自动数据挖掘算法①

郑盼丽, 戴牡红

(湖南大学 软件学院, 长沙 410082)

摘 要: 研究了一种基于文法引导遗传编程(GGP)的自动数据挖掘算法. 规则归纳算法是一种典型的数据分类方法. 采用文法引导的遗传编程对规则归纳算法进行改进, 从而提出了一种规则自动提取的算法. 最后结合电视购物项目, 给出了基于文法引导的遗传编程自动提取规则的实例.

关键词: 数据挖掘; 分类; 规则归纳; 遗传编程; 文法

Automatic Data Mining Algorithms

ZHENG Pan-Li, DAI Mu-Hong

(School of Software, Hunan University, Changsha 410082, China)

Abstract: An automatic data mining algorithms based on grammar guided genetic programming(GGP) is studied. The rule induction algorithm is a typical data classification method. The grammar guided genetic programming is used to improve the rule induction algorithm, and then proposed an algorithm of automatic extraction rules. Finally, associated with the TV shopping programs, and gives examples of automatic extraction rules based on the grammar of genetic programming.

Key words: data mining; classification; rule induction; genetic programming; grammar

近年来数据挖掘技术引起了信息产业界的极大关注, 其主要原因是随着数据库技术和计算机网络的广泛应用, 加上使用先进的自动数据生成和采集工具, 人们所拥有的数据量急剧增大. 激增的数据背后隐藏了许多重要的信息, 如何从大量的数据中提取并找到有用的信息以指导决策, 是迫切需要解决的问题. 通过研究数据挖掘技术, 为决策者提供了重要的、极有价值的信息或知识, 带来不可估量的效益[1]. 其主要表现在它为大量数据的利用提供了有效工具, 将数据坟墓转换为知识“宝藏”. 规则归纳挖掘是数据挖掘研究中的一个重要分支, 是与大多数人想象的数据挖掘过程最为相似的一种数据挖掘形式, 即在大型数据库中“淘”金[2]. 随着大量数据不停地收集和存储, 许多业界人士对于从他们的数据库中挖掘出有兴趣的规则越来越感兴趣. 在此基础上, 本文提出了一种基于文法引导的遗传编程[3]对规则进行自动挖掘的算法. 最后, 结合电视购物项目, 给出了基于文法引导的遗传编程对规则进行自动挖掘的实例.

1 相关技术和理论

1.1 数据挖掘

数据挖掘是从大量的数据中挖掘出有用的信息, 即从大量的、不完全的、有噪音的、模糊的、随机的实际应用数据中发现隐含的、规律性的、人们事先未知的信息, 但又是潜在有用的并且最终可理解的信息和知识的非平凡过程[4]. 数据挖掘是从一个新的角度将数据库技术、机器学习、统计学等领域结合起来, 从更深层次发掘存在于数据内部的、有效的、新颖的、具有潜在效应的乃至最终可理解的模式. 数据挖掘能预测未来趋势和行为, 使商务活动具有前瞻性, 有助于企业做出基于知识驱动的决策.

① 收稿时间:2012-03-08;收到修改稿时间:2012-05-04

1.2 遗传编程

遗传编程(GP), 是一种从生物进化过程得到灵感的自动化生成和选择计算机程序来完成用户定义的任务的技术. 由 Stanford 大学的 Koza[5]教授在 20 世纪 90 年代提出.

遗传编程开始于一群由随机生成的千百个计算机程序组成的“人群”, 然后根据一个程序完成给定的任务的能力来确定某个程序的适应度, 应用达尔文的自然选择(适者生存)确定胜出的程序, 计算机程序间也模拟两性组合, 变异, 基因复制, 基因删除等代代进化, 直到达到预先确定的某个中止条件为止.

GP 比遗传算法(GA)适用的范围更广. GP 是对 GA 的一次突破性发展, 是利用自然进化的原理进行程序的进化. 它与 GA 最大的区别在于其个体是独立的可执行的程序而不是 GA 中固定长度的二进制字符串[6], 克服了传统 GA 在表示方法上的局限, 采用了更为灵活的可变分层结构. 根据对问题的求解要求, GP 采用分层的描述方法, 自动生成解决问题的程序, 因此, 它是一种不局限于某一领域的“遗传或进化搜索”技术.

