东北大学

硕士学位论文

基于强化学习算法的电梯调度系统的研究

姓名：马黎华

申请学位级别：硕士

专业：控制理论与控制工程

指导教师：刘建昌

20070201

基于强化学习算法的电梯调度系统的研究

摘 要

随着高层建筑的不断增多和智能建筑的快速发展，人们对建筑物内的客流运输设备电梯提出了越来越高的要求。为了提高电梯的运输能力和服务质量，实现多台电梯的有效控制——电梯调度，正越来越成为人们研究的热点问题。

本文主要研究了电梯群组调度问题。电梯交通流随时间变化呈现出规律各异的多种交通模式，所以本文在处理交通流分析问题的过程中，将聚类思想引入电梯交通流分析，并提出了一种新型的人工免疫聚类算法。对电梯交通流进行交通模式的聚类分析和模式识别，突破了传统的四种交通流模式：空闲、随机层间、上高峰和下高峰模式的局限，将交通流细分为 7 种模式，便于提高后续进行电梯动态调度研究的准确度，同时计算出各模式对应的浓度和聚类中心，方便对交通流的在线模式识别。

在解决电梯群组调度这种大规模动态优化问题时，本文采用强化学习方法作为在与环境的交互过程中学习最优策略的方法。以马尔可夫决策过程（MDP）为背景，模型化电梯群组调度问题，综合考虑平均等待时间、平均乘梯时间和电梯停靠次数这三个因素，计算出的综合成本作为性能评价的综合指标。采用随机行为选择策略和前馈神经网络分别解决强化学习在应用中所存在的探索问题和值函数的泛化问题。并将两者与 Q-learning 的值迭代算法结合，共同构成基于强化学习的电梯群组调度算法。最后，构建了基于泊松分布的大楼客流交通模型，并设计了基于强化学习的电梯群组调度系统，利用 MATLAB 仿真，结果证明，在对交通流进行反复训练时，Q 值曲线间的误差是逐渐减小的，说明了本文使用的强化学习方法对 Q 值函数的逐渐逼近作用。在 5 种不同的交通流条件下，基于强化学习的动态调度方法与其它方法比较呈现出一定的优越性，尤其在上高峰模式和随机层间模式下有更好的适用性，充分说明了基于强化学习的电梯群组动态调度算法的有效性和可行性。

关键词：电梯交通流；人工免疫聚类分析；电梯群组调度；马尔可夫决策过程；强化学习

Research of Elevator Group Scheduling System based on Reinforcement Learning Algorithm

Abstract

With the continuous increasing of high buildings and the rapid development of intelligent buildings, more and more demands have been brought forward to the elevators that transport the guests in the buildings. In order to improve the transportation efficiency and the service quality, and nowadays people pay more and more attention on how to control the elevators validly-elevator group control so it has become a hot issue in research recently.

This thesis is mainly about the research in elevator group scheduling problem. With the clustering analysis on elevator traffic flow, Elevator group scheduling system based on reinforcement learning algorithm is proposed. During the process of the analysis on elevator traffic flow, based on diversity of elevator traffic with time, Clustering idea is introduced to solve analysis of elevator traffic flow and Artificial Immune Clustering Algorithm (AICA) is proposed to recognize and cluster elevator traffic flow. The elevator traffic is divided into seven traffic pattern particularly, which improves the accuracy of scheduling and also avoids the limits of four traditional traffic patterns: off-peak, inter-floor, up-peak and down-peak patterns. The proportion and center of each traffic pattern are calculated to recognize each traffic pattern with online.

Reinforcement learning, as a method learning optimal policy from interaction with the environment, is suitable for large-scale dynamic optimization problems such as elevator group scheduling. The problem of elevator group scheduling is formulated by the framework of Markov Decision Process (MDP) and then the elements are defined according to the specified field. These three factors, the average waiting time, the average riding time and the times of starting-stopping for elevators, are synthetically considered and the integrated cost is defined as basic of evaluating performance of system. Reinforcement learning is applied, and the stochastic action-selected policy and feed forward neural network are used to handle the problems of exploration and generalization of value function respectively, which are integrated into the value iteration algorithm called Q-learning to construct the whole algorithm for elevator group scheduling. Finally, a simulator model of transportation based on Poisson distributing is established, also the elevator dynamic scheduling system based on reinforcement learning algorithm dispatching algorithm is simulated in the text with

MATLAB. The simulation result of this elevator group control system shows gradual decrease on the errors between the Q function graphs and Q function is batter approximated with reinforcement learning method. With five different elevator traffic flow used for simulating and training of algorithm, the experimental results demonstrate the good learning ability, good performance and the adaptability for different traffic flows of scheduling algorithm in contrast to other ones, especially in up-peak traffic pattern and inter floor traffic pattern. It shows great superior and feasibility for elevator dynamic scheduling system based on reinforcement learning algorithm.

Keywords: Elevator Traffic Flow; Artificial Immune Clustering Algorithm; Elevator Group Scheduling; Markov Decision Process; Reinforcement Learning

独创性声明

本人声明所呈交的学位论文是在导师的指导下完成的。论文中取得的研究成果除加以标注和致谢的地方外，不包含其他人已经发表或撰写过的研究成果，也不包括本人为获得其他学位而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示诚挚的谢意。

学位论文作者签名：

签字日期：2007.3.1

学位论文版权使用授权书

本学位论文作者和指导教师完全了解东北大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人同意东北大学可以将学位论文的全部或部分内容编入有关数据库进行检索、交流。

（如作者和导师同意网上交流，请在下方签名；否则视为不同意）

学位论文作者签名：　　导师签名：

签字日期：2007.3.1　　签字日期：2007.3.1

第一章 绪 论

1.1 引言

随着人口的城市化和社会经济的发展，越来越多的高楼大厦拔地而起。高层建筑特别是智能建筑已从几万平方米，增加到十几万甚至几十万平方米，楼层也由原先的几层、几十层发展到几百层，电梯的交通问题变得越来越复杂，对高效率的垂直交通工具的需求急剧增长。大厦中大量客流、物流的垂直输送将是一个突出的问题，单台电梯显然并不能应付全部的客流，因此，多台电梯并排设置就成为必然。同时，随着智能建筑的发展和社会的进步，人们对电梯也提出了越来越高的要求。对电梯要求已从简单的响应楼层、轿厢召唤，发展为能耗尽量少、候梯时间尽量短、乘梯时间尽量短、长时间候梯率低、轿厢舒适度等多个要求。如何对多台电梯进行有效的调度管理，提高电梯服务效率和质量是人们不得不面对的一个重要课题。这样就提出电梯群控系统（Elevator Group Control System，简称 EGCS）这个概念，为了缩短人们的候梯时间，减少能量损耗，需要合理安装多台电梯并进行集中统一的控制的优化调度系统。

如今，对电梯群控系统的研究层出不穷。一个完整的现代电梯系统不仅包括组成电梯系统所需的轿厢、厅门、拽引电动机、操作面板、指示装置等，还有实现电梯高效服务的控制策略以及执行这些控制策略的微处理器。虽然电力电子技术、数字电子电路和计算机技术的发展都促进了电梯系统的发展，但是电梯给人带来的安全感、舒适感和满意度才是电梯系统进一步发展的内在动力。

乘客的舒适感和满意度始终是电梯及电梯群控研究的一个热点，这主要体现在电梯的运行曲线控制、电梯的减振研究以及电梯群的调度研究等，因为乘客的舒适感和满意度不仅涉及乘客的生理需求还跟乘客的心理需求相关，乘客在乘坐电梯时的生理需求主要是指电梯在垂直平面内运行时，产生不舒服的感觉，即所谓的“重力加速度效应”。而乘客的心理需求是指乘客对电梯服务水准的认可程度。而这两个需求存在着认识上和量化的困难，最大限度地为乘客提供舒适感和满意度构成了控制领域的专家和学者们强有力的挑战。

电子科学和信息技术的快速发展，以及国民经济发展对高层建筑的呼唤，已成为推动电梯技术向前发展的不竭动力。20 世纪 80 年代以来，人工智能理论得到了进一步的发展，为电梯群的协调控制与优化调度奠定了坚实的理论基础[1]。探索人工智能研究的

新成果、新方法，以及将这些成果和方法应用于电梯系统，推动电梯技术不断向前发展，是控制的当务之急。正是在上述因素的启发下，我们才有了本论文的研究课题——电梯群组动态调度的研究。

1.2 电梯群控系统的发展

电梯起源于公元前236年的古希腊，古希腊阿基米德设计的一种人力驱动卷扬机是电梯的最初雏形。事实上，早在公元前我们的祖先也使用了这种称为“轱辘”的人力升降机。1889年，奥的斯公司推出了第一台电力驱动，齿轮直接传动的电梯，也是第一台真正意义上的电梯[2]。

纵观电梯控制系统的发展史，从最初的继电器并行控制方式发展到今天的计算机人工智能群控调度阶段。其发展情况如表1.1所示。

表 1.1 电梯群控的发展

Table 1.1 The development of the elevator group control system

阶段	电梯群控系统	分派系统	控制系统
第一阶段(1941-1941)	自动模式选择系统	区间分配系统	时间间隔控制
第二阶段(1971-1975)	集成电路	区间分配系统	等待时间预测控制
第三阶段(1975至今)	计算机控制（最短候梯时间控制、综合评价系统、应用人工智能）	呼梯分配系统	带有学习功能的控制

1941~1971年，电梯群控系统的最初阶段使用的是继电器顺序控制，也称自动模式选择系统。它根据特定时间段内的交通模式选择与之对应的运行方式。交通模式分为上行高峰期、下行高峰期、非高峰期等。控制方式采用时间间隔控制。时间间隔控制是指：为使梯群中的轿厢沿井道高度均匀分配，特别是在繁忙的交通需要期间，电梯群控系统在适当的时间间隔从层站楼层发出轿厢，就像公共汽车一样运行，群控系统从响应需求分配轿厢的意义上说，不依赖层站呼梯信号而工作，而是按照程序从层站楼层分派轿厢，这种控制方式的缺点是：轿厢在层站需要花费相当多的时间等待分配间隔周期，停在顶层的层站常常是无用的，而且轿厢在等待分配是闲置着。另外，轿厢也频繁地在层站楼层间进行无目的地进行，且在交通量轻时过多地运行。轿厢分派方式为区间分派方式。这种系统的缺点是硬件复杂，可靠性低，维修困难，效率低，不能进行较复杂的逻辑推

理。

1971~1975 年，电梯群控系统的第二阶段，硬件上采用集成电路使系统结构简化，大大提高了系统的可靠性，并且更主要的是能从事和进行较为复杂的逻辑运算。这一时期使用呼梯派梯方式，即针对每个具体的呼梯信号派梯，当一个层站呼梯信号登记后，系统就会根据交通情况和各轿厢的状态，选择一部轿厢为之服务。这一派梯方式成为以后各代群控系统的派梯方式。控制方式是候梯时间控制，即根据各轿厢对呼梯信号的响应时间来决定分派哪一部轿厢为此呼梯信号服务。但是，此系统对于预测候梯时间所必需的较复杂的数值计算还不完善。

1975 年至今，电梯的现代群控阶段，其标志是 1975 年计算机开始应用于电梯群控系统。在此之前的两个阶段主要应用数理统计的方法来研究电梯群控中客流量的特性，称之为统计特性。自从计算机应用于群控系统后，开始使用模糊控制、专家系统和人工神经网络等人工智能技术来描述电梯群控的特性，从而提高电梯群控系统的整体服务性能，完成电梯交通整体配置。计算机应用于电梯群控系统中，完全发挥了计算机所具有的进行复杂的数值计算，逻辑计算及数据记录的能力，极大满足了电梯群控系统中复杂的数值计算和逻辑推理的需求。计算机控制能直接完成控制算法参数的在线修改，通过将新程序输入计算机，并且不需要重新布线，就能很快实现控制算法的完全改变。一套有效的仿真程序能在安装好的系统上离线计算，有利于选择合理的参数。仿真技术也是可以离线评价一种新的算法，这大大增加了改变控制算法的方便。计算机控制的另一个优点是其数据记录功能。在计算机中能记录和分析交通状况和目的地数据，记录和分析轿厢的运行和电梯系统的整体性能，关门时间，故障部分检测记录数据的保存等。可以实现这些数据的远距离查询，并实时监测任何故障的发生。根据这些数据可以改进控制算法参数，实时适应建筑物的需求。计算机的应用也为人工智能技术等高技术在电梯群控系统中的应用提供了硬件基础。

1.3 电梯群控调度算法的研究现状

1949 年，联合国大厦首次使用继电器逻辑组成的电梯群控系统，同时国际上各大电梯公司相继推出与该公司群控系统相适应的控制算法。在我国，电梯群控的研究最早见于 1990 年，主要研究电梯群控系统和人工智能。我国电梯群控技术与国外相比，技术明显落后，最重要的是调度算法方面的差距很大[3]。

电梯的群控调度主要分为电梯群的交通流模式识别技术和电梯群呼叫分配调度算

法两个部分。如今，随着人工智能的不断发展，国内外主要存在以下一些交通模式识别技术与电梯群控调度算法：

1.3.1 电梯群控的交通模式识别技术

（1）基于神经网络的交通流模式识别[4]

人工神经网络是依据人类和动物大脑的工作方式建模的，通常由输入层、隐层、输出层3层节点组成，每层中有一定数量的神经元。每个神经元实际上是多输入、单输出的信息处理单元，它接受与其相连神经元的输出作为其输入信号，执行一个预先定义的数学操作，然后产生单个输出值。不同的神经元之间连接权的大小不同，对于连接权值的调整过程就是神经网络的学习过程。

神经网络学习的主要优点在于它可以通过调整网络连接权来得到近似最优的输入-输出映射，使用的神经网络结构就会相当庞大而且难以确定，训练样本要求多且准确。网络的离线学习或在线学习的时间都会较长，不能提供一个明确的用于网络知识表达的框架[3]。

（2）基于模糊神经网络的交通流模式识别[6]

鉴于模糊逻辑和神经网络的优缺点，把两种方法结合起来，取其长而避其短，形成模糊神经网络(FNN)。它具有一个表达框架，一方面提供用于解释和推理可理解的模型结构，通过它可以以一种清晰的方式描述知识，另一方面它具有知识获取和学习能力。

采用两个模糊神经网络分两步进行识别，第一个网络根据单位时间段的总客流量、进门厅客流量和出门厅客流量辨识出上高峰、下高峰、空闲和层间交通模式的比例；当层间模式占较大比例时，第二个网络根据单位时间段最大楼层客流量和次大楼层客流量辨识出两路、4路和平衡的层间交通模式的比例。

（3）基于统计规律的交通流模式识别[6]

虽然电梯群控系统具有随机性，但对于任何一幢大厦，都有一定的工作周期，在不同周期的同一时刻客流交通状况会存在相似的状况，因此可以依据对电梯所在建筑物长时间的客流交通情况的统计得出其客流变化的规律，从而形成一定的交通模式识别规则，即基于统计规律的交通模式识别方法。

基于统计规律的交通流模式识别的规则制定过程即交通模式识别方法的学习过程。通过对建筑物长时间地观察、记录，可以得出一定周期内客流交通状况的统计规律。在此基础上，加上专家的指导意见就可以得出交通模式的识别规则，采用基于统计规律的

