未知病毒检测技术的研究

Research of Unknown Virus Detection

赖英旭[1]　李　征[2]

(北京工业大学计算机学院　北京 100022)[1]　(北京工业大学应用数理学院　北京 100022)[2]

Abstract　With the development of computer science, more and more computer viruses come out, which seriously compromised the security of the computer world. Current virusscanner do not generalize well to detect unknown viruses. The paper promotes an unknown virus detection technology based on Bayes method and explores the extraction of features and machine learning. The paper also gives an unknown virus detection framework. We provide a new algorithm-half increment algorithm, which inherits virtue of Naïve Bayes' and complex Bayes'. The evaluations and results are given in this paper.

Keywords　Naïve Bayes, Machine learning, Virus detection

1　引言

计算机的广泛应用为病毒的滋生提供了良好的条件,新生病毒层出不穷,病毒的花样不断翻新,编程手段越来越高,防不胜防。特别是 Internet 的广泛应用,促进了病毒的空前活跃,网络蠕虫病毒传播更快更广,Windows 病毒更加复杂,带有黑客性质的病毒和特洛伊木马等恶意代码大量涌现,给计算机的运用带来了很多灾难性的后果。基于特征检测法病毒检测技术的基本原理是:提取已知病毒样本的特征,并将此特征数据添加到病毒特征库中,在病毒检测时通过搜索病毒特征库查找是否存在相匹配的病毒特征来发现病毒。这种检测办法只能用于检测已知的病毒,对于新出现病毒的检测无能为力,并且在升级病毒库的过程中有很大一部分是重复工作,这并不利于反病毒技术的发展。

因此,现有的特征值查毒法对于已知病毒能实现快速、高效的查杀,但对于未知病毒检测却有点力不从心。

Fred Cohen 博士在 20 世纪 80 年代指出了"计算机病毒检测的不可判定性",即"精确检测计算机病毒是不可判定的"。此后 Diomidis Sp inellis 证明了"有界长度病毒的可靠检测是一个 NP 完全问题"。因此长期以来,各反病毒研究机构都在努力探讨病毒检测的近似算法。在 IBM 病毒研究中心,曾经成功地将神经网络用于判断引导型病毒,M. Schultz 等曾提出用数据挖掘的算法如朴素贝叶斯算法等检测未知恶意代码。基于上述思想的启发,本文设计了一未知病毒检测模型,对朴素贝叶斯算法、复合贝叶斯算法进行分析研究,提出一种改进算法——半增量贝叶斯算法,该算法可以对可疑文件进行分析评判,最终实现对未知病毒的识别。最后我们对模型的功能进行了实验测试,结果达到了很好的实际应用效果。

2　系统结构设计

根据上述分析和研究,实现未知病毒检测系统的结构如图 1 所示。

2.1　特征提取

算法只有通过特征项的描述才能建立检测模型。对于存储在计算机上的文件其状态有两种:活动状态(正在被执行),静止状态(没有被执行)。基于这两种状态产生了两种特征提取方式:行为特征提取,静态特征提取。静态特征提取主要方式是提取文件中包含的字符串。在一个文件中包含的字符串有许多,但是真正反映出文件功能的字符串是有限的,所以如何提取到有用的信息将是这一模块需要解决的问题。

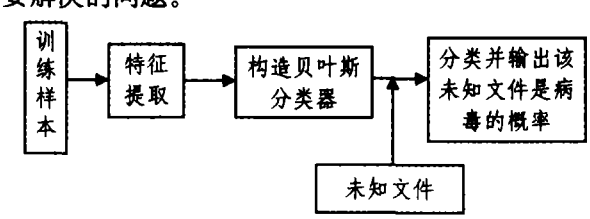


图 1　系统结构图

(1)特征项描述方式　特征项可以有字符串描述方式、特征资源描述方式、二进制描述方式,它们各有特色。在计算机里,任何的应用程序都是通过一定的编码来实现某种功能的,计算机病毒也不例外。而无论是哪一种编码都需要申请系统资源来完成自身的功能,这就在所有的应用程序之间找到了一个共性:需要通过调用系统底层的 API 函数来实

现自身的功能，而这些 API 函数无疑会在该应用程序文件中出现。因此我们采用字符串描述方式作为文件特征的描述方式，这些字符串可以反映出该文件的特性。

(2)特征项精简　在文件中可以找到的字符串并不是都包含有用信息，而且数量是相当大的，因此向量空间的维数也相当大，可以达到几万维，需要一个算法来找到含有用信息的字符串。对于 API 函数来说，都是一些有着特殊含义的类似英文单词的字符串。基于这一事实，用模糊匹配英文单词和词根的方法可以有效地屏蔽不含有用信息的字符串。

通过上面的分析，得到如下的特征提取方法：采用字符串描述方式作为文件特征的描述方式，通过模糊匹配英文单词和词根的方法过滤无用字符串，从而得到文件的特征。

2.2　构造贝叶斯分类器

基于上一模块——特征提取模块所得到的原始信息为从文件中提取到的有用字符串，这种提取特征的方式类似于文本分类的特征提取方式，问题形式也类似于文本分类问题，则将原问题抽象为文本分类问题。

综合朴素贝叶斯算法和复合贝叶斯算法的优缺点，本文提出了一种更有效的结合上述两种算法的优点，降低这两种算法缺点的影响，并且依然是基于贝叶斯定理的算法——半增量贝叶斯算法。

半增量贝叶斯算法将先验知识进行增量式学习，并对先验知识进行保存形成先验知识库。先验知识即为对原始信息的统计得到的特征文本出现的概率。对于朴素贝叶斯算法和复合贝叶斯算法都是通过学习直接产生知识库，而不是先验知识库。

半增量贝叶斯算法就是采用复合贝叶斯算法的方法，从独立的小样本空间中取得先验知识，然后通过复合概率公式叠加到已有的先验知识库当中去，这样对先验知识的增量学习也达到了整个模型增量学习的目的。再根据新的先验知识库对知识库中需要更新的知识采用朴素贝叶斯模型的方法进行更新。

此算法集合了朴素贝叶斯算法与增量贝叶斯算法的优点，并且有效地解决了这问题：学习知识不够，分类效果不好。但是可以通过增量学习，不断更新知识库，两个算法中出现的问题。但在样本量较小的时候，会遇到与复合贝叶斯算法同样的从而使分类效果不断提升，其最终效果应与理想中的朴素贝叶斯模型相同。

2.3　测试数据和实验结果

实验主要是为了检验上述模型对于未知文件的分类识别能力。我们采用江民新科技术有限公司提供的病毒文件作为病毒样本，将病毒样本分为两部分，一部分占总病毒样本的 80%，作为学习用样本，另一部分占总病毒样本的 20%，作为待分类样本。正常文件采用江民公司无毒服务器的 C 盘中的文件作为正常文件学习样本，另外取江民公司无毒服务器 D 盘下文件作为待分类文件。

实验的样本数据如表 1 所示。样本空间中样本总数为 3672，分为正常程序与染毒程序。正常程序从操作系统平台中选取，选用江民公司无毒服务器的全部 PE 文件共 3017 个。江民公司提供的病毒程序 655 个。

表 1　样本集

	样本空间	学习样本集	待分类样本集
正常文件	3017	2600	417
病毒文件	655	524	131
共计	3672	3124	548

在此基础上测试系统实现的分类算法，并对其效率和结果进行比较分析。

(1)耗时比较　朴素贝叶斯算法：在进行系统测试时，朴素贝叶斯算法由于遇到维数灾难问题，学习过程耗时四天，已经大大超出了可以应用的限度，所以在结果测试中并没有提供朴素贝叶斯算法的结果。

复合贝叶斯算法：样本分组大小为 10 个文件，学习过程耗时 35 个小时左右。

半增量贝叶斯算法：以一个文件为一个增量样本空间，学习过程耗时 24 个小时左右。

从三种算法在学习过程所耗费的时间可以明显看出，半增量贝叶斯算法有着明显的优势，并且由于半增量贝叶斯算法自身的特点，使得其可以支持最小单位为一个文件的增量学习方式，这样就不会导致运行过程中由于出现意外终止而全部重新学习的现象发生。无论是复合贝叶斯算法，还是朴素贝叶斯算法都存在这个问题，特别是朴素贝叶斯模型必须全部重新学习。

(2)分类结果比较　由于朴素贝叶斯算法在测试过程中遇到维数灾难问题，导致学习时间过长。这里没有做朴素贝叶斯算法的效果测试。在理论上半增量贝叶斯算法的分类结果与朴素贝叶斯算法的分类结果是相同，这一点从半增量贝叶斯模型与朴素贝叶斯模型所产生的知识库可以得到证明，其产生的知识库近似相同。因此，这里并没有列出朴素贝叶斯模型的测试结果，只对复合贝叶斯算法与半增量贝叶斯算法进行了对比分析，结果如表 2 所示。

通常我们将错误类型分为两种，即：(1)将正常程序判断为病毒，称为 False Positive；(2)将病毒程序判断为正常程序，称为 False Negative。准确率为文件被正确分类的概率；查全率为病毒文件被正确识别为病毒类的概率；误报率为待分类样本集中的

(下转第 313 页)

基于 Agent Grid 的 GBODSS 系统构建*)

Construction of GBODSS Based on Agent Grid

吴铁洲　陈学广　迟嘉煜

(华中科技大学控制科学与工程系　武汉 430074)

Abstract This paper puts forward that the Agent technology can be used in Grid Based Open Decision Support System(GBODSS), and describes the inner construction of Agent Grid and the architecture of GBODSS based on Agent Grid.

Keywords Decision support system, Grid, MAS

1 引言

随着社会信息化进程的进一步加快，管理者所面对的决策问题和决策环境日益复杂，传统的DSS[1]已经不能满足不断变化的环境和实际应用的需求。网格[2]技术的出现与成熟，为决策支持系统的发展带来了巨大的机遇。为实现 Internet 上的决策资源共享提供了一种理想的途径，能够很好地解决目前 DSS 发展中遇到的许多问题。因此提出了基于网格的开放式决策支持系统(Grid Based Open Decision Support System, GBODSS)的概念和思想，并对 GBODSS 的一些基础理论和关键技术进行了探讨和研究，尝试建立了 GBODSS 研究的理论框架和一种新型的基于 Agent Grid 的 GBODSS 系统结构。

2 GBODSS 体系结构

基于网格的开放式决策支持系统是建立在分布式全球网络上的用来进行决策支持和建模支持的电子环境。根据网格的开放式网格服务体系结构(Open Grid Service Architecture, OGSA)[3]，提出了一个基于网格环境的开放式决策支持系统(Grid-Based Open Decision Support Systems)构成框架，如图 1 所示。

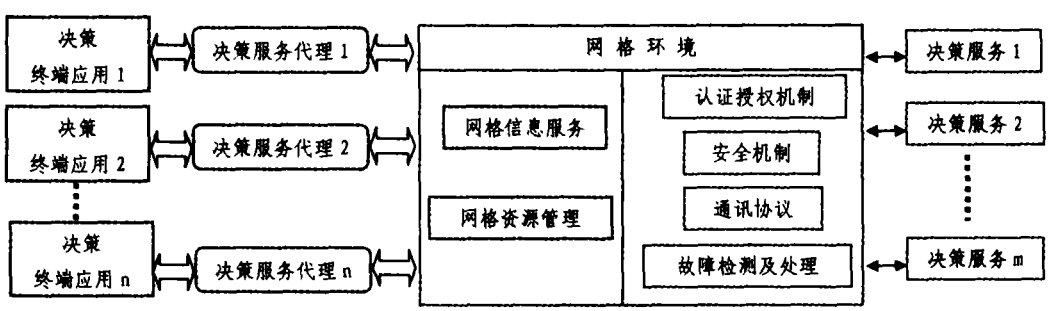


图 1　基于网格的开放式决策支持系统构成

1)决策服务：OGSA 系统模型以决策服务为中心，通过决策服务来整合网格平台上的各种决策资源。OGSA 定义了"网格服务"的概念，网格服务是 Web Service 的扩展，组织和管理服务等更高层次的网格服务。它能够支持临时服务实例的动态创建和删除。几个网格服务还可以以不同的方式聚集起来形成一个更大的网格服务，完成更复杂的任务。网格服务的这些特性使它非常适合用于构建决策服务，这里的决策服务都由网格服务来实现。

2)网格环境：网格环境由两层组成。网格环境的底层由认证机制、安全机制、通信协议和故障检测及处理等网格基础服务构成，为网格环境提供基本的服务支撑；网格环境的第二层由网格信息服务和网格资源管理服务构成，为用户提供各种信息和资源。

网格信息服务能够实现网格上所有资源(并不仅限于信息资源)的注册、发布；网格信息服务可以用来发现、管理和规划网格上的资源。网格资源管

*)国家教育部博士点基金资助，基金号(20040487076)。

理服务能够利用网格信息服务提供的信息把各个需要完成的任务分配到合适的网格资源上，并组织和协调各个任务的完成。网格环境能够为决策者发现并提供决策所需的各种决策服务信息，并组织和协调这些决策服务来共同完成决策任务。

3)决策服务代理模块：决策服务代理模块是决策者和网格环境的中介，为决策者进行决策服务查询、决策服务交易和执行决策服务集成任务，是基于网格的决策支持系统模型提供决策支持的中心部件。决策服务代理模块实现传统 DSS 中的问题处理及求解系统的功能，不同的是它在一个开放的环境中处理和求解并行、分布的问题。

决策服务代理模块接受决策者通过决策者终端应用下达的决策服务查询请求，调用网格环境提供的网格信息服务，完成决策者的查询请求。

4)决策者终端应用模块：决策者终端应用是决策者与 GBODSS 交互的人机接口，决策者通过决策者终端应用向决策资源代理模块发送决策资源查询请求和决策任务；决策服务代理模块通过决策者终端应用向决策者反馈决策服务查询结果和决策任务执行结果。

2 MAS 技术用于 GBODSS

Agent 能够清晰地刻画 GBODSS 的个体特性，MAS[4]可以很好地反映 GBODSS 的整体特性，MAS 还能够为 GBODSS 的分析和设计提供有力的理论基础、方法支持以及强大的技术支撑的 Agent Grid 平台。

Agent Grid 平台充分考虑了 GBODSS 的特殊性，通过定义 Agent Grid 服务和规范，为网格中的 Agent 提供了良好的生存和运行环境，方便了网格中的异构、开放和大规模的 Agent 的管理和维护，为将 MAS 技术用于 GBODSS 系统建模提供了良好支撑平台。

Agent Grid 的结构模型如图 3 所示。该 Agent Grid 由网格基础设施(Grid Infrastructure)、网格社区(Grid Community)、网格成员(Grid Member)组成。它通过提供众多的网格服务(Grid services)来集成网格上的各种分布、异构的软件组件，包括对象、Agent、遗留系统等。

3 基于 Agent Grid 的开放式 DSS 系统结构

3.2 Agent Grid 为中心的网格层次结构

Agent Grid 为中心的网格层次结构包括如下五个层次：

1)计算层，对应于计算网格，提供计算力资源；2)数据和信息层，对应于数据网格和信息网格，提供数据、模型方法和知识等信息资源；3)服务层，对应于服务网格，包括组件、服务、遗留系统等软件资源；4)Agent 层，即 Agent Grid；5)用户层。

每个层次都包含相应的资源，并提供对各自资源的管理服务，各个层次之间需要紧密地协同工作来完成遇到的问题。这五个层次的语义级别和功能性逐渐增强。

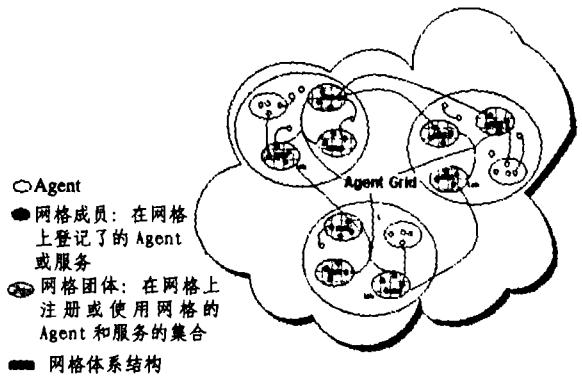


图 3 Agent Grid 结构示意图

Agent Grid 是这种网格层次的中心，在整个网格系统中具有非常重要的作用。一方面，它作为和用户直接交互的一个层次，它所包含的 Agent 及 Agent 系统能够独立地担负起系统总体功能中的各个子功能，实现总体任务的分解和综合，并且需要具有良好的与用户交互的接口，方便用户使用网格中的各种资源来完成其所遇到的问题；另一方面，Agent Grid 需要承接其下的三个网格层次(即计算层、数据和信息层、服务层)，通过下层网格提供的接口来实现对其下的三个网格层次中资源的调用和协调管理，实现各个层次之间的紧密配合，来共同完成具体的任务。

3.2 基于 Agent Grid 的 GBODSS 系统结构

新型的基于 Agent Grid 的开放式 DSS(Agent Grid Based Open DSS，AGBODSS)系统结构，如图 4 所示。

Agent Grid 处于该系统的核心部件，它作为用户和其下层网格层次的中介，起到了承上启下的作用。Agent Grid 由注册的多个 Agent 团队和一些网格服务组成，每个 Agent 团队包含多个实体 Agent。Agent Grid 通过提供众多的 Agent Grid 服务来为其所包含的 Agent 提供管理、交互和集成等功能。目前，Agent Grid 服务包括注册服务、代理服务、日志服务、安全服务和可视化服务等。

针对实际的决策问题，构造者能够快速地集成 Agent Grid 中的多个 Agent 来构建不同的 DSS。DSS 能够通过 Agent Grid 屏蔽网格中各种资源的异构性，方便地利用整个网格系统中的各种资源。

服务网格也是该模型的一个重要部件，它通过构建在其上的决策服务电子市场来管理网格上的各

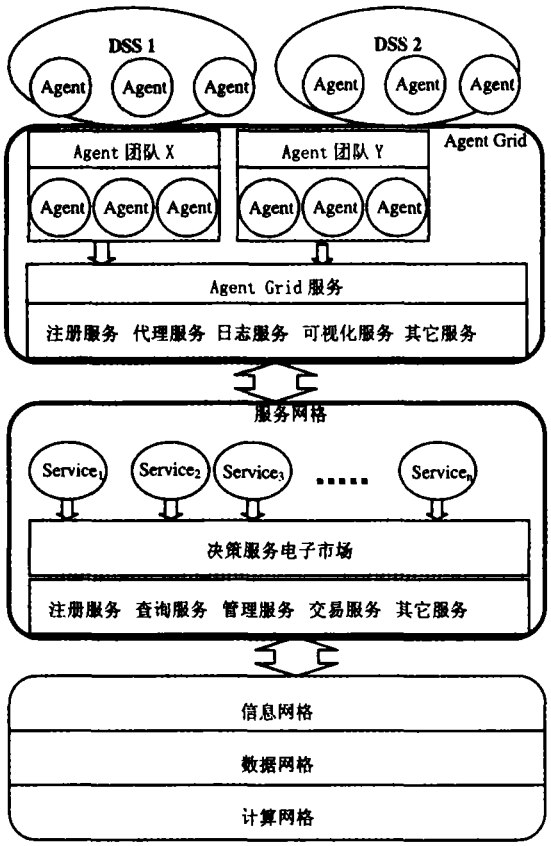


图 4 基于 Agent Grid 的开放式 DSS 系统

种决策服务，利用市场的资源配置和竞争机制来帮助决策服务的使用者获得合适的决策服务。决策服务电子市场可以提供注册服务、查询服务、管理服务、交易服务、认证服务和安全服务等许多服务。Agent Grid 中的实体 Agent 可以通过注册到决策服务电子市场，来发现和使用网格上的各种决策服务。

结束语 基于网格的开放式 DSS 作为网格技术在 DSS 中的应用研究，目前在全球范围还处于起步阶段，在我国还是一个崭新的研究领域，随着网格技术的逐渐成熟和网格应用的逐渐增多，GBODSS 的研究将成为今后决策支持系统研究的一个重要方向。本文对 GBODSS 作了一些有益的探索，提出了 GBODSS 的基本概念和理论框架。但是，这一领域是一个崭新的交叉领域，从功能层面到技术层面涉及的研究问题非常多，还需要不断吸收其它学科领域的优秀成果，其理论探索和应用研究还有待于进一步加强和深入。

参 考 文 献

1 仰炬，张朋柱. 决策支持系统开发技术综述. 工业技术经济，2004，23(3)：59～85

2 谢储晖，郭达志. 网格技术综述. 闽江学院学报，2003，24(5)：7～13

3 刘建新，阎保平. OGSA-DAI 体系结构及其关键技术研究. 计算机应用，2004，24(11)：81～87

4 郭红霞，吴捷，等. 多 Agent 技术研究进展. 河南科学，2004，22(2)：242～246

(上接第 296 页)

4.3 实验结果

本文实验环境为 Matlab6.5，计算机配置为 Pentinum(R)4，CPU2.8 GHz，内存 512 M。下表为实验结果。

表 1 算法改进前后用于识别模型的识别率对比

算法 \ 音种	理想语音	加高斯白噪声语音 (SNR：20db)	电话语音
改进前算法	93.5%	91.4%	83.6%
改进后算法	95.7%	92.6%	86.1%

结论 本文提出了一种基于核主成份分析和支撑向量机的文本无关的说话人识别改进算法。该算法通过先对样本的核主成份分析的预选取，然后进行边界样本选取使训练数据类别平衡，克服了分类结果的倾向性，与未改进前算法相比提高了确认系统的识别性能，纯净语音、电话语音等错误率分别提高了 1.2%、2.5%；同时也证明了子带倒谱具有鲁棒性。

参 考 文 献

1 Campbell J P. Speaker Recognition: A Tutorial. Proceedings of the IEEE, 1997, 85(9)

2 Vapnik V 著. 张学工译. 统计学习理论的本质[M]. 北京：清华大学出版社，2000

3 Burges C. A Tutorial on Support Vector Machines for Pattern Recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121～167

4 Scholkopf B, Smola A. J Learning with kernels. MIT Press Cambridge, MA, 2002

5 Lin H T, Lin C J. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods: [Technical report]. Department of Computer Science and Information Engineering, National Taiwan University. Available at http: //www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf

6 H Wei-Wen, W Hsiao-Chuan. On the Use of Weighted Bank Analysis for the Derivation of Robust MFCCs. IEEE Signal Processing Letters, 2001, 8(3): 70～72

基于代理的分布式入侵检测系统模型的研究

Research on Distributed Intrusion Detection System Based on Agent

王新生　李彦辉　张　颖

(燕山大学信息科学与工程学院　秦皇岛 066004)

Abstract This article analyzes concentrated and distributed type intrusion detection system's advantages and disadvantages first, proposes on this foundation that a kind of partial centralism frame model, and carries on detailed elaboration to this model structure, then elaborates this frame security feature with emphasis.

Keywords Intrusion detection system, Agent, Communication, Security

1　引言

近几年,关于入侵检测技术的研究发展很快,出现了许多入侵检测系统。但是,随着网络高速化发展,分布式、多元化、多服务、多应用、多用户的环境下,更缺少一个有效的入侵检测体系,而且新的攻击方法的不断出现,尤其是一些互相协作的入侵行为的出现,给入侵检测领域研究带来了新的课题。早期的集中式入侵检测系统不能有效适应这些新的问题,完全分布式的系统又过于复杂难以管理,针对这一现状,本人提出一个基于代理的分层入侵检测系统框架。

2　入侵检测技术的分析

从入侵检测系统各个模块运行的分布方式的不同可分为集中式和分布式入侵检测系统。

2.1　集中式入侵检测系统框架

过去的入侵检测系统多数都采用单一体系结构,即所有的工作包括数据的采集、分析处理都由单一主机上的单一程序来完成,而一些分布式的入侵检测系统只是在数据采集上实现了分布式,数据的分析、入侵的发现还是由单个程序完成的,这样的结构有如下的缺陷:

性能瓶颈。由于所有工作都是由单一主机执行,数据过多会造成其过载,从而导致系统因为来不及处理过量的数据或丢失网络数据包而失效。

单点失效。当数据分析处理程序因受到攻击或其它原因不能正常工作时,会影响到整个系统。

系统缺乏灵活性和可配置性。当系统需要加入新的模块和功能时,系统就需要修改和重新安装。

2.2　完全分布式系统框架

文[2,3]中提出了无控制中心的思想,该模型引入了代理(Agent)的思想,代理(Agent)的研究起源于人工智能领域,它是模拟人类行为和关系,具有一定智能并能够自主运行和提供相应服务的程序。该模型的主要思想是在每台机子上部署不同类型的代理来监控本机安全,这样每台机子上的各类代理可以相互通信以便更好地发现针对本机的入侵,同时,整个网络的各个主机间的各个代理也可以通过TSA(通信服务代理)代理通信交流协作发现分布式协作攻击。这种模型虽然避免了控制中心的瓶颈现象,但是出现了新的问题:

(1)多个代理在大规模网络环境下的通信协作问题。例如,某主机内一个代理需要多个代理的协作才能判断入侵,它需要等待其它多个代理的响应和数据的传输。这些都增加了判断入侵行为的时延。

(2)多代理间通信协调的复杂性也会降低系统的效率和可维护性。

(3)无控制中心的系统中,对确定的入侵一般采取广播的方式向全网报告,这不可避免地给网络带来了大量的通信流,影响网络质量。对于易受攻击的网络可能产生更严重的后果。例如一次恶意的针对整个网络的扫描可能使网络陷入瘫痪。

综上所述,完全分布式的入侵检测框架虽然在判断分布式协作攻击方面有很强的识别能力,避免了集中式框架的一些缺陷,但是从实用角度来看系统的复杂性和可维护性是个严重的问题。

3　分层式入侵检测系统

为了更好地理解入侵检测系统,深入地剖析集中式和分布式系统的实质,将入侵检测系统概括为三个主要功能模块:数据采集模块,数据分析处理模块,报警及警报响应模块。一般入侵检测系统,无论

王新生　教授,研究方向为计算机网络及网络安全理论;李彦辉　硕士,研究方向为网络安全;张　颖　硕士,研究方向为网络安全。

是集中式还是分布式都会有此三个主要功能模块。仔细分析以上两种类型的系统框架，可以看到，将网络中所有分析处理功能全部交给一个模块处理太过集中，容易造成性能瓶颈、网络拥塞和单点失效，而无控制中心思想导致实时性和效率的降低而且系统太过复杂。

针对这两种类型的入侵检测系统框架存在的缺陷，本文提出了将集中式和分布式融合一体的思想，从而将两种类型系统的优点结合起来避免其缺点。在此基础上提出了一个基于代理的分层的入侵检测系统模型，该模型将整个网络分成不同的子网或网段，将原有的控制中心的分析处理功能分解到各个网段，每个网段有一个综合分析本网段的代理，该代理能够针对本网段的情况派遣合适的移动代理到重点主机分析其日志并能够监控本网段的网络流信息，对本网段的入侵有清晰的判断，从而大大减轻了全局层的工作量，降低了性能瓶颈问题。整体模型框架如图 1 所示。

