第11卷 第19期 2011年7月
1671—1815(2011)19-4639-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 11 No. 19 July 2011

数据挖掘模型选择的通用建模研究

范广玲[1] 李春生[2] 高雅田[2]

(东北石油大学数学科学与技术学院[1];计算机与信息技术学院[2],大庆 163318)

摘 要 现阶段数据挖掘模型的选择与专家的经验密切相关,有经验的专家就会选择良好的、优质的模型,使得挖掘工作高效、准确;反之,就会浪费时间,或得不到理想的结果。因此数据挖掘模型的设计和选择是挖掘工作的关键。要建立一个模型,实现挖掘目标的特征集与挖掘算法集合间的对应关系。应用该模型,用户可以得到最佳的挖掘方法,应用这种挖掘方法就可最好地实现挖掘目标。

关键词 数据挖掘模型 建模 专家经验

中图法分类号 TP231.3; **文献标志码** A

现阶段数据挖掘(Data Mining)模型的选择与专家的经验密切相关,有经验的专家就会选择良好的,优质的模型,使得挖掘工作高效、准确;反之,就会浪费时间,或得不到理想的结果,因此数据挖掘模型的设计和选择是挖掘工作的关键。

现阶段数据挖掘模型的选择是:针对某些具体的任务和领域选择一些已经成型的模型,如分类模型、回归模型、时间序列模型、聚类模型和关联规则模型[1]。传统数据挖掘系统的建立强调人工主动参与,循环测试可能有效的挖掘技术,最终得出相对可行的系统结构,导致了数据挖掘的手工化,挖掘数据处理复杂化。

本文要建立一个模型,实现挖掘目标的特征集与挖掘算法集合间的对应关系。用户只需提供如挖掘目标、数据类型等必要的挖掘信息,具体地选择哪种算法,都可通过该模型实现,最后,用户可以得到最佳的挖掘方法,应用这种挖掘方法就可最好地实现挖掘目标。

该模型实际应用在“油田开发压裂措施选井”系统,研究油田开发领域业务需求,获取油田开发压裂措施设计的业务特征及数据特征,得到了很好的效果[2]。

2011年4月6日收到

第一作者简介:范广玲(1967—),女,黑龙江省大兴安岭地区人,东北石油大学数学科学与技术学院博士研究生,研究方向:人工智能与数据挖掘。E-mail:fanguangling@126.com。

1 数据挖掘方法的建模

传统的数据挖掘技术往往一次需要处理大量数据,还可能因为低准确性而造成多次重复操作,并且需要大量的手工参与,这使得系统运行效率低,浪费了用户的时间和精力。如何设计适应性好、操作方便、扩展灵活的DM模型是各个方法论无可回避的重点内容,为具体的挖掘任务选择最佳算法配置是DM建模的重要目标。

1.1 传统的建模方法

传统的DM模型设计是一个多步骤的、循环的、非线性的处理过程,完成从源数据中发现有价值的知识的过程,可以概括为:首先,明确能够有效挖掘的数据源,并且将其组织成为适合挖掘的数据形式;然后,根据建立挖掘模型的基本常识,设计可能有效的挖掘模型,即选择相应的挖掘算法及各种算法的搭配组合来处理业务数据,初步建立挖掘模型,通过挖掘模型获得满足业务需求的知识与信息;最后,对挖掘模型进行评估,并且建立适合开发目标的挖掘系统,为应用部门部署应用,通过反馈可能进一步调整系统。这样,DM可以定义为问题分析、数据抽取、数据预处理、DM模型设计、模型评

估等基本阶段[2]。

挖掘模型设计本身也是一个往复的过程。这种方式不仅需要操作人员具有应用领域的专业知识而且需要对 DM 技术有广泛并且深刻的认识。随着 DM 技术研究的不断发展,针对各种不同的应用问题催生出各式各样的 DM 算法,如神经网络、支持向量机、粗糙集等。但这些算法有各自的假设、适用对象及目标,并需要设置各种参数,在没有足够 DM 专业知识情况下,很难确定哪些算法或者算法组合能够有效完成挖掘任务。从某种程度上,传统的挖掘模型设计方法不但降低了 DM 模型建立的效率,而且面对应用领域,专业的建模知识只掌握在少数专业工程师手中,对数据挖掘技术的推广和应用带来一定的难度。