交通模式识别方法具有操作简单，可靠性高，同时适合编程实现等优点。采用根据专家知识制定的样本，用 2 步混合学习算法对两个网络进行训练，测试表明能准确地辨识出各种交通流模式所占的比例。

1.3.2 电梯群控的调度算法

（1）基于专家系统的电梯群控调度算法

电梯群控系统是一个具有大量不确定和不完整信息的复杂的非线性系统。这样一个复杂的系统工程问题很难完全用数学方法精确地描述，而要靠经验的、尚未形成科学体系的知识。专家系统就是研究用于处理复杂系统工程技术问题的知识表示、使用和获取的方法。它具有启发性，能利用专家的知识和经验对不确定或不精确的问题进行启发式的推理，对解决电梯群控系统具有一定的优越性。

文献[7]采用模糊理论与专家系统结合的方法，用专家经验知识建立模糊规则，提出以减小等待时间和乘梯时间的模糊专家系统调度方法，有效地利用这些知识优化调度，显然专家知识的日趋完善及知识的可靠性决定了电梯调度方法的性能。同时，为了满足服务要求多样性，控制规则不可避免增加很多，却还不一定能较为全面地反映问题。而且利用规则比较各种可能的电梯动作，从而选出最佳路线，使乘客运输能力增大而等待时间减小，它适用于层间模式，但不适用于上高峰，因为层间交通下计划最佳路线的范围大，但它没有预测轿厢的加减速时间。

（2）采用模糊控制的电梯群控调度技术

模糊逻辑是一种精确解决不精确不完全信息的方法，其最大的特点是可以比较自然地处理人类的概念。由于模糊控制具有鲁棒性强、不需要建立精确模型等特点，较适合于电梯群控这种离散随机性较强的系统，因此国内外专家在这方面做了大量的工作。文献[8]从呼梯分布的空间位置角度出发，制定模糊规则来确定呼梯分配方案。在电梯群控系统中决定电梯负载偏差的区域权是一个与电梯性能紧密相关的控制参数，文献[9]提出了一个基于模糊模型的调度方法来决定区域权，通过对区域权特征的研究结合专家知识建立了一个两级模糊推理模型。区域权由两组模糊规则组成。第一组以上行交通量和下行交通量为输入输出区域权α'，第二组以平均等待时间、能耗和长候梯率为输入，其输出作为第一组输出α'的修正，两组输出相加决定最后的区域权。这些文献所提出的调度方法虽然控制目标和方法不同，但都是应用模糊理论，建立模糊规则并根据模糊规则的计算结果进行调度，与传统的调度方法相比这些调度方法的系统在性能上有所改进。但

模糊控制的许多规则难以确定，依据专家知识，其自身的特点使得它难以实现控制目标要求下的最优调度。由于模糊控制本身不具有学习能力，使控制系统有以下缺陷：无法跟上交通的变化；其性能极大地依赖于专家知识的好坏；规则库的修改非常困难。由于模糊控制自身的特点使得采用模糊控制的调度方法难以实现控制目标要求下的最优调度，这影响了它在电梯群控系统中的广泛应用。

（3）基于神经网络的群控调度方法[10,11,12]

专家系统与模糊逻辑等人工智能技术在电梯群控系统中的应用极大地提高了梯群的服务质量和效率，但它们的致命弱点是不能通过学习来改善控制算法。因此当客流发生变化时，系统就不能完全适应这种变化，由此引出的问题有：

如果建筑物的实际情况与专家的假设不同，那么基于此假设的规则就不能带来较好的效果；系统的性能由专家的知识、经验决定，具有一定的局限性；调整模糊度隶属函数困难，要做很大仿真；一旦规则集成到系统中去，要想改变就要花费大量时间和精力，且只能由人工来完成。

为使电梯群控系统适应条件的变化，做出实时的自动调整，在各种情况下提供更佳的服务，使系统具有自学习能力，人工神经网络被引入到电梯群控系统中。神经网络技术引入电梯群控系统中的优点有：当专家设想的建筑物条件与实际建筑物不同时，带有神经网络的电梯群控系统具有学习能力。利用非线性和学习方法建立适合的模型，进行高速推理，对电梯交通可以进行短、长期预测。

（4）模糊神经网络的调度方法[13]

模糊神经网络把神经网络与模糊逻辑相结合，吸取各自的优点，它克服了人工神经网络结构难以确定，以及模糊逻辑无自学习功能的缺点。使得模糊神经网络一方面具有知识获取和学习的能力，网络学习的过程就是模糊逻辑规则的优化过程。通过学习调整网络权值，每一个模糊规则中各个前提、模糊集合的相对重要度以及每个规则的相对重要度，来实现规则的优化。另一方面，它提供用于解释和推理可理解的模型结构，可以以一种清晰的方式描述知识。

模糊神经网络的建立和使用分以下步骤：利用专家知识粗略地形成模糊模型以及一些模糊规则和模糊推理方法；基于这一模糊模型构成神经网络；训练神经网络。采集一些特定的交通状况作为样本，采用相应的算法进行学习，调整神经网络必要的权值，以获得优化的模糊规则，然后进行网络应用。

（5）基于遗传算法的群控调度方法

遗传算法(Genetic Algorithm，简称 GA)是以不确定型、非线性、时间不可逆为内涵，以复杂问题为对象的科学新范式。由于遗传算法能有效地求解组合优化问题及非线性多模型、多目标的函数优化问题，故求解复杂问题最优解不现实，寻求满意解，GA 是最佳工具之一。文献[14,15]提出根据客户要求确定控制目标并用遗传算法根据预测到的呼梯产生及分布等数据优化调度评价函数的参数。提出的调度方法虽采用遗传算法优化评价函数的参数，但其优化效果在很大程度上依赖于预测的外呼梯产生时间和分布的精度。由于遗传算法生物基础鲜明，数学基础不够完善，目前还存在搜索效率及其时间性问题，而电梯调度是一种实时性方法，因此目前还存在有实时性问题。

1.3.3 电梯群控研究的意义

综上所述，对电梯群控系统这个复杂的非线性动态离散系统，人们已经提出了各种控制和调度方案，解决了一些实际问题，但同时又发现了新的问题，需要面对新的困难和挑战，从以上分析可知，电梯群控调度方法的发展前景应该研究一种理论上能够保证系统控制与决策的完备性和充分性，同时又是比较容易实现的控制方法。为此，电梯群控的方法以其理论研究及其向着工程实际转化的极大可能性成为我们研究的重点。

从现实社会对电梯系统的控制要求来看，尽管电梯群控算法日趋复杂，但是随着微电子技术和计算机技术的发展，其在电梯群控系统的实际应用将日益成为可能或者更加高效。所以对电梯群控调度算法进行的研究有助于衍生对新技术的需求。同时，现代电梯的群控调度算法的主要理论基础是人工智能，这为人工智能的控制策略提供了一个良好的实验和检测对象，通过对电梯群控这个特殊的对象可以进一步检验人工智能策略的有效性，在实践中推动着人工智能方法的进一步发展。

就目前的研究状况还存在如下的几个问题需要解决：

（1）电梯交通流是电梯状态的一个概念，是影响电梯群控算法性能的一个重要因素，对其研究非常必要。

（2）实际上，仅用一种固定不变的电梯调度方法显然不能适应一天中建筑物内所有交通模式，不同建筑物用途不同，交通特征也不同，而同一建筑物内的交通模式在一天的不同时间也是变化的，可能发生的交通模式有上高峰、下高峰、随机层间、空闲期等。由于人们的作息时间的规律性，在一栋大楼中一天内不同时间的交通流的强度、分布、流向都有所不同。在不同的交通流状况下，应该以不同的方式去为乘客分配电梯，才能够产生较好的控制效果。为了充分满足用户要求提高服务质量，电梯群控系统必须

辨识出交通模式的变化，在高峰、层间、空闲时采用不同的控制策略。

1.4 本文的工作内容

（1）对电梯交通流进行分析及仿真，建立复杂的交通流分析模型，并利用 MATLAB 进行仿真，得到交通流数据；

（2）在电梯交通流仿真的基础上，采用新兴的人工免疫聚类算法对电梯交通流进行交通模式的聚类分析和在线识别，突破了传统的四种交通流模式的局限，将交通流细分为 7 种模式，便于理解和改善后续电梯群控调度策略使用的准确度，同时计算出各模式对应的浓度和聚类中心，完成了电梯交通流的定性分析，体现了人工免疫聚类方法的优势和先进性。

（3）阐述强化学习基本概念、思想和原理，追踪国外强化学习研究的历史进程。通过分析大量关于强化学习的理论与应用的文献，研究强化学习的分类方法学，系统地划分强化学习方法的种类，得到强化学习问题及算法的全貌。研究得出强化学习解决实际问题的一般过程以及过程中需要注意的关键问题。

（4）以随机序贯决策的观点分析电梯群组调度问题，将其抽象为无限时间折扣型的马尔可夫决策过程，建立相应的强化学习模型，研究使用强化学习求解电梯群组调度问题的方法；提出了综合考虑平均候梯时间、平均乘梯时间和电梯停靠次数这三个因素的综合成本，并作为调度和评价系统性能的依据；使用 Gibbs 分布构建随机行为选择策略和利用前馈神经网络逼近值函数来分别解决探索问题和泛化问题，以及与值迭代算法结合构成电梯群组调度算法；在交通流模式识别的基础上，设计了基于强化学习算法的电梯动态调度系统。

（5）在 MATLAB 开发环境下，在对电梯交通流人工免疫聚类分析的基础之上，设计了基于强化学习算法的电梯调度系统的仿真环境，并验证了该算法的可行性和有效性。

第二章 基于人工免疫的电梯交通流分析

电梯交通流分析一直是电梯群组控制与交通疏导的一个重要前提。典型的办公大楼由于工作人员的作息时间和流动特性使得交通流的变化是有规律的，一般存在这么几种交通模式：上行高峰交通模式、下行高峰交通模式、两路交通模式、四路交通模式、随机层间交通模式、空闲交通模式。电梯群控系统中的群控器是优化调度电梯的关键，它对交通模式进行识别，在不同交通模式下采用最合适的电梯群控调度方法，对提高整个电梯系统的服务质量和系统的运行效率有着重要意义。一个对当前的交通流模式毫无知觉，不管客流的强度大小、客流密度的分布、客流流向的趋势，完全按同一种调度方法进行调度的电梯群控系统，其性能不能在各种交通流模式下满足系统控制指标的要求。

在以往的电梯群控算法中，都是根据专家的先验知识，通过模糊推理来判断电梯交通流的当前模式，然后根据判断所得的客流模式，调用针对不同客流情况的电梯群控算法，以实现电梯群的最优质服务。同时，我们也注意到，这种电梯交通流模式的确定缺乏科学依据和理论支持。聚类分析是研究分类的一种多元统计方法，通过对电梯交通流数据的聚类分析，可以达到认识和区分交通流模式的目的。然而，电梯交通流数据量较大，这对任何一种聚类分析方法都是一种挑战。同时，我们注意到这些交通流数据存在着较大的冗余，如能消除这些冗余数据，再进行聚类分析，将大大减少聚类分析的运算量。基于以上难点，本文利用人工免疫的高度并行、分布、自适应和自组织的特点，提出了一种以人工免疫聚类算法进行电梯交通流模式识别和分类的方法，对交通流进行定性分析，将电梯交通流细分为 7 种交通流模式，并确定各个模式的聚类中心，为在线模式识别奠定了基础。

人工免疫系统是新兴的人工智能分支，吸引了国内外众多研究者的研究兴趣。众人拾柴火焰高，经过不懈的努力，人工免疫系统已经广泛地应用于模式识别[16]、聚类分析[17]、计算机安全[18]、故障诊断[19]、控制工程[20]、函数优化[21]等信息处理领域，向人们展示了其巨大的发展潜力。

本章首先介绍电梯群控系统的客流发生模型，为交通流分析和群控算法的研究提供客流数据。再利用人工免疫系统理论对电梯交通流进行聚类研究和分析，为电梯交通流模式的定义和识别提供强有力的理论依据。

2.1 电梯群控系统模拟交通流的产生

2.1.1 电梯交通流分析

描述电梯交通流，主要依据两点：交通流流量和交通流流向。交通流流量是指单位时间内要求服务的乘客到达数，交通流流向是指要求服务乘客的流动方向。只有按照某种方法，从实际交通流数据中找出描述电梯交通流这两种特征的统计规律，才能从整体上描述电梯交通流。

为了建立电梯交通流的概率仿真模型，依据大楼内交通流的流向，将电梯交通流分为三种模式，如图 2.1 所示。

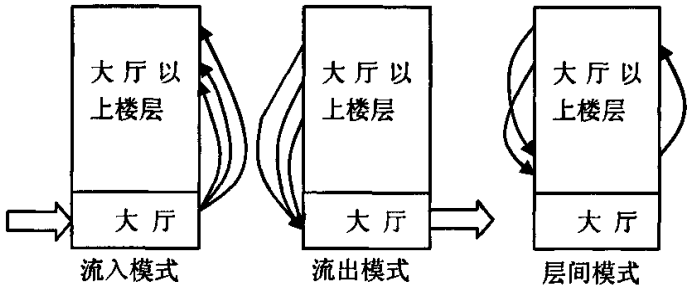

图 2.1 三种基本电梯交通模式

Fig 2. 1Three basic elevator traffic pattern

（1）流入交通模式：大楼内乘客全部从门厅运行到大楼内各层。

（2）流出交通模式：大楼内乘客从大楼内各层向下运行到门厅。

（3）层间交通模式：大楼内乘客的起始层和目的层都不是门厅，乘客在非门厅的任意楼层之间运行。

在实际中，电梯交通流处于单一交通模式的情况极少，大部分情况下，是这三种基本交通模式以不同比例的复合。

2.1.2 模拟交通流的产生

产生模拟交通流，需要确定所有要求服务乘客的到达时间、起始楼层和目的楼层。为此，需要以下三个步骤来产生模拟交通流。

（1）首先确定每位乘客的到达时间。乘客的到达规律符合完全随机分布或泊松分布[22,23]，为此，需采用相应的方法确定乘客的到达时间。

乘客到达符合完全随机分布规律。取交通流持续时间中任意时刻，作为乘客的到达时间。

乘客到达符合泊松分布规律，即乘客到达时间间隔符合式(2.1)

$$P\{\tau_{i+1}-\tau_i<t\}=\begin{cases}1-e^{-\lambda t}, t\geq 0\\ 0, t<0\end{cases} \tag{2.1}$$

式中 λ——乘客到达率(人/s)；

t——乘客到达时间间隔最大值(s)。

所以，采用下面方法来确定乘客的到达时间：首先产生0~1之间的伪随机数p，由式(2.1)的反向推导式(2.2)计算出乘客到达时间间隔t，并通过与前面乘客的到达时间累加，以确定该乘客的到达时间。

$$t=-\frac{\ln(1-p)}{\lambda} \tag{2.2}$$

（2）根据大厦内人们的工作时间规律，确定属于三种基本交通流模式的要求服务乘客数。

（3）根据三种基本交通流模式的特点和这三种基本交通流模式下要求服务乘客数，确定每名乘客的起始楼层和目的楼层。具体地说：

流入交通流模式是指乘客从门厅来到大楼内各层，这样，流入交通流模式下乘客的起始楼层是门厅。根据门厅处要求服务乘客上行百分比，可计算上行乘客数。假定乘客去大楼各层的概率相同，可按照平均分布确定每名乘客的目的层。