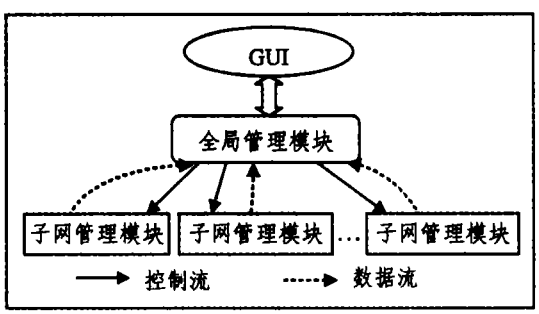


图 1　整体模型框架

该模型分为两层，上层为全局层，下层为局部子网层，所有的子网管理模块向全局管理模块汇报入侵情况。该模型充分利用移动代理的移动特性来保证整个系统以及局部管理代理的安全，具体策略是：由一个代理专门负责全局代理和局部子网的管理代理的安全，它根据 Agent 资料总库的信息按照一定的策略遍历各个局部子网的管理代理并与之通信，如果管理代理受到攻击不能工作或崩溃则返回失败信息并将受攻击代理的安全状态改为非安全状态并呈献给管理员，由管理员通过 Agent 管理系统进行删除并重新分派代理。此外，各个局部管理代理可利用自身的移动特性躲避针对自身的攻击，移动范围限制在本逻辑子网，并且在移动到新主机后向全局代理发送新主机的具体信息并修改 Agent 资料库。下面从该模型的局部和全局结构进行阐述。

3.1　子网管理模块

子网管理模块的系统框架示意图如图 2 所示。

其中局部模块可为一个守护代理，负责监控本子网的安全状况。该局部模块将基于主机和基于网络的入侵检测技术结合一体、其中数据的采集分为主机审计日志和网络数据包，主机审计日志由分布在重点主机上的代理负责收集。整个工作流程可具体描述如下。

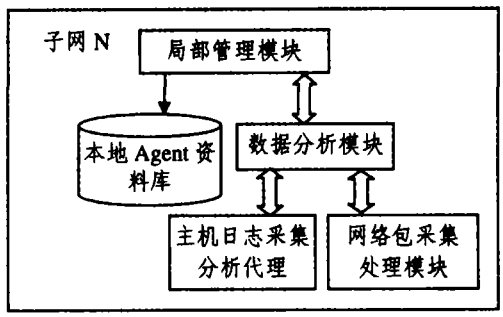


图 2　局部系统框架示意图

(1)网络包采集处理模块以基于特征匹配的方式收集可疑警报上交给数据分析模块并保存在库以备查用；由分布在本子网各个重点主机上的代理监控本主机系统审计日志及应用程序日志的异常事件并上报。

(2)数据分析模块接受两种类型的事件，一种为 Net-alert 另一为 Host-alert，对网络包采集处理传递的报警(Net-alert)进行降冗处理后上报；对主机异常日志的报警(Host-alert)则结合网络报警信息进行验证并上报，如果网络报警信息中有相对应的报警，则向可确认本次入侵，否则可视为可疑入侵。举例来说，某子网内有一台安装 Linux 的主机，运行 httpd 对外提供 Web 服务，在该子网内有两种类型的感应器，一种为网络包分析处理代理(A1)，通过截获网络数据包并进行模式匹配来发现入侵行为；另一种为重点主机日志分析代理(A2)，驻留在该 Linux 主机上，通过监控系统调用来检测分析重要守护程序的各种行为。假设发生了某次针对 httpd 守护进程的缓冲区溢出(buffer overflow)攻击，A1 通过攻击者发送的数据包中检测到了对应该种类型缓冲区溢出攻击的特征代码，向数据分析代理上报事件 E1；同时，A2 在当前 httpd 所产生的系统调用序列与历史序列的比较发现了异常，收集异常调用序列事件 E2 向数据分析代理上报，数据分析代理在对事件 E2 发生的时间、攻击类型及目标 IP 对缓冲区中的网络报警信息进行查询，得出缓冲区溢出攻击事件 M1，就可以确认本次攻击行为。

(3)局部管理模块将数据分析模块上报的警报送交全局管理模块并监控本子网范围内的代理的安全。如发现针对本身的攻击，采取以下流程处理：

a. 先向全局管理模块发送 Attack-agent 类型的消息(包括本代理的相关信息)；

b. 根据本地 Agent 库撤销本代理所在主机的相应代理；

c. 根据一定策略选出要迁往的主机并将自身发

送到该主机；

d. 在新主机上创建相应代理并记录在库，并将本代理现在的相关信息上报全局管理模块；开始工作。

3.2 全局管理模块

全局管理模块的系统框架示意图如图 3 所示。全局框架的主要功能就是负责全局性的报警分析以期发现全局性的或分布式协作的攻击，另外负责跟管理员的接口，具体功能模块如下描述。

(1)全局管理模块。该模块为一个守护代理，该代理接受各类型的消息并根据相关策略进行处理。例如，接收各个子网局部管理模块上报的两类消息(Alert 和 Attack-agent)，对 Alert 消息直接提交给 GUI 并交由全局关联分析模块记录在入侵资料总库；对 Attack-agent 消息则根据其内容修改 Agent 资料总库中相关代理的安全状态，然后发送 Refresh 消息通知安全检测代理；接受全局关联分析模块的 Globol-alert 类型的消息并上交 GUI；接受 GUI 的命令进行系统初始化等相关操作。

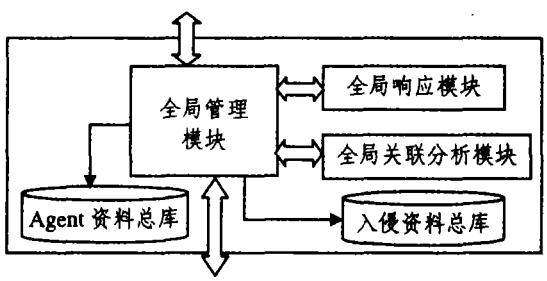


图 3 全局系统框架示意图

(2)全局关联分析模块。该模块对每条全局管理模块接收到的 Alert 进行全局关联分析处理并插入入侵资料总库，关联得出的报警及时通知管理员。

(3)全局响应模块。该模块也通过全局管理模块与 GUI 模块接口，在接到响应命令时派出相应的移动代理到各个子网处理。

3.3 安全检测代理

在每个子网中，局部管理模块起着承上启下的作用，一旦遭到破坏将使该子网的入侵检测功能丧失，故专门设计了一种用来监控各个局部管理模块安全状态的代理(SecurityCheckAgent)。该代理在创建后首先读取 Agent 资料总库中每个子网管理代理的相关资料，然后根据 Agent 资料来循环遍历各个子网局部管理模块并安给定策略访问局部管理代理，成功则访问下一个；失败则返回全局管理模块所在的主机将该局部管理代理在 Agent 资料总库的安全状态改为 false 并修改自身所带的相关资料，同时刷新 GUI 界面(通知管理员)，继续访问下一个。需要注意的是，当 Agent 资料总库中的资料经由全局管理模块修改后，全局管理模块代理需要向安全检测代理发送 Refresh 消息，安全检测代理接到该消息后重新读取 Agent 资料总库的资料，继续遍历。

总结 整个系统通过采用这种分层的入侵检测机制，充分利用移动代理技术构架了一个强大的入侵检测系统。与传统的入侵检测机制相比，能够避免大量的误警。同时整个入侵检测系统具有很强的自我保护机制。整个框架总体采用分布式结构，某一子网代理的故障或崩溃不会影响其他的子网和代理。

本文给出的这个入侵检测模型框架具有很好的可扩充性、很强的安全性和实用性以及比较低的误警率。目前这项技术正在进行开发，并且原型的初步测试表现出了较好的准确性和可靠性。

参 考 文 献

1 杨海松，李津生，洪佩琳. 分布开放式的入侵检测与响应架构—IDRA. 计算机学报，2003，26(9)

2 马恒太. 基于 Agent 分布式入侵检测系统模型. 软件学报，2000，11(10)：1312～1319

3 张勇，张德运，李胜磊. 基于分布式代理的网络入侵检测技术的研究与实现. 计算机学报，2001，24

4 Bernardes M C, Dos Santos Moreira E. Implementation of an Intrusion Detection System Based on Mobile Agents. IEEE, 2000

5 Helmer G, Wong J S K, Honavar V, Miller L, Wang Y. Lightweight agents for intrusion detection. Journal of Systems and Software, 2003, 67(2): 109～122

(上接第 299 页)

3 Homepage of UMBC Agent Web. http://www.cs.umbc.edu/kqml/

4 Homepage of IKF. http://wiki.di.uminho.pt/wiki/bin/view/IKF/WebHome

5 Brian J G, Dickson L. Knowledge Fusion. In: Proc. of the 7th Annual Workshop on Conceptual Structures: Theory and Implementation, Springer-Verlag Published, Heidelberg, 1992. 158～167

6 Bomberger N A, Waxman A M, Pait F M. Synchronization of Dynamic Networks for Knowledge Representation and Higher-level Fusion. In: Proc. of the 7th International Conference on Information Fusion, Stockholm, 2004. 227～234

7 Sawaragi T, Umemura J, Katai O, et al. Fusing Multiple Data and Knowledge Sources for Signal Understanding by Genetic Algorithm. IEEE Transactions on Industrial Electronics, 1996, 43(3): 411～421

8 Gregoire E. Syntax and Semantics in Knowledge Fusion: A Mixed Approach. In: Proc. of SPIE International Conference on Sensor Fusion: Architectures, Algorithms, and Applications, Orlando, 2002. 60～64

一种改进的基于差别矩阵的知识挖掘方法*)

A Method of Knowledge Mining Based on Discernibility Matrix

杨莉萍　陈仪香

(上海师范大学数理信息学院　上海 200234)

Abstract Knowledge mining is one of the most important problems in information system. In this paper, we advance a method of knowledge mining based on Rough Set Theory, and provide an improve algorithm which produces the Skowron Discernibility Matrix. After using this algorithm, we can abstract the important messages from decision table quickly and conveniently.

Keywords Knowledge mining rough set, Decision rule, Discernibility matrix

1 引言

粗糙集(Rough Set)是一种新的处理不精确、不完全和不相容知识的数学理论[1,2]。其中属性约简是重要的研究内容之一，而很多属性约简都是从核开始的，于是求核就成了属性约简的关键步骤。

为了求出决策表中的核属性，人们采用依次去掉决策表中的条件属性的方法，但是却总是被复杂的计算量所困扰，于是 HU 等提出了一种基于差别矩阵的求解核属性的方法[3]，该方法可以有效地减少计算量，提高求解核属性的效率，但是在某些情况下却不能得到正确的核，于是叶东毅等人改进了 HU 的差别矩阵[4]，并证明其求核方法是正确的，但该方法在定义差别矩阵中的每个矩阵元素时又增加了计算的复杂度。

于是本文中，我们对 HU 的差别矩阵生成过程进行改进，并提出相应算法。具体做法为：先将原始决策表分为完全相容和完全不相容决策表；其次分别利用本文所提出的算法对这两张子表进行分析，从中提炼出约简的决策规则，使得从决策表中挖掘出重要信息变得更快更方便。这个算法已成功地用在自适应智能形成的实现[5]。

2 相关概念

定义 1[2,3]　$S=(U,A)$为一信息系统，且 $C,D\subseteq A$，是两属性子集，分别称为条件属性和决策属性，且 $C\cup D=A, C\cap D=\phi$，则该信息系统称为决策表。

定义 2[1,2]　差别矩阵是粗糙集理论中一个非常重要的概念，用于信息表的属性约简。差别矩阵的定义如下：

设 S 为决策表，论域 $U=\{x_1,x_2,\cdots,x_m\}$是研究对象，条件属性 $C=\{c_1,c_2,\cdots,c_n\}$，决策属性 $D=\{d_1,d_2,\cdots,d_m\}$。定义差别矩阵元素为

$$m_{ij}=\begin{cases}① \ a\in C\ a(x_i)\neq a(x_j)\wedge D(x_i)\neq D(x_j)\\ ② \ \phi \text{ 上述条件不满足时}\end{cases}$$

决策表 S 的差别函数定义为

$\Delta=\prod_{(x,y)\in U\times U}\sum a^*(x,y)$

差别函数 Δ 的极小析取范式中的所有合取式是 C 的所有 D 约简。

其中的核属性定义如下：

$Core_D(C)=\{a^*\in C\mid f(x,a)=\{a\}$，其中于 $x,y\in U\}$

利用差别矩阵来表达知识有许多优点，它将信息表中关于属性区分的信息浓缩进一个矩阵中，可以用于信息表的属性约简。

3 改进的差别矩阵生成方法

通过分析，我们发现，如果根据差别矩阵计算后得到的知识约简不止一个的话，那么由不同的知识约简所得到的规则约简在大多数情况下是不同的。因为当存在不止一种知识约简，而人们又选择其中一种知识约简形式时，实质上已经人为删去了某些条件属性，即此时已经添加了人为的因素在其中，而在通常情况下，我们总是希望能尽量减少人为干预，最大限度地挖掘出原始数据中的重要信息。而在原来的知识约简方法中，除非将在不同的约简属性集基础上得到的决策规则集合进行合并，否则丢失某些信息是必然的，而这样就又会增加不少的运算量。

于是在本文中，我们设计了针对 Skowron 差别

*)本文得到了国家自然科学基金(60273052)以及上海市教委研究基金(05DZ06)的资助。杨莉萍　硕士研究生，研究方向：知识智能处理。陈仪香　博士生导师，研究方向：软件理论，知识智能处理。

矩阵生成过程的改进，希望能从原始的数据中挖掘出尽可能多的信息，而又尽量不丢失重要信息。

在大多数的文献中，差别矩阵的生成是用于知识约简，对规则的约简则无能为力，但是如果我们对差别矩阵做进一步深入的分析，却发现我们完全可以类似地利用差别矩阵进行属性约简的方法应用于属性值的约简，思路为将差别矩阵中第 i 行进行属性约简处理，而后将属性用第 i 行记录对应的属性值替换，即得属性值约简，该方法与基于差别矩阵属性约简的差别在于：属性约简在整个差别矩阵中找属性的最小组合，而属性值约简是在差别矩阵的每一行中找属性最小组合，从而保证了约简后的规则与所有其它对象的规则不产生冲突。

由于当差别矩阵应用于不相容决策表时可能会造成核属性求解错误，于是本文中，我们先判断：若原始决策表为不相容决策表，则首先将决策表分为完全相容和完全不相容决策表。我们先分析完全相容决策表。

在此基础上，我们经过认真分析与实践，研究出具体的算法如下：

算法 Reduction′：

输入：决策表，设有 n 条记录，m 个条件属性）；

Step1：令 i=0；

Step2：依次比较第 i 条记录的决策属性值与第 j(j=i+1,…,n)条记录的决策属性值，若决策属性值相同，则 aij =ϕ，否则转 Step3；

Step3：依次比较第 i 条记录的条件属性值与第 j(j=i+1,…,n)条记录的条件属性值，若对应的条件属性值不同，则记录下该属性，直到比较完所有的条件属性，将该元素记为 aij，转 Step4；

Step4：比较 aij 是否包含于第 i 行中的某个元素 amn，若是，则删除 amn，若第 i 行中的某个元素包含于 aij，则删除 aij，若都不是，则保留 aij，转 Step5；

Step5：若 j<n，则令 j=j+1，转 Step1，生成第 i 条记录的下一个元素，否则转 Step6；

Step6：若 i<n，则令 i=i+1，转 Step1，生成差别矩阵的下一行元素，否则，转 Step7；

Step7：差别矩阵的所有元素生成结束；

Step8：删除所有行中元素均为空的行；

Step9：令 i=1(指向第 1 行)，Di=ϕ；

Step10：若第 i 行中含有度为 1 的元素，则将其加入 Di 中，转 Step11；否则，转 Step12；

Step11：检测在第 i 行中是否还有其它元素，若没有，转 Step13，若有，则转 Step12；

Step12：统计第 i 行中其余属性在该行中不同元素中出现的次数，找出出现频率最高的属性，使其加入 Di 中，并删除该行中出现该属性的元素，若此时第 i 行的所有元素为空，则转 Step13，否则，转 Step12；

Step13：则第 i 条记录即可以由 Di 来表示；

Step14：若 i 没有指向最后一条记录，则令 i=i+1，Di=ϕ，转 Step10，否则，结束；

输出：决策表中已经经过约简的决策规则。

此算法中，我们只要生成一次差别矩阵，就能实现单条规则的所有约简。对于一些不必要的元素马上删除，这样就避免了无意义的比较，节省了大量的时间和存储空间，而对结果的正确性却没有丝毫影响。

以我们的一自适应智能体形成为例[6]，有如下一决策表。

U	狼数量(a)	总诱惑力值(b)	总压力值(c)	HP(d)	成功(e)
1	5	10	20	40	否
2	5	10	20	20	否
3	0	10	20	20	是
4	0	10	20	35	否
5	5	60	20	20	是
6	2	25	20	30	是
7	2	25	20	30	是

从差别矩阵的生成过程我们可以看出其是对称矩阵。所以人们一般采用上三角或下三角矩阵的形式来表述，为了论述方便，我们将差别矩阵填充完整，如下所示。

U	1	2	3	4	5	6
1	ϕ	ϕ	a,d	ϕ	b,d	a,b,d
2	ϕ	ϕ	a	ϕ	b	a,b,d
3	a,d	a	ϕ	d	ϕ	ϕ
4	ϕ	ϕ	d	ϕ	a,b,d	a,b,d
5	b,d	B	ϕ	a,b,d	ϕ	ϕ
6	a,b,d	a,b,d	ϕ	a,b,d	ϕ	ϕ

接下来，我们采用本文改进的算法 Reduction′，得最终的差别矩阵，见下表。

U	1	2	3	4	5	6
1	ϕ	ϕ	a,d	ϕ	b,d	ϕ
2	ϕ	ϕ	a	ϕ	b	ϕ
3	ϕ	a	ϕ	d	ϕ	ϕ
4	ϕ	ϕ	d	ϕ	ϕ	ϕ
5	ϕ	B	ϕ	ϕ	ϕ	ϕ
6	a,b,d	ϕ	ϕ	ϕ	ϕ	ϕ

我们由上表，可直接得到原始决策表的所有决策规则的约简形式为：

狼数量(5) ∧ 总诱惑力值(10)→成功否

HP(40) ∨ HP(35)→成功否

狼数量(0) ∧ HP(20)→成功是

总诱惑力值(60) ∨ 总诱惑力值(25)→

(下转第 318 页)

基于确定性理论的不确定推理方法研究

Research of Uncertain Reasoning Based on Certain Theory

叶育鑫[1] 钟绍春[1] 赵瑞清[1,2]

(东北师范大学理想信息技术研究院 长春 130024)[1] (吉林大学计算机科学与技术学院 长春 130012)[2]

Abstract Uncertain Reasoning is a very important research area of artificial intelligence. So far, the research about uncertain reasoning is mainly used to build reasoning model which is based probability theory, fuzzy mathematics or rough theory. Certain theory is the best convenient and practical among all of them. Based on certain theory, evidence representation and knowledge representation about uncertain reasoning are both improved. And in this paper, a new reasoning arithmetic is given based on the above representation of evidence and knowledge. It is used to resolve some problems which exited in old reasoning arithmetic.

Keywords Uncertain reasoning, Certain theory, Important degree, Positive and negative proportion

1 引言

不确定推理是人工智能系统开发实现中的核心问题之一，是解决传统的专家系统知识的脆弱性和推理的单调性的一个重要手段，也是知识工程研究领域的重中之重。对不确定推理的研究一直以来就是计算机学科的一大热门课题。

目前流行的方法有 Bayes 网络(BN)、证据理论等，它们有各自的优缺点。BN 有严密的数学基础，可以精确地计算出推理结果的不确定性，但代价是必须获得关于推理过程以及推理中间信息的全部精确的相关知识；而且在推理过程中计算量非常大，是一个关于推理节点个数的 NP 类问题，在大规模推理网络中处理起来相当困难[1]。证据理论有极强的理论基础，可以表示主、客观信息，区分不确定和不知道，方便地定义各种问题，处理概率、模糊等不确定类型。但是，由于证据理论在不确定性推理中留下的空间太大，并且解释不一，从而得到各种不同的结论。它对于命题规则和命题的不确定性合成问题解决得不够完善，计算复杂度相当高。同时，证据理论推理中的信息合成结果的信任函数取值递减，且不易标准化。其推理过程由于缺乏推理节点之间的精确关联知识，因此在实际应用中尚有困难[2]。本文基于确定性理论给出了一种新的推理算法。它不但灵活简便，实用性强，同时也解决了原确定性理论自身的不足。

2 相关知识

E. H. Shortliffe 等人于 1975 年提出了基于确定性理论的推理模型，属于随机不确定性的一种。该理论在 MYCIN 医疗诊断系统中得以很好的应用[3]。它以可信度度量证据，用产生式作为知识的表示形式，并引入证据前件对结论的支持度与不支持度的思想来传递和更新不确定性。

2.1 证据表示

在精确推理中，前提为真或假，不允许不真不假的情况出现[4]。而在不确定推理问题中，前提或证据本身是不确定的，介于完全的真和完全的假之间。为了描述这种不确定性的程度，引入证据的可信度。

定义 1 证据 A 的可信度用 $CF(A)$ 表示，规定 $-1 \leqslant CF(A) \leqslant 1$。实际使用时，初始证据的 CF 值由专家根据经验给出，其他证据的 CF 通过规则进行推理计算得到。

2.2 知识表示

在逻辑推理过程中，常以 $A \rightarrow B$ 表示规则。其中 A 表示前件，B 表示结论或推论，是在 A 下的直接逻辑结果。确定性因子 $CF(B,A)$ 定义为 A 对 B 的信任与不信任二者之差[5]。

定义 2 $CF(B,A)=MB(B,A)-MD(B,A)$

其中，CF 是由证据 A 得到的假设 B 的确定性因子。MB 是由证据 A 得到的假设 B 的信任增加度量。MD 是由证据 A 得到的假设 B 的不信任增加度量。

2.3 不确定性的更新与传播

以 A 为证据，$A \rightarrow B$ 为规则的系统中，求 $CF(B)$；

I. 当 $CF(B)$ 的先验值未由专家给出时，设定值为 0，其更新值由式(1)给出：

$$CF(B)=\max(0,CF(A))\times CF(B,A) \quad (1)$$

II. 当 $CF(B)$ 的先验值已由专家给出时，要根据证据 A 的值，分情况计算 $CF(B)$ 的更新：

叶育鑫 硕士研究生，研究方向：人工智能、知识工程及不确定性推理。钟绍春 博士生导师，教授，研究方向：智能 Agent 系统，信息融合等。赵瑞清 教授，研究方向：人工智能，专家系统等。

(a)当 $0<CF(A)\leqslant 1$ 时，即 A 必然发生($CF(A)=1$)或 A 可能发生时：

$$CF(B|A)=\begin{cases}CF(B)+CF(A)\times CF(B,A)(1-CF(B)), & CF(B)\geqslant 0,\ CF(B,A)\geqslant 0\\ CF(B)+CF(A)\times CF(B,A)(1+CF(B)), & CF(B)<0,\ CF(B,A)<0\\ \dfrac{CF(B)+CF(A)\times CF(B,A)}{1-\min\{|CF(B)|,|CF(A)\times CF(B,A)|\}}, & \text{其他}\end{cases} \tag{2}$$

(b)当 $CF(A)<0$ 时，即 A 不可能发生时：

规则 $A\rightarrow B$ 不使用，即认为不可能发生的事件对结果 B 没有影响，此时 $CF(B)$ 的值保持不变。

III. 如果有两个规则同时分别作用于同一假设时，对假设的更新采用如下公式：

$$CF(B)=\begin{cases}CF_1(B)+CF_2(B)-CF_1(B)\times CF_2(B), & CF_1(B)\geqslant 0,\ CF_2(B)\geqslant 0\\ CF_1(B)+CF_2(B)+CF_1(B)\times CF_2(B), & CF_1(B)<0,\ CF_2(B)<0\\ \dfrac{CF_1(B)+CF_2(B)}{1-\min\{|CF_1(B)|,|CF_2(B)|\}}, & \text{其他}\end{cases} \tag{3}$$

当有 n 个规则同时作用时，反复使用上式进行计算。

3 问题提出

确定性理论很好地解决了引言中其它推理算法存在的一些问题，它简单直观，容易理解，使用方便；并且该方法有线性的空间和时间复杂度，推理的近似效果较好[6]。但就其本身还有以下两点不足：

(1)当证据是多个条件“与”关系时，没有考虑各个证据之间的权重，默认为它们之间是同等重要的。这是不符合实际情况的[7]。

(2)一个不支持假设证据的影响可以压倒多个支持证据的影响，反之亦然。这是不符合我们直觉的。

如何解决上述两个问题，是确定性理论急需改进的地方。下面分别从证据的不确定性表示、知识的不确定性表示来改进，并提出一种新的算法来解决不确定性的更新与传播。

4 问题解决

4.1 证据的不确定性度量

定义 3 若一个规则的前件为复合命题，其中 and 关系的命题个数为 n 个，为每个命题分配一个 [0,1]之间的实数 IA，称为该命题的权重。它满足：$\sum_{i=1}^{n} IA_i=1$。

该定义理解为：IA_i 为在前件复合命题里，第 i 个命题在 n 个 and 关系中的重要度。例如：如果 $IA_i>IA_k$，表示该规则的结论要成立，则第 i 个命题比第 k 个命题要重要。它是在建立知识系统时，由领域专家给出的。

如图 1 给出，在规则中，前件由 E_1、E_2、E_3 三个命题的复合命题组成，并且它们之间是“and”关系。其中：三个命题的权重分别设定为 $IA_1(E_1)=0.2$；$IA_2(E_2)=0.2$；$IA_3(E_3)=0.6$。

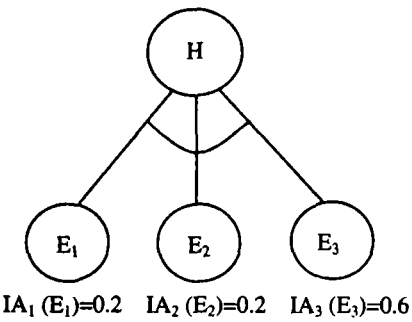


图 1

满足定义 1 中的 $\sum_{i=1}^{n} IA_i=1$。在图 1 中，E_3 的权重大于 E_1 和 E_2。

定义 4 用二元组 $(CF(A),IA(A))$ 来表示证据。其中：$CF(A)$ 为证据 A 的信任度，$(-1\leqslant CF(A)\leqslant 1)$；$IA(A)$ 为该证据在复合前件中的权重。当证据为“and”关系中的一个命题时，$IA(A)$ 的值得确定满足定义 3，否则证据在“or”关系中或为单一命题，此时设置 $IA(A)$ 的值为 1。

4.2 规则的不确定性度量

为解决在多个规则对同一假设合成中，当结论支持与不支持的命题比例失调时，结论可信度受扰动较大的问题，我们在规则的度量上，引入两个计数器，其中一个用于记录对假设支持的规则个数，一个用于记录对结论不支持的规则的个数。

定义 5 用三元组 $(CF(B,A),C_{MB},C_{MD})$ 来表示一条规则。其中：$CF(B,A)$ 表示当证据 A 为真时，规则的可信度；C_{MB} 表示对假设 B 支持的规则个数，C_{MD} 表示对假设 B 不支持的规则的个数，它们都是正整数。

特别地，当在知识系统中，所有的 C_{MB}，C_{MD} 的初始值均设置为 0，在系统每次运行时，由系统自动更新 C_{MB}，C_{MD}。这样动态更新 C_{MB}，C_{MD} 两个参数，就避免了在知识库做必要的规则修改时，同时要做这两个参数的统计静态修改。它在实际应用中，降低的系统的工作难度，提到了工作效率。