1.3 文法模型简介

由于 GP 自身的局限性, 限制了它只能应用于非限定类型的问题领域. 但是实际上有许多的问题是属于限定类型的问题域, 比如数据挖掘中的决策树要求其属性符合某些特定的类型[7], 这样 GP 解决这类问题的过程当中会产生许多无效的个体树, 经过交叉等遗传操作后, 无效的树会在整个进化过程中一直存在, 这无疑就增加了找到正确解的难度.

解决这个问题的一个方法是用正式的文法来限制遗传编程种群中的个体树的生成, 这就是基于文法引导的遗传编程. 到目前为止已经可用于遗传编程中的文法主要有五种: 上下文无关文法(CFG)、属性文法(AG)、确定子句文法(DCG)、确定子句转换文法(DCG)和随机文法(SG). 在本文中主要使用的是 CFG.

1.4 上下文无关文法

在计算机科学中, 可以看作是由一个四元组构成的, 可以表示为 G={N,T,P,S},这里 N 是一个非终止符集, T 是一个终止符集, P 是一个产生规则集, S 是一个开始符号集, 其中集合 N 和集合 T 的交集为空集, S 是 N 的子集. 产生规则集合的元素的形式如 x→y 所示, 这里 x 属于集合 N, y 属于集合 N∪T, 产生规则规定了非终止符结点如何产生它的分支, 直到表达式最终只有终止符结点.

BNF(巴克斯-诺尔范式)经常用来表达上下文无关文法. 在双引号中的字代表着这些字符本身, 而 double_quote 用来代表双引号. 在双引号外的字代表着文法部分. 尖括号(<>)内包含的为必选项. 方括号([])内包含的为可选项. 大括号({})内包含的为可重复 0 至无数次的项. 竖线(|)表示在其左右两边任选一项, 相当于“OR”的意思. ∷=是“被定义为”的意思. 下面的例子描述的就是为了表示数学表达式 x+2 而定义的文法模型, 及该表达式通过自顶向下和自左向右的顺序产生的派生树. 如图 1 所示:

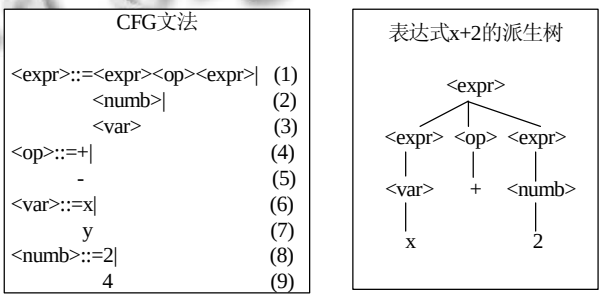

图 1 表达式 x+2 的文法模型以及派生树

1.5 规则归纳

规则归纳可以描述如下: 规则归纳是数据挖掘的一种主要形式, 可用于无监督学习系统的知识发现, 也可用来预测. 它是一种由一连串的「如果…/则…(If / Then)」的逻辑规则对数据进行细分的技术.

本文选择规则归纳算法的原因为: 一般进化思想常用于自动数据挖掘算法. 然而, 为了使自动数据挖掘算法易于处理, 我们不得不考虑一种新型的数据挖掘算法. 本文中我们选择规则归纳算法, 主要是因为该算法采用易于理解的 if/then 语句来发现知识. 其中, if 语块主要描述的是条件部分, then 语块主要描述的是预测部分.

2 规则归纳算法的改进

本文采用文法引导的遗传编程对规则归纳算法进行改进, 提出了一种规则自动挖掘算法. 该算法的总体流程如图 2 所示.

在图 2 中, 首先, 根据文法随机产生个体, 并且表示树形结构, 这些个体作用于训练集中的数据; 接着根据一些约束条件对种群初始化, 防止出现无效个体

树；然后根据适应度函数对种群中的个体进行评估，评估后的个体按照一定的复制、变异和交叉比例，最终形成新的一代；如果满足终止条件最后返回最好的个体用于预测测试集中的新的数据.

