

数据挖掘遗传算法在最优客户分析中的应用

段素花, 揭金良, 梅占勇

(成都理工大学 信息工程学院, 成都 610059)

摘要: 基于遗传算法的数据挖掘技术在求解最优客户方面具有较强的处理能力。阐述数据挖掘和遗传算法的含义, 给出遗传算法的实现流程, 详细论述基于数据挖掘的遗传算法在最优客户分析中的应用。

关键词: 数据挖掘; 遗传算法; 最优客户群; 应用

中图分类号: TP39

文献标识码: A

Application of genetic algorithm of data mining to analysis of optimal customer

DUAN Su-hua, JIE Jin-liang, MEI Zhan-yong

(College of Information Engineering, Chengdu University of Technology, Chengdu 610059, China)

Abstract: College of Information Engineering, Data mining technology based on genetic algorithm had a strong ability in search of the optimal clients. It was explained the meaning of data mining and genetic algorithm, gave implementation process of genetic algorithm, detailed discussed the application in analysis of optimal customer.

Key words: data mining; genetic algorithm; optimal customer base; application

计算机网络与数据库技术的迅速发展和广泛应用导致各单位拥有的数据量与日俱增。而传统的数据库管理系统 (DBMS) 仅提供数据的存取和查询, 如何从信息海洋中提取有价值的知识。进一步提高信息的利用率, 从而为企业的业务决策和战略发展服务, 使一门新兴的自动信息提取技术—数据挖掘逐渐形成了一个热门研究方向。

1 数据挖掘中遗传算法概述

数据挖掘 (DataMining) 是从大量的、不完全的、有噪声的、模糊的、随机的数据中, 提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程, 提取的知识表示为概念、规则、规律和模式等形式。数据挖掘中的常用方法:

神经网络方法是模拟人脑神经结构, 以MP模型和HEBB学习规则为基础, 主要有3种神经网络模型: (1) 前馈式网络; (2) 反馈式网络; (3) 自组织网络。神经网络的知识体现在网络连接的权值上, 是一个分布式矩阵结构; 神经网络的学习体现在神经网络权值的逐步计算 (包括反复迭代

或累加) 上。

遗传算法是模拟生物进化过程的算法, 由3个基本算子 (或过程) 组成。(1) 繁殖 (选择): 即从一个旧种群 (父代) 选出生命力强的个体, 产生新的种群 (后代) 的过程; (2) 交叉 (重组): 即选择2个不同个体 (染色体) 的部分基因进行交换, 形成新个体的过程; (3) 变异 (突变): 即对某些个体的某些基因进行变异 (K 变 C 或 C 变 K), 形成新个体的过程。这种遗传算法可起到产生优良后代的作用。这些后代需满足适应值, 经过若干代的遗传, 将得到满足要求的后代 (即问题的解)。遗传算法已在优化计算和分类机器学习方面发挥着显著的作用。

决策树方法是利用信息论中的互信息 (信息增益) 寻找数据库中具有最大信息量的属性字段, 建立决策树的一个结点, 再根据该属性字段的不同取值建立树的分支, 在每个分支子集中重复建立树的下层结点和分支的过程。在数据挖掘和知识发现中应用人工智能技术, 邻近搜索方法集合论的粗集方法、规则推理、模糊逻辑和公式发现等。

遗传算法将问题域中的可能解看作是群体的一个个体或染色体, 并将每一个体编码成符号串

形式,模拟达尔文遗传选择和自然淘汰的生物进化过程,对群体反复进行基于遗传学的操作(遗传,交叉和变异),根据预定的目标适应度函数对每个个体进行评价,依据适者生存,优胜劣汰的进化规则,不断得到更优的群体,同时以全局并行搜索方式来搜索优化群体中的最优个体,求得满足要求的最优解。

基本遗传算法(Simple Genetic Algorithms, 简称SGA, 又称简单遗传算法或标准遗传算法), SGA算法由编码(产生初始种群)、适应度函数、遗传算子(选择、交叉、变异)、运行参数组成。SGA采用二进制编码方式产生初始种群,遗传操作采用单点交叉,基本位变异,选择算子采用轮盘赌选择方法。SGA的运行参数有:种群规模 M 、遗传运算的终止进化代数 T 、交叉概率 $P(c)$ 和变异概率 $P(m)$ 。这4个参数影响了遗传算法的优化性能, Schaffer建议的最优参数范围是: $M=20 \sim 100$ 、 $T=100 \sim 500$ 、 $P(c)=0.4 \sim 0.9$ 、 $P(m)=0.001 \sim 0.01$ 。

遗传算法的优化处理性能较强,提供了一种求解复杂系统优化问题的通用框架,不依赖于问题的具体领域,广泛应用于:函数优化、组合优化、自动控制和生产调度问题等领域。还应用于数据挖掘技术中规则开采,从大型数据库或数据仓库中提取隐含的、未知的非凡的及有潜在应用价值的信息或模式;由于遗传算法的特点,遗传算法可用于数据挖掘中。

2 最优客户分析

随着互联网的发展和信息的急剧增长,企业的竞争需要以更低成本,更快的速度做出及时而准确的决策。面临这样的世界态势,一个组织或企业要想有活力,并取得丰厚的效益,就要获得有效的数据信息,并让他们利用更加得心应手,顺应决策者逻辑思维和他的主导作用的工具做出正确的决策。这些有效的信息既是企业根据其可以获得更大的利润,得到更好的发展,顺应时势的重中之重。

