978-), 博士, 讲师. 研究方向为机器学习、模式识别.

(上接第 12 页)

[2] Chen WunHwa, Shih JenYing. A study of taiwan's issuer credit rating systems using support vector machines [J]. Expert Systems with Applications, 2006, 30(3): 427-435.

[3] 黄双福, 陈贤富. 基于改进 SVM 主动学习算法的入侵检测 [J]. 微电子学与计算机, 2010, 27(3): 75-77.

[4] CRISP-DM. Cross industry standard process for data mining [EB/OL]. (2000) [2007-7-28] <http://www.Crisp-dm.org>, 2007.

[5] DMG. PMML version 3.1, in data mining group [EB/OL]. (2005) [2007-7-28] <http://www.dmg.org/index.html>, 2005.

org/index.html, 2005.

[6] 孙强. 信息安全风险评估模型的定性与定量对比研究 [J]. 微电子学与计算机, 2010, 27(6): 92-96.

作者简介:

高武奇 男, (1975-), 博士生, 讲师. 研究方向为数据挖掘、分布交互仿真和 SOA 技术.

康凤举 男, (1947-), 博士生导师, 教授. 研究方向为计算机仿真和控制系统仿真技术.

钟联炯 男, (1938-), 博士生导师, 教授. 研究方向为人工智能与系统仿真研究.