“软件人”感知系统的协同分类模型研究*)

米爱中[1,2] 徐国章[1] 曾广平[1] 涂序彦[1]

(北京科技大学信息工程学院 北京 100083)[1] (河南理工大学计算机科学与技术学院 焦作 454000)[2]

摘 要 “软件人”是计算机网络世界中的一类软件人工生命,是一种网络中的“虚拟机器人”,它具有拟人结构。作为对人的模拟,其感知系统应该像人的感知系统一样,具有区分感知对象的能力。基于人体感知系统的启发,运用大系统分析的“分解-集结”方法,本文提出了一种“软件人”感知系统的协同分类模型。该模型通过多分类器系统模拟人的不同感觉,从而实现“软件人”对感知对象的多感觉协同分类。仿真实验结果初步验证了本文提出模型的可行性。今后的工作是研究如何构造实际系统并将其应用于数字气田建设的数据处理工作中。

关键词 “软件人”,感知系统,协同分类,多分类器系统

Research on the Cooperative Classification Model in the Perception System of SoftMan

MI Ai-Zhong[1,2] XU Guo-Zhang[1] ZENG Guang-Ping[1] TU Xu-Yan[1]

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)[1]

(School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000)[2]

Abstract SoftMan, the virtual robot in network environment, is a kind of software Artificial Life living in computer networks. It has a humanoid structure, and simulating human, its perception system should be able to recognize the perceptive object. Enlightened by human perception system and the “disassemble-integration” method in analyzing large systems, a cooperative classification model in SoftMan's perception system is proposed. In the model, different humanized senses are simulated by multiple classifier systems. Consequently, SoftMan can perform multi-sense cooperative classification on its perceptive objects. A simulated experiment validates the basic feasibility of the model. The future work is how to construct a practical system and apply it to the data processing work in building digital gas fields.

Keywords SoftMan, Perception system, Cooperative classification, Multiple classifier systems

1 引言

在作为国家能源的支柱型企业的石油天然气行业中,油气勘探开发是本行业的主体。从世界石油天然气勘探开发技术进步来看,信息技术已成为推动石油工业飞速发展的重要内在动力。数字气田是以气田资源数字化为基础,以现代多种通讯网络为依托,以信息技术为手段,以推动勘探开发科研创新、优化生产运行、规范经营管理、提高科学决策水平为目的的复杂大系统。计算机网络是数字气田的基础设施,因此应用先进的网络技术是数字气田建设的客观要求。

“软件人”(SoftMan, SM)[1]技术融合分布式人工智能、并行分布式系统、移动 Agent 和人工生命技术,是计算机网络时代的一项崭新的关键技术。研究“软件人”的目的主要是为当前网络中存在的许多问题或不足提供一种新的有效的解决方式。把“软件人”的技术和思想应用于数字气田建设,可以提高数字气田大系统的智能性、准确性和有效性。

2 “软件人”及其感知

2.1 “软件人”的概念

随着信息技术的飞速发展,数据已逐渐取代程序成为软件研发人员的首要考虑,面向对象的方法也几乎取代了面向过程的方法。而一种新的设计思想——面向 Agent 的方法正越来越受到重视,这种思想的主要特点是把软件像人一样分派到大量数据的地方去“工作”,而不是用传统的方法把数据发给软件处理。“软件人”即是应用这一思想,从广义人工生命观点出发,为了延伸、扩展人的生命而提出的,具有类似于

*)国家自然科学基金项目(60375038,60503024)。**米爱中** 博士生,主要研究方向:人工智能,计算机网络。**徐国章** 博士生,副教授,主要研究方向:人工智能、计算机网络;**曾广平** 博士,教授,博士生导师,主要研究方向:网络、通信、人工智能及应用;**涂序彦** 教授,博士生导师,主要研究方向:人工智能、大系统控制、智能管理、人工生命。