这些有效信息中,最优客户群是企业获得最大利润的来源。

企业的最优客户群,是指企业可以从中获取

最大利润的,且有潜在更大利润的客户群。

最优客户群分析:可从客户的收入水平、家庭人口、消费层次等特征来识别最优客户群。

3 基于遗传算法的预测最优客户分析

根据客户数据条件,可以从3个方面定义客户类型:客户的消费层次、收入水平、家庭人口状况。其中环境适应函数是客户积累购买商品的金额乘以利润率并减去手续消费表示。

(1) 初始种群的产生:按照客户消费层次高、一般、低3个基因,按照家庭收入的很高、高、中、低、很低设计5个基因,按照人口的多、一般、少设计3个基因。将这11个染色体基因转换成二进制数,即进行遗传编码,产生随机生物个体,按照种群规模 M 产生一群随机的生物个体。然后计算这些生物个体的环境适应性,并利用遗传算法的竞争、杂交算法进行生物的繁殖、杂交、变异的处理。

将所有的生物个体按顺序排放在一起,按照就近原则进行繁衍。

(2) 竞争复制:将每个生物与其隔壁邻居进行比较,考察其环境适应度,如果隔壁邻居的环境适应度大,就用隔壁邻居的遗传信息替代自己的遗传特性;否则就使用自己的遗传特性。最后适应度大的生物个体进行了复制,即竞争优胜劣汰的结果。

(3) 杂交繁殖:对竞争后的生物个体随机交叉繁殖,按照已定的交叉概率,交换部分染色体基因信息,完成全部生物个体的杂交繁殖。

(4) 变异:对新生代个体基因信息随机变异,按照已定的变异概率随机选择生物个体,随机改变基因信息的某一染色体位置值。

改变新生代个体顺序,重复(2)、(3)、(4)过程,即进行下一代的竞争、杂交、变异。再进行了 T 代的进化后,即可得到优种的生物群。也就是经过了生物的繁衍后,生物群逐渐收敛于收益最大的客户特征群,即最优客户群。

以上是基于3个方面、11个染色体进行了遗传迭代。可以在寻找与客户消费有关的其它方面,按照分类组成染色体基因进行生物繁衍,使生物群最终收敛于收益最大的客户特征群,即最优客户群。

基于 MAX267 带通滤波器的设计

李 莉

(青岛四方车辆研究所有限公司 半主动悬挂项目部, 青岛 266031)

摘 要: 利用 MAX267 滤波芯片滤掉半主动悬挂系统采集信号中的干扰噪声和车辆过曲线时的离心加速度。介绍 MAX267 滤波器芯片的使用方法, 说明如何确定芯片 F 与 Q 系列引脚值。在理论计算的基础上, 利用将滤波芯片级联的方法设计出中心频率较小、带宽较窄的 8 阶契比雪夫带通滤波器。实验结果证明该滤波器的滤波效果是可以达到设计要求的。

关键词: 带通滤波; 中心频率; 品质因数; 设计

中图分类号: TP39 : U270.331 **文献标识码:** A

Design of band pass filter based on MAX267

Lili

(Semi-active Suspension, Qingdao Sifang Rolling Stock Research Institute Co., Ltd. Qingdao 266031, China)

Abstract: Filter chip MAX267 was used in order to filter out the interference noise and centrifugal acceleration from the collected signals of the Semi-active Suspension System. The using method for the MAX267 filter chip was described. And how to determine the value of the F and Q was explained. Based on theoretical calculation chebyshev band pass filter was designed used the method of chip cascade. The characteristic of the band pass filter was that the center frequency was smaller and the bandwidth was narrower. What's more, the test had verified that the filter could meet the design demand.

Key words: band pass filter; center frequency; quality factor; design

半主动悬挂系统, 采集信号进入控制装置后, 首先需要对其进行滤波处理。必须过滤干扰噪声信号及车辆在曲线运行时产生的离心加速度信号,

收稿日期: 2009-04-20

作者简介: 李 莉, 工程师。

才能得到对系统有用的横向振动加速度信号。根据《铁道机车动力学性能试验鉴定方法及评定标准》TB/T 2360-1993, 以总加权加速度均方根值作为评价平稳性指标, 选用的低通截止频率为 80 Hz。车辆在曲线上运行时, 车体横向加速度为

4 结束语

随着数据仓库的出现, 其面向主题集成、不可更新、随时间变化等特性可以更好地支持决策分析的数据集合。传统数据库中的数据用以支持事务处理, 属操作型数据而数据仓库中的数据是用以支持分析处理的分析型数据。数据仓库正是为了进行快速而有效地决策分析产生的。数据仓库的出现, 原有的数据库工具已无法满足需求, 数据挖掘正好可以帮助决策者在数据仓库中寻找数据间的潜在关联, 发现被忽略的要素, 提取隐藏其中的信息以便进行分类和优化。辅助决策者进行趋势预测及行为决策。通过运用数据挖掘的模型和技术分析最优客户群进行, 其结果是管理层做决策的基础, 也有利于企业的经营目标的实现。但是

最优客户群分析的前提是具有以客户为主题的数据仓库, 用以支持大量的数据存储, 以便能更准确的算出与最优客户群相关的某些条件和方面。

参考文献:

- [1] 康小东. 基于数据仓库的数据挖掘技术[M]. 北京: 机械工业出版社, 2004.
- [2] 陈京民. 数据仓库原理、设计与应用[M]. 北京: 中国水利水电出版社, 2004.
- [3] 陈 安, 陈 宁, 周龙骧. 数据挖掘技术及应用[M]. 北京: 科学出版社, 2006.
- [4] 范 明, 范宏建. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2006.
- [5] 武 森. 数据仓库与数据挖掘[M]. 北京: 冶金出版社, 2003, 9: 281-325.