流出交通流模式是指乘客从大楼内各层来到门厅，这样，流出交通流模式下乘客的目的楼层是门厅。根据除门厅以外各楼层要求服务乘客数及下行百分比，确定各楼层下行乘客数，从而确定各楼层产生目的层可能是门厅的厅层召唤的概率分布。据此概率分布，可确定流出交通流模式下乘客的起始楼层。

层间交通模式是指乘客在大楼内各层随机运动。这样，在确定了流入交通流模式和流出交通流模式下乘客的起始楼层和目的楼层后，计算剩下的未确定的各楼层要求服务乘客数和上行下行百分比，并据此计算各楼层上行下行乘客数。乘客的起始楼层是有厅层召唤的楼层，乘客的目的楼层是其余各楼层中任意的一层。

经过上述步骤，可以确定交通流持续时间内要求服务乘客数、所有乘客的到达时间、起始楼层和目的楼层，即可产生出与实际交通流情况相符的模拟交通流。

2.2 聚类分析

2.2.1 聚类的概念

一个复杂的系统可视为多个子系统的某种组合。那么如何将复杂系统划分为多个子系统，换言之，原全局系统用多少个子模式来表述并确定各自所在的有效区间就成为首先要解决的问题。因此，人们提出通过聚类分析进行系统的多模式划分。

聚类就是将数据对象分组成为多个类或簇，在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。在这一过程中没有教师指导，因此是一种无监督的分类。聚类分析主要根据事物的特征对其进行聚类或分类，用数学方法研究和处理所给定对象的分类，主要用于辨认具有相似性的事物，并根据彼此不同的特性加以“聚类”，使同一类的事物具有高度的相似性。说得简单一点，就是把事物按其相似程度进行分析，并寻找不同类别事物特征的分析工具。聚类分析可以对变量进行聚类，也可以对样本进行聚类。当要分析的数据缺乏描述信息，或者是无法组织成任何分类模式时，利用聚类分析可以自动将数据按某些特征分成几类。

过去，人们主要靠经验和专业知识进行分类，存在很大局限性，于是数学被逐渐引进到分类中，形成数值分类法。现有的数值聚类方法可以分为两类：结构化聚类方法和非结构化聚类方法。常用的结构化聚类有最短距离法、最长距离法、中间距离法，以及其他对这些算法的改进算法。非结构化聚类方法，又称为动态聚类方法，其特点是先给定一个粗糙的初始分类，然后按照某种预先定义的原则反复进行修改，直至分类较为合理为止。这类算法包括均值法、密度法等，实际中应用广泛的 K-均值法，也属于这一类聚类方法。

动态聚类方法中又可分为传统聚类法与模糊聚类法。传统的聚类分析是一种硬划分，它把每个待辨识的对象严格地划分到某个类中，具有非此即彼的性质，因此这种分类的类别界限是分明的。而实际上大多数对象并没有严格的属性，它们在状态和属性等方面存在着中介性，适合进行软划分。1965 年美国数学家 Zadeh L A 提供模糊集理论，标志着一门新的数学分支——模糊数学的诞生，该理论为这种软划分提供了有力的分析工具，人们开始用模糊的方法来来处理聚类问题，并称之为模糊聚类分析。由于模糊聚类得到了样本属于各个类别的不确定性程度，表达了样本类属性的中介性，即建立起了样本对于各个类别的不确定性程度，能更客观地反映现实世界，从而成为聚类分析研究的主流，应用也最广泛[24]。

模糊聚类方法的原理是按相似性将数据集划分为几个类别以表征系统的不同性征，同时各类别间应满足彼此间最小的重叠，以避免聚类的重复，即各个聚类中心彼此之间应该在包含足够多相似样本的基础上最不相似，换句话讲，同一聚类中的样本应尽可能的靠近，而不同聚类中心之间的距离应尽可能的远。聚类算法也就是寻找若干包含一组与其相似的样本的最不相似样本中心，才能最大程度的代表系统的不同特征，同时各个聚类中心应包含足够数目的样本来保证以尽可能少的聚类中心表达系统。

聚类分析可以分为以下一些步骤：

（1）数据收集并且收集相应的变量；

（2）产生一个相似矩阵；

（3）决定把目标总体细分为几类，及其对每一种类别相应的定义；

（4）实施聚类分析；

（5）产生结果。

2.2.2 现有聚类方法存在的问题和改进

现有的各种模糊聚类方法都是先根据一定的经验准则确定初始聚类个数、选定初始聚类中心，然后根据距离最近的原则对各样本进行依次训练调节，直至目标函数及各样本到相应聚类中心距离平方和收敛到最小为止。由于在聚类算法中自变量（待聚类的坐标值）与目标函数都是离散量，存在着许多局部极值，因此当初始聚类中心在整个样本空间中分布不平衡时，很难把这种不平衡纠正过来，最终结果对初始聚类个数和中心有很大的敏感性。

为了有效地处理局部极值问题，经常采用的对策是用若干不同的初始中心分别进行聚类，然后选择最满意的一个作为最终聚类结果，一般要事先给定聚类个数c，这往往成为制约聚类效果的一个重要因素。目前较常见的c确定方法有两种：比较法和融合法。比较法利用某种度量指标来评价聚类的质量，即将样本集进行若干次聚类，$c\in[2,N]$，N表示样本个数，其中对应于最小度量指标的聚类个数即被视为最佳的聚类数目；融合法从某个较大的聚类个数$c_{\max}$开始聚类，$c\in[2,c_{\max}]$，陆续合并相近或相似的聚类中心，以减少聚类数目。由于聚类个数本质上依赖于系统所呈现的非线性程度，一般只能采用试凑比较的方法来逐步确定聚类个数，这无疑使得初始聚类个数选取具有盲目性同时增加了计算负担。这种对策应用在大型数据分析中，不仅工作量巨大，而且不能保证聚类结果的最优性。

为了克服以上缺点，近年来一些学者将遗传算法引入模糊聚类算法当中以达到全局最优聚类结果。遗传算法作为一类随机搜索算法在组合优化、自适应控制、机器学习等许多领域获得了成功的应用，在聚类分析中运用遗传算法随机产生每一个样本的所属类号，采用遗传算子获得较优聚类，使得算法对初始聚类中心的选取以及样本的输入次序没有任何要求。该算法与一般的模糊聚类法结合后的确取得了不错的效果。但是，正如遗传算法也存在其固有的缺点，在进化过程中不可避免地产生了退化的可能，导致了进化后期的波动现象使迭代次数过大和聚类准确率不很高。

在基于脑神经系统原理的人工神经网络和基于遗传算法机制的进化算法广泛应用于各种知识领域以后，以生物免疫系统为基础的人工免疫系统（Artificial Immune System，AIS）已成为目前国内外计算智能领域一个新的研究热点[25]。近年来，研究者们不断从生物免疫系统中抽取隐喻机制，用于 AIS 的模型设计、算法实现和工程应用[26]。AIS 在信息安全、模式识别、智能优化、机器学习、数据挖掘、自动控制和机器人学等诸多工程领域的初步应用，显示出其强大的信息处理和问题求解能力以及广阔的研究前景[27]。

因此，本文采用了一种借鉴生命科学中免疫概念和理论而发展起来的新兴算法——人工免疫算法，并涵盖了免疫细胞的进化机理，提出了一种聚类的新方案——人工免疫聚类算法。该算法有效地减弱了遗传算法中后期的波动现象，具有更好的整体收敛性能和全局搜索能力，能够使全局模式识别和聚类效果达到最优。

2.3 基于人工免疫系统的聚类方法

2.3.1 人工免疫系统基本原理

生物免疫系统是一个由细胞、分子和器官组成的复杂系统，其主要功能是限制异物对肌体的侵害[28]，免疫系统抵御外部入侵，使其机体免受病原侵害的应答反应称为免疫。外部有害病原入侵机体并激活免疫细胞，诱导其发生反应的过程称为免疫应答。诱导免疫系统产生免疫应答的物质称为抗原（antigen，简称 Ag），能与抗原进行特异性结合的免疫细胞称为抗体（antibody，简称 Ab），抗原与抗体的结合强度用亲和度（affinity）度量。免疫应答主要由分布在生物体内部的免疫细胞实现。人工免疫系统主要设计 T 细胞和 B 细胞的相关免疫特性。B 细胞的作用是识别抗原和分泌抗体，T 细胞能够促进和抑制 B 细胞的产生与分化。两种淋巴细胞共同作用并相互影响和控制对方的功能，形成了机体内部高度规律的反馈型免疫网络。免疫系统的基本结构如图 2.2 所示。

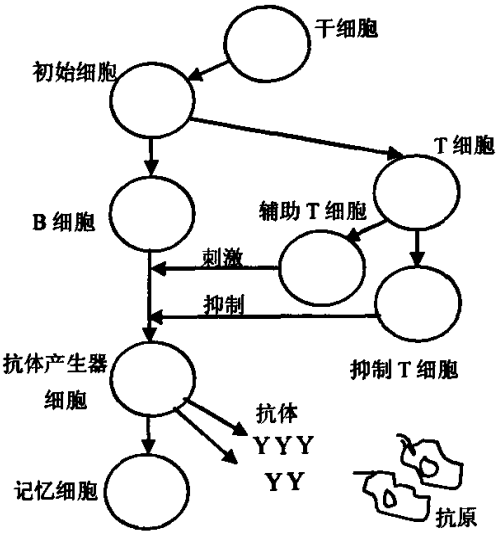

图 2.2 免疫系统的基本结构

Fig 2.2 The basic structure of immune system

免疫细胞主要从骨骼和胸腺中形成，其从产生到成熟，进入免疫循环，需要经历一系列复杂的变化。基于生物免疫构建的人工免疫细胞模型，主要包括自体耐受、克隆、变异、记忆以及死亡等过程。图 2.3 显示了免疫细胞的进化过程。

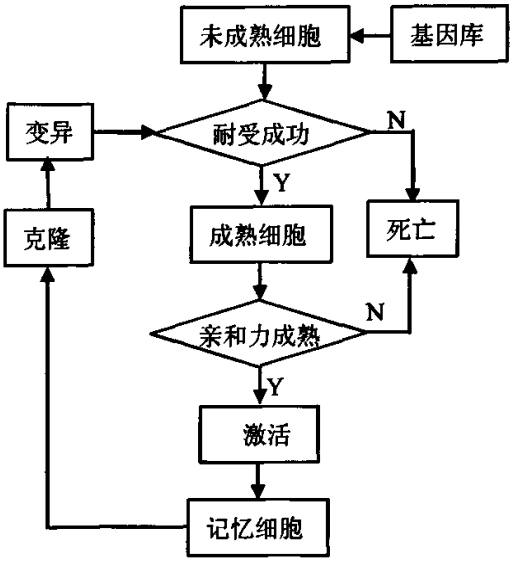

图 2.3 免疫细胞的进化过程

Fig 2.3 The evolution process of immune cells

免疫细胞在骨髓中产生（如图 2.3 所示），需要经历一个耐受过程（免疫细胞包含了 B 细胞、T 细胞和抗体的特性）。如果在耐受期内与自体发生匹配，就会死亡并被新的免疫细胞所替代。免疫细胞经过耐受期后成熟，成熟的免疫细胞被排出骨骼，进入免疫循环。在免疫循环过程中，若遇抗原产生匹配，且累积足够的亲和力（亲和力成熟），则被激活转变为记忆细胞，并进行克隆扩增，产生大量类似的免疫细胞，以抵御更多抗原的入侵，若成熟的免疫细胞在其生命周期内未能累积足够的亲和力，则走向死亡，并被新的成熟的免疫细胞所取代。该过程和未成熟免疫细胞的死亡机制确保了免疫细胞的多样性，保证了其对抗原空间的持续搜索能力，并能保留那些最好的免疫细胞。同时，对一个成熟的免疫细胞，协同刺激也是需要的，这样可能降低错误肯定率。记忆细胞具有更长的生命周期，记忆细胞在再次匹配抗原后就会被再次激活并克隆自己，产生一个再次应答，在这个过程中系统能对以前识别过的抗原做出迅速的反应。克隆生成的新细胞加入成熟细胞集。一部分符合条件的免疫细胞还能进行变异，使系统具有学习进化的能力。变异产生的免疫细胞需要耐受才能成熟。

人们从生物免疫系统的运行机制中获得灵感，开发面向应用的免疫系统计算模型——人工免疫系统，用于解决工程实际问题。从信息处理的观点看，免疫系统是与遗传系统、神经系统并列的人体三大信息系统之一，它具有如下特征或功能：大规模信息并行处理能力、强大的学习能力、记忆能力、识别能力、自适应和鲁棒性、自组织自调整能力和保持抗体多样性的能力。正因为免疫系统具有上述众多的有利用潜力的特性，引发了各个领域众多研究者对免疫系统强烈的研究兴趣，研究者们按照自己的理解和解决现实问题的需要出发，从免疫系统中抽取一个或者几个有用的特性，提出和建立了众多基于免疫系统机理的模型和方法去解决复杂的现实问题，这些基于免疫原理的模型、方法和算法等统称为人工免疫系统。

人工免疫系统是一种并行搜索方法。产生新类型抗体或寻找最合适的抗体来抵御入侵到体内的抗原，是免疫系统所具有的基本能力。免疫系统通过反复试验产生大量类型的抗体来抵抗无数类型的未知抗原。实现抗体类型的多样性是抵抗环境中外来病毒和细菌的基本自适应性。免疫系统通过免疫网络对自身进行调解，保持免疫系统及肌体内环境的稳定，是免疫系统自组织自调整功能的表现[29,30]。

在免疫网络中，抗原与抗体、抗体和抗体存在着激励和被激励、抑制与被抑制的作用，抗原与抗体、抗体与抗体组成一个动态平衡的系统。长期得不到激励的抗体，则既不能识别抗原也不能识别别的抗体的那些抗体将消亡，而那些能够识别抗原或别的抗体

的抗体，将得以在免疫系统中长期存在而称为长命细胞。当相同或相似的抗原再次入侵时，免疫系统就产生所谓的"二次应答"过程，这比"一次应答"更快、更准确、更高效地消除抗原。免疫系统的双重进化机制都是学习和记忆的过程，这为许多工程上的问题提供了有益的启示。

2.3.2 人工免疫聚类算法

自然免疫系统具有个体特异性和整体多样性的双重特性，具备学习、记忆、自我调整、模式识别和特征提取能力。自然免疫系统的抗原识别多样性、抗体的免疫记忆功能和免疫抑制机制，可用来解决数据聚类分析等问题。自然免疫系统同人工免疫系统模型的对照描述如表 2.1 所示。

表 2.1 模型对比（自然的和人工的）

Table 2.1 The competition of natural model and artificial model

生物免疫系统	人工免疫系统模型
抗原	待聚类的原始数据
抗体	记忆数据
B 细胞	初始化数据
激励	满足要求
抑制	不满足要求
记忆细胞	记忆数据集

在数据的人工免疫聚类分析方面，de Castro 和 von Zuben 做了大量的工作[29]。本章采用的免疫算法框架正是得益于他们的免疫机制构想。

为了描述算法的方便，定义如下变量：

Ag：输入抗原，即待聚类的输入数据集合，或研究问题的领域：

$$\begin{cases} Ag=[Ag_1,Ag_2,\cdots Ag_N], \\ Ag_i=[Ag_{i1},Ag_{i2},\cdots Ag_{ip}]\in R^p, i=1,2,\cdots N \end{cases} \tag{2.3}$$