1.2 模型化的建模方法

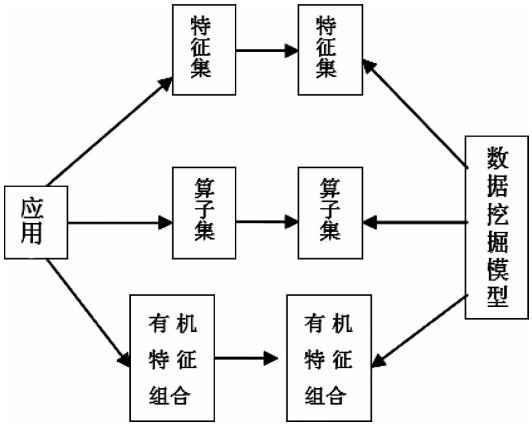
由于每一种数据挖掘技术方法都有其自身的特点和实现步骤,数据挖掘与具体应用问题的密切相关性。因此,成功应用数据挖掘技术以达到目标的过程本身就是一件很复杂的事情。

针对传统的建模方法存在的局限性,本文研究一种模型化的建模方法,其思想是将专业建模人员需要完成的分析、选择、判断等任务由程序自动完成,这样呈现给用户的就是相对简洁的过程,提供具体挖掘的任务目标、数据特征的详细描述,通过程序就可获得所需要的挖掘方法。

2 数据挖掘方法的选择

2.1 数据挖掘方法模型的描述

从数据分析角度,数据挖掘模型包括两类:描述式数据挖掘及预测式数据挖掘,描述式以简洁概要的方式描述数据,并且提高数据的有趣的一般的性质。预测式数据挖掘分析数据建立一个或者一组模型,并试图预测新数据集的行为。从功能角度而言,数据挖掘模型可以分为特征化描述和区分、关联、分类、预测、聚类和演变分析。根据数据挖掘的任务,有不同的模式,但如何发现这些模式需要依赖于具体的方法来实现。



系统结构示意图

一个实际的应用可以主要按任务特征和数据特征分,任务特征:挖掘对象、目标和挖掘类型;数据特征:数据类型、连续性、归一性、指导类型及量化性。对这些特征的具体描述就构成一个数据挖掘的实际应用的特征集合。

应用特征集:由多个集合构成,每个集合即为一类特性的具体特征,则所有集合的并集即为应用的所有特征。各个集合间属于并列关系,每个集合的各元素是互斥关系。如:挖掘对象集合:{工业,农业,林业,…},目标集合:{个体,集体,机关,…},挖掘类型集合:{分类,预测,聚类,关联规则,特征化描述和区分},数据类型集合{文本,数字},连续性集合{离散,连续},归一性集合{归一,不归一},指导类型集合{有师,无师},量化性集合{可量化,不可量化}

应用算子集:包含集合的并集、交集、子集、幂集、笛卡尔积运算。每种运算均符合规范的集合运算规则。

表 1 算子集

算子名称	运算符号	运算变量类型	运算表达式	运算结果
并集	∪	集合	$A \cup B$	$C=\{x \mid x \in A \text{ 且 } x \in B\}$
笛卡尔积	×	集合	$A \times B$	$C=\{(x,y) \mid x \in A, y \in B\}$
选择	select	集合	Select A for 条件	$A=\{x \mid x \text{ 满足条件}\}$

应用有机特征集:是应用特征集中的特征按照某些运算规则组合而生成应用特性集合,每一个集合对应一个实际的应用。生成规则:由笛卡尔积算子作用于所有应用特征集,生成多个所有特性的各种有机组合,使得实际中的应用可以对应于应用有机特征集的某一个元素。同时也可以用有条件的交集、并集和幂集运算得到相应的结果。

数据挖掘算法特征集:集合,包含数据挖掘所有的算法,同时每个挖掘算法所适用的范围和所具有的特性都可表示。

数据挖掘算子集:选择算子,交集,并集。

表 2　实际应用特征库

A(m,n)	层 1	层 2	层 3	特性 4
分类	1	3	1	0
连续	2	2	2	0
特征 3	1	1	0	1

表 3　数据挖掘算法特征库 B(s,t)

算法名称	挖掘类型	数据类型	连续性	归一性	指导类型	量化性	兼容 1	兼容 2
决策树	131	2	1	1	1	1		
线性回归	132	2	2	1	2	1		
神经网络	131	2	1	1	1	1	133	

