# 数据流的预测与分类研究*)

刘耀宗 王 湛 张 宏 刘凤玉

(南京理工大学计算机科学与技术学院 南京 210094)

**摘 要** 数据流的预测和分类技术在网络入侵发现、系统性能分析等应用中具有重要的应用。作者对近年来有关数据流预测和分类的进展做了总结,并提出了一个数据流的预测和分类的通用模型,可用于系统性能的实时预测与异常检测。

**关键词** 数据流,预测与分析,性能分析

## Research on Prediction and Classification over Data Streams

LIU Yao-Zong WANG Zhan ZHANG Hong LIU Feng-Yu

(Department of Computer Science, Nanjing University of S&T, Nanjing 210094)

**Abstract** Prediction and classification over data streams for actionable insights has become an important and challenging task for a wide field of applications including network intrusion detection, system performance analysis etc. The author emphasizes the improvements of prediction and classification over data streams in recent years, then a data stream of prediction and classification normal model is proposed. It can be used in real-time prediction and outlier detection of system performance.

**Keywords** Data streams, Prediction and classification, Performance analysis

## 1 引言

近年来,在传感器网络、系统性能保持等数据流应用领域中浮现出许多新的预测和分类的需求。数据流中蕴涵大量信息,可以用来作为智能决策的依据。预测和分类是两种基本数据分析形式,可以用于预测未来的数据趋势或提取描述重要数据类的模型,一种典型应用是高可靠性计算系统的稳定性评价。由于影响复杂计算系统稳定性的因素极为复杂,采集的各种度量计算系统的性能参数属于大规模非线性时间序列数据流,很难在时变的参数流上预测稳定性级别。如果能够自适应评价参数的变化,动态调整分类模型的结构,就能够提高预测精度,实现系统性能的在线检测和量化,得到最优的系统调控依据。对系统性能数据流进行预测和分类的主要目的是监测和判断系统老化现象的存在,对观察到的实际系统运行的参数变化数值进行量化,估计计算系统性能衰退趋势的变化率。

## 2 数据流挖掘算法

目前数据流挖掘算法研究包括计算数据流的典型趋势、决策树[1]、预测[2]、k 中值聚类[3]、最近邻居查询[4]、回归分析[5]、相似性检测、模式匹配、传感器数据挖掘的预测[6]等。数据挖掘领域中将未来连续值本身取值的估计称为预测,而将预测未知连续值或离散值所属的类别称为分类。在数据流挖掘研究领域,通常预测即是指分类,两个概念是等同的。在本文中特别地对它们加以区分,明确地将研究分为两个方向:数据流值的预测和数据流数据的分类预测。

### 2.1 数据流的预测研究进展

目前数据流研究领域中已有的趋势分析理论和方法一般关注相似性、异常或模式差异的预测,如文[7]在基于异常模式求取的基础上,提出了利用回归分析中最小二乘法进行异常模式趋势监测方法。有关预测数据流值本身的文献较少,文[8]以车辆跟踪信息、电力负荷、网络流量数据为例,提出采用 Kalman 滤波对变化的数据流值进行预测的方法,在文[9]的基础上降低通讯代价;文[10]提出使用回归分析和插值技术预测未来 $w$ 步长的瞬时数据流值。文[11]提出了一种数据流上未来值的连续查询,称为连续预测查询,采用数理统计的方法给出了带有 AVG 聚集函数的连续预测聚集查询。

人工智能方法适合预测数据流值中周期性稳定成分,其预测精度虽高但速度较慢。而回归分析预测法速度快,却难以预测随机变化的非线性成分,预测精度较低[17]。它们应用于在线预测时都存在缺陷。尽我们所知,无论在传统的时间序列领域还是在新浮现的数据流领域,目前已经提出的预测方法的预测步长都是固定的。由于多数应用只需要获知固定间隔的预测值,这种等间隔的预测策略固然有其存在的必要性,然而其缺陷是无法适应流值在不同时段其波动情况不相同的特点,不适合在线自适应资源管理和最优决策的需要。

由于数据流也被看作时间序列,来自于控制理论的一些思想越来越多地被引入到近似计算[12]和时间序列数据流预测[13]研究领域。文[14]采用了不同的方法预测小时级别和 5min 级别的负荷流值,采用了等间隔的多项式插值技术,但是无法适应负荷流的变化。