Ab：抗体，即 q 个 p 维的向量组成的初始抗体群：

$$\begin{cases} Ab=[Ab_1,Ab_2,\cdots Ab_q], \\ Ab_j=[Ab_{j1},Ab_{j2},\cdots Ab_{jp}]\in R^p, j=1,2,\cdots q \end{cases} \tag{2.4}$$

M：记忆数据集：

$$\begin{cases} M=[m_1,m_2,\cdots m_k], \\ m_k=[m_{k1},m_{k2},\cdots m_{kp}]\in R^p, k<<N \end{cases} \tag{2.5}$$

现对提出的免疫聚类算法做如下的简要概述：

Step 1 输入抗原{Ag}，做标准化处理。随机产生(0,1)间的 N 个初始化抗体{Ab}。

Step 2 对每一个抗原 Ag_j 操作如下：

Step 2.1 分别计算每个抗体 Ab_i 同抗原 Ag_j 的亲和力 a_{ij}；

Step 2.2 选择亲和力最高的 m 个抗体作为网络细胞；

Step 2.3 对 m 个被选择的网络细胞，做克隆操作。亲和力越高的网络细胞，克隆数 N_c 越多；

Step 2.4 对克隆后的细胞应用方程 $C = C - \alpha(C - X)$ 进行变异操作，C 是克隆抗体细胞，X 是克隆抗原细胞，α 是变异率；

Step 2.5 重新计算变异操作后的 C 的亲和力；

Step 2.6 选择其中亲和力最好的 ζ%，作为部分记忆细胞数据集 M_p；

Step 2.7 淘汰 M_p 中相似度大于阈值 σ_s 的个体；

Step 3 合并 M_p 到已经得到的记忆数据集 M 中。

Step 4 淘汰 M 中相似度大于 σ_s 的个体。

Step 5 随机产生 N 个(0,1)间的个体，替换亲和力差的个体，同上次免疫计算得到的记忆数据 M 作为下一代免疫计算的抗体，并返回 Step 2，直到达到网络收敛。

算法终止条件为，迭代次数超过设定的最大进化代数或相邻两代免疫网络细胞的距离均值和网络规模不再变化或者变化足够小。

2.4 电梯交通流数据的聚类分析

电梯交通流的人工免疫聚类分析分为两个个阶段进行：第一阶段，将电梯交通流数据与人工免疫系统中的各个元素一一对应，视电梯交通流训练数据集为免疫系统的抗原集，采用人工免疫算法提取交通流抗原特征，得到电梯交通流的免疫记忆数据集；第二阶段，根据免疫记忆数据集分类结果，计算各类数据集的聚类中心点，作为该类别的标志，当在线进行电梯群控时，周期采集交通流特征数据，分别计算与各类中心点的欧式距离，并将此刻的交通流模式归为距离最短的那类模式。

2.4.1 电梯交通流模式分类的人工免疫模型

在人工免疫聚类算法中，有两种方法可用来实现数据训练和免疫记忆：匹配算法和进化算法。相关研究表明，进化算法较匹配算法具有更好的性能[31]。

模型将位于不同统计时间段的客流量数据作为抗原，随机产生适当数量的抗体与之相匹配，然后选择亲和力最好的几个抗体作为记忆细胞，这样的过程作为免疫记忆。

首先对交通流量$X(t)$进行统计，由于论文主要针对一般的办公大楼的交通进行研究，根据办公大楼的特点：交通情况与工作作息安排密切相关，所以交通流主要发生在统计时间为7:00~19:00，共十二个小时，每五分钟为一个统计时间间隔，故一天内的客流量统计为一个具有144个数据点的离散时间序列，即$N=144$，而这个实验用的交通流就是通过前面的模拟方法得到的。为了构成数据训练网络，共统计了连续12个工作日的电梯交通流数据，每一天对一代，则构成拥有$I=12$代的交通流数据训练网络，然后进行免疫聚类分析：

（1）电梯交通流数据标准化。能够代表交通流模式特征数据$X(t)$是指单位时间间隔内上行人数x_1，下行人数x_2，总的流通人数x_3，这样$X(t)=\{x(t_i)\};x(t_i)=[x_1(t_i),x_2(t_i),x_3(t_i)]^T$,其中$i=1,2,\cdots 144$。由于一天内客流量数据变化范围比较大,为了统计方便首先对各代的144个交通流数据进行标准化运算,如公式(2.6)所示，使得处理后的电梯交通流数据在[0,1]范围内，构成各代的抗原集Ag:

$$Ag_i=\frac{x(t_i)-\min\{X(t)\}}{\max\{X(t)\}-\min\{X(t)\}} \quad i=1,2,\cdots,144 \tag{2.6}$$

（2）电梯交通流数据的初始抗体群的生成。以第一代的电梯交通流数据为初始抗原群体Ag^1，按照2.3.2节的算法所述，随机生成q个[0,1]范围内的数据元素组成初始抗体群：$Ab=[Ab_1,Ab_2,\cdots Ab_q]$。这里由于对电梯交通流的深刻了解，并通过实验测试，设定$q=30<<144$，并且相应的$[Ab_1,Ab_2,\cdots Ab_{30}]$分别选取任意的[0,1]范围内的数字，作为交通流初始抗体集，即交通流初始类型。这样操作主要是经过实验认定选取$q=30$，聚类效果最佳，而且训练时间符合要求，该算法中后续的相关参数设定均是在对电梯交通流进行多次聚类分析实验后得到的理想的参考参数。

（3）电梯交通流数据的亲和力和相似度的计算。亲和力的定义如式（2.7）所示：

$$a_{ij}=D(Ag_i,Ab_j)=\left(\sum_{k=1}^{p}\frac{(Ag_{ik}-Ab_{jk})^2}{Ag_{ik}^{\ 2}+Ab_{jk}^{\ 2}}\right)^{-\frac{1}{2}} \tag{2.7}$$

式中，p是抗体和抗原的维数，由于交通流特征数据由三部分组成，所以在这里$p=3$。

按照式（2.7）计算每一个交通流抗原数据与交通流抗体数据的亲和力a_{ij}，以及两两交通流抗体数据之间的相似度s_{ij}。亲和力a_{ij}表明了被聚类的交通流抗原数据和现有已经确定的交通模式类别核（即交通流抗体数据）之间的差距，由式（2.7）可知在抗体对抗原进行识别时，亲和力越大说明该交通流抗原数据距离该交通流抗体数据越近，越

容易被该抗体识别，即被归属为该交通模式类别，相反亲和力越小，越容易被该模式类别排除。相似度s_{ij}与亲和力的计算公式一样，表明两两交通流抗体之间的差别即分辨率，同样根据式（2.7）可知在新创建交通流抗体时，需要判断新的抗体与原有抗体的相似程度，如果相似度越大表明分别代表两个交通流模式的两个交通流抗体之间距离很近，过于相似会导致两个交通流抗体集产生交集，使得模式分类不清，这样就会拒绝新抗体的创建，相反如果正在进行识别的抗原数据所有的相似度都很小，说明该交通流抗原数据与已有的交通流抗体数据相差很远，不能被已有的交通流抗体识别，即该交通流数据无法归属为现存的任何交通流模式，那么就需要生成新的交通流抗体，以满足对所有交通流抗原的识别，即以一个新的交通流数据为抗体建立了新的交通流模式类别，那么，这个新建立的交通流抗体就要参加之后的抗原识别运算。

（4）对已识别的交通流抗体数据进行克隆操作。从抗体集中选择亲和最高的$m=\beta q$个交通流抗体数据构成交通流数据克隆源，其中β为克隆选择率$0<\beta\le 1$，一般由经验给出，在这里，由于初始抗体数据为30个，根据测试取$\beta=0.1$效果比较好，这个运算操作过程主要体现了人工免疫聚类算法的择优特性，使得算法朝着最佳分类结果的方向前进。然后对克隆源内的交通流数据进行克隆操作，即复制，体现了优化特性，克隆操作是按照相应的克隆复制率进行的，因此，首先要计算克隆源中各交通流抗体数据$Ab_l(l=1,2,\cdots,\beta q)$的浓度：

$$c_l=\frac{1}{q}\sum_{l=1}^{q}ac_{lj}\text{，其中}ac_{lj}=\begin{cases}1 & s_{lj}\ge T\\ 0 & \text{其他}\end{cases}\text{，}l=1,2,\cdots,\beta q \tag{2.8}$$

式中，T为预先确定的阈值。再根据浓度计算可克隆源中各抗体$Ab_l(l=1,2,\cdots,\beta q)$的期望克隆复制率：$e_l=\lambda a_{lj}/c_l$，$\lambda$是一个比例叙述，可见克隆数的大小和亲和力成正比，和抗体本身的浓度成反比。同样由经验给出，这里令$\lambda=0.7$，c_l为交通流抗体数据本身的浓度，可根据式（2.8）计算得到，其中$T=0.26$。接下来，对每个选取的交通流抗体数依据其期望克隆操作率e_l进行克隆，克隆数的计算如式（2.9）所示：

$$N_C=\sum_{l=1}^{INT(\beta q)}INT(e_l) \tag{2.9}$$

其中INT()为取整函数，N_C个克隆抗体构成了交通流数据的克隆群C。

（5）对交通流数据克隆群C进行变异和抑制运算。变异运算旨在产生具有更高亲和力的抗体，体现了人工免疫算法的进化特性，具体变异操作为对克隆群体中的抗体c_k

以概率P_k进行变异操作，如式（2.10）所示：

$$c'_k = c_k + P_k(Ag_i - c_k), k = 1,2,\cdots,N_C, c'_k \in C' \tag{2.10}$$

其中，P_k为突变率，与亲和力成反比。得到进化群体C'然后再进行抑制操作，主要是去除不满足事先确定的相关阈值的抗体，抑制操作需要重新计算抗原Ag_i与进化群体C'中抗体的亲和力a'_{jk}，选择其中亲和力最高的$\eta\%(\eta=10)$个抗体作为克隆记忆参考集，$\eta\%$称为再次选择率。然后从中去除亲和力低于阈值$\sigma_d(\sigma_d$=1)的抗体，形成m'_k，σ_d称为自然死亡率。再对m'_k进行克隆抑制，重新计算m'_k中抗体间的相似度s_{ij}，删除$s_{ij} > \sigma_s(\sigma_s = 0.105)$对应的记忆抗体，得到克隆即以抗体集$m_k$，$\sigma_s$为免疫抑制阈值，最后将$m_k$合并为交通流记忆数据集M。

（6）电梯交通流网络抑制和自调节运算。对当前交通流记忆数据集M进行网络免疫抑制操作，删除M中所有记忆抗体间的相似度$s_{ij} > \sigma_s$对应的记忆抗体；然后随机产生d(d=2)个[0,1]之间的数据作为新抗体替换原抗体集中亲和力较低的抗体，以体现人工免疫聚类算法的自组织自调节功能。

（7）循环判断操作。判断聚类算法是否达到终止条件，这里终止条件有多重含义，一是看是否达到迭代次数I=12，二是看算法是否收敛到事先设定的收敛精度。如果不满足终止条件则换代，选择下一天的交通流数据继续进行前面的训练操作；如果满足终止条件，则输出结果即为最终电梯交通流数据聚类集合。

2.4.2 客流仿真及人工免疫聚类计算分析

一个大楼，在不同的时期客流情况可能完全不同，即其具有不同的交通模式。以一个典型办公大楼的客流情况为例：在早上上班前的几分钟，职员到达办公大楼乘梯的时间很集中，5分钟内到达的乘客量一般占建筑物总人数的11%-25%，他们到达后从基站乘梯上行到达各楼层上班，这时的乘客到达率很高，对应上行高峰交通模式；上班后到达率下降，其它的交通模式随之开始，直到中午午餐时间到来前，大楼存在少量的层间随机流动的客流，这时的乘客到达率较低，对应层间均衡交通模式；在中午午餐时间内，大楼内人员先从各楼层到餐厅，就餐后再返回各楼层，这时存在不均匀的层间流动客流，对应非均匀层间交通模式；下班时间到达前同样是少量的层间客流，对应均衡层间交通模式；下班时间来到后的几分钟内，大多数乘客从各层下行到基站后离开大楼，这时对应下行高峰交通模式。那么了解实际电梯交通流的情况后，再利用2.1节提出的模拟交

通流实现方法进行仿真。

上/下行交通流数量正数表示上行人数；负数表示下行人数)