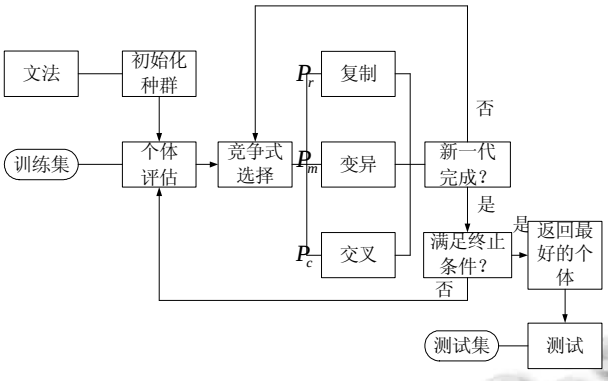

图 2　算法总体流程

该算法主要包括以下几个部分:

(1)个体的表示方法. 本文中的个体是由一组上下文无关文法产生，并且表示成树形结构. 该上下文无关文法如表 1 所示:

表 1　上下文无关文法

序号	规则
1	S → If-S
2	If-S → if Exp|Exp and ... then Stmt

其中，Exp 表示条件部分，Stmt 表示预测部分. 在规则归纳算法中，规则的前件可以包含一个或多个条件，即 if 部分可以包含一个 Exp 或多个 Exp，而后件一般只包含一种情况，即 then 部分只包含一个 Stmt.

(2)个体树的评估. 适应度是用来评估个体树在优化计算中有可能达到或接近于或有助于找到最优解的优良程度. 适应度的高低决定了个体遗传到下一代的概率的大小. 度量适应度的函数被称为适应度函数. 本文中个体树的评估过程，如图 3 所示。

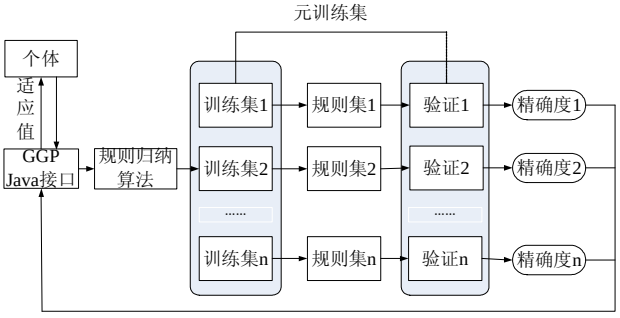

图 3　个体树的评估过程

其中，种群中每个个体树被编译为 java 代码. 规则归纳算法对元训练集中的每个训练集进行规则提取，并且通过有效性判断，得到有效的规则集及相应的精确度. 适应度函数可以通过精确度直接计算出来.

(3)选择和复制操作. 通常这两种操作都是以种群中个体树的适应度为基础，即个体的适应度越大，他被选择的概率就越大.

(4)交叉[8]操作. 进行交叉操作前，先根据种群中个体树的适应度的大小随机选择两棵树，然后分别随机从选择出来的个体树上找出一个非终端结点，这里要注意的是被选择的结点一定要是一样的，然后再将这两个结点下面的子树进行交换操作. 交叉操作如图 4 所示。

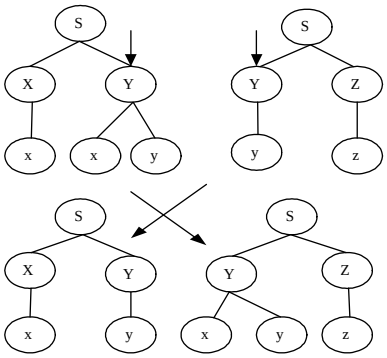

图 4　交叉操作

(5)变异操作: 变异操作是发生在一棵个体树中随机选择的一个内部结点上的，我们将这个结点叫做变异结点. 发生变异操作时，如果变异结点是非终端结点，就保留该结点，同时将以它为根结点的子树删掉，然后根据文法规则在变异结点重新生成子树；如果变异结点是终端结点，就删除该终端结点，用随机产生的另一个终端结点来代替它. 变异操作如图 5 所示:

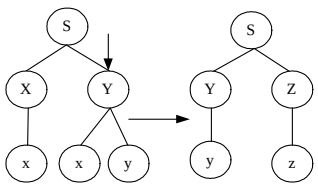

图 5　变异操作

(6)重要参数. 基于语法的遗传编程的重要参数包括: 种群的规模 M、树的最大深度 D、交叉概率 P_c、变异概率 P_m、最大进化次数 T.