文[15]通过小波分解将某些非平稳时间序列分解成多层近似意义上的平稳时间序列,然后采用自回归模型对分解后的时间序列进行预测,最终得到原始时间序列的预测值。然

*)国家自然科学基金(No. 60273035)资助。刘耀宗 博士生,主要研究方向:数据流挖掘;王 湛 博士生,主要研究方向:系统性能保持;张宏 教授,博导,主要研究领域:数据挖掘与软件工程;刘凤玉 教授,博导,主要研究领域:信息安全与人工智能。

而，该文没有讨论如何将本文的方法普遍应用于非平稳时间序列的预测，尚须理论上和算法上的进一步研究。文[16]提出了一种基于多分辨率小波分解与重构的预测方法，实现了对分钟级的交通流量的精确预测。然而文中没有分析这种算法的时空复杂度，无法得知是否适合在线预测。时间序列数据流研究具有许多传统时间序列研究未曾涉及的新内容，比如资源受限、只能一遍扫描、数据项具有易失性、在线式计算等等，需要研究新的理论和方法。

### 2.2 数据流的分类研究进展

数据流数据的分类是机器学习和统计模式识别的理论与方法在数据流领域的具体应用。在技术体系上，数据流分类仍然采用对静态数据进行处理和分析的思路。但在具体的实现环节和操作过程中，由于流式数据源和分类结果等方面具有的特殊性，分类过程变得更加复杂。典型的数据流分类处理可分为分类预处理、分类判别和分类后处理三个阶段，如图1所示。

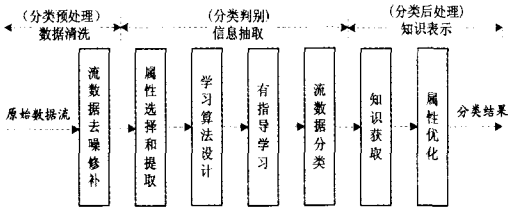


图1 数据流分类的基本过程

分类预处理是数据流分类过程中必不可少的阶段，而且预处理结果的好坏、性能优劣直接关系到分类的精度和处理速度。理想的流数据应该尽可能反映数据源信息特征，而实际应用中得到的数据流都不可避免和实际情况存在差异，如采集、感测、传输及编码等过程都可能造成数据质量下降，导致分类数据的失真和畸变。分类预处理阶段的工作是对原始流数据进行去噪和校正，减少数据失真对分类精度的不利影响。

分类判别阶段是数据流分类的核心部分，与其他领域的监督分类处理十分相似，包括属性选择与提取、分类模式学习算法设计、有指导学习和流数据分类四个步骤。在属性选择时，连续型和离散型信息是两个基本的属性类别。利用设计的学习算法对训练样本集进行学习，即可获得一个监督分类器。最后利用该分类器对未知类别的流数据进行分类和识别。与适用于静态数据的分类器不同，训练数据与待分类数据是变化的，而且分类器需要不断地重新构造。

分类后处理阶段主要是为了让领域专家或用户理解流数据输入属性与输出类别之间的关系，从而将分类器的分类行为表示成易于理解的形式。根据这些知识表示，人们可以获取训练样本集隐含的内在规律，并可进一步指导分类判别阶段的属性选择。

增量（在线）学习算法是从连续数据流中抽取模式的主要方法。增量归纳的基本思想是，每当接收到一个实例时，更新一个已存在的模式比创建一个新的模式的代价低许多。然而，增量算法具有几个不可避免的缺陷，比如对训练样本顺序的高敏感性、较非增量（成批处理）方法需要更长的训练时间等。在处理以每秒几千个速度抵达的事务对象，当每个新事例到来就更新的纯粹的增量方法是不切合实际的。纯粹增量学习算法包括 COBWEB[18] 和 ITI[19] 等。这些方法重点在于如何从固定的数据中产生分类模式的有效方法。

数据流研究领域中有关流数据的分类问题研究比较活跃，近几年出现了许多研究成果。Wang 等提出一个通用框架用于挖掘概念漂移数据流[20]。他们发现，迄今为止提出的数据流挖掘算法没有解决在进化数据上的概念漂移问题。他们提出使用加权分类器组合来挖掘数据流，模型中旧数据的过期特性依赖于数据分布。

Ganti 等开发了一个在插入和删除数据记录时维护模型的算法[21]。此算法能够应用于任何增量数据挖掘模型，这种模型被描述为一种通用的框架，在两个数据集中根据其产生的数据挖掘结果进行变化检测。上述技术被形式化为两个通用算法：GEMM 与 FOCUS，已经应用于决策树模型和频繁项集模型。