13 Bates J. The Role of Emotion in Believable Agents[J]. Communications of ACM, 1992, 37(7):122~125

14 Ortony A, Clore G, Collins A. The Cognitive Structure of Emotions[M]. Cambridge: Cambridge University Press, 1988

15 张智星,孙春在,水谷英二. 神经-模糊和软计算[M],西安:西安交通大学出版社,2000

16 Price D D, Barrell J E, Barrell J J. A Quantitative-Experiential Analysis of Human Emotions[J]. Motivation and Emotions, 1985, 9(1):19~38

17 Conati C, Zhou X. Modeling students' emotions from cognitive appraisal in educational games[A]. In: Cerri S A, Gouarderes G, Paraguacu F, eds. Intelligent Tutoring Systems, LNCS 2363, 2002. 944~954

18 Chaffar S, Frasson C. Using an Emotional Intelligent Agent to Improve the Learner's Performance[A]. In: Workshop on Emotional and Social Intelligence in Learning Environments, International Conference of Intelligent Tutoring System (ITS), Maceio, Brasil, 2004

19 O'Regan K. Emotion and E-Learning [J]. Journal of Asynchronous Learning Networks, 2003, 7(3): 78~92

人的活性的、生存并活动于计算机网络世界中的一类软件人工生命，是一种网络中的“虚拟机器人”，喻示软件像人一样在它所在的“网络社会”中处理各种事务。与 Agent 相比，“软件人”不仅具有全面的拟人智能、拟人行为和功能，而且具有环境识别和自主决策能力及自由意志。“软件人”能够在网上自由迁移，采用“信息推拉技术”自动处理某些指定的任务，充当某类职员角色，如网上软件邮递员（网络通信）、软件资料员（数据采集）、软件医生（反病毒）、软件卫士（防黑客）、软件售货员（电子商务）、软件清洁工（清除垃圾）等。

2.2 “软件人”的感知

“软件人”的基本构造特征是结构分区，即按照人的基本构造对“软件人”进行模块划分。“软件人”由虚拟脑、虚拟眼、虚拟耳、虚拟鼻、虚拟嘴、虚拟手和虚拟脚等客体构成[2]。从智能控制的角度来看，“软件人”作为一个智能系统，应具有感知环境、做出决策和控制动作的能力，这里所说的环境是指可以与“软件人”发生直接联系的网络环境。人对外界环境的感知是由大脑和五感（视觉、听觉、嗅觉、味觉和触觉）在相互协作中共同完成的。“软件人”作为具有拟人结构的软件体，如何从软件的角度上实现其拟人化的感知系统目前还是有待深入研究的问题。感知虽然是属于人在生存方面的低层次上的认知活动，但却是非常复杂的过程。不同领域的学者们尝试从不同的角度、不同的观点对其加以研究和解释，目前仍没有定论。不过，从人能够区分事物并解决生活中遇到的分类问题来看，分类应该是人体感知系统的一项重要功能。多分类器系统（Multiple Classifier Systems，MCS）是近年来模式识别领域中的研究热点，用它可以模仿人体感知系统的分类过程，因此本文提出用多分类器系统来研究“软件人”感知系统的分类问题。

3 多分类器系统

3.1 概述

传统的模式识别系统常常只用一个分类器进行识别，因此要求这个分类器必须在所有的样本特征上都有很好的区分能力，这往往难以实现。近年来在研究中发现，不同分类器的误识集合并不一定重合，这表明不同的分类算法之间存在着互补信息，从而可以利用这种互补信息来提高识别性能。在此背景下，产生了多分类器系统的概念，通过组合多个分类器的结果，以提高模式识别系统的性能。多分类器系统的思想产生之后，研究人员逐渐发现不同分类器之间的互补性很强。在这一发现的驱动下，多分类器组合（multiple classifier combination）技术受到了越来越多的关注，并迅速发展成为模式识别领域的一个研究热点。大量的实验和应用证明：将多个分类器的决策结果结合在一起，往往可以得到比单个分类器更好的性能。目前，基于多分类器系统进行模式识别已经在许多应用领域获得了广泛应用[3,4]。