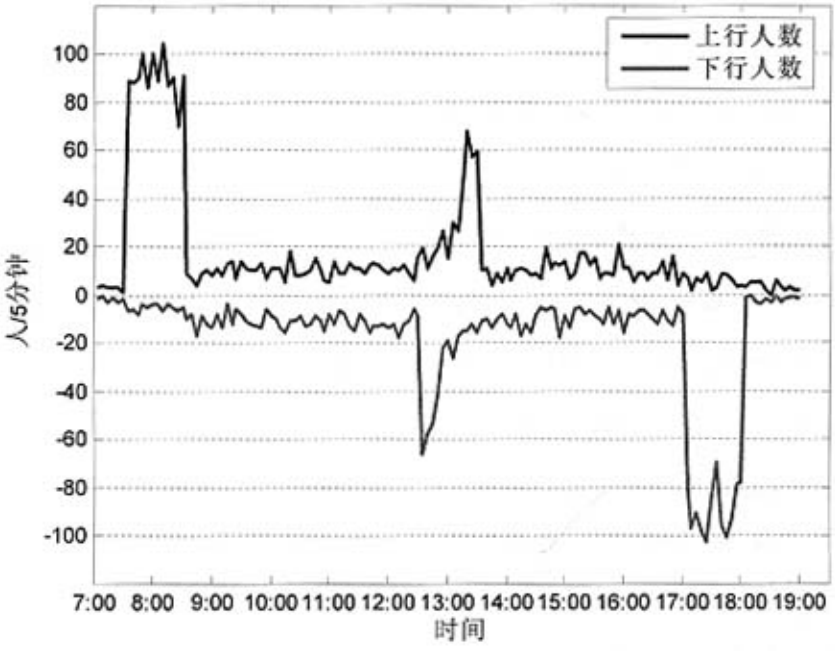

图 2.4 一天中上、下行交通统计图

Fig 2.4 The statistic figure of up and down traffic pattern in one day

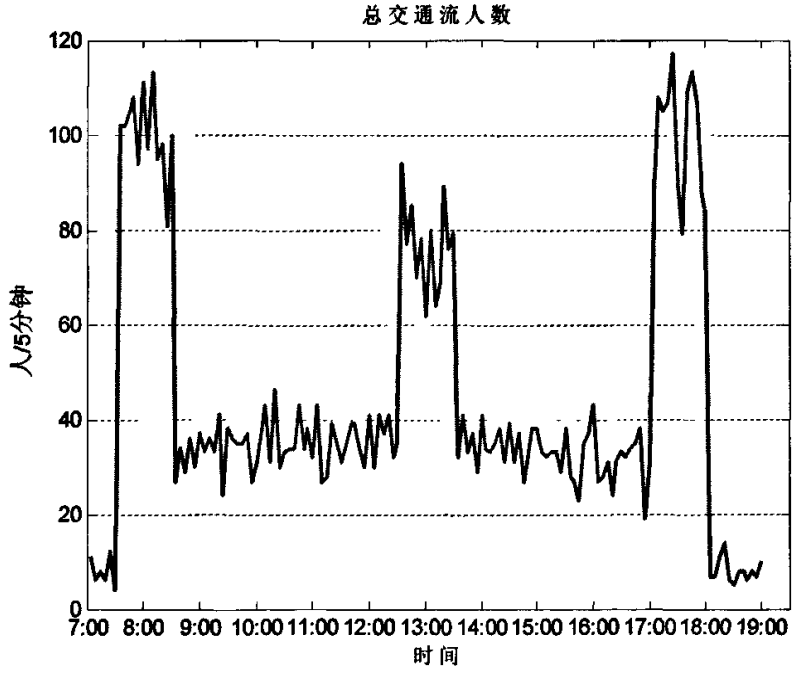

图 2.5 一天中总的交通流统计图

Fig 2.5 The statistic figure of the total traffic in one day

在 Matlab7.0 软件环境下随机产生仿真所需的交通流数据，统计时间从 7：00 到 19：00，每 5 分钟统计一次交通流量，模拟了整个办公大楼白天工作时段的交通状况。并分成上行交通流、下行交通流和总的交通流，如图 2.4 和 2.5 所示。在图 2.4 中，交通流数据的正负号表示客流的方向，正数为向上运行的客流量，负数为向下运行的客流量。将从 7:00 到 19:00 的 144 个交通流采样数据应用到上面介绍的人工免疫聚类算法中。具体的计算过程按照 2.4.1 节的步骤进行，在迭代计算时，发现免疫抑制阈值σ_s是影响整个算法结果至关重要的参数，免疫抑制阈值σ_s越大，得到的免疫记忆数据越多，分类密度越大，但数据的相似性较高；σ_s越小，得到的免疫记忆数据越少，分类密度越小，但所得数据的相似性较低。要得到既能充分反映原始数据特征、数据集规模又比较适中的免疫记忆数据集，σ_s的设置是关键。合理的σ_s取值可以避免相似抗体的大量存在，提高免疫记忆数据的多样性。因此，当样本数据分布稀散时，可适当减小σ_s；当样本相对集中时，可以增大σ_s。本文在电梯交通流样本数据免疫计算中，计算第一组抗原时，初始抗体是用生成的随机数表示的，因此，相似度比较大，所以在进行免疫抑制的时候，设置σ_s=5；在以后的抗原与抗体识别计算中，相似度区域正常，设置σ_s=1.05，较好了反映了交通流分布特点。

表 2.2 人工免疫聚类算法对电梯交通流计算过程

Table 2.2 The calculation of elevator traffic flow with artificial immune clustering algorithm

迭代次数	1	2	3	4	5	6
平均亲和力	0.0359	0.0279	0.0373	0.0338	0.0342	0.0299
网络数据总数	30	26	25	24	23	23
聚类模式个数	20	18	14	10	8	9
迭代次数	7	8	9	10	11	12
平均亲和力	0.0346	0.0382	0.0358	0.0413	0.0427	0.0517
网络数据总数	23	21	22	20	20	20
聚类模式个数	8	7	6	6	6	6

表 2.2 是电梯交通流数据的人工免疫聚类分析计算过程，共迭代了 12 代。由该表可知，免疫记忆数据集的规模和交通流数据之间的平均亲和力都呈稳定趋势，平均亲和力随着迭代次数的增加逐渐增大，说明交通流抗体数据已经逐渐逼近大楼的原始交通流抗原数据。聚类分析结果最终得到 20 个电梯交通流记忆数据作聚类分析，客流量数据可以显著地分为 7 个类。为了直观起见，将 20 个数据点清晰地分成了 7 个类，即 7 种交通流模式，显示在图 2.6 中。

电梯交通流特征数据聚类分析示意图

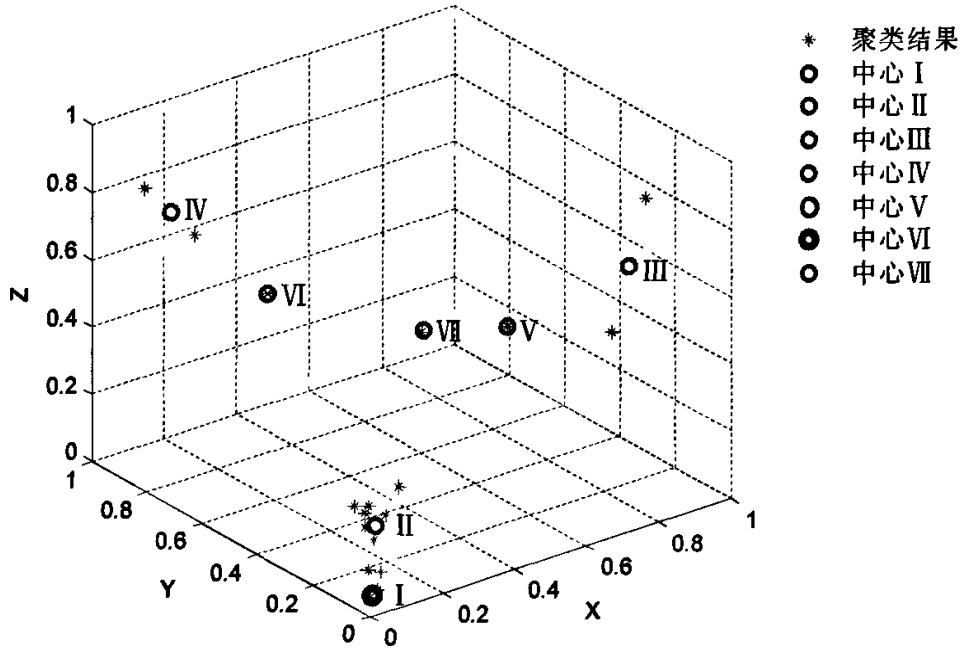

图 2.6 电梯交通流聚类结果

Fig 2.6 The cluster result of elevator traffic flow

X—归一化的上行客流；Y—归一化的下行客流；Z—归一化的总客流

I 类数据点代表的客流量无论是上行方向还是下行方向都明显比其它各类代表的客流量小很多，说明该类型代表了**空闲交通模式**，对应图 2.4 中 7:00~7:30 和 18:00~19:00 这段时间大楼内的交通状况，主要是由于大楼内各单位还没到上班时间或者下班时间已过，因此大楼内几乎没有客流，相应的模式浓度（模式浓度是指该模式所占的比例）为 4/20。

II类数据点所占比例比较多，而且上行客流和下行客流在数量上呈现近似 1:1 的关系，总体客流量较小，这种交通模式出现于一天中的大部分时间，体现了电梯交通流的**随机层间模式**。对应于图 2.4 中 8:30~11:30 和 13:30~16:30 这两大时间段的交通流，这是由人们在大楼中的正常活动而产生的，无论是外来办事人员还是大楼内部单位的工作人员，在大楼内部的活动结束后都会离开大楼或者回到他们原来所在的楼层，使得上下客流量相当，相应的模式浓度为 9/24。

III类数据点的特点是上行客流量远远大于下行客流量，表征了电梯交通流上**高峰模式**，对应于图 2.4 中 7:30~8:30 这一时间段的交通流。这是因为一般办公大楼单位的上班时间都是 8:00 或 8:30，因此在这两个时间点之间形成了明显的上行高峰，相应的模

式浓度为 2/24。

Ⅳ类数据点是下高峰模式，体现了下午下班时的客流情况，对应于图 2.4 中 17:00~18:00 之间的交通状况，因为大多数单位的下班时间是 17:00 和 17:30 这两个时间点，因此在这两个时间前后都会形成一个下高峰模式，相应的模式浓度为 2/20。

Ⅴ类数据点表现出上行客流大于下行客流，而且上行客流的大小仅次于上高峰模式，表征了午饭后上行模式，对应于图 2.4 中 12:30~13:30 期间的交通流，这是因为午饭期间，大楼内许多单位的午休时间是 12:00 到 13:00，工作人员离开大楼就餐完毕，力争在 13:00 前回到工作岗位，因此又形成了一个以上行客流为主要的交通模式，但此时的上行客流量较上班时的上高峰有很大降低，这是因为很多工作人员在午休期间并没有离开办公室，相应的模式浓度为 1/24。

Ⅵ类数据点正好与Ⅴ类数据点对称，说明是午饭前下行模式，该模式的下行客流量的大小仅次于下行高峰模式，对应于图 2.4 中 11:30~12:00 期间的交通流，这是因为大多数员工还是在办公室外就餐或午休，而且由于午休时间不长，从而导致短时间内形成了以下行为主要模式的交通状况，尤其是在 12:00 这一时刻，形成了一个下行尖峰，这是由于一些办公室内为方便员工中午就餐，在大楼下层内部设有食堂，而食堂的开放时间一般都定在 12:00，因此在这个时刻下行流量突增，但总体来说，该模式的下行客流要小于类型Ⅳ的下高峰模式，相应的模式浓度为 1/24。

Ⅶ类数据点的特征是上行和下行客流都比较大，但总客流量较小，这属于复杂层间模式，其实是随机层间模式中比较特殊的一种客流模式，相应的模式浓度为 1/24。

得到分类结果之后，分别计算各类的聚类中心点为(0.0170,0.0149,0.0459)、(0.0604,00621,0.2163)、(0.7554,0.0450,0.7544)、(0.0373,0.7633,0.8344)、(0.4626,0.1068,0.6496)、(0.1052,0.5010,0.6895)、(0.7252,0.6859,0.2426)。

在对电梯群组进行调度时，每隔五分钟统计一次交通流特征数据，求其与各聚类中心的欧几里德距离，如式（2.11）所示。

$$d=\sqrt{\sum_{i=1}^{3}(traffic(i)-center_k(i))^2} \qquad k=1,2,\cdots 7(k\text{表示模式类型}) \qquad (2.11)$$

式中 traffic——当前的交通流特征数据；

center——在各种模式下的聚类中心点。

然后根据距离大小，判断当前情况下的交通流属于哪种模式

2.5 小结

为了获得电梯交通流模式定义和识别的理论依据，对电梯交通流进行聚类研究，基于免疫细胞的进化机理，提出了一种新的人工免疫聚类方法，利用人工免疫系统的免疫激励、免疫抑制和免疫记忆机制，对电梯交通流的 5 分钟客流数据进行特征提取，得到电梯交通流特征数据集，作为抗原数据，进行免疫聚类，最终得到的免疫记忆数据集作为聚类结果，并对聚类结果进行分析和讨论。研究结果表明，本章提出的聚类分析方法不依赖交通流抗体数据的初值，具有较好的聚类结果。

第三章 强化学习方法的基础理论

强化学习是机器学习的一种，是在随机序贯决策问题中学习最优控制策略的理论和技术的整体。所谓强化学习是指从环境状态到动作映射的学习，以使动作从环境中获得的累积奖赏最大[32]。强化学习是目前机器学习研究领域的热点，并同时引起了决策理论、运筹学、控制工程、心理学和神经科学等领域的广泛注意，这是因为强化学习可作为系统正常运行的一部分，而不需要特殊的监督或训练信息。强化学习可被应用于任何涉及采取序列行为的任务，主要集中在有限资源调度、机器人行为规划、各种控制场合、棋类游戏、通信网络和金融等。本章主要介绍强化学习的基本原理、理论模型和算法分类等。

3.1 强化学习的基本原理

3.1.1 强化学习的基本概念

（1）学习、学习系统和外部环境

学习是指一个计算机程序能够通过自身经历提高它在执行某个任务中的性能水平。

学习系统也可以看作是智能体（agent），智能体可定义为处于某个环境一部分的持续自主的运行实体，它感知环境并作用于环境，从事于自己的议程或目标序列以影响其将来可以感知的东西。智能体的一个主要特征是能够适应未知的或变化的环境，学习能力是智能体的关键属性之一。具有学习能力的智能体就被认为是学习系统，在后面的阐述中，智能体和学习系统这两个概念是等价的。

外部环境是指与智能体交互的对象，在控制论中也被称为被控对象。

（2）状态、行为、策略和强化信号

状态（state）是指外部环境（或被控对象）的状态。环境状态的集合称为环境的状态空间。

行为（action），也可称为控制或决策，表示当过程处于某阶段的某个状态时，学习系统选择的作用于外部环境的控制行为，很多研究人员又称之为动作。在实际问题中，行为变量的取值往往根据当前状态而限制在某一范围内，称为允许行为空间。

策略（policy）是各阶段的决策按顺序构成的序列，也称为决策序列。决策可看成为是状态空间到行为空间的映射。

强化信息（reinforcement signal），又称报酬信号（reward signal），是外部环境回馈学习系统的对行为的奖赏（或惩罚）。

3.1.2 强化学习的基本原理和发展历史

强化学习通常包括两个方面：一是将强化学习作为一类问题；二是解决这类问题的一种技术。如果将强化学习作为一类问题，目前的学习技术大致分为两类：一类是搜索智能体的行为空间，以发现智能体最优的行为，通常可以通过遗传算法等搜索技术实现；另一类是采用统计技术和动态规划方法来估计在某一环境状态下动作的效用函数值。

3.1.2.1 强化学习系统的结构及基本原理

标准强化学习系统基本结构如图 3.1 所示，单智能体（控制器）作为学习系统，获取外部环境（被控对象）的当前状态信息 s，对环境采取试探行为 u，并获取环境反馈的对此动作的评价 r 和新的环境状态 s'。如果智能体的某动作 u 导致环境正的奖赏（强化信号），那么智能体以后产生这个动作的趋势便会加强；反之，智能体产生这个动作的趋势将减弱。在学习系统的控制行为与环境反馈的状态及评价的反复的交互作用中，以学习的方式不断修改从状态到动作的映射策略，目的是优化系统性能。

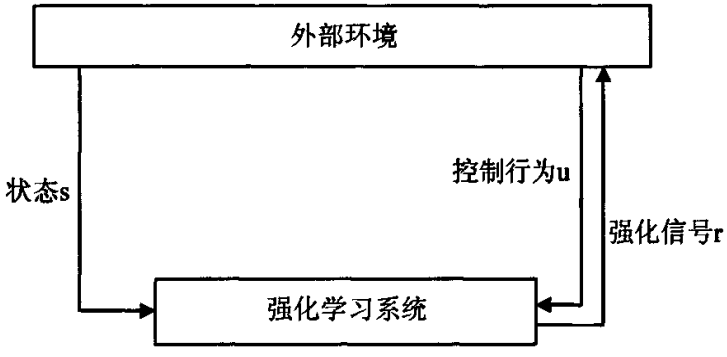

图 3.1 学习系统原理图

Fig 3.1 Theory figure of reinforcement learning system

据此可得到强化学习的目标：学习从环境状态到行为的映射，使得智能体选择的行为能够获得环境最大的奖赏，使得外部环境对学习系统在某种意义下的评价（或整个系统的运行性能）最佳。也就是说，强化学习要解决这样的问题：一个能够感知环境的智能体如何通过学习选择能达到其目标的最优动作。