Domingos 等人开发了 VFDT[22]。这是一个基于 Hoeffding 树的决策树学习系统，它使用当前最佳属性，考察的依据是所用的检验数据项的个数满足 Hoeffding 边界的统计测度，此算法降低最近可能选择叶的活性并且删除了非潜在的属性。

Mania Papadimitriou 等研究者们提出 AWSOM（Arbitrary Window Stream mOdeling Method）用于在传感器网络中发现感兴趣的模式[23]。他们开发了一遍扫描的增量更新模式，其方法只需要 $O(\log N)$ 存储代价，其中 $N$ 为序列的长度。他们使用小波系数作为压缩信息表示和相关结构检测，然后在小波域上应用线性回归模型。

Aggarwal 等采用 On-Demand 分类中 CluStream 的微簇思想，获得了很高的分类精度[24]，此技术通过使用在每个簇中类分布的统计信息对数据进行分类。

Last 提出可以适应概念漂移的在线分类系统[25]。此系统使用信息模糊技术构建模型，使用信息理论计算窗口尺寸，使用最近的样本重建分类模型，使用误差率作为概念漂移的指导，调整模型构建的频度和窗口尺寸。

Ding 等开发了基于 Peano Count Tree 数据结构的决策树[26]，用实验证明了算法快速且适用于流应用。Gaber 等开发了 Lightweight 轻权值分类 LWClass 模型[27]，是 LWC 的变种，需要基于 AOG 的技术，LWClass 模型使用 K 近邻思想更新指定数据流特征类的发生频度。在输入数据流和存储的事例概要相冲突的情况下降低频度，当频率降为零时，从内存中释放所有标记为此类的事例。

传统数据分类算法对于数据流处理有两方面的挑战：无限的数据流与概念漂移。需要多次扫描数据集的方法无法处理无限的数据流，增量算法连续使用流上的新数据反复提炼模型。为解决概念漂移，这些方法以特定的频度重新修正模型以消除旧样本的影响。对于增量决策树分类器，修正模型意味着丢弃，重新生成子树，在某个节点创建一个新树[28]，这往往使算法变得极为复杂，因此需要研究怎样使最新的学习方法适应无限的概念漂移的流环境。除了上述缺陷，增量方法的预测精度一般不是很理想，因为不管样本是否代表概念发生了变化，以固定的速度丢弃旧样本，训练出的模型仅支持当前相对少量的数据快照，这通常会导致很大的预测偏差。为避免过渡拟合和概念漂移问题，旧数据的过期时间长短应当与数据的分布相关而不是数据的到达时间。