4.3 证据的更新与传播算法

在推理过程中，首先不可避免地要对证据前件进行综合计算。这主要是指对复合前件的合成问题。其中包括对复合前件中“与”、“或”、“非”关系的

计算。

1)“与”计算：

$A_1$and $A_2 \rightarrow B$

$CF(A_1 \text{and } A_2)=$

$\min\{CF(A_1)\times IA_1, CF(A_2)\times IA_2\}$

2)“或”计算：

A_1or $A_2 \rightarrow B$

$CF(A_1 \text{or } A_2)=$

$\max\{CF(A_1)\times IA_1, CF(A_2)\times IA_2\}$

3)“非”计算：

$CF(\sim A)=-CF(A)$

通过上述计算，将规则中的复合命题前件计算出单值可信度，然后将其视为单一证据前件，用本文第2部分中的式(1)和式(2)，按情况计算出假设的更新值CF(B)。

在求出多条规则的对同一假设的更新可信度之后，我们还要考虑如何合并多个更新值，得到最终可信度的度量。其方法如下：

1)将多条规则按如下二叉树结构表示：

根结点为最终合成的可信度值。将假设值为正值的规则放置在左侧分支树的节点中，将假设值为负值的规则放置在右侧分支树的节点中。每一层的两个亲子节点中，左侧放置一个规则的可信度值，右侧放置下一层传递上来可信度更新值，并设置初始值为零。每分配一个规则的可信度值给二叉树，计数器C_{MB}，或C_{MD}的值减1，直到C_{MB}，与C_{MD}两个计数器的参数值都为零时，分配完毕，二叉树构建完成。

2)从叶子结点向上分别计算正值和负值的可信度，一直计算到根结点的下一层为止：

①根结点左侧的正值分支树自叶子结点向上，应用如下公式进行计算：

$$CF_{父}(B)=CF_{左}(B)+CF_{右}(B)-CF_{左}(B)\times CF_{右}(B) \quad (4)$$

特殊地，当右结点为0时，公式简化为$CF_{父}(B)=CF_{左}(B)$

每应用公式计算一次，同时将计数器C_{MB}数值加1。

②根结点右侧的负值分支树自叶子结点向上，应用如下公式进行计算：

$$CF_{父}(B)=CF_{左}(B)+CF_{右}(B)+CF_{左}(B)\times CF_{右}(B) \quad (5)$$

特殊地，当右结点为0时，公式简化为$CF_{父}(B)=CF_{左}(B)$

每应用公式计算一次，同时将计数器C_{MD}数值加1。此阶段计算完毕时，计数器C_{MB}和C_{MD}累计回初始统计值。

3)计算根结点的最终可信度。

定义7 在规则的合成中，将假设可信度值为正值的条件个数与假设可信度值为负值的条件个数的比值称为规则合成的正负比率γ。

其中$\gamma=C_{MB}/C_{MD}$；它反映了在知识库中对同一假设，所有支持规则和不支持规则的比例。在定义中值得注意的是，这里的正负比率中的“正负”分别代表的是“支持规则和不支持”的含义，而并不是传统意义上的“正数和负数”，所以其结果也不可能出现负数。它的值域为$(0,+\infty)$，是一个非负数。

计算γ值，当正负比率大于等于1时，应用如下公式进行计算：

$$CF_{根}(B)=CF_{左}(B)+CF_{右}(B)\div(C_{MB}/C_{MD}) \quad (6)$$

当正负比率小于1时，应用如下公式进行计算：

$$CF_{根}(B)=CF_{左}(B)+(C_{MB}/C_{MD})+CF_{右}(B) \quad (7)$$

特殊地，当正负比率γ等于1时，计算公式简化为：

$$CF_{根}(B)=CF_{左}(B)+CF_{右}(B)$$

综上，得出根结点的信任度。该方法有效地解决了复合证据“与”关系中，各证据的权重不同的问题；以及同时证据中存在一个不支持假设证据的影响可以压倒多个支持证据的影响问题。

结束语 本文基于确定性理论给出了一种新的推理算法，解决了原确定性理论自身的不足，该算法使用方便，易于理解。确定性方法的宗旨并不是理论上的严密性，而是处理实际问题的可用性，这一点在本文的算法中充分体现出来。但是，该方法无法解决确定性理论本身，当证据不存在时或对证据的不确定性不知道时，证据对假设的影响就无法确定。这是确定性理论本身无法逾越的。为此，我们要研究比满足确定性理论更弱的公理系统——证据理论。证据理论的理论性强于本算法，但实用性远远不够，有关它的研究正在探索中。

参 考 文 献

1 Boyen X, Koller D. Approximate learning of dynamic models. In: Kearns M S, Solla S A, Cohn D A, eds. Proceedings of the 11th Annual Conference on Neural Information Processing Systems (NIPS'98). Cambridge, MA: MIT Press, 1998. 396～402

2 Zhang Yao-ting, Du Jin-song. The Probability Approach inArtificial Intelligence. Beijing: Science Press, 1998(in Chinese)

3 Weiss S M, Kulikowski C A. 专家系统设计实用指南. 宫雷光，陈守礼译. 长春：吉林大学出版社

4 Sowa J F. Knowledge Representation. Beijing: China Machine-Press, 2003

5 蔡自兴，徐光. 人工智能及其应用. 北京：清华大学出版社，2004. 102～104

6 刘洁，陈小平，等. 不确定信息的认知结构表示、推理和学习. 软件学报，2002，13(4)：649～651

7 赵瑞清. 广义规则表示及其推理算法. 计算机学报，1992. 120～127

带重要度可信度框架规则知识表示及其模糊推理算法

Frame and Rule-based Knowledge Representation with Important Degree and Credible Degree and its Fuzzy Reasoning

李俊玲[1] 周东岱[1] 钟绍春[1] 赵瑞清[1,2]

(东北师范大学理想信息技术研究院 长春 130024)[1] (吉林大学计算机科学与技术学院 长春 130012)[2]

Abstract A general knowledge representation can be represented by rules and frame, but the two methods have some defaults. A new method is a combination of rule-based and frame-based representation. In this paper, it introduces two concepts: important degree and credible degree, and it can be called frame and rule-based knowledge representation with important degree and credible degree. It also gives fuzzy reasoning.

Keywords Knowledge representation, Fuzzy reasoning, Important degree, Credible degree

1 引言

知识表示是人工智能中一个重要研究内容,它不仅是知识工程的关键问题,而且还是其他 AI 研究中的重要问题[5]。人类之间互相交往都是用自然语言来描述和表达的,要让计算机来理解和推理,就必须将自然语言知识化,变成计算机能使用的形式,因此,如何准确地表达人类的知识是知识表示的首要要求。

目前的知识表示方法有规则表示法,框架表示法、语义网络表示法等等,但是这些表示方法都有不足之处。例如规则表示法中,由于它是建立在因果关系的基础上的,它表示的知识有一定的格式,且规则之间不能直接调用,因此那些具有结构关系或层次关系的知识不易用它来表示。又如框架表示法,虽然它能够表示具有结构关系或层次关系的知识,但是它的最主要的问题是缺乏形式理论,没有明确的推理机制保证问题求解的可行性。文[3]把这两种知识表示方法结合起来,提出了框架规则知识表示法。

在框架规则知识表示法的基础上,本文考虑到在实际应用中,前提中每个框架作用不是同等重要的,例如在图 1 中,体积和密度对物体重量的影响,往往可能密度对重量影响的程度比体积对重量影响的程度大,比如相同体积的泡沫塑料和钢材,当然是钢材重了。另外对于每条知识,它的可信程度往往是不相同的。基于以上所述,本文提出了带重要度可信度框架规则知识表示法,在其中提出了"重要度""可信度"等概念,并给出其相应的模糊推理算法。

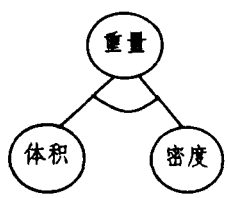


图 1

2 重要度和可信度

定义 1 一框架规则的前提条件中 and 的个数 n 称为分支数,如图 1 中分支数为 2。

定义 2 每个分支赋以一个实数 $IM(0<IM\leqslant 1)$,称之为该分支的重要度。它满足

$$\sum_{i=1}^{n} IM_i=1$$

其中 n 为该规则的分支数。

它们的意义是:IM_k 表示该框架规则的第 k 个分支在前提中所占的重要程度。例如,如果 $IM_i>IM_k$,表示该框架规则的结论要成立,则第 i 个分支比第 k 个分支重要。它是在建立该知识系统时,由领域专家给出的。

例如图 2,其分支数为 4。而在这一规则中,条件 E_3 的重要度比 E_1、E_2 及 E_4 的重要度大,即在此规则中,条件 E_3 对结论 H 的成立比 E_1、E_2 及 E_4 要重要一些。

定义 3 在专家系统中知识的不确定性一般由领域专家给出,通常是一个数值,它表示相应知识的不确定性程度,称之为知识的静态强度,一般用 RC 表示。

定义 4 证据的不确定性通常也用一个数值表

李俊玲 硕士研究生,研究方向:人工智能、知识工程及不确定性推理。周东岱 教授,研究方向:信息融合、GIS 开发及应用。钟绍春 博士生导师,教授,研究方向:智能 Agent 系统,信息融合。赵瑞清 教授,研究方向:人工智能,专家系统。

示。它代表相应证据的不确定性程度，称之为动态强度，一般用 CF 表示。

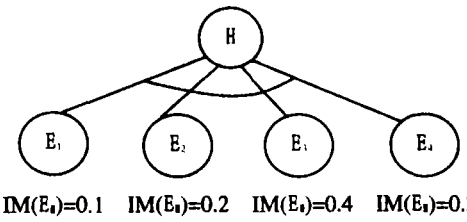


图 2

3 知识表示

我们将采用如下方法来表示知识：

$$Rule\left(\begin{matrix}E_1/u_1/P_1/IM_1/CF_1,E_2/u_2/P_2/IM_2/CF_2,\cdots,\\E_n/u_n/P_n/IM_n/CF_n,H/u_h/CF_h/RC/\lambda\end{matrix}\right)$$

Rule 代表框架规则，其中，$E_i(i=1,2,\cdots,n)$表示框架规则的前提条件(证据)，它是本框架规则内的一个槽名，也可以是其子框架的框架名；$u_i(i=1,2,\cdots,n)$表示对应前提的发生程度，是模糊量词，取值于[0,1]之间；$P_i(i=1,2,\cdots,n)$表示框架规则的前提条件之间的关系，值为 1 时，表示该前提与框架规则中其他前提条件是 AND 关系，值为 0 时表示 OR 关系；$IM_i(i=1,2,\cdots,m)$表示 $P_i=1$ 的那些框架规则前提条件在框架规则中所占的重要程度，当 $P_i=0$ 时 IM_i 默认为 0；$CF_i(i=1,2,\cdots,n)$表示该框架规则前提条件的可信程度；H 表示规则的结论；u_h 表示对应结论的发生程度，是模糊量词，取值于[0,1]之间；CF_h 表示结论的可信程度；RC 表示在前提条件下结论成立的可信程度，是规则的静态强度；λ 表示规则的可用阈值，一般由专家给出。框架规则中每个前提条件及其结论在图中可用节点来表示。

例如图 2 之框架规则可以表示为

$$Rule\left\{\begin{matrix}E_1/0.6/1/0.1/0.5,E_2/0.5/1/0.2/0.6,\\E_3/0.8/1/0.4/0.8,E_4/0.9/1/0.3/0.4,\\H/0.6/0.8/0.6/0.7\end{matrix}\right.$$

其中，$E_i(i=1,2,3,4)$是框架规则的前提条件；$u_i(i=1,2,3,4)$表示对应前提条件的发生程度，分别是 0.6、0.5、0.8、0.9；$P_i=1(i=1,2,3,4)$表示框架规则的前提条件之间是 AND 关系，且它们的重要度分别是 0.1，0.2，0.4，0.3，可信度是 0.5，0.6，0.8，0.4；H 为框架规则的结论，其发生程度为 0.6、可信度为 0.8，静态强度是 0.6、规则的可用阈值为 0.7。

4 推理机制

4.1 综合匹配度的计算

(1)首先根据以下公式分别计算框架规则前提条件的语义距离：

设 E 与 E' 分别是论域 $U=\{u_1,u_2,\cdots,u_n\}$ 上表示相应模糊概念的模糊集，则它们之间的语义距离定义为：

$$d(E_i,E'_i)=\frac{1}{n}\sum_{i=1}^{n}|\mu_{E_i}(u_i)-\mu_{E'_i}(u_i)| \quad (1)$$

其中，$\mu_{E_i}(u_i)$是隶属函数，表示元素 u_i 隶属于 E_i 的程度；

(2)分别计算出相应前提条件的匹配度 $\delta_{match}(E_i,E'_i)=1-d(E_i,E'_i)$ (2)

(3)则其规则的综合匹配度采用取“极小”方法，为

$$\delta_{match}(E,E')=\min\{\delta_{match}(E_i,E'_i)\mid i=1,2,\cdots,n\} \quad (3)$$

(4)若 $\delta_{match}(E,E')\geqslant\lambda$，则框架规则的前提条件与证据可匹配；否则，不可匹配。

4.2 动态 CF 的传播

(1)AND 结点的 CF 传播算法

定义 5 称 CM 为带重要度的可信度，计算方法如下：

$$CM=CF\times IM$$

总可信度为：

$$CF(H)=\delta_{match}(E,E')\times RC\times\sum_{i=1}^{k}(CF_i\times IM_i)=\delta_{match}(E,E')\times RC\times\sum_{i=1}^{k}CM_i \quad (4)$$

(2)OR 结点的 CF 传播算法

$$CF(H)=\delta_{match}(E,E')\times\Big(\sum_{i=1}^{m}CF_i-\sum_{i\neq j}CF_i\times CF_j+\sum_{\substack{i\neq j\neq k\\ i\neq k}}CF_i\times CF_j\times CF_k-\cdots+(-1)^{m-1}\prod_{i-1}^{m}CF_i\Big) \quad (5)$$

从式(5)可知，对于框架规则只有两个前提条件的 $CF(H)=CF_1+CF_2-CF_1\times CF_2$

对于更多前提条件的按式(5)进行类推即可得到。

4.3 实例分析

模糊知识库假设有如下两条规则：

$$Rule1\left(\begin{matrix}E_1/0.6/1/0.4/0.7,E_2/0.7/1/0.6/0.8,\\H_1/0.6/0.7/0.6/0.6\end{matrix}\right)$$

$$Rule2\left(\begin{matrix}E_1/0.7/1/0.6/0.8,E_2/0.3/1/0.4/0.7,\\H_2/0.8/0.5/0.7/0.6\end{matrix}\right)$$

已知证据 $E'_1/0.5/1/0.7/1,E'_2/0.4/1/0.3/1$

推理过程如下：

(1)①对 Rule1 根据式(1)和(2)分别计算 $d(E_i,E'_i)$和 $\delta_{match}(E_i,E'_i)$为

$d(E_1,E'_i)=0.1$　$\delta_{match}(E_1,E'_1)=0.9$

$d(E_2,E'_2)=0.3$　$\delta_{match}(E_2,E'_2)=0.7$

根据式(3)则其框架规则的综合匹配度为

$\delta_{match}(E_1,E'_1)=\min(0.9,0.7)=0.7$

(下转第 324 页)

模拟细胞生命活动的电子细胞*)

E-Cell to Simulate the Life Activity of Cells

卢欣华[1,2] 孙吉贵[1,2]

(吉林大学计算机科学与技术学院 长春 130012)[1] (符号计算与知识工程教育部重点实验室 长春 130012)[2]

Abstract The present paper provides a presentation of the definition of the E-Cell, its research significance and the existing models of E-Cell blueprints and makes out that E-Cell technique is a valid method to study the life activity inside the cell. With regard to the sketch manifestation, the existing models of E-Cell blueprints have disadvantage in the observation of the biochemical reactions inside the cell. In an attempt to cope with the problem, we have described Analog-Cell, a new model of E-Cell blueprint to simulate genetic expression, with the related biological theories briefly discussed. The paper points out that the advantage of Analog-Cell lies in having abundant picture information, and simulating the gene mutation on the molecule level. At last, the paper predicts the future research direction.

Keywords E-Cell, Genetic expression, Artificial life

1 引言

电子细胞(E-Cell, Electronic Cell)亦称虚拟细胞(Virtual Cell)、人工细胞,是人工生命的重要研究内容之一。它在计算机上构建一个虚拟的细胞体系,模拟真实细胞的结构、物质组成、生命活动的动力学行为和生命现象,通过图表或虚拟现实的方式进行人机交互,以便研究者构造细胞结构和其内外部环境物质组成,考察、记录细胞实验现象和功能,再现细胞生命活动,最终发现新的生物学现象和规律[1]。

电子细胞技术的研究及模型的建立有着重要的意义。由于基因组计划的实施,积累了大量的生物学数据,包括人类基因组约 30 亿个碱基对,以及其它上百个物种的基因序列。与此同时,分子生物学的发展使人们对细胞内的一些生物化学过程了解得极为透彻,从基因表达到跨膜信号传递,从细胞的能量产生与消耗到细胞周期的循环规律和细胞凋亡。但是无论对这些具体过程了解得多么透彻,人们仍然不能明白这些过程如何组合为一个整体在进行运作,对于细胞的生命活动规律仍然认识得很少。在这样的背景下,借助于人们对细胞已有知识的掌握和对积累的大量生物学数据的计算机分析,研究和构造细胞的计算模型,并对模型加以模拟、检测及修正,是深入研究并发现细胞生物机制的一种有效途径[1]。

本文描述的 Analog-Cell 模型是国内第一个电子细胞图形模型。它以真核细胞为对象,针对现有电子细胞模型图形显示薄弱、不利于观察细胞内生物化学反应的问题,以生动、形象的图形模拟基因表达的全部过程。该模型成功模拟了 DNA 序列根据酶促聚合作用原理,转录得到 mRNA,然后得到的 mRNA 又根据三联体密码子对应特定氨基酸的生物学原理,在 tRNA 转运氨基酸的前提下产生相应的多肽链,该多肽链是各种酶或其它调控因子的前体蛋白质,它将会最终影响整个细胞的活动。Analog-Cell 较之国外其它电子细胞模型,其优势一方面在于 Analog-Cell 模拟了其它模型所没有的基因突变,而基因突变是生物进化的重要手段之一;另一方面在于 Analog-Cell 含有更丰富信息的细胞活动影像,用户会清楚看到 DNA 转录为 mRNA、mRNA 翻译为多肽链的全部生物化学过程,有利于研究者观察细胞内的生命活动,总结与发现其生物学现象和规律。

2 电子细胞的研究现状

目前国外有三种较成熟的电子细胞模型,而国内只有本文描述的一种电子细胞模型 Analog-Cell。最早的是日本 Keio 大学 Masaru Tomita 教授等人设计的原核细胞能量代谢模型 E-CELL[1]。E-CELL[2,3]是一个反映细胞内生物化学反应的模型及仿真环境,它对细胞内与代谢过程密切相关的物质与能量建立了数学模型。它以原核生物支原体(目前已知拥有最小的染色体)中与糖代谢有关的

*)基金项目:国家自然科学基金(批准号:60473003)、教育部高等学校博士学科专项科研基金(批准号:20050183065)、吉林省科技发展计划项目(批准号:20040526)。卢欣华 博士研究生,从事人工生命、电子细胞的研究。

127 个基因为对象，成功模拟了细胞内基因转录与翻译、能量产生及消耗、糖与磷脂的合成和代谢等生化机制。用户通过计算机的模拟，能够观察到蛋白质、蛋白质复合物及细胞内其它化学反应物的浓度变化[4]。

第二个较成熟的模型是美国康涅狄格州州立大学 Leslie M. Loew 和 James C. Schaff 等学者设计的真核细胞钙转运模型 Virtual Cell[5,6]。Virtual Cell 是一个用户可以自定义细胞组成结构和内部动力学模型的虚拟细胞应用演示软件。利用该虚拟细胞可以完成离子通道的钙离子流、RNA 的转运、线粒体的作用和细胞核膜的作用等一系列关于真核细胞的生物学活动和功能[1]。Virtual Cell 与 E-CELL 的不同之处除了以真核细胞为研究基础之外，还在于它通过简单的彩色图形模拟钙离子流的移动过程，可以对细胞进行结构学和形态学上的研究。

第三个较成熟的模型是由美国印第安那大学化学系的 Peter J. Ortoleva 教授等人开发的一套模拟细胞对外界刺激反应、物质运输及基因组变化的虚拟细胞模型 CyberCell[7]。该模型阐明了细胞核、线粒体等特定功能区域的生化过程及它们之间的物质运输过程[8]。CyberCell 主要应用于药物的研发与治疗手段的优化、细胞基础理论研究和生命起源的探索等方面。

第四个模型 Analog-Cell 是国内第一个电子细胞图形模型。它成功模拟了 DNA 序列转录得到 mRNA，mRNA 翻译得到相应的多肽链这一基因表达过程，该多肽链是各种酶或其它调控因子的前体蛋白质，它将会最终影响整个细胞的活动。Analog-Cell 的优势在于含有丰富的细胞活动影像，并且模拟了其它模型所没有的基因突变。

3 基因表达的生物学原理

电子细胞技术综合了细胞生物学、分子生物学、生物化学、计算机科学、数学等多门学科的理论知识，并且一定是以生物学的理论、数据为基础和前提，所以在对细胞内基因表达过程模拟以前，必须先对基因表达的生物学基本原理进行深入的了解。

3.1 DNA 转录为 mRNA

为了使细胞能更好地利用其基因组中包含的生物学信息，代表信息单位的各个基因必须协调表达。这种协调表达决定了细胞中 RNA 的组成，后者限定蛋白质组的性质，并最终决定细胞行使的功能[9,10]。

DNA 转录为 mRNA 是基因表达的第一步。以 DNA 为模板，根据互补方式合成 RNA，把 DNA 上所带的遗传信息传给 RNA，这一过程称为转录[10]。任何一个基因的转录过程都可分为两个阶段（图 1）。

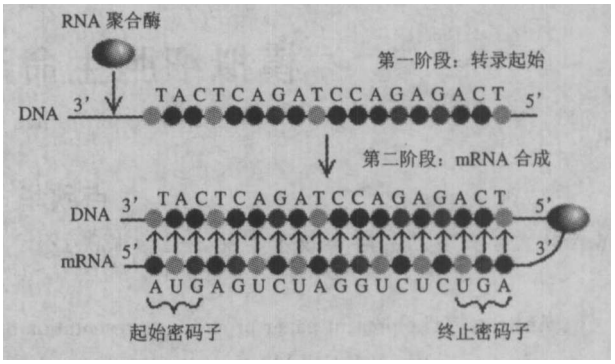


图 1 转录的两个阶段

第一阶段是转录起始，指蛋白复合物在基因上游的组装，蛋白复合物以 RNA 聚合酶为主要成分，负责将 DNA 转录成 RNA。基因能否真正表达由该步骤决定，也就是说 RNA 聚合酶不结合到 DNA 基因的上游区域，不形成蛋白复合物，该基因无法表达。这里 RNA 聚合酶对每个基因的表达调控进行了活性调节。

转录的第二阶段是依赖 DNA 模板（A、T、C、G 序列）的 RNA 合成，指 RNA 聚合酶沿基因移动合成 mRNA，即基因的直接拷贝。核糖核苷酸（A、U、C、G）被一个一个地加到 RNA 转录物的 3'末端，形成由核糖核苷酸构成的 mRNA 链。哪个核糖核苷酸进行连接由碱基配对原则决定：A 与 T 或 U 配对；G 与 C 配对。该基因转录完毕后，RNA 聚合酶脱离 DNA 和 mRNA，mRNA 在细胞核内游动，最后游出细胞核，在细胞核外即胞浆中准备合成多肽链，进行基因表达的第二步——翻译过程。

3.2 mRNA 翻译为多肽链

mRNA 翻译为多肽链是基因表达的第二步。以 mRNA 为模板，根据三联体遗传密码合成蛋白质的过程称为翻译。而三联体遗传密码是指从 mRNA 的 5'末端开始，相连的三个核苷酸决定一个特定的氨基酸[10]。因此由遗传密码来决定 mRNA 序列被翻译成何种多肽链。遗传密码表[9,10]在相关的生物学书籍上都可以找到。翻译过程可分为起始、延伸、终止三个阶段。

起始阶段首先是核糖体结合在 mRNA 的 5'末端的起始密码子（AUG）处。核糖体类似于加工工厂，给翻译过程提供了空间。根据遗传密码表，起始密码子对应甲硫氨酸（Met），则负责转运 Met 的 tRNA 把 Met 送到核糖体内。tRNA 是 mRNA 与正在合成的多肽链的信息及物理连接者，确保了氨基酸顺序通过遗传密码与 mRNA 的核苷酸序列相对应。为了携带氨基酸，tRNA 会被氨酰化，即与相应氨基酸进行化学连接。这样多肽链上的第一个氨

基酸 Met 被送到了核糖体内，起始阶段结束。

延伸阶段首先读取起始密码子后的第二个密码子，然后根据遗传密码表找到对应的氨基酸，负责转运该氨基酸的 tRNA 将此氨基酸送到核糖体上，并与第一个氨基酸(Met)形成肽键，同时与 Met 连接的 tRNA 完成使命从核糖体中脱离。核糖体读取完毕两个密码子，向 3'方向移动三个核苷酸，准备读取下一个密码子，延伸过程一直持续下去，一个一个氨基酸通过肽键连接逐渐形成多肽链，直到核糖体移动到终止密码子的位置，进入最后一个阶段——终止阶段。

当遇到三种终止密码子(UAA、UAG、UGA)之一时，蛋白质合成即告结束。翻译过程进入终止阶段。此时进入核糖体的不是 tRNA，而是释放因子这一蛋白质(eRF)。释放因子发出终止信号使多肽链从核糖体中释放出来，核糖体也从 mRNA 上脱离，进入胞浆中，等待下一次翻译再被启用，翻译的终止过程结束。

3.3 mRNA 的降解

mRNA 的降解是基因表达的最后一步。翻译成多肽链的 mRNA 已完成使命，为防止它再次被翻译为多肽链，在胞浆中该 mRNA 会被核糖核酸酶识别并降解，还原为游离状态的核糖核苷酸。至此，一次完整的基因表达过程结束。

4 Analog-Cell 的模拟结果

Analog-Cell 是一个在分子水平上模拟细胞内基因表达过程的图形显示模型，它已完成上一节叙述的基因表达的全部过程，采用 C^{++} 面向对象的方法实现计算机上的模拟，并参照 TimJ. Hutton 的 Squirm3 程序界面[11]，设计并完成细胞内基因表达的模拟过程。Analog-Cell 主要由顺序执行的三个过程组成：转录、翻译和降解。模型运行以后，一个细胞诞生于窗口内，整个窗口代表一个细胞，一条 DNA 模板链在具有棕色核膜的细胞核内游动，在胞浆中还有游离的核苷酸、核糖体、已连接氨基酸的氨酰 tRNA、核糖核酸酶等物质。