可见，强化学习是通过与外部环境的不断交互，从自身经验学习优化策略的方法。

3.1.2.2 强化学习方法与其他方法的比较

目前在人工智能领域，认为强化学习属于机器学习的研究范畴。在机器学习中与强化学习相提并论的，包括无监督学习（Unsupervised Learning）和有监督学习（Supervised Learning）。所谓监督，就是明确地存在一个教师的指导，向学习系统提供清楚的关于期望响应的指导信息，具体说就是，向学习系统提供的数据集中包含要由系统复现的期望输出。三者的区别主要表现在：教师提供的训练方式不同。

教师是学习中训练数据的提供者。无监督学习方法不依赖于外部教师去引导学习过程，而是将教师内置于学习方法中。与无监督学习不同的是，监督学习和强化学习都需要外部的教师提供训练信号而引导学习过程。

在监督学习中，教师向学习系统为每个给定输入提供期望的输出。学习过程要通过最小化系统的实际输出和期望输出间的差异来记忆这些期望的输出。可见，监督学习中的教师信号是指导性的，监督学习时从教师提供的样本（显式的输入输出对）中学习。虽然这是一个重要的学习方式，但不太适用于与环境的交互学习。强化学习中的教师信号是评估性的，不需要事先提供训练例子。强化学习中的教师（又称为评估准则）根据某种性能测量，向学习系统提供对一个给定行为的执行所做的标量评估。因此，强化学习可以是一种在线学习技术，可被用于通过提供对系统性能的合适评价去学习未知的期望的输出问题。

强化学习不同于自适应控制技术。自适应控制中，系统模型必须可以从统计数据中估计，而且系统动态模型的结构是固定的。因此，自适应控制本质上是一个参数估计问题，而且可以通过统计分析进行假设估计，而在强化学习中并没有这些限制。

强化学习也不同于规划技术，区别在于规划需要构造复杂的状态图，而强化学习系统只需要记忆所处的环境状态和当前的策略知识；其次，规划技术总假定环境是稳定的，系统与环境的交互作用能够通过某种搜索过程来预测，由于规划技术并没有真正地考虑行为如何适应环境的问题，所以只适用于系统完全了解并可控的环境。相反强化学习强调与环境的交互作用，因此强化学习技术比规划技术适用面更广[33]。

3.1.2.3 强化学习方法的历史发展

强化学习是人工智能领域中既崭新又古老的课题，其发展历史可粗略地划分为两个阶段：第一阶段是 50 年代至 60 年代，可以成为强化学习的形成阶段；第二阶段是 80 年代以后，可以成为强化学习的发展阶段。

在第一阶段，“强化”和“强化学习”这些术语由 Minsky 首次提出并出现在工程文

献上[34]。当时数学心理学家探索了各种计算模型以解释动物和人类的学习行为。他们认为学习是随机进行的，并发展了所谓的随机学习模型[35]。Widrow、Hoff 和 Rosenblatt 这些神经网络先驱们，以及心理学家 Bush 和 Mosteller 等都研究过强化学习。他们利用了“奖励”和“惩罚”这样的术语，但他们的研究系统越来越趋向于监督学习[36]。在控制理论中，由 Waltz 和付京孙于 1965 年分别独立提出这一概念[36]。在应用方面，最早的应用例子是 Samuel 的下棋程序，该程序采用类似值迭代、瞬时差分和 Q 学习的训练机制，来学习用线性函数表示的值函数[37]。Widrow 及其同事们在研究监督学习时，认识到监督学习和强化学习之间的不同，并于 1973 年 Widrow、Gupta 和 Maitra 改正了 Widrow-Hoff 监督学习规则（常称为 LMS 规则）。新规则可实现强化学习，即根据成功和失败的信号进行学习，代替原来的使用训练样本进行学习。他们用“有评价的学习”一词代替“有教师的学习”[38]。Saridis 把强化控制系统的控制器看成一个随机自动机，首次系统地提出了采用强化学习来解决随机控制系统的学习控制问题的方法[39]。

在六七十年代，强化学习研究进展比较缓慢。进入 80 年代以后，随着人们对人工神经网络的研究不断地取得进展，以及计算机技术的进步，人们对强化学习的研究又出现了高潮，逐渐成为机器学习研究中的活跃领域。Richard S. Sutton 在 1988 年提出的瞬时差分算法（Temporary Difference）和英国的 Chris Watkins 在 1989 年提出的 Q-learning 算法。这些研究扩展并整合了强化学习研究的所有主线中的前期工作。之后，随着对强化学习的研究和应用日益开展起来，它越来越成为目前机器学习研究领域的热点之一。

强化学习的一个更大的流行趋势是人工智能和其他工程学科间的进一步联系。现代的人工智能研究者接受了统计学和控制理论算法。人工智能和传统工程间曾被遗忘的领域已变得非常活跃，比如神经网络、智能控制和强化学习。在强化学习中，扩展了来自最优控制和随机逼近的思想来达到人工智能的更广阔、更远大的目标。

3.2 强化学习的理论背景与基本模型

强化学习模型有简单复杂之分，最简单的情况是外部环境只有单状态，行为只影响立即报酬，对后续状态以及后续报酬没有影响；较为复杂的情况是行为不仅影响当前报酬，并通过影响后续状态进而影响后续报酬，这一特点在强化学习研究领域成为延迟报酬。后者是目前主要研究的情况，本节针对后者重点阐述强化学习模型的理论背景，并在此基础上说明强化学习在延迟报酬情况下的基本要素，这些要素构成了强化学习的基本模型。

3.2.1 强化学习的理论背景

强化学习模型的理论背景主要是马尔可夫决策过程（Markov Decision Process，简称MDP）和半马尔可夫决策过程（Semi-Markov Decision Process，简称SMDP）。在多数学者的研究中，强化学习模型是基于马尔可夫决策的[40,41]。但需要澄清的是，强化学习不受限于MDP，只是MDP是随机序贯决策问题的基本模型，其离散时间有限状态的特点为研究强化学习的基本算法和特点提供了最简单的框架[42]，学术界经常使用有限状态MDP作为强化学习的典型背景，研究强化学习算法框架，测试强化学习算法有效性。

马尔可夫决策过程又叫马尔可夫决策规划或马尔可夫决策问题，是解决随机性序贯决策问题的重要分支学科[43]。随机序贯决策问题是一类多阶段决策问题，也就是说，在系统运行过程中，在一系列观察时刻点上都要做出决策，并且系统状态的转移是事先无法确切知道的随机事件。在每个观察时刻，决策者首先根据观察所得的系统状态，从其所有的被选方案中选择一个方案（做出决策）执行，然后再观察下一时刻系统出现的状态，据此再作新的决策，如此一步一步地进行下去。各阶段决策构成一个决策序列，使整个过程达到某种意义上的最优。整个过程为前后关联具有链状结构的多阶段过程。如果在序贯决策过程中，系统状态的转移与系统以前的发展历史无关，则称为这类系统具有无后效性（或马尔可夫性），并称解决这类序贯决策问题的数学模型为马尔可夫决策过程。马尔可夫决策过程中，系统状态的转移概率与选用的方案两者交互作用决定了系统的发展进程。

用数学描述就是：在一系列阶段中的每个阶段，智能体观测环境的状态s，这个状态属于外部环境的有限状态空间S。然后从有限行为空间U_s中，选中一个行为u并执行，智能体立刻收到一个期望值是$R(s,u)$的奖赏$r(s,u)$，且外部环境以概率$P(s'|s,u)$进入下一阶段的状态s'。$R(s,u)$和$P(s'|s,u)$构成了强化学习研究者所谓的“行为u的一步模型”。智能体的策略是状态空间到行为空间的映射，表示为函数$\pi: S \to U_s$。为了评价在外部环境处于状态s时，采用策略π产生行为u所产生的影响的好坏，就需要定义一个指标函数来表明从长期的观点看，确定什么是最优的行为。因此，在任意状态下定义一个返回值，这个值是在任意状态s为初始状态时，采用策略π在后续有限或无限的时间区间中获得报酬的期望值的和，这一返回值也叫做状态的值，用$V^{\pi}(s)$表示。更准确的说，$V^{\pi}(s)$是策略π下状态s的值，相应地V^{π}叫做策略π的值函数。

由此看见，MDP的六个基本要素可用下面的六元组表示：

$$\{S, \bigcup_{s\in S} U_s, R: S\times U \to R, P: S\times U\times S \to [0,1], V^{\pi}\}$$

式中，五元组$\{S, \bigcup_{s\in S} U_s, R: S\times U \to R, P: S\times U\times S \to [0,1], V^{\pi}\}$是已知的，目标就是求解这个五元组而获得策略$\pi^*$，使得每个状态的值$V^{\pi}(s)$最大。$\pi^*$被称作最优策略，相应的值函数$V^*$为最优值函数，即：$V^*(s) = \max_{\pi} V^{\pi}(s)$。最优策略可以有很多种，但均对应唯一的最优值函数。

值函数$V^{\pi}(s)$的定义有不同的方式，主要包括以下三种：

（1）有限时间区间型

值函数定义为在后续 h 步的报酬和的期望。

$$V^{\pi}(s) = E(\sum_{i=1}^{h} r_{t+i} \mid s_t = s, \pi) \tag{3.1}$$

式中，s_t和r_{t+1}分别表示在阶段t时环境状态和在阶段t时采取行为而获得的立即报酬。这种类型的 MDP 目前研究的不多。

（2）无限时间区间折扣型

值函数定义为在后续无限时间区间内报酬折扣和的期望。

$$V^{\pi}(s) = E(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid s_t = s, \pi) = E(\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i} \mid s_t = s, \pi) \tag{3.2}$$

式中，γ是折扣因子。在这种$V^{\pi}(s)$定义下的 MDP 又被称为折扣型 MDP。

（3）平均报酬型

值函数定义为在无限时间区间内，每个阶段平均的报酬值。

$$V^{\pi}(s) = \lim_{h\to\infty} \frac{1}{h} E(\sum_{i=1}^{h} r_{t+i} \mid s_t = s, \pi) \tag{3.3}$$

在这种$V^{\pi}(s)$定义下的 MDP 被称为平均报酬 MDP。

目前研究比较广泛的是后两者，折扣型 MDP 和平均报酬 MDP。而由式（3.1）（3.2）（3.3）可见，无论是哪一种 MDP 的模型，不同点只是在于状态值函数的不同定义。相应的，不同 MDP 的动态规划求解方法或强化学习求解方法间的差异也主要来源于状态值函数的差异。

3.2.2 动态规划与强化学习

3.2.2.1 动态规划

动态规划是运筹学的一个分支，是解决多阶段决策过程最优化的一种数学方法[44]，是考察问题的一种途径，而不是一种特殊算法。动态规划可用于解决最优路径问题、资源分配问题、生产调度问题、库存问题、装载问题、排序问题、设备更新问题和生产过程最优控制问题等等。动态规划是求解某类问题的一种方法，因此，不像线性规划那样有标准的数学表达式和明确定义的一组规划，而必须具体问题具体分析处理。

动态规划是求解 MDP 的基本方法。由于目前折扣性 MDP 的研究最为广泛，而且后续的电梯群控系统所用也是这种模型，所以下面主要阐述动态规划求解折扣性 MDP 的方法。对于平均报酬 MDP，美国 Massachusetts 大学的 Sridhar Mahadevan 在文献[45]中有详细描述。

在动态规划中，折扣型 MDP 的最优值函数式（3.2）满足 Bellman 方程形式：

$$V^*(s)=\max_{u\in U_s}[R(s,u)+\gamma\sum_{s'\in S}P(s'|s,u)V^*(s')] \quad \forall s\in S \tag{3.4}$$

由式（3.4）可见，Bellman 方程就是最优值函数V^*的递归定义。动态规划通过解 Bellman 方程得出最优值函数V^*，进而根据式（3.5）找到最优策略π^*。

$$\pi^*(s)=\arg V^*(s)=\arg\max_{u\in U_s}[R(s,u)+\gamma\sum_{s'\in S}P(s'|s,u)V^*(s')] \tag{3.5}$$

3.2.2.2 动态规划的求解方法

动态规划求解值函数的常用算法有两种：值迭代和策略迭代。

（1）值迭代

寻找最优策略的一种方法就是寻找最优值函数。值迭代通过迭代计算逐步更新每个状态的值，逐渐逼近最优值函数。

在每个迭代步 k，为每个状态应用式（3.6）的操作来更新V^*的逼近$\hat{V}_k$。对每个状态s的值的一次更新称为回溯（backup），对所有状态值的回溯构成一次搜寻（sweep）。所谓一次搜寻就是一次迭代（iteration）。

$$\hat{V}_{k+1}(s)=\max_{u\in U_s}[R(s,u)+\gamma\sum_{s'\in S}P(s'|s,u)\hat{V}_k(s')] \tag{3.6}$$

每一次回溯都是用当前状态可能的后续状态的值的估计来更新当前状态的值。在计算的初始，人为给定初始V_0，通过反复的迭代产生的迭代序列$\{V_k\}$将收敛于V^*。找到了最优值函数，那么任何产生使式（3.6）等号右边的值为最大的行为的策略（称为贪婪

策略）就是最优策略。

（2）策略迭代

策略迭代是动态规划求解方法的另外一种。策略迭代直接操纵策略，而不是通过最优值函数间接寻找优化策略。步骤如下：

首先给定一个初始策略π_0，然后根据式（3.7）计算当前策略相应的值函数，再根据计算得出的当前策略的值函数，利用式（3.8）更新策略，重复这个过程直至前后两次更新的策略相同。

$$\hat{V}_{\pi}(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s,\pi(s)) \hat{V}_{\pi}(s') \tag{3.7}$$

$$\pi'(s) = \arg\max_{u \in U_s} [R(s,u) + \gamma \sum_{s' \in S} P(s' \mid s,u) \hat{V}_{\pi}(s')] \tag{3.8}$$

策略迭代和值迭代都是同步的，也就是说，算法的操作是在每个迭代步，对整个状态空间作了一次耗尽式的搜寻。在更新状态值的时候，仅利用了其他状态的旧值。

3.2.2.3 动态规划在应用中存在的问题

MDP 可以效率较高得使用值迭代或策略迭代来解决。然而，动态规划虽被认为是求解一般随即优化控制问题的有效方法，但其可应用性是受限制的，存在如下三个问题：

（1）对于大规模 MDP，过大的状态空间使得式（3.6）中对每个状态的值的更新计算和搜寻一次的迭代计算不可行。这就是所谓的 Bellman“维数的诅咒”（the curse of dimensionality）。

（2）未知的动力学方程。这里所谓的动力学方程是指环境的模型。在 MDP 中，环境模型是指状态转移概率函数和期望报酬函数，即

$$R: S \times U \to P: S \times U \times S \to [0,1]$$

动态规划的求解技术需要获得系统状态转移概率和期望报酬函数，但是在多数状况下是不可获得的。

（3）难以显式的计算出环境模型。即使面对的外部环境是完全可知的，但仍然很难获得式（3.6）所需要的显式表达的状态转移概率和期望报酬函数。

这三个问题制约着动态规划在求解随机序贯决策问题中的实际应用。其中，大状态空间的问题是主要问题，也是工程上近似动态规划方法的背景主题。

3.2.2.4 强化学习的优势

现代强化学习来源于动态规划，被看作动态规划的近似方法。虽然已有很多方法被建议用来近似求解 MDP，但是强化学习新颖地融合了蒙特卡罗（Monte Carlo），随机逼