## 3 数据流预测与分类的模型与应用

### 3.1 数据流预测与分类模型

我们通过对目前数据流的预测与分类的研究的现在技术

进行分析，结合我们的系统性能分析的长期研究成果，提出了一个通用的数据流预测与分类模型。该模型也可用于网络入侵检测系统或系统性能衰退预测与异常检测，如图2所示。

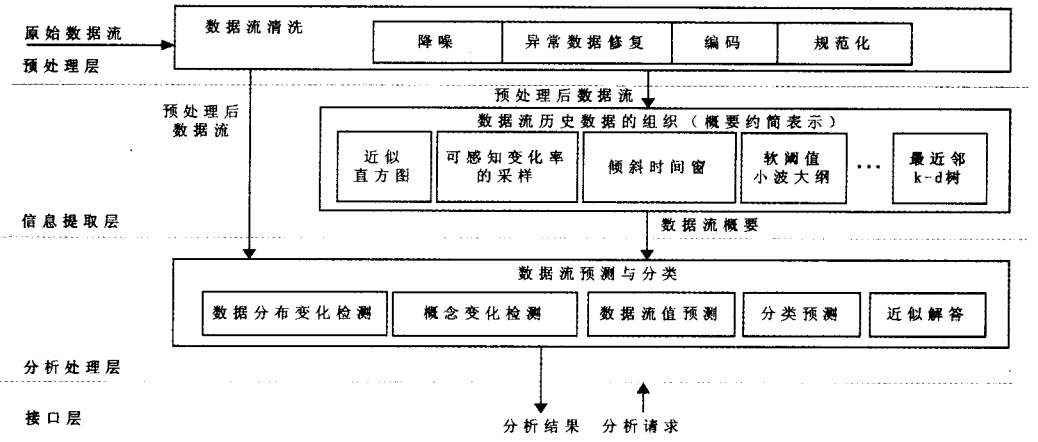


图2 数据流预测与分类通用模型

数据预处理层主要是对数据流进行加工以改善数据流的质量，为流数据的连续查询和复杂分析打下基础，包括对流数据进行去噪、压缩编码、修正以减少存储空间和传输时间。

信息提取层主要是对数据流中感兴趣的目标进行检测和识别，该层处理的对象具有基本的语义单位（比如关系数据库中的元组），研究适合于数据流分类与预测的概要表示方法，如近似直方图，可感知数据流变化的采样，多重时间粒度的倾斜时间窗，可调整阈值的软阈值小波大纲，能够提高近邻距离判断效率的k-d树等。处理后的结果是对数据流的特点与性质进行描述的概要数据，如类别符号、小波系数、分位数或其他统计信息等。

分析处理层主要完成预测和分类等复杂分析。在源源不断流入的经过预处理的数据流数据上，利用信息提取层生成的历史概要信息，对数据流进行数据分布变化的检测、概念变化的检测及预测与分类。这些数据流变化检测的结果既可以作为最终分析结果提交给其他应用（如入侵检测、性能监控），也可以为数据流值预测和分类预测提供模型重建的依据。

接口层主要完成预测与分类请求的提交，请求的发出者可以是应用程序或最终用户。同时接口层也负责将变化检测或预测及分类结果形成易于理解的知识表示返回给用户。在实际应用中，我们可以做成一个警告提示的反馈系统，如当系统性能衰退到警界线或出现异常现象时通知管理员。

### 3.2 基于数据流的系统性能分析

#### 3.2.1 基于数据流的系统性能分析模型

数据流模型在位于性能分析的在线监测点上对所到达的每个性能数据项进行拷贝，然后再进行本地分析是不现实的，只能对其进行在线状态的无保留监测，这就要求突发高频检测算法具有低时间复杂度。数据流模型是最近几年才出现的一种新的数据处理应用模型。数据流模型的最大优势就是能够在不保留或者少保留数据的前提下对在线数据进行实时一次性的单遍处理。

#### 3.2.2 系统性能衰退的实时预测

光有系统状态的历史数据和实时数据对于性能抗衰是不够的，这是因为系统未来的负载信息和性能状态对于抗衰决策的制定同样有着重要的影响，需要根据它们做出合理的抗衰决策，从而降低抗衰成本，同时提高可用性。目前这方面的研究工作主要集中在采用线性预测方式预测各种系统资源耗尽的时间估计，这种方式往往难于刻画真正的衰退趋势，误差较大，因此有必要开展这方面的研究工作，提出符合精确性更高的预测方式，我们下一步准备借鉴混沌时间序列理论，结合数据流的值预测可以有效解决预测这个问题。

#### 3.3.3 系统性能的在线异常检测

系统性能异常检测属于异常检测的范畴，目前这方面的研究主要是通过检测系统资源的占用和释放情况、服务的响应时间和响应率来验证性能异常的出现，多数都是基于离线的环境，不能实时地检测出异常，很难实现系统性能的在线检测，而数据流的分类技术能解决传统异常检测的不足。

**小结** 本文对数据流的预测与分类近年来的研究进行比较详细论述，建立了一个通用的数据流预测与分类模型，并提出可应用于系统性能的实时预测与异常检测。将数据流处理技术运用到系统性能保持上来，是一种很有前景的研究思路。下一步，我们将考虑如何对数据流预测进行深入研究，进一步提高预测的准确性；对数据流的在线分类技术进行改进，从而提高系统性能检测的实时性。

## 参考文献


1 Hulten G, Spencer L, Domingos P. Mining Time-Changing Data Streams. In: Proc. ACM Int Conf. on Knowledge Discovery and Data Mining, 2001. 97～106

2 Yi B K, Sidiropoulos N, Johnson T, et al. On-line Data Mining for Co-evolving Time Sequences. In: Proc. Int Conf. on Data Engineering, 2000. 13～22

3 Guha S, Mishra N, Motwani R, et al. Clustering Data Streams. In: Proc. IEEE Symp on Foundations of Computer Science. 359～366

4 Korn F, Muthukrishnan S, Srivastava D. Reverse Nearest Neighbor Aggregates over Data Streams. In: Proc. Int Conf. on Very Large Data Bases, 2002. 814～825