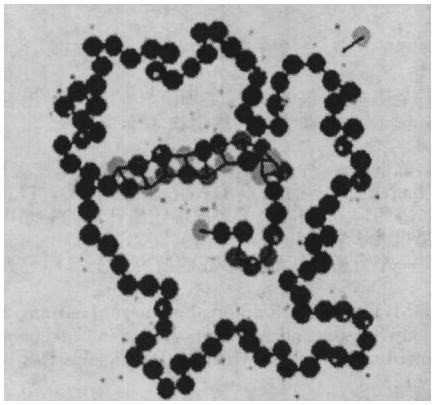


图 2 转录过程：mRNA 链正在产生

当在 RNA 聚合酶的作用下，基因上游结合到核膜上时引发转录的产生。按照前面介绍的转录过程原理，依据碱基配对原则，以 DNA 序列为模板，mRNA 链逐渐产生于细胞核内(图 2)。

当以 DNA 序列为模板的 mRNA 完全转录完毕后，DNA 和 mRNA 形成的双链分开。DNA 模板链继续在核内游动，而 mRNA 则需要游动出细胞核，在细胞核外即胞浆中准备合成多肽链，进行下一步的翻译过程。按照前面介绍的翻译过程原理，mRNA 在胞浆中与核糖体结合引发翻译的产生。随着核糖体的移动，依据三联体遗传密码识别出来的氨基酸通过肽键连接成多肽链，直到核糖体移动到终止密码子处为止(图 3)。

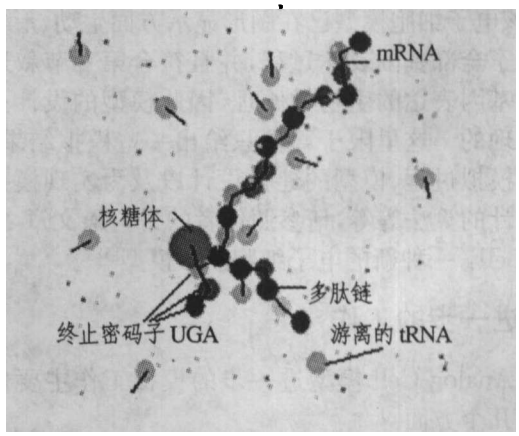


图 3 翻译过程：多肽链基本形成

核糖体移动到 mRNA 的 3'末端，识别出终止密码子，释放因子发出终止信号，使核糖体、多肽链、运送最后一个氨基酸的 tRNA 都与 mRNA 脱离。这时胞浆中可以看到完整的多肽链、已完成翻译的 mRNA、游离的 tRNA、独立的核糖体等物质(图 4)。

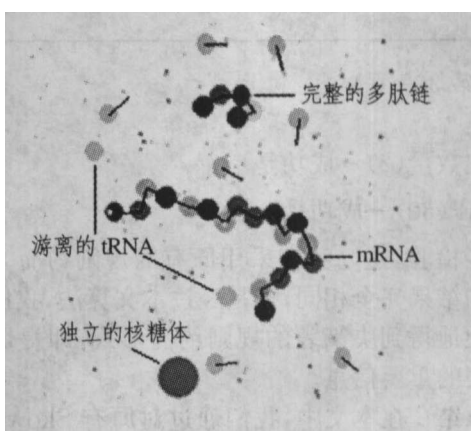


图 4 多肽链形成，翻译过程结束

基因表达过程模拟的最后一个步骤是 mRNA 的降解。翻译成多肽链的 mRNA 已完成使命，在胞浆中被核糖核酸酶识别并降解。当核糖核酸酶遇到

mRNA 时，会导致 mRNA 链的断裂，断裂片断逐渐消失，重新变成游离的核苷酸(图 5)。

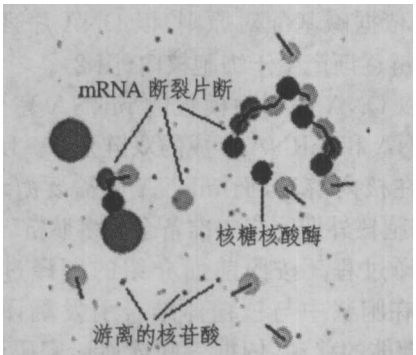


图 5　mRNA 被核糖核酸酶降解

由上述的 Analog-Cell 的基本执行过程可以看出，该电子细胞模型已在图形显示方面生动、形象地模拟了全部基因表达过程，并且符合第三节叙述的关于基因表达的生物学原理，因此模型的设计是完全合理的。这里限于篇幅只给出一些模拟结果，具体的模拟过程、模型的建模设计以及为实现模拟过程设计的算法等等，请参照作者的另一篇文章 Analog-Cell：一种新的电子细胞图形模型[12]。

5　进一步的工作

Analog-Cell 模型进一步的模拟工作主要包括以下几个方面：

(1)实现对基因表达过程中物质与能量产生及消耗的数据统计。目前为止，Analog-Cell 只是完成了基因表达图形方面的模拟，而电子细胞技术研究的最终目的是要通过模拟细胞内的生化过程来发现生物学的新规律，或者用细胞生命活动体现出来的生命特征来帮助计算机实现智能。那么对与生命现象相关的生化过程进行数据统计，并以曲线的形式直观地体现在窗口中供学者们研究，就变得格外重要。

(2)加入多种酶及调节因子，真实地模拟它们在真核细胞基因表达过程中如何控制反应的进行。目前 Analog-Cell 已实现一些主要酶及因子的调节作用，但仍然有许多因子没有加入到模型中来。如果能在模型中反映出所有与基因表达有关的因子的调节作用，对于进一步了解基因表达过程的调控将起到重要作用。

(3)实现用户接口。为了使 Analog-Cell 模型真正地可以用来观察细胞内某个基因表达的过程并发现其中的某些规律，应当实现用户接口使用户可以输入从生物学数据库提取的真核生物基因序列；可以自定义酶及调节因子的类型、数量与浓度；可以自定义基因突变发生的概率，考察因为基因突变导致基因表达为完全不同的蛋白质，进而对细胞产生的影响。

完成上述工作，Analog-Cell 就不再只是一个图形模型，而成为真正意义上的电子细胞，这将为 Analog-Cell 模型进一步模拟细胞内其它生物化学反应、发现生物学新规律提供一定的可能性。

参 考 文 献

1　赵明生，尚彤，孙冬泳，等. 电子细胞的研究现状与展望. 电子学报，2001，12，29(12A)：1740～1743
2　Tomita M. Whole-cell simulation：a grand challenge of the21st century. Trends in Biotechnology，2001，19(6)：205～210
3　http://www.e-cell.org/software/ecellsystem
4　Tomita M，Hashimoto K，Takahashi K，et al. E-CELL：software environment for whole-cell simulation. Bioinformatics，1999，15(1)：72～84
5　http://www.nrcam.uchc.edu/technology/modeling_process.html
6　Slepchenko B M，Schaff J C，Choi Y S. Numerical Approach to Fast Reactions in Reaction-Diffusion Systems：Application to Buffered CalciumWaves in Bistable Models. J Comput Phys，2000，162：186～218
7　http://cybercell.biochem.ualberta.ca/Research_New/Overview_New.html
8　陈源，李朝军. 虚拟细胞. 细胞生物学杂志(Chinese Journal of Cell Biology)，2004，26(3)：231～234
9　Brown T A. 袁建刚，等译. 基因组. 北京：科学出版社，2002
10　刘祖洞. 遗传学(下册). 第二版. 北京：高等教育出版社，1991
11　http://www.swarmagents.com/thesis/detail.asp? id=143
12　卢欣华，孙吉贵. Analog-Cell：一种新的电子细胞图形模型. 已投电子学报

(上接第 309 页)

成功是

狼数量(2)→成功是

HP(30)→成功是

经检验，这与我们采用原有的规则约简方法所得到的结果完全相同，于是经过本文算法，我们可以很方便地得到决策表的规则约简形式，即提炼出了决策表的重要信息。

总结　在本文中，我们通过对原有 Skowron 差别矩阵的生成过程进行改进，不仅最终同样能够达到决策表的知识约简和规则约简生成的目的，而最关键的是，我们通过对原有算法的改进，节省了大量的运算时间和数据存储空间，既又不丢失决策表中的任何信息，又保证了最终从决策表中所提炼信息的正确性。

对于完全不相容决策表，提取约简的决策规则的方法与完全相容决策表一致，在此我们就不多做赘述。

参 考 文 献

1　张文修，吴伟志，等编著. 粗糙集理论与方法. 科学出版社，2001
2　杨善林，倪志伟. 机器学习与智能决策支持系统. 北京科学出版社，2004
3　尹旭日，陈世福. 一种基于 Rough 集的缺省规则挖掘算法. Journalof Computer Research & Development，2000，37(12)
4　叶东毅，陈昭炯. 不相容决策表属性约简计算的一个可辨识矩阵方法. 福州大学学报(自然科学版)，2005，33(1)
5　杨莉萍. 一个自适应智能体形成的研究：[上海师范大学硕士学位论文]. 2006，5
6　Porta J M，Celaya E. ReinforcelearnernI Learning for Agents with Many Sensors and Actuators Acting in Categorizable Environlearnernts. Journal of Aritficial Intelligence Research，2005，23：79～122
7　Kirchner F. Q-learning of complex behaviors on a six-legged walking machine. Robotics and Autonomous System，1998，25：253～262

基于模型推理的系统修复与重用设计研究

System Repair and Redesign Using Model-based Reasoning

李占山[1,2] 王 涛[3] 孙吉贵[1,2] 寇飞宏[1,2]

(吉林大学计算机科学与技术学院 长春 130012)[1]

(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)[2]

(长春工业大学计算机科学与工程学院 长春 130012)[3]

Abstract Based on the Model-based diagnosis theories, this paper analyzes the relationship among the diagnoses and conflicts in the system being diagnosed and the replaced component system, investigates the replacement repair and reconfiguration, proposes the concepts of replacement repair and reconfiguration for the system being diagnosed, then, makes the best use of the effects of component replacement upon the observations of the system being diagnosed to characterize the repairs and reconfigurations. Based on those results described as above, the paper further explores the application of model-based diagnosis to product design, proposes the concepts of system redesign diagnosis and system redesign and so on.

Keywords Model-based diagnosis, Component replacement, Repair, System redesign

1 引言

基于模型的诊断是近年来人工智能领域一个十分活跃的研究方向之一,其正确性、完备性及可维护性等特点能够克服传统诊断系统的缺陷,因而显示出充满生机的诱人前景[1~3]。近年来,国外的一些研究人员把基于模型的诊断方法应用于重新配置问题[4]。其主要思想是把对系统新的功能要求看作约束条件或诊断中的观测值,利用已建立的系统模型求解需要改进的部分,但是这些工作局限于系统内部属性值的重新赋值,不能处理系统元件或结构需要改变的情况。本文使用元件替换的方法进行系统修复与重新配置问题研究,得到了一些结果。另外,制造领域的产品设计时,常常存在一个某种程度上类似要设计的产品作为进行新的设计的基础,因此设计过程常常是一个重用设计的过程。对于此问题虽有相关文献进行了描述[4],但并没有进行形式化描述工作,本文在诊断的框架下对此问题进行了探索,得到一些结果,提出了重用诊断、重用设计等相关概念,希望对产品的设计工作有所帮助。

2 系统的冲突与诊断

系统诊断的最终目的是恢复或重新配置系统,由于替换的双重作用,我们可以把系统的修复、替换测试与重新配置结合起来进行研究。

定义 2.1 设(SD, COMPS)是一系统,σ 为(SD, COMPS)的一替换,$\Delta \in HYP$ 是对(SD, COMPS, OBS)的候选诊断。当 $\Delta=\sigma$ 时,我们称此替换为相对 Δ 的完全替换。当 σ 只含有一个元素时,我们称为单元替换。

定义 2.2 设(SD, COMPS)是一系统,σ 为(SD, COMPS)的一替换,$\Delta \in HYP$ 是对(SD, COMPS, OBS)的候选诊断。如果相对于 Δ 的替换测试(σ, o)的结果是 o,那么我们就把替换 σ 称为相对于 Δ 的修复替换。

定义 2.3 对(SD, COMPS, OBS)的候选诊断 Δ 预测替换测试(σ, o)的结果是 o 当且仅当 $SD^{\sigma} \cup IN \cup \{AB(c) \mid c \in \Delta-\sigma\} \cup \{\neg AB(c) \mid c \in COMPS^{\sigma}-(\Delta-\sigma)\} \models o$,

也就是,假设 Δ 中的元件是系统所有故障元件,而系统中所有其他元件正常,在相同输入下替换系统的输出行为 o 一定成立(由于空间限制以下定理证明过程省略了)。

定理 2.1 如果 $\Delta \subseteq COMPS$ 是预测替换测试(σ, o)的结果是 o 的候选诊断,那么 $\Delta-\sigma$ 也是换元系统$(SD^{\sigma}, COMPS^{\sigma}, o)$的候选诊断。

定理 2.2 令 Δ 是对(SD, COMPS, OBS)的候选诊断。那么 Δ 预测替换测试(σ, o)的结果是 o 当且仅当 $\Delta-\sigma$ 与每一个从替换结果 o 生成的对换元系统$(SD^{\sigma}, COMPS^{\sigma}, o)$的极小冲突有交。

定理 2.3 如果 $\Delta \subseteq COMPS$ 是预测 $\neg o$ 的对(SD, COMPS, OBS)的候选诊断,那么 $COMPS-\Delta$ 是对$(SD^{\sigma}, COMPS^{\sigma}, o)$从 o 推出的一个冲突。从 o 推出的对$(SD^{\sigma}, COMPS^{\sigma}, o)$的每个极小冲突是 $COMPS-\Delta$ 的一个子集,其中 Δ 预测换元系统$(SD^{\sigma}, COMPS^{\sigma})$的输出结果是 $\neg o$ 的对(SD,

COMPS,OBS)的候选诊断。

定理 2.4 令 P 是对(SD,COMPS,OBS)的极小冲突集。当 $P\cap\sigma\neq\phi$ 时,若与 P 相关的系统输出 $o_P\subseteq o$ 异常,那么 P 是从 o 推出的对(SD$^\sigma$,COMPS$^\sigma$,o)某极小冲突的严格超集;当 $P\cap\sigma=\phi$ 时,P 是对(SD$^\sigma$,COMPS$^\sigma$,o)的极小冲突集;若与 P 相关的系统输出 $o_P\subseteq o$ 正常,那么 $P\cap\sigma=c_i$ 是原系统中故障元件。

定理 2.5 对(SD$^\sigma$,COMPS$^\sigma$,o)的任意极小冲突集是对(SD,COMPS,OBS)的极小冲突集或是从 o 推出的对(SD$^\sigma$,COMPS$^\sigma$,o)的极小冲突集。

给定定理 3.4 和 3.5,我们就会想知道对(SD,COMPS,OBS)的极小冲突集和对(SD$^\sigma$,COMPS$^\sigma$,o)的极小冲突集之间存在的关系。

推论 2.1 令 C_1 是对(SD,COMPS,OBS)的极小冲突集族,C_2 是对(SD$^\sigma$,COMPS$^\sigma$,o)的极小冲突集族。那么 C_1 可划分为 C'和 C_{11},C_2 可划分为 C"和 C_{21},其中 C"是 C'的子集;C_{21} 是从 o 推出的对(SD$^\sigma$,COMPS$^\sigma$,o)的极小冲突集族且对任意的 $c\in C_{11}$,c 是 C_{21} 中某 c'的严格超集。

定义 2.4 设(SD,COMPS)是一系统,$\Delta\in$ HYP 是对(SD,COMPS,OBS)的候选诊断。σ 是相对于 Δ 的完全替换。如果相对于 Δ 的替换测试(σ,o)的结果是 o,那么我们就把此替换 σ 称为相对于 Δ 的完全修复替换。

3 系统的重新配置

前面我们讨论了当系统发生故障时,通过替换测试能够使系统恢复正常。但在实际生产过程中,由于条件限制或者修理成本等因素,我们希望系统进行修复后能够实现特定的目标,这时我们有必要研究系统的重新配置问题。与修复类似,重新配置问题中也是在 Δ 涉及的元件行为反常这一假设下,通过替换测试(σ,o)找到一个满足预期行为的极小集合 σ。其中相对于 Δ 的替换测试(σ,o)中的 o 是换元系统的预期行为或可接受行为的一阶谓词公式的集合,因此可以如下定义相对于 Δ 的重新配置替换。

定义 3.1 设(SD,COMPS)是一系统,σ 为(SD,COMPS)的一替换,$\Delta\in$ HYP 是对(SD,COMPS,OBS)的候选诊断。我们把 σ 称为相对于 Δ 的重新配置替换,当且仅当 $\sigma\subseteq\Delta$ 是极小的替换集合使得 SD$^\sigma\cup$IN$\cup\{AB(x)\mid x\in\Delta-\sigma\}\cup\{\neg AB(x)\mid x\in COMPS-\Delta\}\cup o$ 是可满足的。

定理 3.1 设(SD,COMPS)是一系统,σ 为(SD,COMPS)的一替换,$\Delta\in$ HYP 是对(SD,COMPS,OBS)的候选诊断。如果 σ 是相对于 Δ 的重新配置替换,那么对于每个 $c_i\in\sigma$,SD$\cup$IN$\cup o\cup\{\neg AB(x)\mid x\in COMPS-\Delta\}\mid=\neg AB(c_i)$。

定义 3.2 设(SD,COMPS)是一系统,$\Delta\in$ HYP 是对(SD,COMPS,OBS)的候选诊断。σ 是相对于 Δ 的重新配置替换,(SD,COMPS,o)的一个冲突集是集合$\{c1,\cdots ck\}\subseteq\Delta$,使得 SD$\cup$IN$\cup o\cup\{AB(c1),\cdots,AB(ck)\}\cup\{\neg AB(x)\mid x\in COMPS-\Delta\}$不可满足。(SD,COMPS,o)的冲突集是极小的,当且仅当它不含冲突真子集。

定理 3.2 设(SD,COMPS)是一系统。$\sigma\subseteq\Delta$ 是相对于 Δ 的重新配置替换,当且仅当 σ 使得 $\Delta-\sigma$ 是(SD,COMPS,o)非冲突集的极小集。

定义 3.3[5] 设 C 是集合族,C 的一个碰集(hitting set)是集合 $\sigma\subset\bigcup_{s\in C}S$,使得对于 $\forall S\in C$,$\sigma\cap S\neq\phi$。一个碰集是极小的,当且仅当它不含碰集真子集。

定理 3.3 设(SD,COMPS)是一系统。$\sigma\subseteq\Delta$ 是相对于 Δ 的重新配置替换,当且仅当 σ 是对(SD,COMPS,o)冲突集的极小碰集。

修复是在假定 Δ 中元件发生故障条件下进行的一般修理,而重新配置是在假定 Δ 中元件发生故障条件下至少修理哪些元件才能使换元系统的输出达到特定的目标。利用上面的结果我们就可以指导系统的修复或重新配置。事实上,我们可以把上述方法进行推广,当系统没有故障时可以把这一思想应用于产品的升级换代设计上。

4 基于模型诊断框架下的产品重用设计

产品设计在企业的生产过程中占有重要的地位,它直接影响到产品的质量和成本,因此如何提高产品的设计质量和效率是设计人员必须面对的问题。当设计人员要设计一个新产品时,常常存在一些某种程度上类似要设计产品的已有产品,这些已有产品可作为进行新设计的基础,因此设计过程又可以看为一个重用设计的过程。那么对已有产品如何进行重新设计从而满足新的功能需求是设计人员应该解决的问题。本节利用前面的方法对此进行刻画。产品的重新配置是一类特殊的重用设计活动[6],对产品的重新配置不仅涉及到零部件的选择,而且涉及零部件之间的连接以及零部件本身的参数(属性)赋值问题。为了借助于诊断方法描述重用配置问题,我们仍然使用诊断中的一些术语。

从诊断的角度来看,元件是提供某些功能的部分系统,如果一个参数值的选择影响到配置的有效性,那么诊断系统可把此参数值看作异常行为的可能原因。这样从诊断的角度来看该参数值就变成了一个元件。

在一般的设计部门这些修改都是通过手工测算实现的,因此问题考虑得很难周全并且效率较低。

本节研究利用基于模型诊断的方法来刻画自动求解系统要修改的部分(问题解)。为了进行问题描述,我们给出一些概念。

定义 4.1 设计问题定义为逻辑句子的偶对(SD,SRS)其中 SD 是系统与领域知识描述,SRS 是与应用相关的原有功能需求的特定规范(specifications)。

定义 4.2 设计是组件 COMPS 与 ATTRS 中的部分文字集合 CONF,使得 SD∪SRS∪CONF 是相容的,并且 SD∪CONF|=SRS。其中 CONF 有形如 type(c_i,t_j)和 val(c_r,a_s,v_t)的文字构成,c_i,c_r∈COMPS。

这个定义表明设计不但要与领域知识和系统功能需求相容,而且必须能够推出系统功能。设计 CONF 是极小的当且仅当不存在其他设计 CONF′使得 CONF′⊂CONF。极小设计可以节约制造成本。

定义 4.3 重用设计问题定义为逻辑句子集的偶对(SD,SRS′),其中 SD 是系统与领域知识描述,SRS′是与应用相关的新的用户需求规范(specifications)。

定义 4.4(重用设计冲突) 对重用设计问题(SD,SRS′)的一个冲突 C 是组件 COMPS 和 ATTRS 中的部分文字的集合,使得 SD∪SRS∪{type(c_1,t_i),type(c_2,t_j),…type(c_n,t_k),val(c_i,a_j,v_k),…,val(c_r,a_s,v_t)}是相容的,并且 SD∪SRS′∪{type(c_1,t_i),type(c_2,t_j),…type(c_n,t_k),val(c_i,a_j,v_k),…,val(c_r,a_s,v_t)}是不相容的。

定理 4.1(重用设计诊断) 令(SD,SRS′)是重用设计问题,CS 是重用设计问题的所有极小冲突集合族。DIAG 是一个重用设计诊断当且仅当 DIAG 是 CS 中所有重用设计冲突的极小碰集。

显然,DIAG 中的文字就是原设计 CONF 中与 SRS′不一致的元件或属性值,于是通过这个定理我们能够确定原有系统要修改的部分(诊断),由于诊断子任务可能找到多个解,那么我们就会相应有多种配置方案。这些配置方案包括属性值修改,元件替换和修改原设计的结构三种。一般来说,三者的求解复杂度是由易到难的,所以当有多个诊断解时,首先试验参数修改,如果参数修改不行再进行元件替换,最后如果元件替换还不行就进行结构修改。但这也不是绝对的,可能属性值修改的复杂性高或不能保证找到最好的解(成本低的),而元件替换相对简单,这是因为规范中某属性值从 3 改为 25,可能需要开发完全新的技术来获得新的参数值,因此要根据实际情况确定最佳配置。

定义 4.5(重用设计) 相对于重用设计诊断 DIAG 的重用设计是 CONF′,使得 SD∪SRS′∪CONF′可满足,且 SD∪CONF′∪CONF\DIAG|=SRS′。

假设已经存在一个广义产品结构树(关于广义产品结构树细节可参阅文[12])和原设计 CONF,使得 SD∪SRS∪CONF 是相容的,但由于出现新的 SRS′,因此 SD∪SRS∪CONF 不可满足。根据前面我们描述的方法能够求出多个重用设计诊断,取其中任意一个 DIAG,由定理 5.1 我们能够知道需要修改的部分为 DIAG 中的文字,那么 CONF\DIAG 就是原设计中相对 DIAG 要保留的部分。

结束语 本文利用基于模型的诊断方法对故障系统与换元系统的诊断和冲突之间的关系进行了分析,研究了系统修复与重新配置问题,利用替换发生后换元系统输出的变化刻画了故障系统的重新配置,提出了系统修复与重新配置的概念,对利用替换给系统输出行为带来的影响刻画了系统的修复与重新配置的生成过程,这为发生故障系统的修复提供了理论基础。在此基础上,我们刻画了基于模型的诊断方法在产品设计中的应用,提出了重用设计诊断、重用设计等概念。

参 考 文 献

1 李占山,姜云飞.基于模型诊断推理的回顾与展望.计算机科学,1998,25(6):54~57

2 Hamscher W,Console L.,de kleer J. Readings in Model-Based Diagnosis. San Mateo:Morgan-kaufmann publishers,1992. 1~28

3 Console L, Friedrich G. Model-Based Diagnosis. Basel-Switzerland,Science Publishers,1994. 1~15

4 Bakker R R,Eldonk S J M,Mars N J I. The use of model-based dagnosis in redesign. In:Cohn,ed. 11th European Conference on ArtificialIntelligence,John Wiley &Sons Ltd,1994. 44~48

5 Reiter R. A theory of diagnosis from first principles. Artificial Intelligence,1987,32(1):57~96

6 Stumptner M. An overview of knowledge-based configuration. AI Communications,1997,10(2):1~15

基于剪枝跳跃技术的最长公共子序列算法*)

An Algorithm for Solving Longest Common Subsequence Based on Pruning and Skipping Rules

刘 维[1] 陈 峻[1,2]

(扬州大学信息工程学院计算机系 扬州 225009)[1]

(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]

Abstract A fast algorithm for LCS problem FAST_LCS is presented. The algorithm first seeks the successors of the initial identical character pairs according to a successor table to obtain all the identical pairs and their levels. Then by tracing back from the identical character pair with the largest level, the result of LCS can be obtained. For two sequences X and Y with length n and m, time complexity of the algorithm is O(L), here L is the number of identical character pair. Pruning and skipping operations are also defined so as to reduce the searching space and accelerate the computation. Experimental result on the gene sequences of tigr database shows that our algorithm can get exact correct result and is faster and more efficient than other LCS algorithms.