数据挖掘算法有机组合集:是算法的分类集合。根据算法对挖掘类型和数据描述的适应性,对算法分类。根据算法是否可独立进行某类挖掘,得到针对每一类挖掘的算法集合;针对算法对数据特性的适用性,得到符合每一类数据特征的混合算法集合。例如适用于分类的{决策树,神经网络,贝叶斯分类,…},适用于预测的{多元线性回归,非线性回归,广义线性回归,…},适用于连续数据的算法集合,适用于归一类型的算法集合等等。

建立应用特征集与数据挖掘算法特征集的某种映射关系:分别建立数据挖掘算法的属性库和实际应用特征库,根据应用特征,实现每个特征与某些算法的一对一或一对多关系,同样可以实现一个算法与多个应用特征的映射关系。

建立应用有机特征集与数据挖掘算法有机组合集的某种映射关系:在应用有机特征集中每个应用都有其具体的应用特征,根据这些应用特征以及数据挖掘算法有机组合集的特征,通过相应的运算可以实现在应用有机特征集中的每一个元素对应数据挖掘算法有机组合集中的一个或多个集合的并集或交集,从而实现完成某种应用所需要的独立算法集和混合算法集的选择。

表 4　具体应用特征库

T(h,l)	应用名称	挖掘类型	数据类型	连续性	归一性	指导类型	量化性	独立	混合
	应用 1	131	2	2	1	1	2		
	应用 2	133	1	1	2	2	1		

2.2　数据挖掘方法选择的模型

2.2.1　已知

赋值变量:m,n,s,t,k;

赋值矩阵:$A(m,n)$, $B(s,t)$, $T(k,t)$;

生成矩阵:$D(n,2)$, $E(m,2)$。

释义:m:应用特征的个数(如从:挖掘对象,挖掘类型,数据类型,连续性来描述应用,则 $m=4$);n:应用特征的层数(取所有特征层数的最大值,如 $n=4$);s:数据挖掘算法的个数;t:挖掘算法所适用的数据特征的个数(如挖掘类型、数据类型、连续性、归一性等) k:具体实际应用的个数。

$A(m,n)$:代表应用特征所处应用中的位置,按所在层位置定义;

$B(s,t)$:代表具体数据算法技术所具备的应用特征,按数据描述及挖掘类型定义;

$T(k,t)$:代表实际应用所具有的特性,根据任务描述和数据描述定义;

$D(n,2)$:集合库,存放临时生成的数据挖掘算法集合;

$E(m,2)$:最终生成的独立数据挖掘算法模型库;最终生成的混合数据挖掘算法模型库。

2.2.2　计算

(1) *if*　$b_{ik}(i=1,2,\cdots,m)=t_{1k}$, 则,$d_{(k-1)1}=\{b_{i1}, i\in(1,\cdots,m)\}$,$(k=2,\cdots,l)$,所以,$e_{11}=\cap d_{k1}$,

$k=(1,\cdots,l-1)$。

即为应用独立算法。

(2) if $b_{i8}(i=1,2,\cdots,m)=t_{12}$,

则 $d_{12}=\{b_{i1},i\in(1,\cdots,m)\}$,

if $b_{ik}(i=1,2,\cdots,m)=t_{1k}$,

则 $d_{k2}=\{b_{i1},i\in(1,\cdots,m)\}.(k=2,\cdots l-1)$。

所以 $e_{12}=\cap d_{k2}\cup e_{11},k=(1,\cdots,l-1)$。即为应用混合算法。

2.2.3 生成经验库

$t_{p8}=e_{p1},t_{p9}=e_{p2}(p=1,2,\cdots,m)$。

注:在该模型中,矩阵 $A(m,n)$ 与 $B(s,t)$ 的数据为固定模式,用户只需要确定矩阵 $T(k,t)$ 即实际应用所具有的特性,即可应用该模型实现数据挖掘模型的自动选取。

2.3 数据挖掘方法选择的应用

在油田开发领域中的具体应用。油田压裂选井选层决策系统项目中,应用数据挖掘方法选择模型,首先确定矩阵 $T(1,t)$,挖掘目的(分类)、数据特点(数值型、连续型、归一、有师、可量化)、任务目标(压裂效果的评价与预测)等方面描述 DM 任务。