(7)该算法可以描述如下:

输入: 训练样本集

输出: 适应度最高的个体

1)设置参数: 主要是确定训练参数, 如最大进化代数 T, 复制概率 p_r, 交叉概率 P_c, 变异概率 P_m 和种群大小 M 等.

2)进化代数 T=1, 根据文法随机产生第一代个体, 初始化种群.

3)读入训练样本集, 对训练样本进行规则归纳.

4)使用适应度评价种群中的各个个体.

5)判断个体的适应度是否达到指定的标准, 是则选择其中适应度最大的个体输出, 结束训练.

6)分别按照复制概率 p_r、交叉概率 P_c 和变异概率 P_m 对个体进行复制操作、交叉操作和变异操作.

7)进化代数 T=T+1.

8)判断是否达到规定代数, 否则转到 4), 是则选择适应度最高的个体输出, 结束训练.

3 应用实例

电视购物是时下比较流行的购物方式之一, 它不仅为人们提供了快捷、方便的购物服务, 而且也为商家提供了更好的销售渠道. 下面收集了通过电视购物渠道消费的顾客资料数据. 根据顾客是否购买计算机对其分类. 利用本文提出的数据挖掘方法对其进行分析, 寻找最有可能购买计算机的顾客, 即信用度比较高的顾客. 顾客的属性(字段)如表 2 所示:

表 2 顾客属性表

字段名	数据类型	说明
id	int	顾客 ID
name	varchar	姓名
age	int	年龄
income	int	收入
credit rating	varchar	信用度

表 2 中, 由于顾客 ID 与我们分析的问题无关, 所以在将数据处理的过程中将略去对顾客 ID 的描述. 其中, 年龄属性值分别为<=30、30~40、>40, 收入属性值分别为高、中、低, 信用度属性值分别为优秀、良好、一般、差.

训练集中数据如图 3 所示:

3.1 参数设置

本例中初始种群的规模 M 选定为 100, 树的最大深度 D 设为 4, 交叉概率 P_c 为 0.8~0.9, 变异概率 P_m 为 0.02~0.06, 最大进化次数 T 为 100.

表 3 数据表

name	age	income	credit rating
张三	<=30	低	差
李四	<=30	中	一般
王五	30~40	高	优秀
陈六	>40	中	良好
……	……	……	……

3.2 规则发现与结果分析

将以上的算法作用于测试集, 我们可以得到下面的规则.

规则 1: if age(<=30) and income(低) then credit rating(差)

规则 2: if age(30~40) and income(高) then credit rating(优秀)

规则 3: if age(>40) and income(中) then credit rating(良好)

……

设 P_g 为规则对测试集覆盖的程度, 则得到表 4:

表 4 规则覆盖情况

规则	P_g
若年龄<=30 且收入为低, 则信用度为差	95%
若年龄 30~40 且收入为高, 则信用度为优秀	98%
若年龄>40 且收入为中, 则信用度为良好	96%
……	……

通过表 4 可以看出, 该算法还是很有效的. 挖掘出了顾客的各个属性与顾客信用度的关系, 为商家挖掘出感兴趣的的顾客提供了有力的帮助.

4 小结

本文采用文法引导的遗传编程对数据挖掘中的规则进行自动提取, 并结合电视购物项目中顾客信用度的调查实例对该算法进行了验证. 实验结果表明, 该算法切实可行, 能够提取数据库中隐含的且有实用价值的信息.

参考文献

1 宋小娜,朱齐亮.遗传算法在数据挖掘中的应用.科技信息, 2008,17:65–66.

2 马昕,孙优贤.由规则归纳系统中发掘感兴趣模式.计算机应

(下转第 193 页)

持 0.48 左右的归一化寿命后网络规模达到阈值(即 10).