近和函数逼近技术。特别地，强化学习具有下列特征和优势：

（1）通过将计算限制于实际的或仿真的多个样本轨迹附近的状态上，而避免了动态规划的耗尽式搜寻。因为计算是沿着样本轨迹进行的，所以这种方法可以最大限度的利用很多状态在实际经历中出现的概率低的情况。

（2）通过采样简化基本的动态规划回溯。通过从合适的分布中采样，而不是通过产生和评估所有可能的下个状态，来估计一次回溯的结果。

（3）通过使用函数逼近方法，比如基函数的线性组合、神经网络或其他方法，来表达值函数或策略函数，比表格表示更加紧凑。

特征（1）和（2）反映了在使用强化学习时所追求的是逼近的本质。强化学习最终得到的反映智能体行为的逼近函数在整个状态空间上并不是一致的。智能体的策略在很少到达的状态上不需要高度精确，而强化学习的逼近本质避免了动态规划因为状态空间大而导致的不可计算的问题。对于第 3 个特征，有很多例子[46,47]证明了函数逼近对于困难问题的有效性。这三个特征说明了强化学习在应用于大规模问题时相对于动态规划的优越性。

3.2.3 强化学习模型

强化学习模型以 MDP 为背景，并与其模型具有极强的相似性。强化学习模型中包括的主要基本元素是：

状态空间S、行为空间U、策略π、报酬函数和值函数，还有一个可选元素是环境模型。这里的状态空间是有限状态集，行为空间是有限行为空间。

强化学习模型的分析可在模型中对值函数定义的基础上得出。下面针对 MDP 背景给出强化学习中的值函数的定义。

在 MDP 背景下，强化学习模型中的值函数可以使用原有 MDP 中定义的状态值函数，也可对 MDP 状态值函数加以改动而定义状态-行为对的值函数$Q^{\pi}(s,u)$。以折扣型 MDP 背景为例，$Q^{\pi}(s,u)$的定义为：

$$Q^{\pi}(s,u)=E(r_{t+1}+\gamma r_{t+2}+\gamma^{2}r_{t+3}+\cdots\mid s_{t}=s,u_{t}=u,\pi) \tag{3.9}$$

相应的Q^{π}表示状态行为对的值函数。最优策略π^{*}对应的状态-行为对的值函数为Q^{*}，称为优化状态-行为值函数。Q^{*}在很多强化学习算法中起着极为重要的作用，Q^{*}表示在任意状态-行为对(s,u)下，使用最优策略在后续阶段报酬的折扣和。Q^{*}与V^{*}关

系是：

$$V^* = \max_u Q^* \tag{3.10}$$

由式（3.4）和（3.10）得出：

$$Q^*(s,u) = R(s,u) + \gamma \sum_{s' \in S} P(s' \mid s,u) \max_{u' \in U_{s'}} Q^*(s',u') \quad \forall s \in S, \forall u \in U_s \tag{3.11}$$

式（3.11）是状态-行为值函数的 Bellman 方程。可见，$Q^*(s,u)$ 是对在状态 s 下执行行为 u 的评价值，与状态最优值函数 $V^*(s)$ 是等价的，即：可通过求解式（3.11）间接寻找最优策略，与通过求解式（3.4）的 $V^*(s)$ 进而得到最优策略是效果相同的。对应于由 MDP 状态值函数的 Bellman 方程进行值迭代的式(3.6)，有求解状态行为值函数 Q^* 的值迭代公式（3.12）：

$$\hat{Q}_{k+1}(s,u) = R(s,u) + \gamma \sum_{s' \in S} P(s' \mid s,u) \max_{u' \in U_{s'}} \hat{Q}_k(s',u') \tag{3.12}$$

在 MDP 背景下，虽然 $Q^*(s,u)$ 的定义以及迭代计算公式中仍然包含状态转移概率函数和期望报酬函数，但是在以后说明的强化学习算法中不再需要环境的模型。

在强化学习模型中，策略定义了学习系统在某个时刻行为的方式（如何做出动作）。大约来说，策略是由环境被感知的状态到在这些状态所采取行为的映射。在一些情况下，策略可能是一个简单的函数或者表格；而在其他一些情况下，可能涉及类似一个搜索过程的大量计算。策略是强化学习智能体的核心，一般地，策略可以是随机的。

报酬函数定义了强化学习问题的目标。粗略地说，是将环境被感知的状态（或状态-行为对）映射到一个数值。强化学习智能体的核心目标是最大化在长期运行中所获得的报酬的总和。报酬是作为改变策略的基础，如果策略选择了一个行为而获得了较低的报酬，那么策略将被变更而去选择其他行为。一般地，报酬函数也可以是随机的。

报酬函数是从立即的角度表明了什么是好的，而值函数是从长远的角度表明什么是好的。由一个状态可能会得到低的立即报酬，但可能会有高的值，因为后续状态能给出高的报酬，相反的情况也能出现。值函数在决策和评估决策时是最受关心的，行为的选择也是在值判断的基础上做出的。被选中的行为都是最大化状态值，而不是最大化立即报酬。但是，确定值比确定报酬更难，因为报酬是由环境直接给出，而值必须由智能体在整个生存期中根据观测序列估计和重估计。因此，几乎所有强化学习算法的最重要部分就是有效估计值函数的方法。

一些强化学习系统中还包括一个可选要素：环境模型。模型是对环境行为的模拟。

比如，给定一个状态和行为，模型能够预测后续状态和奖励。模型是用于计划的，计划是指在实际运行前，通过考虑可能的未来状况来决定行为序列。模型、计划与强化学习系统的结合是相对新的发展。然而，研究者已渐渐意识到强化学习方法与传统动态规划方法是密切相关的。

3.3 强化学习的经典算法

目前各种类型的强化学习算法都是经典的标准强化学习算法的衍生，因此重点研究标准强化学习中的经典算法，对于进一步深入研究来说是基础的一步。

这些典型算法的核心都是对随机序贯决策问题中的状态值函数V^*或者状态-行为值函数Q^*的迭代逼近，迭代的基本算式为：

$$V(s) \leftarrow (1-\alpha)V(s)+\alpha(r(s,u)+\gamma V(s')) \tag{3.13}$$

式中，α是学习率，随时间衰减。状态s的值或状态-行为值的更新是根据当前状态值、立即报酬和后续状态值做出的。

（1）TD 算法

TD（Temporal Difference）学习是强化学习中最主要的算法之一。TD 学习是蒙特卡罗和动态规划思想的结合，也就是说，一方面 TD 算法不需要系统模型就可以从学习系统的经验中学习；另一方面 TD 算法和动态规划一样，迭代估计最优值函数。

TD 算法由 Richard S. Sutton 于 1988 年提出，最简单的 TD 算法是 TD(0)算法，即一步 TD 算法。也就是说，学习系统得到的瞬时奖励仅向后回退一步，只修改了前一步状态的值。迭代公式为

$$V(s_t) \leftarrow V(s_t)+\alpha(r_{t+1}+\gamma V(s_{t+1})-V(s_t)) \tag{3.14}$$

学习过程为：首先初始化V值；然后智能体在状态s_t下根据当前策略确定行为u_t，进而得到经验$<s_t,u_t,r_{t+1},s_{t+1}>$；利用此经验根据式（3.14）修改状态值，当智能体访问到目标状态，算法结束一次迭代循环。算法继续从初始状态开始新的迭代，直至学习结束。

式（3.14）中的更新操作类似于值迭代中的回溯计算，唯一的不同就是样本的抽取来自于真实世界，而不是来自于模拟一个已知的模型。

TD(0)算法在学习率因子α递减的情况下，能够保证收敛到最优值函数。

（2）Q-learning 算法

Q-learning 算法是一种离策略（off-policy）TD 算法，Q-learning 算法的提出是强化学习研究的突破性进展。不同于 TD 算法，Q-learning 在迭代中使用状态-行为对报酬和状态-行为值函数。Q-learning 算法的最简单形式是一步 Q-learning。其值函数迭代规则为：

$$Q(s_t,u_t) \leftarrow Q(s_t,u_t)+\alpha[r_{t+1}+\gamma \max_{u\in U_s} Q(s_{t+1},u)-Q(s_t,u_t)] \tag{3.15}$$

式中行为值函数 Q 是对期望的最优行为值函数 Q^* 的估计，其独立于所采用的策略。这极大的简化了算法的分析。虽然策略决定了被访问和被更新的状态行为对，但是只要所有的状态行为对持续不断地被更新，算法就能正确收敛。

学习过程为：主观初始化 $Q(s,u)$；在每一个时间段中，初始化状态 s；在该时间段内的每个时间步，在状态 s_t 下，使用由 Q 得来的策略来选择行为 u_t，执行 u_t 并观察 r_{t+1} 和 s_{t+1}，得到经验 $<s_t,u_t,r_{t+1},s_{t+1}>$，根据式（3.15）更新 $<s_t,u_t>$ 的值；当访问到目标状态时，结束一次迭代循环。在下一时间段中，开始新的迭代循环。

（3）Sarsa 算法

Sarsa 算法最初被称为改进的 Q-learning 算法，它仍采用 Q 值迭代，一步 Sarsa 算法可用下式表示：

$$Q(s_t,u_t) \leftarrow Q(s_t,u_t)+\alpha[r_{t+1}+\gamma Q(s_{t+1},u_{t+1})-Q(s_t,u_t)] \tag{3.16}$$

智能体在每个学习步，首先策略根据当前 Q 值确定行为 u_t，得到经验 $<s_t,u_t,r_{t+1},s_{t+1}>$；然后策略根据当前 Q 值确定 s_{t+1} 状态下的行为 u_{t+1}，这样得到五元组 $<s_t,u_t,r_{t+1},s_{t+1},u_{t+1}>$，并根据式（3.16）更新 $Q(s_t,u_t)$；之后将行为 u_{t+1} 作为智能体的下一行为执行。显然，Sarsa 算法与 Q-learning 算法的差别在于，Q-learning 时使用后续状态的最大行为值来更新当前状态-行为值，而 Sarsa 算法使用实际的后续状态-行为值进行迭代。

（4）R-learning 算法

R-learning 算法由 Schwartz 提出，是一个离策略（off-policy）强化学习算法，其适用的强化学习问题背景是平均报酬型 MDP，目标是获得每个时间步上最大的平均报酬。策略 π 的值函数与该策略下单个时间步的平均期望报酬 ρ^π 有关。ρ^π 的定义是：

$$\rho^\pi = \lim_{h\to\infty}\frac{1}{h}\sum_{t=1}^{h} E\{r_t^\pi\} \tag{3.17}$$

假设整个过程是各态历经的（在任何策略下，任何状态由任意其他状态以非零概率到达），因此ρ^{π}不依赖起始状态。从任何状态起始，在长期运行看来，平均报酬是相同的，但是存在一个瞬态。从某些状态起始，在这一小段时间内接收到的报酬多于平均报酬，从其余状态，接收到的报酬少于平均报酬。由此，定义状态的值为

$$\tilde{V}^{\pi}(s)=\sum_{k=1}^{\infty}E(r_{t+k}^{\pi}-\rho^{\pi}\mid s_t=s) \tag{3.18}$$

状态-行为对的值类似的是

$$\tilde{Q}^{\pi}(s,u)=\sum_{k=1}^{\infty}E(r_{t+k}^{\pi}-\rho^{\pi}\mid s_t=s,u_t=u) \tag{3.19}$$

$\tilde{V}^{\pi}(s)$和$\tilde{Q}^{\pi}(s,u)$被称为相关值，因为其与当前策略下平均期望报酬相关。求解的目标是寻找使ρ^{π}最大的策略，这个策略是最优策略。R-learning 中包括对状态-行为值的迭代和对单时间步平均报酬的迭代。状态-行为值的迭代公式为

$$Q(s_t,u_t)\leftarrow Q(s_t,u_t)+\alpha[r_{t+1}-\rho+\max_{u}Q(s_{t+1},u)-Q(s_t,u_t)] \tag{3.20}$$

平均报酬值的更新公式为：

$$\rho\leftarrow\rho+\beta[r_{t+1}-\rho+\max_{u}Q(s_{t-1},u)-\max_{u}(s_t,u)] \tag{3.21}$$

学习过程为：首先初始化ρ和Q；在状态s_t下策略根据 Q 值选择行为u_t，然后获得经验$<s_t,u_t,r_{t+1},s_{t+1}>$，根据式（3.20）更新当前状态行为对的 Q 值；如果更新后的 Q 是状态的最大行为值，那么根据式（3.20）更新ρ。

（5）Dyna-Q 算法[32, 48]

Richard S. Sutton 在 1991 年提出 Dyna 学习结构[49]。Dyna-Q 算法是有模型的折扣型强化学习算法，不同于无模型的方法的是：充分利用实际获得的经验于环境模型的建立中。算法同时做三件事情：利用实际中的经验执行强化学习，将同样的强化学习用于模型产生的经验，利用实际经验更新系统模型[50]。

Dyna-Q 算法的运行是在智能体与环境的交互循环中执行的。在与环境的一次实际交互中，获得实际经验$<s_t,u_t,r_{t+1},s_{t+1}>$，进而执行下面操作：

首先，利用式（3.15）更新当前状态的 Q 值；然后根据实际经验，利用概率统计技术更新对外部环境模型的估计，即状态转移函数$\hat{P}(s'\mid s,u)$和期望报酬函数$R(s,a)$；之后根据更新后的模型，随机模拟确定 k 个$<s_t,u_t,r_{t+1},s_{t+1}>$经验样本，进行值迭代；最后，

智能体选择执行状态s_{t+1}下的行为u_{t+1}，一次与环境的交互完成。

（6）H-learning 算法

H-learning 算法由美国俄勒冈州立大学的 Prasad Tadepalli 和 DoKyeong OK 在 1997 年提出[51]，是基于模型的平均报酬强化学习算法，H-learning 是 Adaptive Real-Time Dynamic Programming(ARTDP)的非折扣版本[52]。不同于 R-learning 和 ARL，H-learning 是基于模型的，也就是说，智能体学习并显式地使用环境的模型。

其学习步骤基本上是：获取经验$<s_t,u_t,r_{t+1},s_{t+1}>$后，根据这个四元组更新当前一步行为模型$P(s_{t+1}\mid s_t,u_t)$和$R(s_t,u_t)$；然后根据当前环境模型，选择使当前状态值最大的行为 u，如果该行为就是实际中选择的行为u_t，那么根据式（3.22）更新平均报酬ρ，其中ρ的定义见式（3.18），并根据式（3.23）更新学习率α；最后根据式（3.26）更新当前状态s_t的值，完成与环境的一次交互。

$$\rho \leftarrow (1-\alpha)\rho+\alpha(R(s_t,u_t)-H(s_t)+H(s_{t+1})) \tag{3.22}$$

$$\alpha \leftarrow \frac{\alpha}{\alpha+1} \tag{3.23}$$

$$h(s_t)=\max_u(R(s_t,u)+\sum_s R(s_t,u)+\sum_s P(s\mid s_t,u)h(s)-\rho) \tag{3.24}$$