Keywords Bioinformatics, Longest common subsequence, Identical character pair

1 引言

在生物信息学中[1,2],对 DNA 的研究实质上就是对生物序列进行比较分析[3,4],通过检测其相似成分来探测其生物特性的相似性。而探测序列相似性的方法之一就是找出序列间的最长公共子序列(LCS)[5]。本文提出一种快速的最长公共子序列的算法 FAST_LCS。对于长度为 n 和 m 的两序列 X,Y,该算法所需的计算时间为 $O(L)$,这里 L 指同字符对的个数。实验结果证明,本文算法与其它经典的 LCS 算法相比,不但能够取得准确的结果,而且在速度、效率上有了很大的提高。

2 同字符后续表及其同字符对

设欲比对的生物序列分别为 $X=(x_1,x_2,\cdots,x_n)$,$Y=(y_1,y_2,\cdots,y_m)$,其中,$x_i,y_i\in\{A,C,G,T\}$。为了找出这两条序列的最长公共子序列,我们首先要对这两条序列分别建立同字符后续表。我们定义'A'=CH(1),'C'=CH(2),'G'=CH(3),'T'=CH(4),将 X,Y 的同字符后续表分别记为 TX,TY,这里,TX、TY 分别为 $4*(n+1)$、$4*(m+1)$的二维数组。我们对 $TX(i,j)$定义如下:

$$TX(i,j)=\begin{cases}\min\{k\mid k\in SX(i,j)\} & SX(i,j)\neq\phi\\ - & \text{otherwise}\end{cases}\quad(1)$$

其中 $SX(i,j)=\{k\mid x_k=CH(i),k>j\}$,$i=1,2,3,4$,$j=0,1,\cdots,n$。"−"表示无定义。由此定义可知,若 $TX(i,j)$不为"−",它表示 X 第 j 个字符后面第一个为 $CH(i)$的字符的位置。

例 1 设 X ="T G C A T A",Y ="A T C T G A T"则 TX 及 TY 分别为:

TX:

i	CH(i)	0	1	2	3	4	5	6
1	A	4	4	4	4	6	6	-
2	C	3	3	3	-	-	-	-
3	G	2	2	-	-	-	-	-
4	T	1	5	5	5	5	-	-

TY:

i	CH(i)	0	1	2	3	4	5	6	7
1	A	1	6	6	6	6	6	-	-
2	C	3	3	3	-	-	-	-	-
3	G	5	5	5	5	5	-	-	-
4	T	2	2	4	4	7	7	7	-

在 X、Y 中,若有 $x_i=y_j$,则记(i,j)为一个同字符对。X、Y 所有同字符对的集合记为 $S(X,Y)$。若(i,j)、(k,l)皆为同字符对,且有 $i<k$、$j<l$ 则称(i,j)为(k,l)的一个前驱,或称(k,l)为(i,j)的一个后继,记为$(i,j)<(k,l)$。若设集合 $P(i,j)=\{(r,s)\mid(i,j)<(r,s),(r,s)\in S(x,y)\}$为$(i,j)$的所有后继同字符对集合,若有$(k,l)\in P(i,j)$,且不存在$(k',l')\in P(i,j)$,使得:$(k',l')<(k,l)$,则称 (k,l)为(i,j)的直接后继,记为$(i,j)<(k,l)$。对没有前驱的同字符对,我们称为初始同字符对。对任意的同字符对$(i,j)\in S(X,Y)$,它的层次号 $level(i,j)$定义为:

$$level(i,j)=\begin{cases}1 & \text{if}(i,j)\text{为一初始同字符对}\\ \max\{level(k,l)+1\mid(k,l)<(i,j)\} & \\ \text{otherwise}\end{cases}\quad(2)$$

*)基金项目:国家自然科学基金(60473012);国家科技攻关项目(2003BA614A-14);江苏省自然科学基金(BK2005047)。刘 维 硕士研究生,主要研究方向为算法优化和并行计算。陈 峻 教授,博士生导师,主要研究方向为算法设计和并行计算。

记 X、Y 最长公共子序列的长度为 $|LCS(X,Y)|$，则 $|LCS(X,Y)|=\max\{level(i,j)|(i,j)\in S(X,Y)\}$。

3 产生后继及剪枝跳跃操作

在我们的算法中，首先对所有初始同字符对利用同字符后续表产生其所有的直接后继，然后再对这些后继并行地产生其所有直接后继。重复这样的操作，直至不能继续产生后继为止。对某同字符对 $(i,j)\in S(x,y)$，由 (i,j) 产生其所有后继同字符对的操作可表示为：

$$(i,j)\rightarrow\{(TX(k,i),TY(k,j))|k=1,2,3,4,TX(k,i)\neq'-'\text{and}TY(k,j)\neq'-'\} \quad (3)$$

即对 TX 的第 i 列、TY 的第 j 列相应元素配成对偶。例如对例 1 中的同字符对(2, 5)，上述操作可表示为：

$$(25)\rightarrow\begin{bmatrix}4&6\\3&-\\-&-\\5&7\end{bmatrix}\rightarrow\begin{bmatrix}(4&6)\\(3&-)\\(-&-)\\(5&7)\end{bmatrix}\rightarrow\begin{bmatrix}(4&6)\\(5&7)\end{bmatrix}$$

即(2, 5)的后继为(4, 6)和(5, 7)。对任一同字符对(i, j)，反复使用上述方法可以产生其所有的后继。$(TX(k, 0),TY(k, 0))$，$k=1,2,3,4$，为 X、Y 的所有初始同字符对。由这些初始同字符对，我们可以得到所有同字符对及其层次值。在利用上述方法产生后继同字符对的过程中，我们可进行剪枝跳跃操作，删除某些明显不能得出最长公共子序列的同字符对，以缩小搜索空间、加快搜索速度。

定理 1 如果在同一层上所产生的同字符对 (i, j)和(k, l)中，有$(k, l)>(i, j)$，则删去(k, l)不影响算法得到最优解。

由于篇幅限制，定理 1 的证明略去。根据定理 1，我们可以在每一层产生同字符对后进行剪枝及跳跃操作。类似于定理 1，还有一些有用的剪枝跳跃操作。这些操作基于如下一些定理和推论。

定理 2 若在同一层中有同字符对 (i_1,j) 及 (i_2,j)，其中 $i_1<i_2$，则 (i_2,j) 可以剪去。

推论 1 若在同一层中，有同字符对 (i_1,j)，(i_2,j)，…，(i_r,j)，其中 $i_1<i_2<\cdots<i_r$，则 (i_2,j)，…，(i_r,j) 可以剪去。

由于篇幅限制，定理 2 和推论 1 的证明略去。

4 算法框架及复杂性分析

在算法中，我们设立一个同字符对表 *pairs*，表中的每一项由四元组 $(k, i, j, level, pred, state)$ 构成，分别表示记录号、同字符对(i,j)、层次号、直接前驱以及当前的状态。算法框架如下：

算法 FAST_LCS(X,Y)

输入：长度分别为 m,n 的序列 X,Y；

输出：X,Y 的最长公共子序列 LCS(X,Y)；

Begin

1. 构造 TX,TY 表；
2. 找出所有的初始同字符对：$(TX(k,0),TY(k,0))$，$k=1,2,3,4$；$level=1$；
3. 将所有初始同字符对$(k, TX(k, 0),TY(k, 0), 1,\phi, active)$，$k=1,2,3,4$ 加入到表 *pairs* 中。
   /* 对所有初始同字符对，赋予层号 1，$pred=\phi$，$state=active$ */
4. while 表 *pairs* 中有处于 *active* 状态的项 do
   4.1 对表 *pairs* 中所有层号为 *level* 的同字符对进行剪枝，将所有冗余的同字符对从表中去掉
   4.2 对表 *pairs* 中所有层号为 *level* 的活动同字符对$(k, i, j, level, pred, active)$并行地进行：
      4.2.1 使用产生后继操作和跳跃操作产生$(k, i, j, level, pred, active)$的所有后继集合 $(k', g, h, level+1, k, active)$，再将该集合插入到表 *pairs* 中；
      4.2.2 将$(k, i, j, level, pred, active)$的状态置为 *inactive*.
   4.3 $level=level+1$；
   end while
5. 计算 $r=$ 表 *pairs* 中的最大层次值；
6. 对于表中的所有同字符对$(k, i, j, r, l, inactive)$并行地进行：
   6.1 $pred=l$；$LCS(r)=x_i$.
   6.2 当 $pred\neq\phi$ 时
      6.2.1 从表 *pairs* 中得到其前驱$(pred, g, h, r', l', inactive)$
      6.2.2 $pred=l'$；$LCS(r')=x_g$.

End

设 X,Y 中同字符对的个数为 L，由于剪枝跳跃技术，本算法对 X、Y 中所有同字符对至多进行一次产生后继操作，因此算法 FAST_LCS(X,Y)串行执行的过程至多为对所有同字符对进行一次产生后继操作的过程，因此算法的时间复杂度为 $O(L)$。由于表 *pairs* 记录所需处理的同字符对，其所占的存储空间为 $O(L)$，同字符后续表 TX,TY 的存储空间分别为 $4*(n+1)$及 $4*(m+1)$。本文算法所需的存储复杂度为 $\max\{4*(n+1)+4*(m+1),L\}$。

5 实验结果及分析

我们从 tigr[6] 数据库中抽取了稻谷基因序列对本文算法 FAST_LCS 进行了实验，并将本文算法分别与目前应用最广泛的 Smith-Waterman 算法[7,8]及 FASTA 算法[9,10]在速度上进行了比较，由于本文算法与 Smith-Waterman 算法都可以得到精确解，因此我们将 FAST_LCS 算法与 Smith-Waterman 算法在速度上作了比较，同样地，我们在计算时间相同的前提下，也将本文算法与 FASTA 算法在精度上作了比较。

我们对不同长度的基因组序列使用本文算法 FAST_LCS 与 Smith-Waterman 算法分别进行了测试，并在计算速度上作了比较。结果如图 1 所示。由图 1 可以看出，对于不同长度的序列，本文算法要明显快于 Smith-Waterman 算法。在输入序列长度为 150 之前，两者速度差距变化较为缓慢，而在序列长度为 150 之后，速度差距变化异常迅速，本文算法

的速度大大超过 SW 算法。由此可见，对于长序列的 LCS 问题，本文算法和 SW 算法虽然都能获得精确的解，但本文算法的速度更快、效率更高。

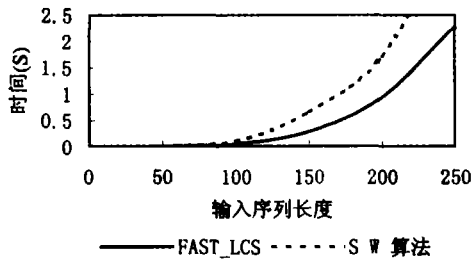


图 1　本文算法 FAST_LCS 与 Smith-Waterman 算法在计算时间上的比较

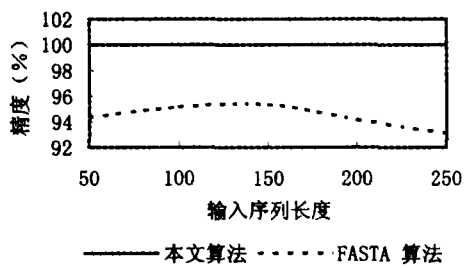


图 2　本文算法 FAST_LCS 与 FASTA 算法在精度上的比较

同样地，我们用本文算法与 FASTA 算法在计算时间相同的情况下，对计算结果精度进行了比较，这里精度是指：精度$=\frac{\text{算法求得的公共子序列的长度}}{\text{正确匹配中的最长公共子序列长度}}$，其结果如图 2 所示。由图 2 可见，不管输入序列长度如何增加，本文算法都能够取得准确的结果，而 FASTA 算法则随着序列长度的增加，精度明显下降。因此本文算法的精度要明显优于 FASTA 算法。

结论　为了在不影响精确度的前提下提高 LCS 计算的速度，本文提出了基于同字符对的求生物序列的最长公共子序列的算法 FAST_LCS。该算法通过对两条序列建立相应的同字符后续表，随后对于所有的初始同字符对，在该表中逐层地搜索其后继同字符对，以得到所有的同字符对及相应的层次值。最后由最大层次值的同字符对进行回溯，依次求得其所有前驱同字符对，最后得到相应的比对结果。对于长度为 n 和 m 的两序列 X，Y，该算法所需的内存为 $max\{4*(n+1)+4*(m+1),L\}$，时间为 $O(L)$，这里 L 指初始同字符对的个数。我们还给出了一些剪枝跳跃操作，以加快计算速度。我们对 *tigr* 数据库中的基因序列进行的实验结果证明，本文算法与其它经典的 *LCS* 算法相比，不但能够取得准确的结果，而且在速度、效率上有了很大的提高。

参 考 文 献

1　郝柏林，张淑誉. 生物信息学手册[M]. 上海科学技术出版社，2000. 171～172
2　李衍达，孙之荣，等译. 生物信息学-基因和蛋白质分析的实用指南[M]. 清华大学出版社，2000. 138
3　Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. ,1970,48(3):443～453
4　Waterman M S. Introduction to Computational Biology, Maps, Sequences and Genomes. London: Chapman & Hall, 1998
5　Aho A, Hirschberg D, Ullman J. Bounds on the Complexity of the Longest Common Subsequence Problem, J. Assoc. Comput. Mach. , 1976,23(1):1～12
6　http://www. tigr. org/tdb/benchmark
7　Smith T F, Waterman M S. Identification of common molecular subsequence. Journal of Molecular Biology, 1990,215:403～410
8　Altschul S F, Gish W, Miller W, Myers E W, and Lipman D J. Basic local alignment search tool, J. Mol. Biol. 1990,215:403～410
9　http://alpha10. bioch. virginia. edu/fasta_www/cgi/
10　http://www. ebi. ac. uk/services/

(上接第 314 页)

且 $0.7>\lambda=0.6$，可以匹配，故框架规则 Rule1 可用。

②计算结论的可信度 $CF(H'_1)$，由于 $P_1=1$，$P_2=1$，因此 E'_1 和 E'_2 是 AND 关系，运用式(4)计算 $CF(H'_1)=\delta_{match}(E,E')\times RC\times\sum_{i=1}^{k}CM_i=0.7\times 0.6\times(1\times 0.7+1\times 0.3)=0.42$

(2)按(1)步骤计算 $\delta_{match}(E,E')$ 和 $CF(H'_2)$

$\delta_{match}(E,E')=0.8>\lambda=0.6$

故框架规则可用 $CF(H'_2)=0.56$

(3)对于相同的证据却得到两种结果，我们采用匹配度大的那个作为最后结论，由于 $0.8>0.7$，故 H'_2 为最后结果，且其可信度为 0.56。

结论　从以上可以看出，这种带重要度可信度框架规则知识表示方法可以较准确地表示知识，它弥补了单独运用规则表示或框架表示的不足。同时本文把模糊理论的思想运用到这种混合知识表示推理中，具有一定的现实意义。在不精确推理中，一般的推理算法，都推到终结点，其动态强度超过规定的阈值，则为成立，否则为不成立，而不能在推理过程中，随时就能判断是否还须往下推，所以，影响了推理速度。本文没有解决这个问题，有待于继续研究解决。

参 考 文 献

1　尹朝庆，尹皓. 人工智能与专家系统[M]. 北京：中国水利水电出版社，2002
2　程伟良. 广义专家系统[M]. 北京：北京理工大学出版社，2005
3　Golding A R, Rosenbloom P S. Improving Accuracy by Combying Rule-based and Case-based Reasoning [J]. Artificial Intelligence, 1996,87(1-2):215～254
4　Zadeh L. A. The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems [J]. Fuzzy Sets and Systems, 1983, 11:199～227
5　赵瑞清，王晖，邱涤虹. 知识表示与推理[M]. 北京：气象出版社，1991
6　赵瑞清. 广义规则表示及其推理算法[J]. 计算机学报，1992.2:120～127
7　钟绍春. Processing of Uncertainty Temporal Relations [J]. 计算机学报(英文版)，1996,11(1)

动态约简抽样分析

Dynamic Reduct Sampling Analysis

王加阳

(中南大学信息科学与工程学院 长沙 410083)

Abstract The paper describes dynamic reduct model and discusses various formal dynamic reducts, and studies the sampling problem in depth. By pointing out some problems about Bazan theory, it presents new method for the sampling computation. The reduct precision coefficient is brought in sampling estimation, which makes computation method suitable to various formal dynamic reducts. Then a complete dynamic reduct framework is constructed.

Keywords Rough set, Dynamic reduct, Sampling computation

1 引言

粗糙集理论研究中对属性的约简已有许多研究成果[1],但这些研究基本都是基于静态约简的方法,对于海量数据集合存在着很大的不稳定性,寻求约简的稳定性是约简研究的一个关键问题。动态约简提供了一种新的思想,用以解决海量数据在静态约简时出现的决策规则不稳定的实际问题,具有其自身的优越性[2]。

2 动态约简

2.1 F族动态约简

Bazan 等人提出了动态约简概念和方法[3],建立了 F 族动态约简、(F-λ)动态约简和广义动态约简思想体系,在理论上为决策信息系统最稳定约简奠定了初步的基础。在 Bazan 的基本思想基础上,对精度系数进行调整,更加突出体现了动态约简的特征,理论体系更为完备,为探讨动态约简的更深层次问题奠定了基础。

定义 1 决策信息系统 $S=(U, C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,且 $B=(U', C\cup D)$,$U'\subseteq U$,则称 B 为 S 的子决策信息系统。S 的所有子决策信息系统构成的集合为 $\rho(S)$,$F\subseteq\rho(S)$ 为 S 的一个子决策信息系统集合,称为 S 的 F 族。

定义 2 决策信息系统 $S=(U, C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,S 的所有约简构成的集合称为决策信息系统 S 的约简集,记为 RED(S),子决策信息系统 B 的所有约简构成的集合称为子决策信息系统 B 的约简集,记为 RED(B)。

一个给定的决策信息系统可能存在多个约简,不同约简对应的规则集不完全相同,很难确定哪一个是最优约简,寻求最稳定的约简即是动态约简的目标,亦即优化的约简。

定义 3 决策信息系统 $S=(U,C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,集合 $F\subseteq\rho(S)$,S 的约简集为 RED(S),动态约简 DR(S,F) 表示如下:

$$DR(S,F)=RED(S)\cap\bigcap_{B\in F}RED(B)$$

DR(S,F)中任一元素称为 S 的 F 动态约简。

定义意味着,S 的一个相对约简是一个 F 动态约简,当且仅当它是 F 的所有子表的相对约简。F 族动态约简将原决策信息系统中所有抽取子决策信息系统约简的交集作为最终约简结果。

2.2 (F-λ)动态约简

F 动态约简概念对数据要求十分严格,为了适应噪声数据的处理,进一步把 F 动态约简的概念泛化,引入(F-λ)一动态约简的概念。

定义 4 决策信息系统 $S=(U, C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,子表族 $F\subseteq\rho(S)$,且 $\lambda\in(0.5,1]$,则(F-λ)动态约简定义为:

$$DR_\lambda(S,F)=\{Q\in RED(S)\mid \frac{|\{B\in F: Q\in RED(B)\}|}{|F|}\geqslant\lambda\}$$

λ 是约简精度系数,λ 趋近 1 时,$DR_\lambda(S, F)$ 接近 DR(S,F)。λ 越小,$DR_\lambda(S, F)$ 包含的约简越少,$DR_\lambda(S, F)$ 与标准动态约简 DR(S,F)相比愈粗糙;λ 值越大,$DR_\lambda(S, F)$ 包含的约简越少,λ 值决定了哪些约简属于 $DR_\lambda(S, F)$。

王加阳 教授,工学博士,研究方向:粗糙集理论与方法。

动态约简为所有子表约简的交集，λ限制在(0.5,1]这样一个区间，就是保证有50%以上的子表含有此约简，若低于这一值，则认为所得到的动态约简具有较大的随机性，稳定性不好，偏离了动态约简概念基本宗旨，不能认作为一个动态约简。

2.3 广义动态约简

根据动态约简定义，若S的子表族F中任意表的相对约简是动态的，那么它也必定是S的一个约简。这一概念限制很不方便，有时所需要的属性集合不一定是S的约简，需要把动态约简概念进一步扩展。

定义5 决策信息系统S=(U, C∪D)，U为论域，C为条件属性集合，D为决策属性集合，F⊆ρ(S)，

且有：

$GDR(S,F)=\bigcap_{B\in F} RED(B)$

则称GDR(S,F)中元素为S的F广义动态约简。

定义说明，S的任一子表是一个广义动态约简，那它须是一个给定子表族F中所有子表的约简。

定义6 决策信息系统S=(U, C∪D)，U为论域，C为条件属性集合，D为决策属性集合，子表族F⊆ρ(S)，且λ∈(0.5,1]，有：

$$GDR_\lambda(S,F)=\{Q\subseteq C\mid \frac{|\{B\in F: Q\in RES(B)\}|}{|F|}\geqslant\lambda\}$$

$GDR_\lambda(S,F)$中元素为S的(F-λ)广义动态约简。

动态约简针从决策表随机抽取若干对象样本组成较小的决策表，把复杂的大型信息表的约简问题转化为若干子决策表的最优约简的交集问题，需要解决的关键问题是抽样计算与分析。

3 抽样分析

动态约简方法的核心思想在于对大型决策信息系统进行多次抽样，把复杂决策信息系统的约简问题转化为若干子决策信息系统的最优约简的交集问题，使得动态约简较静态约简方法更具有处理大型数据集的能力，其关键在于抽样出的子表族F的确定。Bazan等人采用统计学原理来估计子表随机抽样得到的数量[4]，在此对该思想方法进行了详细讨论，分析了其中存在的问题，给出动态约简子表族大小的计算方法和详细分析。

3.1 F族计算

动态约简集本质是那些频繁出现在全局决策信息系统的所有子表中的约简集合，Bazan等人以此为出发点，提出了动态约简子表族F大小的确立方法，从理论上为动态约简奠定了基础。

动态约简R在全局决策表的所有子表中出现的概率为$P_G(R)$，假设抽取全局决策表的一个子表为一个随机变量，R在某个子表中出现的概率为一个(0,1)分布函数，证明动态约简的稳定系数值是$P_G(R)$的极大似然估计值，将所有子表作为是一系列的随机变量，独立且都服从(0,1)分布，采用独立同分布中心极限定理，通过正态分布的区间估计，推出子表族F的大小。

然而Bazan的上述思想方法存在着缺陷，一方面其极大似然估计最大误差假设的概念，推出的不一定是所有动态约简的抽样值；另一方面没有考虑到约简精度系数对抽样的影响。下面根据正态分布区间估计推算对所有动态约简都适应的F族抽样值，特别是把约简精度系数纳入了F族抽样范畴。

(1)极大似然估计$P_G(R)$

设动态约简R在全局决策信息系统W=(W,C∪D)的所有子表G=ρ(W)中出现的概率为$P_G(R)$，定义为$P_G(R)=(G_R|/|G|$，其中$G_R=\{B\in G: R\in RED(B,D)\}$。在实际中，全局决策信息系统W并不能确定，子表族G和概率$P_G(R)$是不可求得的。

对于决策信息系统S，它是全局决策信息系统W的一个样本，R是通过对S的子表族F计算得到的。通过对样本各项参数的研究，可以反映总体的特征，即F族的特征反映了G族的特征。

对于任何$B\in G$，R是否为子表B的一个约简构成(0,1)分布，分布函数表示为，$X_G^R(B): G\to\{0,1\}$，有：

$$X_G^R(B)=\begin{cases}1 & R\in RED(B,D)\\ 0 & R\in RED(B,D)\end{cases}$$

记$G^1=\{B\in G: X_G^R(B)=1\}$和$G^0=\{B\in G: X_G^R(B)=0\}$。(0,1)分布中成功概率为$P[X_G^R(B)=1]=P_G(R)$，失败概率为$P[X_G^R(B)=0]1-P_G(R)$。

记$MLE(P_G(R))$为$P_G(R)$的极大似然估计，由参数估计理论，(0,1)分布概率$P_G(R)$的极大似然估计是所有子表X_G^R的一个算术平均值：

$$MLE(P_G(R))=\frac{\sum_{B\in F}X_G^R(B)}{|F|}=\frac{\sum_{B\in F\cap G^1}X_G^R(B)+\sum_{B\in F\cap G^0}X_G^R(B)}{|F|}=\frac{|F\cap G^1|}{|F|}$$

可见，$P_G(R)$的极大似然估计值与(F-λ)广义动态约简R的稳定系数相等。

(2)正态分布区间估计F族

每个子表可看成是一个随机变量X_i，则F族的子表为一系列随机变量$X_1, X_2, \cdots, X_n$，动态约简在某个子表中的出现是一个(0,1)分布，发生的概率为$P_G(R)$，假设这一系列的随机变量独立服从同一分布，且具有数学期望和方差，由独立同分布的中心

极限定理：

$$\lim_{|F|\to\infty} P\left\{\frac{\sum_{i=1}^{|F|}X_i-E(\sum_{i=1}^{|F|}X_i)}{\sqrt{D(\sum_{i=1}^{|F|}X_i)}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

其中，$\sum_{i=1}^{n}X_i$ 表示 F 族总体，$E(\sum_{i=1}^{n}X_i)$表示总体的期望，$D(\sum_{i=1}^{n}X_i)$表示总体的方差。

对于(0,1)分布而言，$E(\sum_{i=1}^{n}X_i)=n\cdot P_G(R)$，$D(\sum_{i=1}^{n}X_i)=n\cdot P_G(R)\cdot(1-P_G(R))$，这里 $n=|F|$，即子表族随机变量总体近似服从标准正态分布，从而有：

$$\lim_{|F|\to\infty} P\left[\frac{MLE(P_G(R))-P_G(R)}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}\leqslant x\right]=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

现在的问题是，要抽取多大容量的 F 族样本，才能使误差 $|MLE(R_G(R))-P_G(R)|$ 小于 $\Delta MLE(P_G(R))$ 的概率不小于 $1-\alpha$，表示置信水平至少为 $1-\alpha$ 的正态分布区间估计范围，即：

$$P[|MLE(P_G(R))-P_G(R)|<\Delta MLE(P_G(R))]\geqslant 1-\alpha$$

$$P\left[\frac{|MLE(P_G(R))-P_G(R)|}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}<\frac{\Delta MLE(P_G(R))}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}\right]^{\geqslant 1-\alpha}$$

根据标准正态分布有，

$$2\cdot\Phi\left[\frac{\Delta MLE(P_G(R))}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}\right]^{-1\geqslant 1-\alpha}$$

亦即：

$$\Phi\left[\frac{\Delta MLE(P_G(R))}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}\right]^{(2-\alpha)/2}$$

设 $\Phi(t_{\alpha/2})=(2-\alpha)/2$，由 Φ 函数的单调性，则

$$\frac{\Delta MLE(P_G(R))}{\sqrt{\frac{P_G(R)\cdot(1-P_G(R))}{n}}}\geqslant t_{\alpha/2}$$

推导变换得到

$$|F|\geqslant\frac{t_{\alpha/2}^2\cdot P_G(R)\cdot(1-P_G(R))}{(\Delta MLE(P_G(R)))^2}$$，这里 $|F|=n$

由于 $P_G(R)$的值无法求得，可用 $MLE(P_G(R)$ 来近似代替上式中的 $P_G(R)$，从而得到表达式：

$$|F|\geqslant\frac{t_{\alpha/2}^2\cdot MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{(\Delta MLE(P_G(R)))^2}\quad(1)$$

3.2 F族分析

从(1)式来分析，$|F|$ 与容许误差 $\Delta MLE(P_G(R))$成反比，表明容许误差大时，可以抽取较少的子表，容许误差小时，误差精度要求越高，需要抽取的子表就越多；$|F|$ 与 $t_{\alpha/2}$ 成正比，表示置信水平要求较高时，需要抽取较多的子表，这些与统计思想是一致的。根据(1)式，作图 1 曲线。

$MLE(P_G(R))\cdot(1-MLE(P_G(R)))$

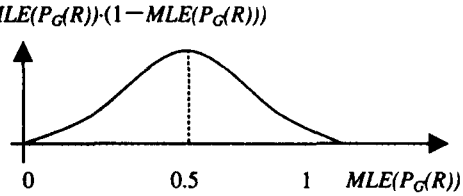


图 1 $MLE(P_G(R))\cdot(1-MLE(P_G(R)))$曲线

由图 1 可知，$MLE(P_G(R))\cdot(1-MLE(P_G(R)))$在 $MLE(P_G(R))=1/2$ 时，取得极大值 1/4。在(0.5,1]区间，它是递减的，表示 $MLE(P_G(R))$越大，R 出现的概率也就越大，抽取较少的子表就可得到其估计值，体现了极大似然估计的统计思想。在 $MLE(P_G(R))\leqslant 0.5$ 时，由前面对动态约简模型描述，说明 R 不是一个动态约简。

(1)对于任意给定的动态约简 R_0，由(1)式有：

$$|F_0|\geqslant\frac{t_{\alpha/2}^2\cdot MLE(P_G(R_0))\cdot(1-MLE(P_G(R_0)))}{(\Delta MLE(P_G(R_0)))^2}$$