3.2 相关研究

多分类器系统的体系结构主要分为串行和并行两类[4]。采用并行结构的 MCS 如图 1 所示，目前研究和应用得最为广泛，大多数的多分类器组合方法都属于这一类。本文也主要针对并行结构讨论 MCS 的相关问题。

多分类器系统研究中的两个根本问题是分类器集合的设计和多分类器组合方法。前者致力于寻求一个最优的分类器集合，该集合通过简单的组合方法也能获得很好的分类性能；后者是对于任意一个给定的分类器集合，寻求一种组合方法，使得分类器集合能够得到最优的分类性能。两个方面的研究结合起来，作为多分类器系统设计中的不同阶段，往往能获得更理想的结果。

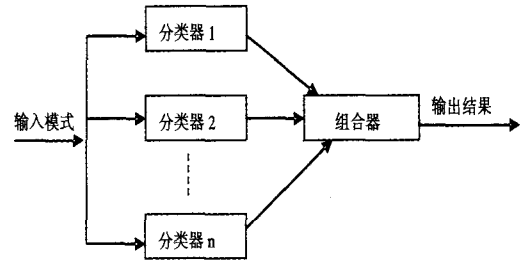

图 1 多分类器并行组合

产生 MCS 中分类器集合的算法，在很多文献中一般称为集成方法（ensemble method）。研究和应用得最为广泛的该类方法如 Bagging、Boosting 和随机子空间方法（RSM）等。

MCS 中多分类器组合方法根据最终输出结果获得方式的不同，分为多分类器融合和多分类器选择两类[5]。所谓多分类器融合（multiple classifier fusion）是指：多个分类器的性能在整个特征空间中被认为近似相同，将这些分类器的输出结果按某种方式结合在一起来达到“共识”，得到最终的分类决策。常用的这类方法如：乘积法、求和法、均值法、最大值法、最小值法和中值法[6]、决策模板法（Decision Templates，DT）[7]和 D-S 理论法等。而实际上，很多情况下各分类器在特征空间中不同区域的性能存在较大的差异，多分类器选择（multiple classifier selection）就是要找出在输入样本周围区域中具有最优局部性能的分类器，并以该分类器的输出作为整个系统的输出结果。这类方法主要有：聚类选择法（Clustering and Selection，CS）[5]和基于局部精度的动态选择法（DCS_LA）等。

4 协同分类模型

4.1 基本思想

“软件人”是活动在网络空间中的纯软件体。在网络环境中，“软件人”面对的感知对象是不同格式的数据，这些数据可以来源于通信消息、网络状态、网络数据包、主机文件、数据库等等。它所要完成的各种任务从根本上说，也主要是对数据的处理。在一定意义上，这些数据对“软件人”都是等同的。可以对哪些数据操作，以及怎么操作完全由“软件人”的内部机制来决定，从而也产生了可以完成不同任务的多种“软件人”角色。“软件人”感知系统对其所感知信息分类的高准确性是“软件人”很好地完成一系列高级认知活动（包括学习、思考、联想记忆及决策等）以及控制和管理功能的前提保证。而且“软件人”的感知系统可以对不同类型的数据正确分类，就可以相应地采取不同的处理过程，进而为一个“软件人”扮演多种角色、完成多种任务提供支持。

人在对感知到的物理世界中的事物进行识别时，事物不同类型的信息分别由不同的感觉器官获得。多种信息送到大脑产生多种感觉。然后由大脑对这些感觉加以整合，做出判断，这是一个多种感觉协同分类的过程。并行结构的多分类器系统可以很好地模仿人体感知系统的分类过程：

（1）人在识别事物时，各种感觉是分别接收不同类型信息的，这本身就是一种并行的工作方式。

（2）人在识别事物时，如果依靠某种感觉就可以做出准确

的判断，那么这一感觉的判断即可作为其最终的判断结果。多分类器选择方法选取最优分类器的结果作为系统最终结果，可以很好地对人的这一过程进行模仿。

(3)人在识别事物时，如果依靠一种感觉不能够做出准确的判断，就需要借助多种感觉共同做出判断。多分类器融合方法通过融合多个分类器的结果来获得系统的最终决策，可以很好地对人的这一过程进行模仿。