5 Chen Y, Dong G, Han J, Multi-Dimensional Regression Analysis of Time-Series Data Streams. In: Proc. Int Conf. on Very Large Data Bases, 2002. 323～334

6 Faloutsos C. Sensor Data Mining: Similarity Search and Pattern Analysis. In Proc Int Conf on Very Large Data Bases, 2002

7 宋国杰，唐世渭，杨冬青，等. 数据流中异常模式的提取与趋势监测. 计算机研究与发展，2004, 41(10): 1754～1758

8 Ankur J, Edward C Y, Yuan-Fang W. Adaptive Stream Resource Management Using Kalman Filters [A]. In: Proc. ACM SIGMOD [C]. Paris, SIGMOD, 2004. 11～22

9 Olston C, Jiang J, Widom J. Adaptive Filters for Continuous Queries over Distributed Data Streams [A]. In: Proc. ACM SIGMOD [C]. San Diego, SIGMOD, 2003

10 Faloutos C. Stream and Sensor data mining [A]. In: 9th International Conference on Extending DataBase Technology [C], Heraklion, Greece, EDBT, 2004. 25～27

11 李建中，郭龙江，张冬冬，等. 数据流上的预测聚集查询处理算法. 软件学报，2005, 16(7): 1252～1261

12 Ahmet B, Ambuj K S. STARDUST: Fast Stream Indexing U-

sing Incremental Wavelet Approximations. In: Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003
13 王永利,周景华,徐宏炳,等. 时间序列数据流的自适应预测. 自动化学报(已录用,待发表)
14 Trudnowski J D, McReynolds W L, Johnson M J. Real-time very short-term load prediction for power-system automatic generation control [J]. IEEE Transactions on Control Systems Technology, 2001, 9(2): 254～260
15 徐科,徐金梧,班晓娟. 基于小波分解的某些非平稳时间序列预测方法[J]. 电子学报,2001,29(4): 266～268
16 贺国光,马寿峰,李宇. 基于小波分解与重构的时间序列预测法[J]. 自动化学报,2002(6): 1012～1014
17 Liu K, Subbarayan S, Shoults R R, et al. Comparison of Very Short-term Load Forecasting Techniques [J]. IEEE Transactions on Power System, 1996,11(2):232～239
18 Fisher D H. Knowledge Acquisition via Incremental Conceptual Clustering, Machine Learning,1987. 139～172
19 Utgoff P E. An Improved Algorithm for Incremental Induction of Decision Trees. In:Proc. of the Eleventh International Conference on Machine Learning,1994. 318～325
20 Wang H, Fan W, Yu P, et al. Mining Concept-Drifting Data Streams Using Ensemble Classifiers. In: 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington DC, USA, Aug. 2003
21 Gehrke J, Rastogi R. Querying and mining data stream: you only get one look. A tutorial. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. Madison: ACM Press, 2002. 635
22 Domingos P, Hulten G. Mining High-speed Data Streams. In: Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, 2000
23 Papadimitriou S, Faloutsos C, Brockwell A. Adaptive, Hands-off Stream Mining. In: 29th International Conference on Very Large Data Bases VLDB, 2003
24 Aggarwal C, Han J, Wang J, et al. On Demand Classification of Data Streams. In: Proc. 2004 Int Conf. on Knowledge Discovery and Data Mining, Seattle, WA, Aug. 2004
25 Last M. Online Classification of Nonstationary Data Streams. Intelligent Data Analysis, 2002,6(2):129～147
26 Ding Q, Ding Q, Perrizo W. Decision Tree Classification of Spatial Data Streams Using Peano Count Trees. In: Proceedings of the ACM Symposium on Applied Computing, Madrid, Spain, March 2002
27 Gaber M M, Krishnaswamy S, Zaslavsky A. On-board Mining of Data Streams in Sensor Networks. S Badhyopadhyays, Maulik U, Haolder L, et al. eds. Advanced Methods of Knowledge Discovery from Complex Data, Springer Verlag
28 Last M, Klein Y, Kandel A. Knowledge Discovery in Time Series Databases. IEEE Transactions on Systems, Man, and Cybernetics, 31(Part B, 1):160～169, 20

(上接第 155 页)

### 2.5 变异操作

变异操作的基本内容是对群体中个体串的某些基因座的基因值作变动。本文变异操作的操作思路如下:

STEP1:在父代染色体对应的服务标的排列上随机选择某一段子排列;

STEP2:使得这个服务标的排列上的子排列顺序产生倒转,这样就形成一个新的服务标的排列;