由此得出了关于动态约简 R_0 的极小抽样值 $|F_0|$，该值具有针对性，对其它的动态约简 $R\neq R_0$ 并不一定适应。

另外，在 $MLE(P_G(R_0))=1$ 时，$|F_0|\geqslant 0$，说明约简 R_0 是出现在所有子表中的，抽样的多少对其是否为动态约简没有影响，稳定系数为 1 的约简 R_0 是恒定的动态约简。

F 族可能存在多个动态约简：$R_1,R_2,\cdots,R_k$，对于每个约简 R_i，用 $MLE(P_G(R_i))$来估计 $P_G(R_i)$ 时，会存在一个估计误差 $|MLE(P_G(R_i))-P_G(R_i)|$，对于 F 族中每一个动态约简 R_i，在容许的估计误差 $\Delta MLE(P_G(R_i))$下，都可计算出相应的 $|F_i|$ 值。

当所有动态约简的误差满足 $MIN[\Delta MLE(P_G(R))]\leqslant\Delta MLE(P_G(R_i))$时，由(1)式可得到满足所有动态约简的一个 $|F|$ 极小值，即对所有的动态约简 R_i 都适应的 F 族抽样值：

$$|F|\geqslant\frac{t_{\alpha/2}^2\cdot MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{(MIN[\Delta MLE(P_G(R))])^2}$$

$MLE(P_G(R))\cdot(1-MLE(P_G(R)))$在 $MLE(P_G(R))=1/2$ 时，得到极大值 1/4，$|F|$ 的整体极小值为：

$$\frac{t_{\alpha/2}^2}{4\cdot(MIN[\Delta MLE(P_G(R))])^2}$$

(2)根据(F-λ)动态约简描述,F 族的抽样与约简精度系数λ应该具有相关性,反映不同约简精度下对抽样的要求。

根据约简精度系数λ的描述,有λ∈(0.5,1],R 的稳定系数值在大于λ时,即 $MLE(P_G(R))\geqslant\lambda$,R 才可被视为一个(F-λ)动态约简,否则 R 只可能是某些子决策表的约简,而不能认作为动态约简。

由图 1,在 $MLE(P_G(R))\geqslant\lambda>0.5$ 时,$MLE(P_G(R))\cdot(1-MLE(P_G(R)))$是递减的,则有:

$$\lambda(1-\lambda)\geqslant MLE(P_G(R))\cdot(1-MLE(P_G(R)))$$

在 R 为一个动态约简的情况下,用λ来代$MLE(P_G(R))$,可得:

$$\frac{t_{\alpha/2}^2\cdot\lambda\cdot(1-\lambda)}{(\Delta MLE(P_G(R))^2}\geqslant\frac{t_{\alpha/2}^2\cdot MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{(\Delta MLE(P_G(R)))^2}$$

对给定的λ值,将得到一个相应|F|值,表示 R 在满足约简精度系数λ条件下的客观抽样率,可取:

$$|F|\geqslant\frac{t_{\alpha/2}^2\cdot\lambda\cdot(1-\lambda)}{(\Delta MLE(P_G(R))^2}$$

F 值在λ增加时,是减少的,即所需抽取的子表数目就越少。这说明,约简精度要求较高时,对应的都是较高稳定性的动态约简。从统计观点看,只有较高稳定性的动态约简,在抽样较少时,可能在子表中出现。当λ=1 时,表示若 F 的任一个子表不含约简 R,则 R 就不是动态约简,从理论上说,可以一次抽样确定。

在给定λ情况下,所有(F-λ)动态约简 R 的稳定系数都要求大于λ,因而对所有(F-λ)动态约简,都有$\lambda(1-\lambda)\geqslant MLE(P_G(R))\cdot(1-MLE(P_G(R)))$成立,则:

$$|F|\geqslant\frac{t_{\alpha/2}^2\cdot\lambda\cdot(1-\lambda)}{(MIN[\Delta MLE(P_G(R))])^2}$$

该式对所有满足约简精度系数λ的动态约简 R 都是适应的。

4 F 族质量

动态约简的抽样计算还须强调另一个重要方面,即样本分布对动态约简的影响,抽样的分布状态对动态约简有效性具有较大程度影响,也体现了 F 族所具有的代表性。一般而言,F 族的抽取采取简单随机抽样方法,这里提出了对 F 族质量进行评价的基本思想,衡量 F 族抽取的均衡性,以反映动态约简的有效性。

定义 7(全样本) 决策信息系统 S=(U,Q=C∪D,V,F),U 为论域,C 为条件属性集合,D 为决策属性集合,F⊆ρ(S),且有:

$$UF=\bigcup\{U_B|B=(U_B\subseteq U,\ C\cup D)\in F\}$$

称 UF 为全样本集合。

定义 8(中心样本) 决策信息系统 S=(U,Q=C∪D,V,F),U 为论域,C 为条件属性集合,D 为决策属性集合,F⊆ρ(S),且有:

$$UCF=\bigcap\{U_B|B=(U_B\subseteq U,\ C\cup D)\in F\}$$

称 UCF 为中心样本集合。

在简单随机抽样前提下,先从整体上来考虑 F 族抽样的均衡性,通过如下描述的二个比值来衡量。

(1)整体抽样率

设定比值$|U_F|/|U|$,表示了对决策信息系统 S 的整体抽样率,可以很直观地看出$0\leqslant|U_F|/|U|\leqslant1$,如果太小$|U_F|/|U|$,则说明抽样样本较为集中,不能反映出原决策信息系统的样本特征,降低了动态约简的可信度。

显然,当$|U_F|/|U|$的值小于 0.5,可以认为 F 族的抽样是不够充分的,对原决策信息系统的覆盖率太低。可以设定一个阈值,当$|U_F|/|U|$大于阈值时,抽样合符要求。

(2)整体均衡度

设定比值$|U_{CF}|/|U|$,表现了决策信息系统的整体重叠抽样率,可以直观地得到$0\leqslant|U_{CF}|/|U_F|\leqslant1$,如果$|U_{CF}|/|U_F|$太大,则说明抽样样本过于重复,不能反映出原决策信息系统的对象特征,动态约简的有效性降低。

显然,若$|U_{CF}|/|U_F|$的值大于 0.5,则认为 F 族的抽样是不够均衡的,样本的覆盖较为集中,而其值越小,抽样分布越均匀,越能更好地体现动态约简的思想。可以设定一个阈值,当$|U_{CF}|/|U_F|$小于阈值时,抽样分布均匀性较好。

仅考虑 F 族整体的抽样性能还是不够的,还必须避免 F 族个别决策表的异常现象,即一个决策表的个别抽样对整体分布产生过大的影响。对于 F 族中的所有抽样子决策信息系统,逐一考察其对中心样本集合 UCF 的包含程度,以及被全样本集合 UF 包含的程度,也通过二个参数来衡量。

(3)局部抽样率

设定比值$|U_B|/|U_F|$,描述了决策子表 B 在全样本中所占比率,同样直观地有$0\leqslant|U_B|/|U_F|\leqslant1$。若 F 中有一定数量决策子表的比值$|U_B|/|U_F|$较高,体现了抽样的不均衡性,即使整体均衡性比值尚可,但存在着局部的样本重叠性过高,降低了动态

(下转第 337 页)

基于 PSC-CEA 的移动 IP 资源分配动态策略触发*)

Dynamic Policy Trigger of Mobile IP Handoff Based on PSC-CEA

魏 达[1,2] 刘衍珩[1,2] 刘雪洁[1,2] 李连登[1]

(吉林大学计算机科学与技术学院 长春 130012)[1]

(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)[2]

Abstract This paper addresses the necessity of dynamic policy triggered by product space algebra of conditional events during the process of mobile IP handoff and presents a dynamic policy trigger model according the characteristic of network administration which based policies, and designs a mechanism , which is used to update dynamic policy according to known information、experience and knowledge to satisfy the QoS of mobile traffic and use current network resource sufficiently.

Keywords Procduct space algebra of conditional events, Dynamic policy trigger, Mobile IP handoff, Resource allocation

1 引言

随着移动 IP 技术的应用,保证移动业务的 QoS,特别是在切换过程中如何为用户提供满意的业务服务质量变得至关重要。而 Shadow Cluster 的改进方案[1]等移动设备位置的跟踪和移动的预测方案的提出,可以确切预测移动用户下一个或几个小区,这使得在相应小区内预留资源,从而提供一致的 QoS 服务成为可能。

基于策略的网络管理 PBNM(Policy Based Network Management)[2]是近年来新兴的研究和应用课题,其基本思想是将网络管理分为策略的制定和策略的执行。策略由特定的服务进行处理,根据当前的网络状态或外部事件的触发,将策略解释成针对网络设备的具体指令,并发送到网络设备进行执行。

将移动 IP 和 PBNM 技术结合,可通过对移动主机移动性及切换时所需资源的预测,形成预测策略,在移动节点到达前触发执行该策略,可实现移动网络的动态管理。但预测信息是不确定的、模糊的,所以基于预测所形成的预测策略也是不确定的,而执行的策略必须是确定的,这就需要根据知识、经验信息来对预测策略进行更新触发使其符合移动主机的 QoS 的要求。本文在预测策略的基础上,根据知识、经验等条件信息,提出使用乘积空间条件事件代数 PSC-CEA 对预测策略进行更新触发的方法,把定量的数值信息和定性描述的经验条件信息紧密结合在一起来对预测策略进行更新和决策,使其很好地满足主机的服务质量要求。

第 2 部分对乘积空间条件事件代数进行介绍;第 3 部分介绍动态策略的事件空间模型和触发流程;第 4 部分对动态策略触发机制进行描述,提出了利用规则相关性度量方法进行动态策略触发的机制;第 5 部分对动态策略的触发进行了应用举例;最后对全文做出总结。

2 乘积空间条件事件代数(PSC-CEA)

PSC-CEA 是一种布尔型的条件事件代数系统,基本思想是:设(Ω,B,P)是对非条件事件 a,b,c,d ……的概率空间,现构建一个扩展的乘积概率测度空间(Ω_0,B_0,P_0),其中:

$\Omega_0=\Omega*\Omega*\cdots\cdots$ 是扩展的样本空间;

$B_0=B*B*\cdots\cdots$ 是扩展的布尔代数或 δ 代数;

P_0 是乘积空间的概率测度。

可以得到:

$$(a|b)=(a\wedge b)\times\Omega_0)\vee(b'\times(a\wedge b)\times\Omega_0)\vee(b'\times b'\times(a\wedge b)\times\Omega_0)\vee\cdots$$

设有映射 $f:B\times B\to B_0$,其中 f 由 $f(a,b)$定义:

$f(a,b)=ab\vee(b'\times ab)\vee(b'\times b'\times ab)\vee\cdots$,

则可得:

$$P_0(f(a,b))=P_0[ab\vee(b'\times ab)\vee(b'\times b'\times ab)$$

*)基金项目:本文受到国家自然科学基金资助项目(No. 60573128)和吉林大学科技创新基金的资助。魏 达 博士研究生,副教授,主要从事计算机网络和人工智能研究;刘衍珩 博士生导师,主要从事计算机通信与网络研究;刘雪洁 博士研究生,主要从事基于策略网络管理、无线网络服务质量研究;李连登 硕士研究生,主要从事计算机网络和人工智能研究。

$\vee\cdots]$

$=P_0(ab)+P_0(b'\times ab)+P_0(b'\times b'\times ab)+\cdots$

$=P(ab)+P(b'\times ab)+P(b'\times b'\times ab)+\cdots$

$=P(ab)(1+P(b')+p(b')^2+\cdots)$

$=P(ab)\times\frac{1}{1-P(b')}=P(a|b)$

实现了(Ω_0,B_0,P_0)对(Ω,B,P)的扩张，且$P_0(f(a,b))=P(a|b)$。

目前，研究人员普遍认为 PSC-CEA 可将定量的数值信息处理和定性描述的自然语言文本信息和经验条件信息的处理纳入一个具有坚实逻辑数学基础体系，进行不确定性、概率性和模糊性推理的数学工具，并其应用是以随机集理论为基础的，可以把许多信息融合和不确定性推理的方法，如人工神经网络、贝叶斯规则、模糊集理论、统计理论、D-S 证据理论等等融入其中，便于与原有不确性推理系统进行有机结合。

本文正是基于这一点，在动态策略触发机制中采用 PSC-CEA 对预测策略进行更新。

3 动态策略的事件空间模型和触发流程

策略是由一组形如 if condition then action 的规则组成，这里的规则不同于专家系统中形如 if b then a 的规则那样要求满足 a≤b，策略规则中 condition 和 action 不可能满足 action≤condition，因为它们不是在同一事件空间上。通过对策略核心信息模型 PCIM 的分析可知，策略中规则的动作只能推出其子动作或其相关的动作，而不能推出其子动作相关的条件，因此策略中规则的条件和动作应相应地划分为条件事件空间和动作事件空间，condition 属于条件事件空间，而 action 属于动作事件空间。

设(Ω,B,P)是一个测度空间，(X,A,ψ)是另一个测度空间，$T:\Omega\to X$是一个映射，则可把策略中的规则 if condition then action 表示为：

$T:\{X|X\in condition\}\to\{y|y\in action\}$

其中：$condition\subset\Omega$，$action\subset X$。

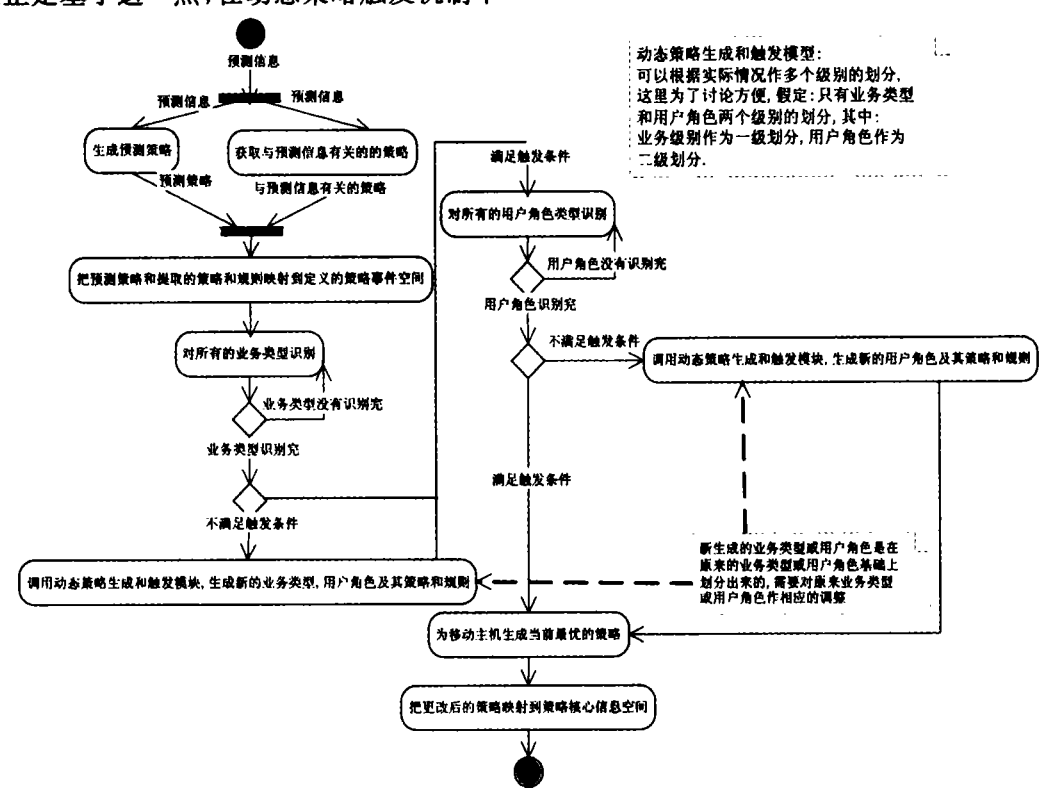


图 1

由于对移动主机预测策略中的规则同样具有 if condition then action 形式，且策略信息库中存储着已执行过的策略（相当于成功的案例），这些策略是不断通过策略执行的反馈信息利用粗糙集理论动态学习获得的。由于在这里触发的策略，已不是在数值层面的经信息融合得到的策略，而是根据网络拓扑结构、网络资源状态、业务的 QoS、知识经验以及策略执行的反馈信息统计等因素的基础上自动执行决策形成的策略，因此可采用 PSC-CEA，把定量的数值信息、定性描述的自然语言文本信息和经验条件信息紧密结合起来进行综合评估和决策。

为了在事件空间中对策略进行处理，必须把策

略映射到事件空间。根据策略核心信息模型的特点，为了提高性能、便于冲突检测，采用了水平划分和垂直划分相结合方法划分策略空间。建立的动态策略触发流程如图 1 所示。

4 动态策略触发机制

4.1 规则的测度空间定义

对规则 if condition then action，根据乘积空间条件事件代数中度量空间扩展方法，从基本的度量空间出发构造一个包含原度量空间的更大的度量空间 (Ω_0, B_0, P_0)，其中：

$\Omega_0=\Omega\times X, B_0=a\times b, \forall a \text{ in } B, \forall b \text{ in } A,$

$P_0(a\times b)=f(P(a),\psi(b)), P_0(\Phi)=0, \Phi\subseteq B_0;$

再以 (Ω_0, B_0, P_0) 为基本度量空间扩展为 PSC-CEA 的度量空间 $(\Omega_{00}, B_{00}, P_{00})$，其中：

$\Omega_{00}=\Omega_0\times\Omega_0\times,\cdots,$

$B_{00}=a\times b\times c\cdots, \forall a,b,cinB_{00},$

$P_{00}(a\times b\times c\times\cdots)=P_0(a)\times P_0(b)\times P_0(c)\times\cdots$

根据上述扩展，策略中的规则(if b then a)可以表示为条件事件(a|b)，其中：$P_{00}(b)>0$ 且 a，b in B_{00}，很显然，其满足乘积空间条件事件代数的($\wedge$，$\vee$，$(\ )'$)操作。

4.2 规则更新机制

为了保证移动主机的服务质量，减少预测策略的模糊性和不确定性，必须根据当前成功执行的策略为范例对预测策略进行修正。为了保证修正的正确性，必须先对预测策略中规则和当前成功执行策略中的规则进行相关性度量。本文提出了一种规则相关性度量的方法，并在此基础上提出一种对预测策略进行更新的方法。

在此我们只考虑预测策略中不确定的规则，而不考虑确定的规则(如：关于用户的认证的规则等)。由于预测策略中规则的条件是不确定的、模糊的，当前已成功执行策略中规则的条件也看成是不确定的、模糊的，如果把规则中的条件用集合表示，则此集合是一个有限集合并且是一个模糊集合。不确定性规则多数是对网络资源配置的规则，其对资源的配置可用一个区间[min，max]表示，其中 min<=max，当 min=max 时，区间就简化为一个点。

假设：预测策略为 $P_p(R_1,R_2,\cdots,R_k)$，当前成功执行的第 i 条策略为 $P_i(R_{i1}R_{i2},\cdots,R_{in})$，其中规则 $R_q=(S_q|X_q), 1\leqslant q\leqslant k \quad R_{ij}=(T_{ij}|X_{ij}), 1\leqslant i\leqslant m, 1\leqslant j\leqslant n$。

A_q 为预测策略中第 q 条规则的条件集合，B_{ij} 为当前已成功执行的策略中第 j 条规则的条件集合，条件事件论域为 (Ω_c, B_c, P_c)，则 $A_q=\subseteq B_c, B_{ij}\subseteq B_c$。

$[D_{ij}\min, D_{ij}\max]$ 为预测策略中第 j 条规则的动作资源配置区间，$[E_{ij}\min, E_{ij}\max]$ 为当前已成功执行的策略中第 j 条规则的动作资源配置区间，其中，$E_{ij}\min\in R^+, E_{ij}\max\in R^+, D_{ij}\min\in R^+, D_{ij}\max\in R^+, i=1,2,\cdots,m, j=1,2,\cdots,n$，m 为策略的个数，n 为规则的个数。

定义 1 预测策略中规则 R_q 与当前执行策略中规则 R_{ij} 之间的相关性度量为：

$$S(R_q,R_{ij})=P_{00}((S_q|X_q)\wedge_{X_q\cup Y_{ij}}(T_{ij}|Y_{ij}))$$

定义 2 预测策略中规则 R_q 与当前执行策略中规则 R_{ij}、规则 R_{ik} 之间的相关性度量为：

$$D(R_q,R_{ij},R_{ij})=P_{00}((S_q|X_q)\wedge_{X_q\cup Y_{ij}}(T_{ij}|Y_{ij})\wedge_{Y_{ij}\cup Y_{ik}}(T_{ik}|Y_{ik}))\quad j\neq k$$

定义 3 当前执行策略中规则 R_{ij} 的平均一致度为：

$$A(R_q,R_{ij})=\frac{S(R_q,R_{ij})-\sum_{q=1}^{j}D(R_q,R_{ij},R_{iq})}{f(R_q)+\sum_{i=1}^{n}S(R_q,R_{ij})-\sum_{i=1,j\neq k}^{n}D(R_q,R_{ij},R_{ik})}$$

$$f(R_q)=\begin{cases}1-\sum_{i=1}^{n}S(R_q,R_{ij})-\sum_{i=1,j\neq k}^{n}S(R_q,R_{ij},R_{ik})\\ \sum_{i=1}^{n}S(R_q,R_{ij})-\sum_{i=1,j\neq k}^{n}D(R_q,R_{ij},R_{ik})\leqslant 1\\ 0,\sum_{i=1}^{n}S(R_q,R_{ij})-\sum_{i=1,j\neq k}^{n}D(R_q,R_{ij},R_{ik})>1\end{cases}$$

用区间 $[R_{ij}\min, R_{ij}\max]$ 表示规则 R_{ij} 中对网络资源的配置行为。

定义 4 预测策略规则中动作的更新方法：

$$R_q\min=R_q\min^* f(R_q)+\sum_{n=1}^{n}(R_{ij}\min^* A(R_q,R_{ij}))$$

$$R_q\max=R_q\max^* f(R_q)+\sum_{i=1}^{n=1}(R_{ij}\max^* A(R_q,R_{ij})), i=1,2,\cdots,m$$

则更新算法的步骤为：1)利用定义 1 计算规则 R_q 和规则 R_{ij} 之间的相关性；2)利用定义 2 计算规则 R_q 和规则 R_{ij}、规则 R_{ik} 之间的相关性；3)对步骤 1 中计算的相关性 $S(R_q,R_{ij})$ 按大小对规则按降序排序；4)根据步骤 3 中得到相关性的先后顺序，利用定义 4 计算规则 R_{ij} 的平均一致度；5)利用定义 4 对预测策略中每一个规则中动作进行更新即可。

本算法与现有的许多相似性度量算法不同，它把规则作整体考虑而不是先人为的分离再组合，很显然这样能充分利用已有的知识、经验信息，使决策更精确。

5 应用举例

假定移动主机已通过 AAA 认证，并根据预测信息生成的预测策略为：

P_1：if(role = high and protocol = tcp)(0.8) then delay time <= 100ms(0.8)(0.8)

P_2:if(role = high and protocol = tcp)(0.8) then guarantee 8% of available BW (0.8)(0.8)

获取的相关策略和规则为一个带宽分配规则组和一个时延规则组和一个核心规则组成。

带宽分配规则组有如下规则组成:

R_1:if (role=normal and protocol = udp)(0.6) then (guarantee 3% of available BW)(0.6)(0.8)

R_2:if (role = high and protocol = udp)(0.8)then (guarantee 5% of available BW)(0.8)(0.8)

R_3:if (role=normal and protocol = tcp)(0.7)then (guarantee 6% of available BW)(0.7)(0.8)

R_4:if (role = high and protocol = tcp)(0.9)then (guarantee 9% of available BW)(0.7)(0.9)

时延规则组有如下规则组成:

R_5:if(role=high)(0.8)then delay time <=200ms(0.7)(0.8)

R_6:if(role=normal)(0.7)then delay time <=400ms(0.9)(0.8)

核心规则为:

R_7:if (guarantee 8% of available BW)(1.0)then (delay time <= 200ms)(1.0)(1.0)

假定规则的条件中采用集合表示方式计算概率,规则的动作中配置区间服从均匀分布,随机变量role的取值中high=2normal。

利用规则更新算法计算出:

$S(P_1,R_5)=0.8$ $S(P_1,R_6)=0.1$

$D(P_1,R_5,R_6)=0.1f(P_1)=0.2A(P_1,R_5)=0.8$

$S(P_1,R_6)=0$ $S(P_2,R_4)=0.9S(R_2,R_3)=3/8$

$S(P_2,R_1)=S(P_2,R_2)=0$

$D(P_2,R_1,R_2)=D(P_2,R_1,R_3)=0$

$D(P_2,R_3,R_4)=4/15$

$A(P_2,R_1)=A(P_2,P_2)=0f(P_2)=0$

$D(P_2,R_1,R_4)=D(P_2,R_2,R_3)=D(P_2,R_2,R_4)=0$

$A(P_2,R_4)\approx0.11,A(P_2,R_3)\approx0.89$

P_1min更新为:0ms,P_1max更新为:180ms

P_2min更新为:8.67% of available BW

P_2max更新为:100% of available BW

因此,规则P_1更新为:

P11:if(role = high and protocol = tcp)(0.8) then delay time <=180ms(0.8)(0.8)

规则P_2更新为:P22:

if(role = high and protocol = tcp)(0.8)then guarantee 8.67% of available BW (0.8)(0.8)

由核心规则R6和规则P22可以推出如下规则:

Fp22:if(role = high and protocol = tcp)then (delay time <= 200ms)(0.8)

再利用规则更新算法中步骤1检查规则P11和Fp22是否发生冲突?若发生冲突,则根据规则触发算法进行更新即可。

由于规则的条件都是相同的即X=Y,在此基础上利用定义1得:S(P11,Fp22)=0.9>0.8,所以,规则P11和P22是无冲突的,并满足核心规则R_7

综上所述:更新后的预测策略为:

NP1:if(role = high and protocol = tcp)(0.8)then delay time <= 180ms(0.8)(0.8)

NP2:if(role = high and protocol = tcp)(0.8)then guarantee 8.67% of available BW (0.8)(0.8)

总结 乘积空间条件事件代数是在确保规则概率与条件概率相容的前提下,把布尔代数上的逻辑运算推广到条件事件(规则)集合中得到的代数系统。条件事件代数理论在自身不断发展完善时,已广泛应用于人工智能和专家系统等领域[7]。本文提出了在数据融合之后形成预测策略的基础上,根据当前已知信息、经验、知识对预测策略更新的一种方法,使生成的动态策略不仅能满足移动业务的QoS,而且能充分利用当前网络资源。

参考文献

1 Su W, Gerla M. Bandwidth allocation strategies for wireless ATM networks using predictive reservation. In: Proceedings of Global Telecommunications Conference (IEEE Globecom), 1998,4: 2245~2250

2 Verma D. Simplifying Network Adminis-tration using Policy based Management . IEEE Network Manazine, March 2002

3 Mahler R P S. Representing Rules as Random Sets, I: Statistical Correlations between Rules. Informatics Sciences, 1996,88: 47~68

4 Goodman I R. Applications of product space algebra of conditional events and one-point random set representations of fuzzy sets to the development of conditional fuzzy sets. Fuzzy Sets ans Systems, 1995,69: 257~278

5 Goodman I R. New Application of Contional and Relational Event Algebra to Fusion of information. IEEE, 1998

6 Goodman I R. Use of Relational and Conditional Event Algebra in Addressing Modeling and Combining of Information in Expert Systems. IEEE, 1995

7 Calabress P C. A Theary of Conditional Information with Application. IEEE Transaction on. System, Man, and Cybernetics, 1994,24(12):1676~1684

基于本体概念结构的SVM多类分类方法及其在本体自动扩充中的应用*)

A Taxonomy-based Method for Multi-class SVM and its Application on Ontology Auto-population

唐晋韬　王　挺

(国防科学技术大学计算机学院　长沙 410073)

Abstract Ontology auto-population with machine learning methods is an important research field in semantic web. In this paper. Through deeply analysis of the characteristic of ontology population and traditional multi-class SVM methods, a new ontology taxonomy-based method for multi-class SVM is proposed.. Auto-populating tests with free text have been carried out in Chinese time domain. Different multi-class classification strategies and kernel functions are compared to verify their impacts on classification effect. Experiment results indicate that 'Taxonomy-based' multi-SVM method is more effective and efficient than other commonly used methods.