网络环境虽然与物理环境不同，但是通常情况下，网络空间中的数据与物理世界的事物相似，其本身特征也存在着天然可分割性[8]，即数据的某些特征能够在某种角度上描绘数据的某种属性。而这些特征不是唯一的，有许多不同的特征能够将同样的属性从不同的角度描绘出来(如全局特征与不同局部的特征)，这样的数据特征集就具有天然可分割性。针对数据的这一特点，可以模拟人体感知系统对事物的分类过程，先对网络环境中数据的不同特征集或用不同方法对相同特征集分别识别，再对多个识别结果加以组合，获得最终决策。这样，不仅可以减少"维数灾(curse of dimensionality)"对高维数据识别准确度的影响，而且能利用多分类器系统相对于单分类器的性能优势，有利于提高"软件人"感知系统对感知对象分类的准确性。

4.2 协同分类模型

基于人体感知系统的启发，运用大系统分析的"分解-集结"方法[9]，本文提出了一种基于多个分类器的"软件人"感知系统的协同分类模型，如图2所示。为了简洁，图中省略了一些功能类似模块中的细节部分。

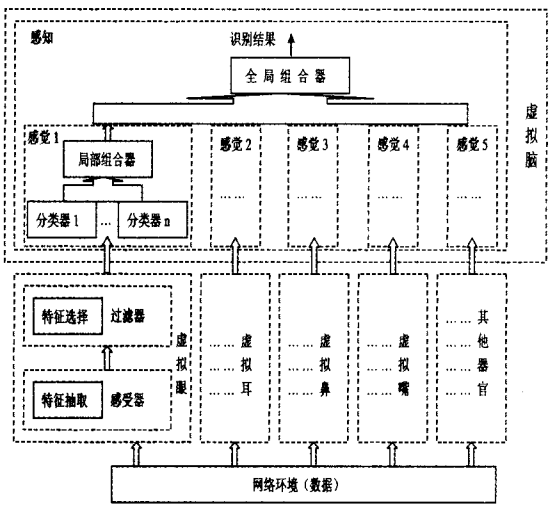

图2 软件人感知系统的协同分类模型

图2中的多种"感觉"只是一种抽象意义上的对人体"五感"的模拟，它们互相之间是等价的，因此我们只以序号加以区别，而没有使用五感的名称(只有当处理图像、声音、气味等多模式信息时，才具有对人体"五感"具体的模拟意义)。需要说明的是，该模型是"软件人"感知系统中用于解决分类问题所涉及到的基本结构，并不是一个完整的"软件人"感知系统。在具体应用中，该模型具有一些灵活性：每种"感觉"中的局部组合器是可选模块，通常在分类器数量和种类较多时使用，可以提高组合的效率。整个系统采用的是一种"集中-分散"相结合的管理模式；五种"感觉"在具体应用中并非总是全部使用，很多情况下仅使用部分感觉，这点也符合人体感知系统的工作特征。另外，作为开放性的结构，模型中根据需要可以增加"感觉"的数量；不同的"感觉"只对高维数据才使用不同的特征集进行识别，特征数较少的或特征不具有明显可分割性的数据仍使用整个特征集合，避免因特征不足或对特征集合的硬性划分而引起的识别率的下降。

5 仿真实验

5.1 实验说明

我们以"软件人"的一种角色——软件卫士(防黑客)为例，使用入侵检测中的数据集在Matlab中进行仿真实验来验证本文所提出模型的可行性。

所使用的样本数据来自KDD'99的入侵检测数据集[10]，这是目前入侵检测领域比较权威的测试数据。该数据集中的元素是把网络上收集的TCP/IP包进行预处理后所获得的连接模式，其中所包含的特征可以分为两大类[11]：网络特征是从TCP/IP包头抽取的，又可进一步分为本质(intrinsic)特征和通信量(traffic)特征。本质特征表示与连接相关的一般信息(如连接类型、服务协议等)，通信量特征是对与当前连接相近的过去连接的统计信息(如在特定时间段内，对同一目的主机或与同一服务相关的连接数等)；内容(content)特征从TCP/IP包的数据部分抽取，是负载中与发现入侵有关的信息(如操作系统的错误报告等)。每个连接模式使用41维特征向量表示，所有连接都属于5个类别中的一个(正常、拒绝服务攻击DoS、远程未授权访问攻击R2L、非授权特权攻击U2R、监视及扫描攻击Probing)。