3 结论

定位技术是 WSN 必不可少的一项关键技术, 其提供的位置信息为事件监测或目标位置信息获取、路由协议、覆盖质量及其他相关研究起到关键性作用. 然而, 节点的定位信息一旦被非法滥用, 必将导致严重位置隐私问题.

但对于位置隐私问题的以往研究都是假设窃听者不能监控整个网络. 相对于良好协调、严重的攻击, 这个假设是无效的. 本文假定存在全局窃听者, 通过 LP 框架, 提出并形式化 FS 和 BF 这两种 sink 隐藏方法. 我们分析并比较了两种方法对网络寿命的影响. 研究结果表明: 保护 sink 不可观测对网络寿命的影响相当大. 同时也表明: BF 实现的网络寿命数量级高于大型网络的 FS.

参考文献

1 姚剑波.基于定向贪心游走的 WMSN 位置隐私.计算机应用与软件,2011,28(3):137−138,165.

2 刘昭斌,刘文芝,顾君忠.位置感知的自适应隐私保护策略.计算机工程与设计,2011,32(3):839−841,1032.

3 陈娟,方滨兴,殷丽华,等.传感器网络中基于源节点有限洪泛的源位置隐私保护协议.计算机学报,2010,33(9):1736−1747.

4 彭志宇,李善平.移动环境下 LBS 位置隐私保护.电子与信息学报,2011,33(5):1211−1216.

5 任丹丹,杜素果.一种基于攻击树的 VANET 位置隐私安全风险评估的新方法.计算机应用研究,2011,28(2):728−732.

6 Yang Y, Shao M, Zhu S, et a1. Towards event source unobservability with minimum network traffic in sensor networks. Proc. of the ACM Wisec'08, USA: Alexandria and VA,2008.77−88.

7 Cheng Z, Perillo M, Heinzelman W. General network lifetime and cost models for evaluating sensor network deployment strategies. IEEE Trans. on Mobile Computing, 2008,7(4): 484−497.

8 Brook A, Kendrick D, Meeraus A, et a1. GAMS: A User's Guide. The Scientific Press, 1998.

(上接第 211 页)

运行, 从而提高了网络的稳定性和可靠性. 该实现机制在大型企业网和运营商网络中具有广泛应用.

参考文献

1 孟华.互联网路由技术及发展前景展望.中国新技术新产品,2011,19(15):19−20.

2 王勐.核心路由器高可用性研究.信息与电脑,2011,23(9):151−152.

3 Coltun R, Ferguson D, Moy J. OSPF for IPv6. IETF RFC2740, 1999.

4 Moy J, Pillay-Esnault P, Lindem A. Graceful OSPF Restart. IETF RFC3623, 2003.

5 Pillay-Esnault P, Lindem A. OSPFv3 Graceful Restart. IETF RFC5187, 2008.

6 孙作聪,王立松,顾宝根.基于 OSPF 的温和重启的触发机制的研究与实现.计算机工程与技术,2006,27(14):2653− 2656.

7 张丹,商云飞,张显峰.基于 OSPF 协议的 Graceful Restart 技术的研究与实现.仪器仪表用户,2007,14(6):21−22.

(上接第 221 页)

用,2003,23(4):26−28.

3 Ratle A, Sebag M. Genetic Programming and Domain Knowledge: Beyond the Limitations of Grammar-Guided Machine Discovery. Parallel Problem Solving form Nature(PPSN 2000), Berlin: Springer, 2000,211−220.

4 朱彦廷.基于遗传算法的关联规则挖掘.西昌学院学报,2010,24(3):60−62.

5 KoZa, John R, Keane MA, Yu J, Bennett FH, Mydlowee W. Automatic creation of human-competitive programs and controllersby means of genetie programming. Genetlc Prograrnming and Evolvable Maehines, 2000,l(1-2):124−164.

6 徐哲,白焰.遗传编程.自动化仪表,2002,23(10):1−6.

7 徐扬,任庆生,戚飞虎.一个基于遗传编程的机器人足球系统.计算机仿真,2005,22(4):178−182.

8 江海燕.关于 GP 的研究与探索.山东教育学院学报,2007,1:96−100.