Keywords Ontology population, SVM, Multi-classification, Ontology taxonomy

1 引言

本体是哲学研究中发展出来的一个概念，指形成现象的根本实体。近十多年来，随着研究的日益深入，本体已经远远超过了哲学的范畴，在信息技术、知识工程等方面都有广泛的应用。尤其是本体在Web上的应用直接导致了语义Web的诞生，语义Web企图解决Web信息共享中的语义问题，而本体为Web信息实现知识的融合和交互提供了词汇和规则共享基础。在知识工程领域，本体目前最得到普遍认可的定义是Neches等人(1991)给出的："一个本体定义了组成主题领域的词汇的基本术语和关系，以及用于组合术语和关系以定义词汇的外延的规则。"作为共享的和通用的领域知识，本体可被视为具有明确语义信息且能被机器处理的数据，使网络信息变得结构化，能够包含更多的机器可处理的语义信息，这给基于内容的信息处理技术进一步发展带来了新的契机。

目前，基于本体的信息处理的一个研究热点是如何在本体的基础上从海量数据中抽取不同主题的实例，建立本体种群(population)。从理论上来说，完整的本体应该包括本体中各种概念的实例，这些实例构成了本体的种群。本体自动扩充是指根据领域本体提供的概念和概念间的关系，设计对语料进行分析处理的方法，从中抽取本体概念实例以及概念实例间的关系信息，并以结构化的形式加以存储，这些本体实例及其相互之间的关系构成了一个知识库，为进一步的信息处理如检索、分类、过滤等，提供了结构化的、易于机器"理解"的信息。

随着统计学习理论(SLT)的发展，V. Vapnik发明的支持向量机(SVM)逐步成为机器学习领域最有前途的统计分类器之一。SVM基于结构风险最小化原理，将算法最终化为二次寻优问题，解出全局最优点，避免了局部最优问题；将实际问题通过非线性变换转换到高维特征空间进行求解，避免了维数灾难问题；最优分界面使其在小训练样本集的条件下也具有较高的泛化能力。SVM的这些优点使得它得到了广泛的应用，成为研究的热点。

SVM是依据两类的距离最大化的原则求一个最优分类超平面，将两类样本分开。但是在许多情况下任务的要求是要将多类样本分开，在这些情况下传统的SVM显然无法胜任，必须对其进行改进，以实现多类分类的目的。很多研究者在这方面做了深入的研究，提出了很多改进方法。

本文构建了一个面向开放领域的本体自动扩充系统，采用有监督的SVM模型对本体实例进行分类。并根据本体自动扩充任务的特点，设计了一个基于本体概念结构的SVM多类分类分解方法，利用本体的概念结构信息将多类分类问题按层次分解成多个二类分类问题，以期降低多类SVM的计算复杂性和训练开销，获得更为准确的分类结果。在实验中，我们测试了几种核函数和不同的多分类器对自动扩充任务性能的影响，实验结果很好地说明了基于本体概念结构的SVM多类分解方法的优越性。

在本文第2部分我们简要回顾了相关的工作。第3部分介绍了SVM的基本原理，以及面向自由文本的本体自动扩充系统的基本结构。第4部分比较了几种SVM多类器构建方法的优缺点，并根据本体自动扩充任务的特点，设计了基于本体概念结

*)本课题得到国家自然科学基金资助(60403050)。唐晋韬　博士研究生。

构的多类 SVM 分类器算法。第 5 部分是实验结果以及对于结果的一些讨论。最后是结论和对未来工作的展望。

2 相关工作

SVM 是目前机器学习领域研究的一个热点，在 John C . Platt 提出序列最小优化算法(SMO)以后，SVM 的收敛速度和学习效率都得到了较大的提高。SVM 从此被广泛应用于文本分类、人脸识别等工作，并取得了很好的效果。在这些多类分类领域的识别工作中，研究者们对改进 SVM 使其适应多类分类的问题也进行了深入的研究。总的来说，这些改进措施可以分为两类：一种方式是将多类分类分解为多个二类分类，分别求解，而后再按一定的方式重构起来成为多类分类的解。这类方法中最常用的方法有“一对一方法”、“一对其它方法”，它们均假设所有概念相互独立。为避免忽略概念之间关系而影响分类性能，Scott Selikoff 提出了采用二叉树结构来组织 SVM 多分类器，Thomas Hofmann 等人在此基础上对二叉树的组织方式进行了研究，在文本分类的任务中通过词典等资源判断概念之间的相关程度，依此组织分类二叉树结构。而 FELIX MANDOUX 等人提出利用 Adaptive Code 算法来组织多类分类结构，通过引入先验知识表示概念之间的关系，一次性求出分类决策，避免了“一对其它”等方法的分类错误叠加放大的情况。构造 SVM 多分类器的另一种方式是修改 SVM 二次规划的形式，对多类分类问题一步求解出来，但这种方法会大大增加二次规划的维数，增加了问题的复杂度，目前这一类方法的研究主要集中在如何使得 SVM 能够高效地进行训练和多类分类。如黄景涛等人将遗传算法与传统的 SVM 算法结合，构造出一种参数最优的进化 SVM(GA－SVM)。这种方法将 GA 引入支持向量分类器的训练过程中，利用 GA 的搜索特性来加快 SVM 的求解。

也有不少研究者对统计学习理论和 SVM 在知识工程领域的应用进行了深入的探索。如 Alexander Maedche 等人于 2000 年提出了一个自动发现概念结构和生成本体的通用体系结构，并在电信领域的词典和自由文本中进行了自动学习、发现本体的实验。ZHOU GuoDong 等人以 SVM 为基础，研究了如何组合词典、语法、语义等知识构造特征向量进行关系抽取。

本文在前人工作的基础上，对 SVM 在本体自动扩充任务中的应用进行了进一步的研究和探索。我们针对本体自动扩充任务的特点，设计了一个基于本体概念结构的多类 SVM 分类算法，进行面向自由文本的本体实例的自动扩充。

3 SVM 简介

支持向量机(SVM)理论的主要思想是通过所构造的非线性函数将输入空间映射到高维特征空间，从而在高维特征空间中基于结构风险最小化原则实现线性分类得到最优超平面分类器。

对于训练向量 $x_i \in R^{(n)}$，$y_i \in \{1,-1\}$，$i \in 1\cdots l$。SVM 问题可以描述为：$\min\limits_{w,b,\xi,\rho} \frac{1}{2} \dot{u}^T\dot{u} + C\left(-v\rho+\frac{1}{l}\sum\limits_{l=1}^{1}\xi_i\right)$

s. t. $y_i(w_T\phi(x_i)+b) \geqslant \rho-\xi_i$，$\xi_i \geqslant 0$，$i \in 1\cdots l$，$\rho \geqslant 0$

其中向量 $\dot{u}$ 和标量 b 决定了分类超平面的位置。$\phi(x_i)$把输入向量映射到高维空间。

我们可以通过拉格郎日优化方法和 KKT 条件解出权重 ω 和阈值 b。最后得到如下的分类函数：

$f(x)=\mathrm{sgn}\{(w.x)+b\}=\mathrm{sgn}\{\sum\limits_{i=1}^{n}a_i^* y_i(\phi(x_i)\cdot\phi(x))+b^*)\}$这就是支持向量机。

但是很多分类问题很复杂，不是一个线性问题，简单的超平面不能达到分类的目的。对于非线性问题，可以通过非线性变换转化为某个高维空间中的线性问题，在变换空间求最优分类面。这个变换可能比较复杂，但这个问题中分类函数只涉及训练样本之间的内积运算(x_i, x_j)，在高维空间实际上只需进行内积运算，而这种内积运算是可以用原空间中的函数实现的，我们甚至没有必要知道变换 Φ 的形式。根据泛函的有关理论，只要一种核函数 $K(x_i, x_j)=(\Phi(x_i)\cdot\Phi(x_j))$满足 *Mercer* 条件，它就对应某一变换空间中的内积。因此，在最优分类面中采用适当的内积函数就可以实现某一非线性变换后的线性分类。此时，分类函数变为：

$f(x)=\mathrm{sgn}\{\sum\limits_{i=1}^{n}a_i^* y_i(K(x_i,x))+b^*)\}$

这就提供了解决算法可能导致的“维数灾难”问题的方法：在构造判别函数时，先在输入空间比较向量，对结果再作非线性变换。这样，大的工作量将在输入空间而不是在高维特征空间中完成。大大降低了计算的复杂性。而且只要定义不同的内积函数，就可以在 SVM 中实现多项式逼近、径向基函数(RBF)等许多现有的机器学习算法。

我们基于 SVM 构建了一个本体自动扩充系统，该系统以 Maedche 等人提出的本体自动挖掘体系结构为基础，包括领域本体分析、文本预处理与特征抽取、多分类器构造、SVM 实现、本体实例填充等相对独立的模块；该系统是面向开放领域的，可以通过更换领域本体，指定不同的语料资源和特征向量，进行不同领域的本体实例挖掘工作；该系统是面向自由文本的，在文本预处理阶段通过加入分词、词性

标注、命名实体识别等自然语言处理模块，对文本进行分析和加工，识别某个领域的所有短语实体，生成用户预定义的特征向量，交给 SVM 进行本体实例和概念的匹配；在 SVM 的实现方面，我们采用了 SMO 算法。SMO 采用启发式的迭代策略，将工作样本集的规模减到最小的两个样本，因此迭代过程中每一步的子问题的最优解可以直接用解析的方法求出来，避开了复杂的数值求解优化问题的过程，大大加快了算法的收敛速度。

4 基于本体概念体系的多分类器

一个领域本体中往往包含多个概念，这些概念之间有继承等关系相互关联。因此我们需要将 SVM 改进为多类分类器来完成本体实例与概念的匹配任务。常用的“一对一”、“一对其它”等多类分类分解方法，均假设需要分类的类别都是相互独立的，将所有的概念层次平板化，忽略了类别之间的关系与相似性。“二叉树”的方法虽然使得分类器拥有了一部分层次信息，但是它对概念是任意分类，这些层次信息并不能真正表示概念之间的关系。“Adaptive code”方法注意到了这个问题，可以通过加权的方式引进概念关系等先验知识，也可以利用学习算法来学习到最佳的概念编码。这个算法虽然为我们加入概念之间的先验知识提供了便利，却没有直观地利用概念的分类层次关系（Taxonomy）来构造多分类分类器。而且它需要使用比其它方法多得多的 SVM 二类分类器，对训练的速度有较大的影响。

本体自动扩充任务是以本体中的概念为模板，对语料中的实例进行抽取、分类和信息填充。本体中的概念具有较好的层次关系，我们可以将这些类别知识利用起来，指导对实例的分类工作。因此，我们设计了一个基于本体概念结构（Taxonomy）的 SVM 多分类器构造方法。这个方法以决策树结构为基础，结合了“一对一”方法和本体中概念的分类先验知识。这个方法的构造算法和训练算法描述如下：

如一个本体的某些概念需要进行实例的自动扩充，依据这些概念之间的层次关系，我们可以通过虚拟节点的添加等方法将这些概念组织成一个树结构，树中所有叶节点和一些内部节点都参加分类。基于这棵树结构的多分类器的构造算法如下：

1）读取树的根节点，设为当前节点。

2）对当前节点进行判断，如果该节点是内部节点并参与分类，则转到 3），如果该节点是内部节点但不参与分类，则转到 4），如果该节点为叶节点则转到 5）。

3）如果只有一个参与分类的子节点，则在此使用一个 SVM 二类分类器，否则将当前节点和它所有参与分类的直接子节点一起，按照“一对一”的方式组织一个多类分类器。转到 6）。

4）如果只有一个参与分类的子节点，则将这个子节点设置为当前节点，转到 2），如果有两个子节点，则对这两个子节点使用一个 SVM 二类分类器。如果有多于两个子节点，则将当前节点的所有子节点按照“一对一”的方式组织一个多类分类器。转到 6）。

5）叶节点是需要分类的概念，返回。

6）依次将当前节点的子节点设置为当前节点，递归调用 2）。

这个算法结合分类知识，将多类分类器组织成一个类似于决策树结构的树形结构。如某个本体的概念有如图（1）的层次结构，概念 C_A 是个抽象概念，其它概念均有本体实例需要归类。按照算法对这棵树进行递归遍历，我们可以得到如图（2）所示的结合“一对一”方法的一个类似于决策树的多分类器分类结构。

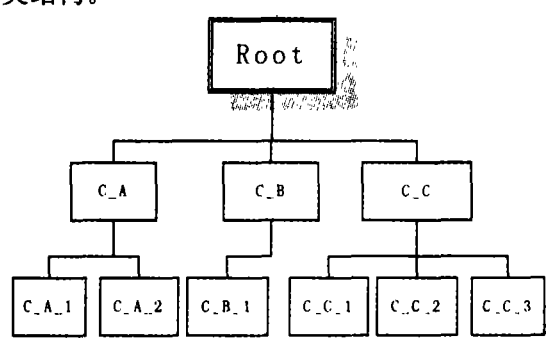


图 1 概念层次树示意图

在这个类似于决策树的多分类器结构中，各个需要分类的类别都变成了叶节点。对这个多类分类器各个 SVM 进行训练的数据集的划分，我们采用“bottom-to-top”的方法进行，为方便直观理解，我们以图 2 所示的多分类器为例来描述数据集的训练算法。

1）将训练数据按所属类别进行划分，例如，我们要将数据集中属于“C_B”这个类别的实例归为一类，命名为“C_B”的实例集。同样的其它几个类别也有自己的实例集，在这里，父节点的实例集不包含子节点的实例，如“C_B”的实例集不包括“C_B_1”的实例。这样就避免了实例集的重叠。因为要分类的类别都是叶节点，我们这一步设定了所有叶节点的实例集。

2）按照深度优先的原则，依次由每个叶节点回溯到上一级节点，称之为父节点。如果父节点有 SVM 二类分类器，以这个 SVM 要区分的两个类别的实例集对其进行训练，一个的实例集都标记为＋1，另一个的标为－1。如果是多分类分类器，则按照

"一对一"分类器的数据训练方法对其进行训练。之后将所有子类别的实例合并，组成父类别的实例集。如"C_A_1"和"C_A_2"的所有实例组成了"C_A"的实例集。

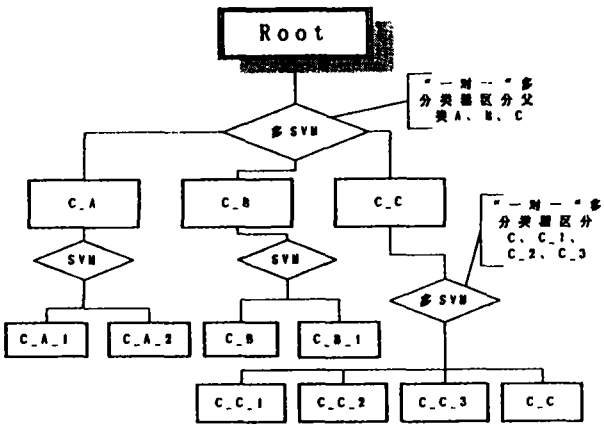


图 2　基于本体概念体系的多分类器构造示意图

3)按照 2)中的方法继续向上回溯，直到根节点的分类器被训练好以后停止。

当一个待分类的样本进入这个网络时，从根节点开始，每一步利用 SVM 或者"一对一"方法进行分类决策，就可以将其归入正确的类别，在此不详细描述这个测试的算法。

5　实验与分析

本文在中文时间领域利用本体自动扩充系统进行了本体自动扩充实验。通过领域分析和概念相似度分析，我们构建了一个包含 15 个概念的中文时间本体，并定义了这些概念的属性和关系。根据自由文本中中文时间描述的特点，我们选取了 110 个能够较好地表示实体的词汇特征和语法特征构成特征向量，抽取时间实体的相关特征用 SVM 多类分类器进行学习和分类。

在对时间描述实体进行分类之前，我们首先通过命名实体识别系统对自由文本进行了处理，识别并标注文本中的时间描述实体，我们的工作是基于这种半标注的语料进行的。在这里，为了不让命名实体识别系统的结果影响到后面分类的结果，我们在完全正确的半标注语料上进行工作。我们使用 2004 年 863 命名实体评测所公布的标注语料进行训练和测试。因为 863 命名实体评测中对时间概念的定义是我们的一个子集，为了覆盖全部分类，我们手工标注了一部分的时间概念。我们在语料库随机挑选了 100 个文本按照时间描述本体的概念层次手工进行了分类，将分类好的文本 50% 进行训练，50%用做测试。这些文本共有 1.12Mb，1832 个被标注的时间实例，覆盖了时间本体中 13 个概念。

对分类效果的量化评测，我们采用和文本分类类似的评估方式。对每个时间描述概念类别的分类结果计算它的查准率，查全率和 F1 值进行量化评估。而对于整个系统的评价则采用宏平均值和微平均值作为量化指标。这些指标包括宏平均和微平均的查准率、查全率和 F1 值。

我们测试了不同的多类分类方法对分类结果的影响，包括："一对一"的方法、"一对其它"的方法、"SVM-Tree"和"基于本体概念结构的多类分类"方法。我们也测试了选择线性核函数和 RBF 形式的核函数对分类结果的影响程度。表 1、表 2 列出了各种情况下的实验结果。

比较不同核函数对分类结果的影响，我们可以发现在同样的条件下，RBF 核函数取得了比线性核函数要好的效果。这说明如何选择核函数对于 SVM 的分类效果还是有一定的影响，但是目前如何选择核函数、如何指定核函数的参数还没有一个科学的方法，需要我们继续深入的研究。

我们的实验也发现，使用不同的多类分类策略对多类分类的性能有着显著的影响，使用不同的构造方法的分类效果最多相差了 2.7 个百分点。"一对一"的方法和"基于本体概念结构"的方法取得了比其它两个方法明显要好的效果，"一对其它"和"二叉树"方法的决策误差会叠加放大可能是其中的一个原因。而我们设计的这个引入先验分类知识的基于本体概念结构的方法在不同核函数下 F-值均要高于"一对一"方法，平均预测时间开销则仅仅是"一对一"方法的 1/3。这可以用语义相似度关系解释，本体概念层次结构中包含丰富的语义信息和类别层次信息，概念在概念层次结构中距离越近，说明这些概念越相似，它们的实例相同的特征就越多。我们在多分类器组织时将其组织在一起，有利于降低训练中噪点的影响，提高分类的正确率。

结束语　本文介绍了在语义 Web 领域越来越受到重视的本体自动扩充工作，使用 SVM 学习算法构建了面向自由文本的本体自动扩充系统。在此

基础上，本文着重研究了如何使用SVM构造多类分类器完成多类分类任务的方法，并根据本体自动扩充任务的特点，设计了基于本体概念结构的多类SVM分类器算法。在实验中我们测试了小样本情况下使用不同的多类分类器和不同的核函数对分类性能的影响。实验结果显示，在相同核函数时，"基于本体概念结构"的多类分类策略要好于现在流行的几种分类策略；而在相同的分类策略下，使用不同的核函数对最终的分类结果也有一定的影响。

目前我们的工作都是在小领域本体和小样本数据的情况下进行，而大规模本体问题所面临的复杂性和难点和小样本情况可能不同。因此，我们下一步工作的重点是面向更大规模的本体进行本体自动扩充，测试不同核函数和不同多类分类策略在大规模数据情况下的性能和效果。针对大规模本体自动扩充的特点，设计更好的多类分类策略，对其它领域进行本体自动扩充。

表1 不同分类策略在线性核函数情况下的分类结果

	SVM个数	宏平均准确率	微平均准确率	宏平均召回率	微平均召回率	宏平均F1值	微平均F1值
一对一	156	89%	90%	86%	90%	87.5%	90%
一对其它	12	86%	88.5%	81%	88.5%	83.5%	88.5%
SVM-Tree	12	85%	88%	82%	89%	83.4%	88.5%
基于本体概念结构	60	89.5%	92%	86%	90.5%	87.7%	91.2%

表2 不同分类策略在RBF核函数情况下的分类结果

	SVM个数	宏平均准确率	微平均准确率	宏平均召回率	微平均召回率	宏平均F1值	微平均F1值
一对一	156	89.5%	91%	87%	91%	88%	91%
一对其它	12	87%	89%	80%	89%	83%	89%
SVM-Tree	12	84%	89%	82%	90%	83%	89.4%
基于本体概念结构	60	91%	93%	86%	90%	88.4%	91.4%

参考文献

1 Neches R, et al. Enabling technology for knowledge sharing. AI magazine, 1991, 12 (3): 36～56
2 Berners-Lee T, Hendler J, &Lassila O. The Semantic Web. Scientific American, May 2001
3 Vapnik V. Nature of Statistical Learning Theory. John Wiley and Sons, Inc., New York, in preparation
4 Dumais S. Decision Theory and Adaptive, Using SVMs for text categorization. In: Marti Hearst, ed. IEEE Intelligent Systems Magazine, Trends and Controversies. 1998, 13(4)
5 徐勋华，王继成．支持向量机的多类分类方法．微电子学与计算机，2004，21(10)
6 Platt J C. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines
7 Keerthi S S, et al. Improvements to Platt's SMO Algorithm for SVM Classifier Design
8 Maedche A, Staab S. Mining Ontologies from Text, EKAW, 2000. 189～202
9 Zhou GuoDong, Su Jian. Machine Learning-based Named Entity Recognition via Effective Integration of Various Evidences. Natural Language Engineering, Cambridge Press, ISSN 1469～8110, 2005, 11(1)
10 Selikoff S. The SVM-Tree Algorithm: A New Method for Handling Multi-Class SVMs
11 Hofmann T, Cai L. Learning with Taxonomies: Classifying Documents and Words. In: proceedings of NIPS, 2003
12 Mandoux F. Multi-Class SVM Learning using Adaptive Code, Master's Degree Project, Stockholm, Sweden, 2004
13 黄景涛，马龙华，钱积新．一种用于多分类问题的改进支持向量机．浙江大学学报(工学版)，2004，38

(上接第328页)

约简的有效性。

(4)局部均衡度

设定比值$|U_{CF}|/|U_B|$，描述了决策子表B对中心样本的包含程度，直观地可有$0 \leqslant |U_{CF}|/|U_B| \leqslant 1$。若F族中有一定数量决策子表的比值$|U_{CF}|/|U_B|$较高，体现了抽样的局部集中性，即使整体均衡性比值可以接受，但却存在着局部的样本堆砌。

上述四个比值参数从整体和局部两个方面对F族进行了衡量，从对各自含义的分析来看，当它们的值处于一个合理取值区间时，就表明F族抽样是均衡合理的，动态约简是有效性。

结论 动态约简实质是进行多次抽样，把复杂大型决策信息系统的约简问题转化为若干子决策表最优约简的交集问题，其子表族抽取是动态约简结果有效与否的关键。文中对动态约简理论中子表族的确定问题进行了详细阐述，指出了Bazan的思想中存在的问题，给出了确定动态约简子表族大小的方法，并讨论了抽样质量评价问题，为进一步的研究工作建立了理论基础。

参考文献

1 Jelonek J, Krawiec K, Slowinski R. Rough Set Reduction of Attributes and their Domains for Neural Networks. Computational Intelligence, 1995, 11(2): 339～347
2 Bazan J, Skowron A, Synak P. Dynamic Reducts as a Tool for Extracting Laws from Decision Tables. In: Methodologies for Intelligent System. Proc. 8th International Symposium ISMIS'94, Charlotte, NG, October 1994, LNAI vol. 869, Springer Verlag 1994. 346～355
3 Bazan J. A Comparison of Dynamic and non-Dynamic Rough Set Methods for Extracting Laws from Decision Tables[C]. In: Polkowski, Skowron, eds. Rough sets in Knowledge Discovery 1: Methodology and Applications. Physica-Verlag, Heidelberg, 1998. 321～365
4 Jan G, Bazan J. Dynamic Reducts and Statistical Inference. In: Proceedings of the Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96), July, Granada, Spain, 1996, (2): 1147～1152

利用阴影信息检测高分辨率遥感图像中的建筑物*)

Applying Shadow Information to Detect Buildings in High-Resolution Remote Sensing Images

强永刚　殷建平　陈　涛　张国敏

(国防科技大学计算机学院　长沙 410073)

Abstract　A new method for detecting buildings in high-resolution remote sensing images is proposed. In this method, we only consider the shadow information of the buildings in remote sensing image. It needn't to know the information of the satellite while it gets the images. Experiments verifies its effectiveness for detecting buildings in high-resolution remote sensing images.