采用与文[11]中类似的实验方法，从两个具有正确类别标号的数据集中，分别抽取与最常用的ftp服务相关的连接模式作为样本数据。对选取的样本数据进行如下处理：

(1)维数消减。在所有41个特征中选取29个特征(其中本质特征4个、通信量特征19个，内容特征6个)，丢弃12个特征(这些特征与其他服务相关，在ftp服务中其值始终为常量)；(2)符号型特征值量化。(3)将所有特征值使用线性规格化方法处理成[0,1]之间的数。

样本处理后，获得A和B两个数据集：A集合由800个样本组成，其中各类别分布如下：正常(375个)、DoS(104个)、R2L(313个)、U2R(3个)和Probing(5个)；B集合由5212个样本组成，其中正常(3821个)、DoS(1042个)、R2L(313个)、U2R(3个)和Probing(33个)。以A集合作为训练样本集，每一类特征分别训练3个k-最近邻(k-NN)分类器(k=1,3,5)，构成一种"感觉"的分类器集合。也就是说，使用3种"感觉"，分别针对本质特征、通信量特征和内容特征进行识别。这里未使用局部组合器，所有分类结果通过全局组合器加以组合。为了进行对比，使用所有的29个特征也训练了三个对应的k-NN分类器(k=1,3,5)。B集合作为测试样本集，分类时不区分4种具体的攻击类型，只把它们作为一个异常类别，从而简化实验为两类对象的识别问题。在具体代码编写过程中，使用了模式识别工具箱"PRTools4"[12]。

5.2 实验分析

首先对各个分类器的性能进行了测试，结果如表1所示。

从表1可以看到，使用部分特征的分类器在错误率和误警率上都要比使用全部特征的分类器高得多。这是意料之中的结果。只采用部分特征进行识别，犯了跟"瞎子摸象"中同样的错误，必然会导致较多的错误判断。

表 1 分类器的性能(%)

分类器 \ 性能		错误率	误警率
本质特征	1-NN	4.029	0.503
	3-NN	0.998	3.539
	5-NN	0.979	3.343
通信量特征	1-NN	3.626	11.474
	3-NN	6.466	19.073
	5-NN	7.655	21.914
内容特征	1-NN	21.546	4.730
	3-NN	1.938	5.084
	5-NN	2.015	5.347
全部特征	1.353	1-NN	0.480
	3-NN	0.883	2.406
	5-NN	0.748	1.506

然后使用比较常用的组合方法将多个部分特征分类器的识别结果加以组合,来获得系统的最终决策,测试结果如表 2 所示。

表 2 部分特征分类器的组合性能(%)

组合方法 \ 性能	错误率	误警率
乘积法	0.806	2.865
均值法	0.499	1.630
最大值法	0.710	2.525
最小值法	0.710	2.525
中值法	0.518	1.700
投票法	0.518	1.700
DT 法	0.480	1.560
CS(3)	0.153	0.501
CS(5)	0.365	1.278
CS(8)	0.365	1.278

表 2 中,聚类选择法(CS)中采用了 k-means 方法进行聚类分析,括号中的数字为聚类数。由于 k-means 方法对初始聚类中心的选择是随机的,因此 CS 方法每次的测试结果会有所不同。通过多次测试发现,其结果仅是几个数值的重复出现,而且这些数值的差别并不是很大。我们给出的 CS 方法的性能是测试中出现次数最多的一组结果。可以看到,将原本性能很差的部分特征分类器的结果组合以后,系统的分类性能得到了很大的提高,大多数情况下的错误率和误警率已经低于所有特征分类器。实验结果说明本文提出的基于多种拟人感觉的协同分类模型是可行的。