STEP3:将新产生服务标的排列按优先适合启发式规则构造一个新的主体(即染色体)。

### 2.6 再生操作

为了提高本文算法的收敛速度,笔者在主体完成演化操作后,对当代主体中适应度最低的主体执行一次再生操作。再生操作建立在群体中个体的适应度评估的基础上,目的是把优化的个体直接遗传到下一代,或者通过配对交叉产生新的个体再遗传到下一代。本算法再生操作就是对当前主体再次进行正交交叉操作和变异操作,希冀能通过再生操作来提高该主体的适应度。

### 2.7 自学习操作

为了提高本文算法的全局优化能力和收敛速度,笔者在主体完成演化操作后,对每个主体都执行一次自学习操作。

该操作的具体流程如下所示。

STEP1:在父代染色体对应的服务标的排列上随机选择某一段子排列;

STEP2:将这个服务标的排列上的子排列中的最后一个服务标插入到子排列的最前面位置,子排列中的其他服务标依次向后移一位,这样就形成一个新的服务标的排列;

STEP3:将新产生服务标的排列按优先适合启发式规则构造一个新的主体(即染色体)。

### 2.8 终止条件

严格地讲,遗传算法的迭代终止条件目前尚无定论,一般通过多次进化逐渐逼近最优解而不是恰好等于最优解,因此需要确定其终止条件,最常用的终止方法是规定遗传的代次。在本算法求解过程中,笔者始终保存历代最优解,同时指定最大演化代数作为停止准则,即选一个大的正整数 Max_Gen 为最大迭代次数,若当前迭代次数 $k$ 大于 Max_Gen ,则停止迭代并输出历史最优解作为最终的结果。

**结论** 本文主要创新点:提出了一种基于正交多主体遗传算法的电信组合服务竞标的竞胜标确定方法,以解决业务规则引擎无法实现这类复杂业务规则求解的问题。该方法以多主体系统为基础,通过每个主体的逐步演化和自学习功能来提高算法的全局优化能力和收敛速度;利用正交试验设计方法产生较好的初始种群和设计正交交叉算子以获得更好的后代。

## 参 考 文 献

1 张渊,夏清国. 基于 Rete 算法的 JAVA 规则引擎. 科学技术与工程,2006,6(11):2
2 Mito M, Fujtta S. On heuristics for solving winner determination problem in combinatorial auctions [J]. Journal of Heuristics , 2002, 10(5): 507～523
3 Sandhoim T, Suri S, Gilpin A, et al. CABOB: A fast optimal algorithm for winner determination in combinatorial auctions. Management Science (S0025-1909), 2005, 51(3): 374～390
4 Sandholm T. Approaches to winner determination in combinatorial auctions [J]. Decision Support Systems (S0167-9236), 2000, 28(1): 165～176
5 Kelly F, Steinberg R. A combinatorial auction with multiple winners for universal service [J]. Management Science, 2000, 46 (4): 586～596
6 Kuyanoglu E, WU S D. On combinatorial auction and Lagrangian relaxation for distributed resource scheduling [J]. IIE Transactions , 1999, 31(9): 813～826
7 陈培友,汪定伟. 用遗传算法求解组合拍卖竞胜标[J]. 东北大学学报(自然科学版), 2003, 24(1): 7～10
8 陈培友,汪定伟. 用改进遗传算法求解组合拍卖竞胜标[J]. 东北大学学报(自然科学版), 2004, 25(4): 349～352
9 Roias E M, Mukherjee A. Multi-agent framework for general-purpose situational simulations in the construction management domain [J]. Journal of Computing in Civil Engineering,2006, 20 (3): 165～176
10 Ota J. Multi-agent robot systems as distributed autonomous systems [J]. Advanced Engineering Informatics,2006, 20(1):59～70
11 Camacho D, Aler R, Borrajo D, et al. Multi-agent plan based information gathering [J]. Applied Intelligence , 2006, 25(1): 59～71
12 Kan H, Shen H. Lower bounds on the minimal delay of complex orthogonal designs with maximal rates [J]. IEEE Transactions on Communications , 2006, 54(3): 383～388
13 何大阔,王福利,毛志忠. 遗传算法在离散变量优化问题中的应用研究[J]. 系统仿真学报, 2006, 18(5): 1154～1156
14 姚钦,史仪凯,夏锐. 多目标交互式遗传算法在测试点确定问题中的应用[J]. 系统仿真学报, 2006, 18(6): 1469～1472