Keywords　High-resolution remote sensing images, Shadow, SUSAN, Building detection

1　引言

在遥感图像中，阴影是理解、分析图像的一个显著特征。在特定的遥感图像中，结合成像时太阳、卫星等的参数，利用阴影信息对建筑物的高度、面积以及形状进行估测已经取得了比较好的结果[1,2]，也可以利用阴影信息对建筑物大致定位，然后利用剪切融合算法结合概率模型对建筑物的轮廓进行识别[3]，还可以利用阴影对图像分割的结果进行验证、细化[4]。在未知遥感图像成像参数的情况下，本文利用阴影信息对遥感图像中的建筑物进行检测，实验结果表明，本文算法是一种有效的建筑物检测算法。

2　问题描述及算法概述

2.1　太阳、卫星、建筑物、图像阴影的几何位置关系假定

相对建筑物来说，太阳光的照射分为垂直照射和斜照射两种情况，对于太阳光垂直照射或近似垂直照射的情况，图像中的阴影信息变得非常少，通过建立屋顶模型提取直线构建矩形来提取建筑物[5,6]，本文考虑太阳斜照射的情况；卫星的拍照角度分为垂直拍照和斜拍照，对于垂直拍照的情况，也可以通过屋顶的模型或两幅遥感图像进行比对检测建筑物，本文考虑卫星拍照时和建筑物有一定角度的情况，它们的位置关系示意图如图 1 所示。在本文中，考虑 $\alpha>0$，$\beta>0$ 的情况。

根据常识可知，建筑物的体积具有一定的范围，也即建筑物在遥感图像中的阴影面积在一定范围内，同时从大量的遥感图像观察我们发现，在有阴影信息的遥感图像中，物体一般都有一个向阳面和一个背阳面，这两个面的交线(在图 1 中用粗的黑线表示)一般和地面垂直或近似垂直，本文就是利用这个结论得到以下检测算法。

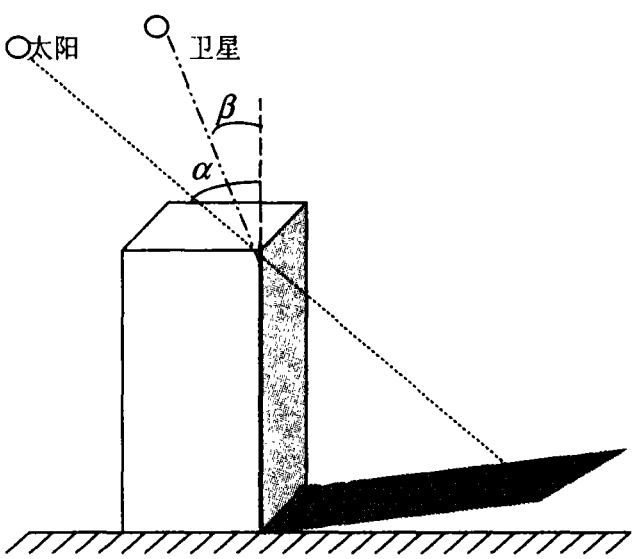


图 1　太阳、卫星以及建筑物的位置假定

2.2　算法概述

根据以上的假定，可以将利用阴影信息检测遥感图像中的建筑物分为以下几个步骤：

step1：识别出遥感图像中的阴影区域，本文通过统计的方法得到一个判定阈值，对图像进行分割，得到候选阴影区域；

step2：利用建筑物的阴影具有一定面积去除候选区域面积过小和过大的区域；

*)基金项目：国家自然科学基金第 60373023 号资助。强永刚　硕士生，研究方向为图像处理，模式识别；殷建平　教授，博士生导师，研究方向为网络算法、模式识别与人工智能、信息安全等；陈　涛　硕士生，研究方向为知识检索，语义网络；张国敏　博士生，研究方向为图像处理，模式识别。

step3：利用 SUSAN 算子对阴影区域进行边缘检测，得到阴影的边缘；

step4：利用形态学方法检测阴影区域的直线，主要检测垂直或近似垂直地面的直线，通过直线聚类抽取长直线；

step5：统计阴影和垂直方向的直线的相对位置，以此来估计太阳照射方向；

step6：沿太阳光的逆方向搜索建筑物，并进行标定。

3 算法实现

3.1 统计计算分割阴影区域阈值

本文通过统计阴影区域的灰度值，采用的抽样方法是在特征点处取它的 3×3 的邻域窗口，求窗口内各点的灰度平均值为该特征点灰度值，在不同的区域各取 100 个特征点。图 2 是统计分析的柱状图，横轴是灰度值，纵轴为特征点个数，其中深灰色的条对应阴影区域特征点的分布，黑色的条对应建筑物的背阳面的特征点的分布，浅灰色的条对应建筑向阳面特征点的分布。

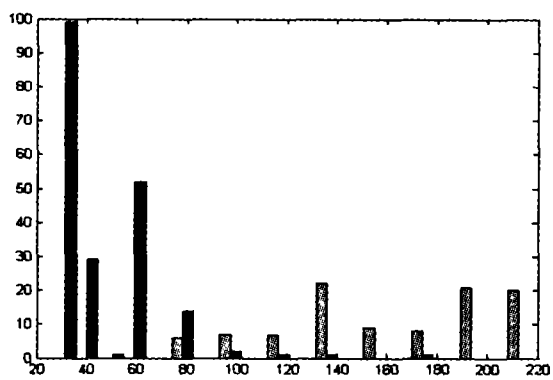


图 2 不同区域的灰度值分布

通过图 2 可以得出，为了分割出阴影区域，检测建筑物向阳面和背阳面的交接线，把阈值设为 80 比较合适。

3.2 利用 SUSAN 算法进行阴影区域边缘检测

SUSAN 算法[8]将圆形窗口模板中心置于图像的每一个位置上，计算窗口中心点 r0 与窗口内其它像素点 r 具有相近亮度的点的个数 n(r0)，以确定该像素是否是图像边缘点。n(r0)按下式计算：$n(r_0)=\sum_r c(r,r_0)$，其中 $c(r, r_0)$表示窗口内点 r 的亮度 I(r)与窗口中心点 r_0 的亮度 $I(r_0)$的相似程度。$c(r, r_0)=e^{-(\frac{|I(r)-I(r_0)|}{t})^6}$，其中 t 是亮度阈值。

SUSAN 算法在对边缘检测不用计算微分，所以 SUSAN 算法对噪声不是很敏感，且它对角点有比较强的响应，结合遥感图像在成像时可能受到噪声的干扰，用其他的检测算子可能会在角点处以及边缘模糊的情况下失效，而 SUSAN 算法只要设定适当的几何阈值就可以比较好地适应检测的需要。

3.3 形态学方法检测直线

在检测图像中特定结构信息时，形态学方法是一种很好的选择，它比其它的空域和时域方法有明显优势，表现在它的运算速度快，可以实现并行算法，而且可以很容易地加入目标的先验信息。假设边缘图像 I_b 中为 1 的像素集合是 A，定义 B={(0, 1),(0,0),(0,−1)}为竖直结构元素，在边缘图像中检测竖直直线的运算就表示为 A∘B，其中∘是形态学开运算[7]。

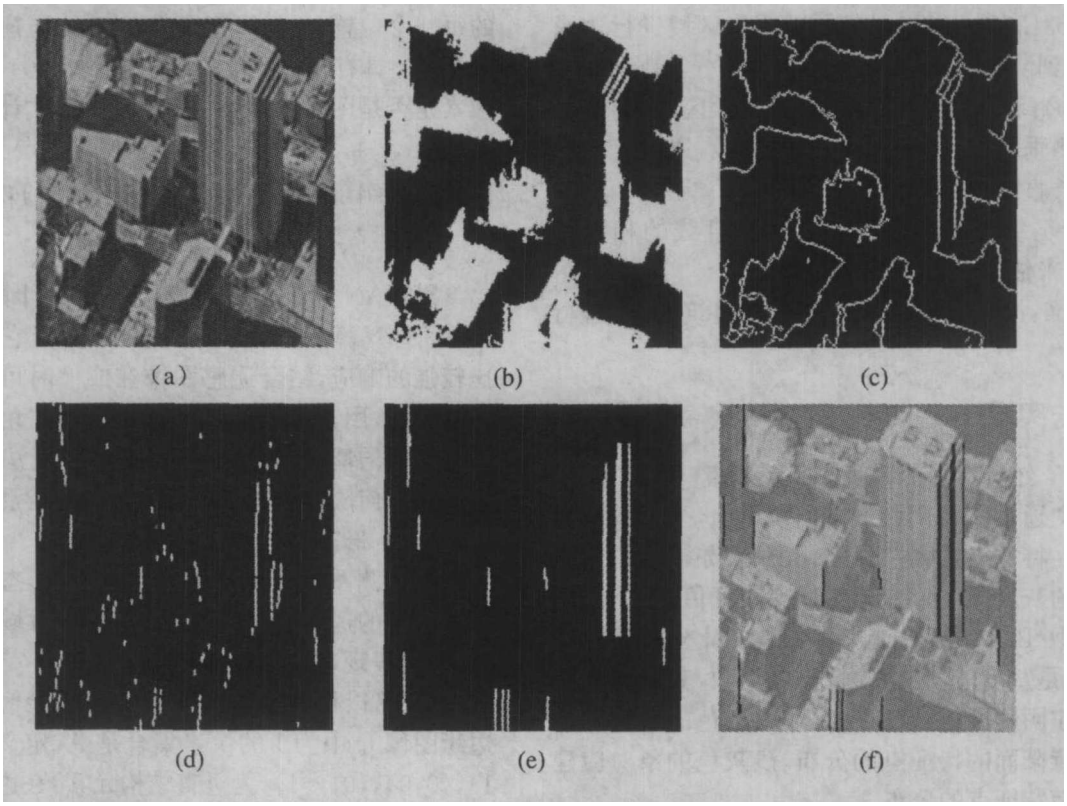
因为噪声对边缘的影响，使得通过形态学计算后的图像中的竖直直线可能会有断裂，本文通过在候选阴影区域附近对竖直直线进行聚类并连接为建筑物的候选边缘线。

3.4 定位建筑物

利用检测出来的直线相对它最近的阴影区域位置，推测阴影的方向，然后逆阴影方向寻找建筑物，这时候可以利用建筑物的区域特性，对建筑物进行定位。

4 实验结果

本文对一幅 Quickbird 拍摄的遥感图像进行了建筑物的定位，实验结果如下：



(a)原始的 quickbird 遥感图像　(b)经过阈值化的候选阴影区域
(c)用 SUSAN 算子进行边缘检测的结果　(d)用形态学检测到的竖直直线
(e)利用知识进行直线聚类并去除短小直线的结果　(f)叠加到原图中

图 3　实验结果

结束语　从算法的结果来看，比较准确地定位了遥感图像中的建筑物，但有些建筑物定位还不是很准，可以结合图像分割、区域一致描述以及建筑物纹理分析等一些处理手段来解决这个问题，这也是我们下一步的工作。

参 考 文 献

1　Tupin T C, Maitre F H. Retrieval of building shapes from shadows in high resolution SAR interferometric images. In: Geoscience and Remote Sensing Symposium, Proceedings. 2004 IEEE International, 2004, 3: 1788～1791

2　何国金，陈刚，何晓云，等. 利用 SPOT 图像阴影提取城市建筑物高度及其分布信息[J]. 中国图象图形学报，2001(6)：425～428

3　刘炜. 基于概率模型的高分辨率卫星图像建筑物识别及变化检测：[硕士学位论文][D]. 中国科学院自动化研究所，2005

4　杨益军，赵荣椿，汪文秉. 航空图像中人工建筑物的自动检测[J]. 计算机工程，2002，28(8)：20～21

5　Liu Z J, Wang J Liu, W P. Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform. Geoscience and Remote Sensing Symposium, Proceedings, 2005, 4(7): 2250～2253

6　汪行，陈学佺，金敏. 线段提取在高分辨率遥感图像建筑物识别中的应用[J]. 计算机辅助设计与图形学学报，2005，17(5)：928～934

7　阮秋琦. 数字图像处理学[M]. 北京：电子工业出版社，2001

8　Smith M, Brady J M. SUSAN-A New Approach to LowLevel Image Processing[J]. Int. Journal of Computer Vision, 1997, 23(1): 45～78

一种多 agent 活动理论

A Theory of Multi-agent Action

刘 越 陈跃新 周丽涛

(国防科技大学计算机学院计算机系 长沙 410073)

Abstract The study of multi-agent action has received a great deal of attention in recent years. A number of theoretical formalization for such multi-agent system has been proposed. However most of these formalization do not have a strong semantic basis nor a sound and complete axiomitizaton. We develop a formal system which combines the BDI logic and dynamic programming logic to assist the cooperation of multi-agent system.

Keywords Joint plan, Multi-agents action

1 介绍

目前对联合活动理论的研究是哲学和 DAI 界广为关注的话题,诸多问题未达成共识,尽管已有许多实用的多 agent 系统,但在数理基础与实践之间尚有甚大间隙。我们的工作是开发一种将 BDI 逻辑与动态逻辑风格的程序逻辑相组合的形式体系为渐近式合作[2](一阶近似)中的 agent 做出形式规范或模型,其中许多细节被抽象或概括掉。也想对联合活动的理论基础做一点探讨。

2 形式框架

本文对 BDI 框架的研究工作[3,4]做了扩充,提出了一种新型 BDIP 逻辑$\mathcal{L}$,较深入地探讨了信念、愿望、意图三种"思想状态"间的关系,按语义对象给出了计划的形式定义以及计划执行的语义,最后给出了语言$\mathcal{L}$的若干重要性质。

2.1 语法

定义 1 语言$\mathcal{L}$含下述符号:谓词符号的可数集 Pred;函数符号的可数集 Fun;变量符号的可数集 Var;算子符号 true、Bel、Goal、Int、Agts、=、∈、A、Pre、Post、Body、Has、Holds、Exec、U、O。

定义 2 类型 σ 可为 A_g、A_c、Π、B、G_r、C 或 U。σ 的项的集合 Termσ 定义为:

a)若 $x\in Var_\sigma$,则 $x\in Term_\sigma$;

b)若 $f\in Fun_\sigma$,arity(f)=n,

且$\{\tau_1,\cdots,\tau_n\}\subseteq Term$,

则 $f\{\tau_1,\cdots,\tau_n\}\in Term_\sigma$,其中

$Term=\bigcup\{Term_\sigma\mid\sigma\in\{Ag,Ac,\Pi,B,Gr,C,U\}\}$$\mathcal{L}$的语法定义如下:

- ⟨ag−term⟩ ::= $Term_{Ag}$的任一元素;
  ⟨Π−term⟩::= $Term_\Pi$ 的任一元素;
  ⟨β−term⟩::= $Term_B$ 的任一元素;
  ⟨gr−term⟩::= $Term_{Gr}$的任一元素;
  ⟨C−term⟩::= $Term_C$ 的任一元素;
  ⟨term⟩::=Term 的任一元素;
  ⟨pred−sym⟩::=pred 的任一元素;
  ⟨var⟩::=Var 的任一元素;
- ⟨state−fmla⟩::=
true|⟨pred−sym⟩(⟨term⟩,…,⟨term⟩)|¬⟨state−fmla⟩|
|⟨state−fmla⟩∨⟨state−fmla⟩|
∀⟨var⟩⟨state−fmla⟩|A⟨path−fmla⟩|
(⟨term⟩=⟨term⟩)|(⟨ag−term⟩∈⟨gr−term⟩)|
(Bel⟨ag−term⟩⟨state−fmla⟩)|
(Good⟨ag-term⟩⟨state−fmla⟩)
|(Int⟨ag−term⟩⟨state−fmla⟩)|
(Has⟨ag−term⟩⟨Π−term⟩)
|(Body⟨Π−term⟩⟨β−term⟩)|(Pre⟨Π−term⟩)|
(Post⟨Π−term⟩)|(Agts⟨β−term⟩⟨gr−term⟩)|
(Holds⟨c−term⟩)|。
- ⟨path−fmla⟩ ::=
⟨state−fmla⟩|¬⟨path−fmla⟩|
⟨path−fmla⟩∨⟨path−fmla⟩|
∀⟨var⟩⟨path−fmla⟩|⟨path−fmla⟩∨⟨path⟩|
O⟨path−fmla⟩|(Exec⟨β−term⟩)。
- ⟨fmla⟩::=⟨state−fmla⟩。

2.2 语义

定义 3 量化域 D 为 $D_{Ag}\cup D_A\cup D_\Pi\cup D_B\cup D_G\cup D_C\cup D_U$,其中:$D_{Ag}$为 agent 的非空集合;$D_A$为原本活动的非空集合;$D_\Pi$ 为计划的非空集合;D_B为计划体的集合;D_G 为 D_{Ag} 上的 agent 团体的集合;D_C 为状态的非空集;D_U 为其它个体(如积木)的非空集。

我们要求:D_B 的元素中的各活动都是 D_A 的元素;D_Π 的元素中的各计划体都属于 D_B;包含在 D_B

元素中的任何计划体也在 D_B 中。

逻辑 $\mathcal{L}$ 允许我们表示可按不同方式演进的系统的性质，系统的演进方式取决于系统内各 agent 做出的选择。我们用二元分支时间关系 $R\subseteq T\times T$ 模拟系统历史的所有可能的进程，其中 T 为时间点集。任一时间点都可通过由某 agent 执行一原本活动跃迁到另一时间点：R 中的各弧都对应这类活动的完成。

定义 4 世界 W 为 (T',R')，其中 $T'\subseteq T$ 为非空时间点集且 $R'\subseteq T'\times T'$ 为 T' 上倒线性分支时间全关系。令 W 为所有世界的集合(在 T 上)。若 $w\in W$，则记 w 中的时间点集为 T_w，分支时间关系记为 R_w。

定义 5 序偶 (w,t) 称为状态，其中 $w\in W$，$t\in T_w$，若 $w\in W$，则 $S_w=\{(w,t)\mid t\in T_w\}$ 为 w 中所有状态的集合。$S=\bigcup_{w\in W}S_w$ 为所有状态的集合。

定义 6 令 $w\in W$，则穿越 w 的有限路径为序列 $(t_0,t_1,\cdots,t_k)$，满足 $\forall u\in\{0,\cdots,k-1\}$，有 $(t_u,t_{u+1}\in R_w$。令 $fpaths(w)$ 为穿越 w 的有限路径的集合。穿越的无限路径为序列 $(t_u\mid u\in N)$，满足 $\forall u\in N$ 有 $(t_u,t_{u+1})\in R_w$。令 $paths(w)$ 为穿越 w 的无限路径的集合。若 p 为路径，则 $p(o)$ 为 p 中第一时间点，$p(1)$ 为第二时间点…。若 p 为路径，$u\in N$，则 $p(u)$ 为由 p 中去掉前 u 个时间点得出的路径。

计划体的结构定义如下：

$\langle sit-set\rangle ::=\wp(S)$ 的任一元素

$\langle plan-body\rangle ::=D_{Ac}$ 的任一元素

$\quad\mid\langle plan-body\rangle;\langle plan-body\rangle\mid$

$\quad\langle plan-body\rangle'\mid'\langle plan-body\rangle$

$\quad\mid\langle plan-body\rangle\mid\mid\langle plan-body\rangle\mid$

$\quad\langle plan-body\rangle *\mid\langle sit-set\rangle?$

其中测试活动的变元为一状态集：$C?$ 将成功，若当前状态为集合 C 的成员。测试条件的更自然的表示可能是语言的公式：$\varphi?$ 将成功，若公式 φ 在当前状态中被满足。语言 $\mathcal{L}$ 对描述计划尚有诸多困难，这里不赘述。

定义 7 计划描述为二元关系 $\delta\subseteq S\times S$，满足若 $((w,t),(w',t'))\in\delta$ 则 $w=w'$。令 Δ 为所有计划描述的集合。若 $\delta\in\Delta$ 描述 $\beta\in D_B$ 的行为，则：

- $dom\delta$ 表示 β 的执行可合法开始的状态集，即 β 的前提；
- $ran\delta$ 表示自 $dom\delta$ 中的一状态开始，执行 β 可产生的状态集，即 β 的后果；

若 $(s,s')\in\delta$，则 s' 为自 s 状态开始执行 β 可能产生的一状态。

定义 8 计划为序偶 (β,δ)，其中 $\beta\in D_B$，$\delta\in\Delta$。$D_\Pi=D_B\times\Delta$。若 $\pi\in D_\Pi$，则令 $\hat{\beta}(\pi)\in D_B$ 指称 π 的体，$\hat{\delta}(\pi)\in\Delta$ 指标 π 的描述。$dom\hat{\delta}$、$ran\hat{\delta}(\pi)$ 分别表示 π 的前题和后果。若 $s\in dom\hat{\delta}(\pi)$，则记 $\hat{\delta}(\pi)(s)=\{s'\mid(s,s')\in\hat{\delta}(\pi)\}$。

若 $\beta\in D_B$，则可用 $agents(\beta)$ 指称完成 β 中各活动可能需要的所有 agent 的集合：

$agents(\alpha)\overset{def}{=}\{Agt(\alpha)\}\ \alpha\in D_{AC}$；

$agents(\beta *)\overset{def}{=}agents(\beta)$；

$agents(\beta\oplus\beta')\overset{def}{=}agents(\beta)\cup agents(\beta')\oplus\in\{;,\mid,\|\}$；

$agents(C?)=\phi$(空集)

用元谓词 $exec$ 说明计划执行的语义：$exec(\beta,u,v,p)$ 只当 $\beta\in D_B$ 在路径 p 上“时刻”$u,v\in N$ 之间被执行时成立。为定义 $exec$ 要用到函数 $Act:R\to D_{Ac}$，它将 R 中的每条弧都与一活动相结合。

$exec(\alpha,u,v,p)$

- $iff\ v=u+1, Aat(p(u),p(v))=\alpha$ $\alpha\in D_{Ac}$
- $exec(\beta,\beta',u,p)\ iff\ \exists k\in\{u,\cdots,v\}$ 使得 $exec(\beta,u,k,p)$ 且 $exec(\beta',k,v,p)$
- $exec(\beta\mid\beta',u,v,p)\ iff\ exec(\beta,u,v,p)$ 或 $exec(\beta',u,v,p)$
- $exec(\beta\|\beta',u,v,p)\ iff\ exec(\beta,u,v,p)$ 且 $exec(\beta',u,v,p)$
- $exec(\beta^*,u,v,p)\ iff\ u=v$ 或
- $exec(\beta;(\beta^*),u,v,p)$ 为不动点方程，描述 β^* 的执行：或者不做什么或者执行 β 一次而后执行 β^*。
- $exec(c?,u,v,p)\ iff\ (w,p(u))\in c$。

实际上 w 总是受囿的，p 为穿越 w 的路径。

假定计划是“正确的”，即，$\forall\pi\in D_\Pi$，$\forall w\in W$，$\forall p\in paths(w)$，$\forall u,v\in N$，若 $exec(\hat{\beta}(\pi),u,v,p)$ 则 $(w,p(u))\in dom\hat{\delta}(\pi)$ 且 $(w,p(v))\in\hat{\delta}(\pi)((w,p(u)))$。

定义 9 L 的模型 M 为结构

$M=(T,R,W,D_{Ag},D_{Ac},D_\Pi,D_B,D_{Gr},D_C,D_U,$
$Act,Agt,PL,BR,DR,IR,F,\Phi)$

- T 为所有时间点的集合；
- $R\subseteq T\times T$ 为 T 上倒线性分枝时间全关系；
- W 为世界的集合，使得 $\forall w\in W$ 有 $Tw\subseteq T$，且 Rw 是由 R 中去掉不全含 Tw 中元素的任何弧得出的关系；
- $Agt:D_{Ac}\to D_{Ag}$；
- $PL:D_{Ag}\times W\times T\to\wp(D_\Pi)$ 给出各 agent 在各状态中的计划库；
- $BR:D_{Ag}\to\wp(W\times T\times W)$ 各 agent 都结合一可继续、传递、欧几里得信念可达关系；
- $DR:D_{Ag}\to\wp(W\times T\times W)$，$IR:D_{Ag}\to\wp(W\times T\times W)$ 各 agent 都结合一可继续愿望可达关系以及可继续意图可达关系。

为描述计划体的结构我们引入中缀形式带引号的函数项，如，对任意 $\beta,\beta'\in Term_B$，则$[\beta,\beta']$指称 $\|\beta\|;\|\beta'\|$。对测试(?)和重复(*)仍按后缀形式但带引号。对测试活动需做更多的说明。令 $c\in Term_C$，c 指称一状态集(即条件)，φ 为状态公式。定义 $(c\equiv\varphi)\stackrel{def}{=}A\Box((Holds\ c)\forall\Leftrightarrow\varphi)$，即 φ 只在 c 指称的那些状态中被满足。若$(c\equiv\varphi)$，我们用 $\varphi?$ 代替 c?。这样路径公式(Exec[(Bel i p)?])应理解为 $\forall c(c\equiv(Bel\ i\ p))\Rightarrow(Exev[c?])$。

2.3 一些性质

下面给出几个有效的公式。

- $\models_p([Exec\ \ulcorner await\urcorner])\Leftrightarrow\Diamond\varphi$
- $\models_p(Plan\ \pi\ \varphi\ \psi\ \beta)\wedge((Plan\ \pi'\ \varphi'\ \psi'\ \beta')\wedge((Exec\ulcorner\beta|\beta'\urcorner)\Rightarrow((Exec\ulcorner(\psi\vee\psi')?;(\beta|\beta')\urcorner)$
- $\models_s(Bel\ i\ (Plan\ \pi\ \varphi\ \psi\ \beta))\wedge((Bel\ i\ (Plan\ \pi'\ \varphi'\ \psi'\ \beta'))\wedge((Bel\ i\ A(Exec\ulcorner\beta|\beta'\urcorner)\Rightarrow((Bel\ i\ \varphi\vee\varphi')$

篇幅所限，$\mathcal{L}$的其它性质不再给出。

结论 用 π 指称 agent 计划库中现有的某计划。各 agent 计划执行的同步、合作中冲突消解等问题文中未做讨论，按文中的形式模型这些问题的处理嵌入到计划的前题、计划体的描述，对要执行的某 π 的“一票否决”或“一致同意”中，更大方面是 agent 的“互相支持”和是“随和的”假设。采用的协调方法类似于文[1]中“承诺与约定”。若某 π 的执行未达到预期后果，则从实际达到的世界状态出发向 φ 前进。若考虑 agent 间是竞争的，模型复杂度剧增，模型会变得无用，团体 g 的形成是棘手问题。PGP 作为一种重要的协调技术与渐近式合作在思路上有共同之处，差别可能是渐近式合作中的联合计划是隐式、回顾性的，随联合活动的进行逐渐呈现出(通过回顾)全局计划以及哪些 agent 为达到目标做了贡献。按我们的形式途径在何程度上能描述 PGP 技术将是有趣的课题。

参 考 文 献

1 Wooldridge M J, Jennings N R. Cooperative Problem Solving. Journal of Logic and Computation, 1999

2 刘越，等. 一种渐进式多 agent 合作理论. 计算机学报，2003(4)

3 Werner E. What can agents do together: A semantics of co-operative ability. In: Proceedings of the Ninth European Conference on Artificial Intelligence(ECAI'90)

4 Wooldridge M, Jennings N R. Intelligent Agent: theory and practice. The knowledge engineering review, 10(2): 115～152

(上接第 301 页)

正常文件被判断为病毒的概率。

表 2 测试结果

	False Negative	False Positive	准确率	查全率	误报率
复合贝叶斯	13	44	89.59%	90.08%	10.55%
半增量贝叶斯	5	30	93.61%	96.18%	7.19%

从表 2 所示测试结果看来，半增量贝叶斯算法效果令人满意。

结论 综上所述，计算机病毒问题已引起人们的广泛关注，并成为计算机安全领域的重点问题之一。目前，存在着多种计算机病毒防治技术，其中对未知病毒的检测技术可以发现未知的病毒，有效地改善查杀总是落后于新病毒产生这一现状。未知病毒查杀技术已成为计算机安全领域的一项重要技术。本文给出了未知病毒检测系统的基本框架，通过对朴素贝叶斯模型、复合贝叶斯模型的分析研究，提出一改进模型——半增量贝叶斯模型，此模型继承了朴素贝叶斯模型和复合贝叶斯模型的优点，有效地避免了这两种模型的缺点。

参 考 文 献

1 Harney H, Muchenh Irn C. Group key management protocol (GKMP) specification [S]. RFC 2093, July 1997

2 Harney H, Muchenh C. Group key management protocol (GKMP) architecture [S]. RFC 2094, July 1997

3 Wong C K, Gouda M, Lam S S. Secure group communication u2sing key graphs [J]. IEEE/ACM Transactions on Networking, 2000, 8(1): 16～30

4 刘淼. 基于贝叶斯算法的未知病毒检测的研究. 湖南理工学院学报(自然科学版)，2005, 18(1): 18～22

5 宫秀军，刘少辉，史忠植. 一种增量贝叶斯模型. 计算机科学报，2002, 25(6): 645～650

6 姜卯生，王浩，姚宏亮. 朴素贝叶斯分类器增量学习序列算法研究. 计算机工程与应用，2004, 14, 57～59

7 张凡. 面向未知病毒检测方法与系统实现技术研究：[硕士学位论文]. 西北工业大学，2003

8 张利军. 数据挖掘系统及其应用研究—用关联特征提高朴素贝叶斯文本分类器的性能：[硕士学位论文]. 西北工业大学，2003

9 蔡志平. 计算机病毒检测技术研究与实现：[工学硕士学位论文]. 国防科学技术大学，2001