表 5 压裂措施应用特征矩阵 $T(h,l)$

应用名称	挖掘类型	数据类型	连续性	归一性	指导类型	量化性	独立 混合
决策系统	131	2	1	1	1	1	

其次,依赖模型化建模方法,选择挖掘方案,求得 e_{12},得到混合算法库:决策树算法结合神经网络算法。生产因素数据经过量化、归一化预处理后,转换为数值信息作为神经网络的输入向量。系统调度路径分别经过了标准 BP 网络,共轭梯度 BP 网络,和动量项惯性系数指数增长的 BP 网络。同时可生成经验库,遇到与此一致的应用即可直接选择上述方法组合进行数据挖掘。

表 6 结果示意图

方案 ID	算法类型	挖掘方案
1	直接分类算法	共轭梯度 BP 算法
2	直接分类算法	决策树方法
3	属性选择-分类算法	决策树-共轭梯度 BP 算法

通过油井压裂措施选井应用研究,设计适应的压裂措施选井系统,该系统能够正常运行,并且较好地完成压裂井选择任务,从应用角度验证了数据挖掘方法模型系统的理论基础与逻辑设计。

在数据挖掘方法模型研究中,已经建立挖掘技术的表达框架与表达形式,使挖掘特征框架更加规范化,能够支撑更广泛的挖掘模型选择。

3 结论

本文提出了一种数据挖掘模型选择的方法,根据该方法对于一个具体的数据挖掘的实际应用,通过应用分析、业务分析和数据分析给出具体应用的特征,应用本文研究的模型化的建模方法,可以根据该应用所具备的业务、应用和数据特征,自动选择一种最优的数据挖掘方法,也许是某种单一的算法,或是多个算法的组合。

参 考 文 献

1 马江洪,张文修,徐宗本. 数据挖掘与数据库知识发现:统计学的观点. 工程数学学报, 2002; 19 (1) : 1—13

2 谭锋奇, 李洪奇, 孟照旭, 等. 数据挖掘方法在石油勘探开发中的应用研究. 石油地球物理勘探, 2010; 45(01) : 85—91

(下转第 4645 页)

很多都是经过计算的,所以需要用 AdvanTrol-Pro 的 FBD 语言命令对数据进行相应地处理。

4　结论

在辽河油田蒸汽辅助重力泄油 2 号注汽站项目实施一年多来,DCS 与其他控制系统之间的通讯从未出现过问题,充分证明了 Modbus 协议的稳定性。项目实施后,工作人员的工作量大大降低,在中控室就可以监控各个车间的生产情况。Modbus 协议作为目前自控领域使用最广泛的通讯语言之一,在本项目应用过程中充分地体现了其侦错能力强、数据传输量大、实时性好等特点,实现了生产的统一监控,统一调度。

参　考　文　献

1　邱公伟. 可编程控制器网络通讯及应用. 北京:清华大学出版社,2000:78—79

2　李肇庆,韩　涛. 串行端口技术. 北京:国防工业出版社,2004:26—62

3　邓志君,梁松峰. 基于 RS485 接口 Modbus 协议的 PLC 与多机通讯. 微计算机信息,2010;26(8):107—108

The Application of Modbus Communication Technology in the Controlling System of Steam Injection Station

ZHU Yu-heng

(Huayou Company, Liaohe Oilfield, Panjin 124010, P. R. China)

[**Abstract**] Based on the communication technologies of Modbus, the hardware of centralized control system had been installed and connected and the software had been loaded and run in order to ensure the steam boilers working normally in thermal recovery oilfield of steam injection with intelligent power supply, water treatment, boiler steam generation and water-steam separator control. The result shows that the controlling system of injection boilers station, water/steam separator, water treatment and the subsystems of intelligent power supply can be integrated very well with satisfactory effect obtained.

[**Key words**] aggregation and dispersion control　logic control　Modbus communication technology　steam injection station

(上接第 4642 页)

Research on Data Mining Model of General Modeling

FAN Guang-ling[1], LI Chun-sheng[1], GAO Ya-tian[2]

(School of Mathematics Science and Technology[1]. School of Computer and Information Technology[2] Northeast Petroleum University, Daqing 163318, P. R. China)

[Abstract]:The present data mining model selection related closely with expert experience, experienced experts will choose good, quality model, making the excavation of highly efficient and accurate, Conversely, it would be a waste of time, or get ideal result, so data mining model design and choice are digging the key. A model is established to realize mining features of a target set and mining algorithm the corresponding relationship between collections. The application of the model, the user can get the best mining method, using the mining method can be best to realize mining goal.

[**Key words**] data mining modeling　the modeling　expent experience