表 3 全部特征分类器的组合性能(%)

组合方法 \ 性能	错误率	误警率
均值法	0.365	0.718
均值法	0.365	0.718
最大值法	0.345	0.789
最小值法	0.345	0.789
中值法	0.422	0.720
投票法	0.422	0.720
DT 法	0.345	0.789
CS(3)	0.480	1.353
CS(5)	0.480	1.353
CS(8)	0.480	1.353

对于特征数较少或特征不具有明显可分割性的数据,本文模型中的不同"感觉"仍然使用整个特征集合进行分类。假设实验数据中的特征是不可分割的,将 3 个所有特征分类器分别作为一种"感觉",同样用表 2 中的组合方法进行实验测试,结果如表 3 所示。

可以看到,组合后的性能比 3 个分类器有所提高,而其中 CS 方法的性能在聚类数不同的时候没有变化。这说明在 3 个分类器中,1-NN 分类器不仅在整个特征空间的性能优于其他两个分类器,在不同局部区域上也是如此,导致总是选择其分类结果作为整个系统的决策。表 3 的实验结果说明本文模型在面对特征不可分割的数据时也是可行的。

结论 本文提出了一种基于多个分类器的"软件人"感知系统的协同分类模型。该模型用多分类器系统模拟人的不同感觉,先对网络环境中数据的不同特征集或用不同分类器对相同特征集分别识别,再对多个识别结果加以组合而获得最终决策。仿真实验验证了本文模型的可行性,而如何构造模型中的多分类器系统,主要是分类器集合和多分类器组合方法的优化设计将是我们下一步的研究工作。

以国家"十五"重大科技攻关项目"数字气田关键技术及应用示范研究"作为基础,在"十一五"期间,我们将继续开展数字气田的信息化建设项目,这为本文的研究工作提供了很好的应用场景。在数字气田建设中,面临的一个艰巨任务是如何对各种资料和数据进行准确且高效的处理,本文提出的分类机制除了为"软件人"担任网络中的不同角色,完成相应的任务提供保障,还使得"软件人"可以与项目中需要研究数据分类问题的应用领域(如数据采集、数据处理、辅助决策、异构数据库集成等)嫁接。进而如何针对气田实际数据实现模型中感受器和过滤器的主要功能(特征提取和选择),将是我们下一步需要开展的另一项研究工作。

参 考 文 献

1 曾广平,涂序彦."软件人".见:中国人工智能进展 2003(中国人工智能学会第十届全国人工智能学术年会论文集).北京:北京邮电大学出版社,2003.677~682

2 曾广平,涂序彦."软件人"的概念模型与构造特征.计算机科学,2005,32(5):135~136,143

3 Rahman A F R, Fairhurst M C. Multiple classifier decision combination strategies for character recognition: A review. International Journal on Document Analysis and Recognition (IJDAR), 2003, 5(4): 166~194

4 Jain A K, Duin R P W, Mao J. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4~37

5 Kuncheva L I. Switching between selection and fusion in combining classifiers: an experiment. IEEE Transactions on Systems, Man and Cybernetics-Part B, 2002, 32 (2): 146~156

6 Kuncheva L I. A theoretical study on six classifier fusion strategies. IAEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 281~286

7 Kuncheva L I, Bezdek J C, Duin R P W. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition, 2001, 34(2): 299~314

8 刘世岳,李珩,张俐,等. Co-training 机器学习方法在中文组块识别中的应用.中文信息学报,2005,19(3):73~79

9 涂序彦,王枞,郭燕慧.大系统控制论.北京:北京邮电大学出版社,2005

10 University of California. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999

11 Giacinto G, Roli F, Didaci L. Fusion of multiple classifiers for intrusion detection in computer networks[J]. Pattern Recognition Letters, 2003, 24(12): 1795~1803

12 Duin R P W, Juszczak P, Paclik P, et al. PRTools4, A Matlab Toolbox for Pattern Recognition. Delft University of Technology, 2004