H-learning 可以看成是 R-learning 和 ARTDP 的交叉产物。

在 H-learning 基础上，改动了探索策略，扩展得出 AH-learning 算法。在用于大规模问题求解时，使用动态贝叶斯网络描述环境模型，使用局部线性回归的函数逼近方法近似值函数，解决了 H-learning 算法在大状态空间上的泛化问题。

以上各种算法，无论是无模型还是有模型，无论是折扣型还是平均型，均是以 MDP 为背景，以动态规划为算法思想，迭代逼近最优状态值函数或最优状态-行为值函数，进而得到隐式的优化策略。其中，Q-learning 和 Sarsa 算法研究较为广泛。

3.4 小结

强化学习是一种理解和使目标导向的学习与决策自动化的计算性方法。它不同于其他计算性方法，强调单体从与环境的直接交互中学习，且不依赖监督或完全的环境模型。当为了获取长效目标而从与环境的交互中学习的时候，强化学习是解决这种情况下所出现的计算性问题的首选。

第四章 基于强化学习方法的电梯调度系统

4.1 电梯群组的调度问题

4.1.1 电梯群组调度系统

大楼尤其是高层建筑，对电梯的使用率很高，一部电梯难以满足需要，因此常常安装多部电梯形成群组。电梯群组系统的运行是多部电梯沿各自井道上下接送乘客。群组系统运行机制如图 4.1 所示。

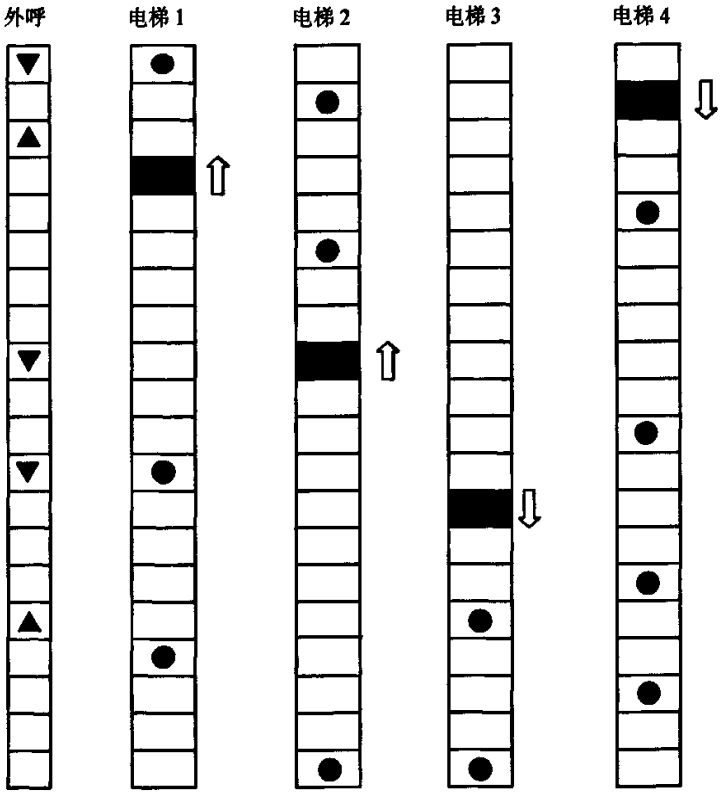

图 4.1 电梯群组运行机制图

Fig 4.1 Running rules of elevator group system

图中，灰色长方框代表电梯，“▲▼”表示外呼，“●”表示内选。每部轿厢由单梯控制器按集选控制规则控制起停，接收外呼层站的乘客，送出内选层站的乘客，外呼层站可以不停靠，而内选层站必须停靠。

多部电梯如果单独运行或简单的并联运行，则运行效率低。随着人们对电梯这种随

机服务系统性能的要求逐渐增高，电梯群组的调度是非常必要的。而电梯群组调度问题是指在乘客与电梯系统的请求/应答服务的交互过程中，如何将外呼合理地分配给可用的电梯，进而使电梯群控系统的性能最优。电梯群控系统是提高电梯群组系统性能的物理系统，基本结构如图 4.2 所示：

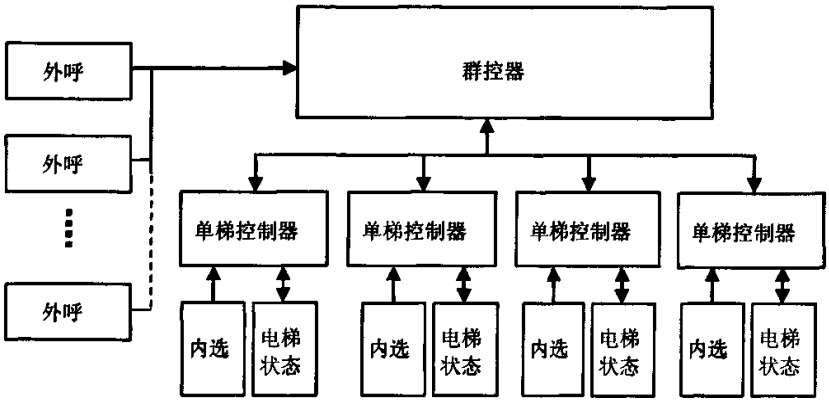

图 4.2 电梯群控系统结构图

Fig 4.2 The structure of elevator group control system

电梯群控系统主要分为三部分：单梯控制器、电梯群控器和信号传输系统。各部分的职责分配为：

（1）单梯控制器：每部电梯都当作一个电梯控制器。单梯控制器负责电梯的启动、加速、减速、停止、开关门等运动控制，另外同时负责轿厢运行是否转向、是否停靠、启动等逻辑控制。

（2）群控器：通过信号传输系统接收乘客的外呼请求信号和单梯控制器传送过来的单梯状态信号（位置、方向、单梯的内选等），并据此决定派梯方案，即决定将外呼信号分配给哪部电梯去响应。

（3）信号传输系统：负责电梯系统中各部分间的通讯，传输来自各层站的外呼信号，向单梯控制器传送派梯信息，同时负责与乘客交互。

其中，群控器是电梯群控系统的核心，调度算法是系统的关键，调度方法的好坏直接决定了系统性能的优劣，是目前研究的热点。群控器内部结构如图 4.3 所示。

群控器中的状态处理单元的职责是接收外来的电梯状态和外呼状态，加以分析后，处理成能够使调度单元满意的信息模式。调度单元的职责是以优化事先定好的性能指标为目的，执行调度算法，发出恰当的派梯信息。

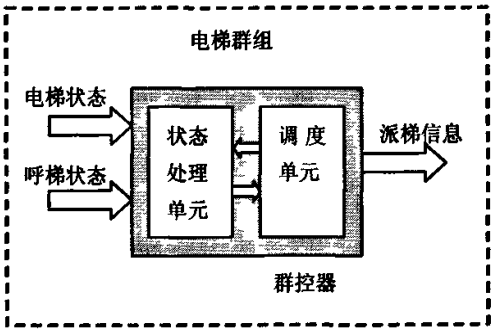

图 4.3 群控器内部结构图

Fig 4.3 Inside structure of the elevator group controller

4.1.2 调度目标

电梯群组调度作为优化问题，必然存在目标。在这个特殊领域，调度目标需要特别的分析。调度的目的主要分为两个方面：

（1）高服务质量。从服务质量角度来讲，电梯系统提供垂直的人员交通服务，乘客希望在电梯外等待的时间和乘梯时间越短越好，轿厢内拥挤程度越小越好。

（2）提高系统运行效率。从电梯系统本身来讲，要求电梯尽量避免空驶，提高载客效率，减少起停车次数。电梯的起停次数也与电梯乘客的生理承受能力密切相关，过多的起停次数必将增加乘客的生理负反应。

通过上述分析，可将电梯群组调度的目标完全根据乘客的需求划分为心理需求和生理需求。心理需求使用乘客候梯时间、乘梯时间和轿厢内拥挤度等指标衡量；生理需求使用起停次数指标衡量。电梯调度就是要满足乘客的心理和生理双重需要，因此可将电梯调度问题看作以满足乘客要求的目的为驱动的控制或决策问题，而以上指标成为对控制或决策性能的评价准则。

4.1.3 调度问题难点

电梯群组调度问题由于在实用中的重要地位已经得到广泛研究，但这一问题是极度复杂的，问题的组合复杂性和随机性为电梯群组调度方法的研究带来困难与挑战。难点主要体现在：

（1）状态空间巨大。因为电梯是在连续时间空间上运行的离散事件动态系统（Discrete Event Dynamic System，简称 DEDS），状态包括所有电梯的位置、运行方向、速度、每个轿厢内的乘客数量及其目的楼层和等候乘梯的乘客数量及其呼梯楼层、呼梯方向（在上高峰交通模式下，候梯乘客的所在楼层和呼梯方向均相同，都是门厅）。当

一个新乘客到达时，好的电梯群控器（调度器）要决定如何为该乘客派梯就必须考虑所有这些状态信息。巨大的状态空间使得电梯群组调度问题是组合爆炸的。

（2）系统运行过程中伴随着极大的不确定性。电梯的运动完全由当前的调度决定，而调度是随新乘客的不断到达而不断改变的，乘客的到达是一个随机过程。乘客的到达具有三种不确定性：到达时间，到达楼层和乘客的目的楼层。而且同一天内大楼内的交通流模式也是不断变化的，不同天内同一时段交通流分布也可能是不同的。

（3）如果调度允许重新访问已有的分派并将外呼重新分派给轿厢，那么在短时间内增加计算量，需要考虑的分派数极大。在西方国家这是典型的操作模式，成为重分派策略，日本恰恰相反，分派结束后不再重新考虑。

总之，根据电梯群组调度问题的定义，通过对电梯群组运行特点、群组调度目的和交通流模式的分析和调度问题难点的阐述，可知电梯群组调度问题在运筹学观点下是随机的组合优化问题，在控制的观点下是随机最优决策问题。

4.2 电梯调度系统的问题模型

问题模型的建立是应用强化学习算法解决电梯群组调度问题的关键环节。由于实际问题的极度复杂性，在理论研究中有必要做出一些假设。在假设的基础上，定义问题模型中的状态、行为、成本以及值函数等要素。

4.2.1 电梯调度的相关假设

假设 1：假设调度单元知道所有乘客的到达，即乘客的到达时间、到达楼层和乘梯方向。因此，可假设电梯的所有状态都是可观测的，因为尤其重要的是，厅外候梯的乘客数和轿厢内乘客均是知道的，这在实际中很难测到。

假设 2：假设不会有两个乘客同时达到，即所有乘客均是按先后次序排队到达。这也是比较符合实际的，因为即使乘客同时到达，由于电梯群控系统中信号系统是串行通信，在群控器看来也是有先后关系的。

假设 3：调度单元决策时只考虑目前系统内所有乘客的服务，未来乘客的到达不被考虑其中。因此，不必假设乘客到达的概率分布。这是符合实际情况的，现实中乘客到达的随机性是最难以规律性表达的。

假设 4：群控器只负责将轿厢分派给请求服务的乘客，不负责单梯的运行控制，单梯按照集选控制规则逐个接收被分派的呼梯乘客和运送乘客到目的层。

假设 5：电梯负载容量有限，满载情况下，不响应任何分派给它的外呼梯信号，需要重新分配。假设由于满载而导致的重分派是调度单元唯一考虑的重分派呼梯的情况。

4.2.2 电梯调度系统的问题模型

4.2.2.1 电梯调度系统的框架背景

电梯群组调度问题在控制的观点下是随机最优决策问题，因此可以考虑使用 MDP 或 SMDP 等随机序贯决策问题模型作为强化学习方法求解电梯群组调度问题的模型框架。

电梯群组系统具有 DEDS 的特点，乘客的到达事件和轿厢的到达事件影响系统的状态不断改变。而电梯群控系统作为电梯群组系统的调度（控制）中枢，对群组系统状态的观测可以分为两种方式：时间驱动型和事件驱动型。时间驱动型是指群控系统主动地定时观测群组运行和呼梯状态，派梯行为的决策发生在状态观测之后，同样是定时机制；而事件驱动型是指外部乘客到达、轿厢到达等离散事件发生之后，群组状态由单梯控制器主动发送，群控系统被动地接收。

电梯群组系统状态观测的两种不同机制影响问题模型的建立。在时间驱动型情况下，可以使用 MDP 作为电梯群组调度问题模型。这是因为 MDP 将时间轴分成多段，在每个时间段的起始时刻观测状态、决策行为，而不考虑时间间隔，适合应用于定时机制条件下。在事件驱动型情况下，可以使用 SMDP 作为模型。因为 SMDP 模型中考虑两个决策点之间的时间量，这个时间量是随机变量，而事件的发生是不定时的和随机的，相应的决策也是如此，将决策过程模型定为 SMDP 是合适的。

本文使用 MDP 作为电梯群组调度问题的模型框架，也就是说，群控器以时间触发机制观测群组系统状态和进行决策计算。这一方面是对问题模型的简化，另一方面也是符合实际情况，因为在电梯群控系统中信号系统的串行性常常使得群控器与单梯控制器和层站呼梯板间的通信机制为轮询的主从方式。

MDP 框架六元组为：

$$\left\{\mathbf{S},\bigcup_{s\in\mathbf{S}}U_s,\pi:\mathbf{S}\to U_s,R:\mathbf{S}\times U\to R,P:\mathbf{S}\times U\times\mathbf{S}\to[0,1],V^{\pi}\right\} \tag{4.1}$$

式中的粗体变量表示向量和相应的向量空间。将电梯群组运行中的电梯状态和乘客状态均视为系统状态$\mathbf{s}$，调度单元对电梯的派梯方案看作决策行为u，整个电梯群组的运行就是在群控器观测状态$\mathbf{s}$，做出决策u的循环过程中经历的；为保证 Markov 性，假

设电梯系统状态转移仅与电梯系统前一时刻状态有关，与系统以前的发展历史无关。

由于强化学习不需要明确的概率模型，因此，在建立问题模型中，只需定义四元组：

$$\left\{\mathbf{S},\bigcup_{s\in\mathbf{S}}U_s,R:\mathbf{S}\times U_s\to R,V^{\pi}\right\} \tag{4.2}$$

由于电梯运行以最小化乘客等待时间、乘梯时间、减少停靠次数为目标的特殊背景，则在电梯群组调度问题中，报酬应为成本，而调度问题的目标是寻求最优的派梯方案，使得定义的值函数最小，即求π^*，使得

$$V^*(\mathbf{s})=\min_{\pi}V^{\pi}(\mathbf{s}) \tag{4.3}$$

下面逐个定义状态、行为、报酬和相应的值函数。

4.2.2.2 电梯状态与调度行为

如前面分析，电梯群组系统的状态空间巨大，状态不仅仅包括离散的各层外呼和内选状态，同时包括电梯速度、位置等连续变量，而且还有人员数量的各层分布情况。在一个20层大楼4部电梯的系统中，粗略地估计，状态数量为：$2\times2^{19}\times2^{20}\times4\times4\times20\times2\approx2^{49}$。

而在不同交通模式下，所需观测的状态数量是不同的，比如在纯上高峰模式下，不需考虑外呼和外呼方向，因为只有一个乘客到达楼层；而在纯下高峰模式下，不需考虑内选状态，因为电梯只有一个目的楼层。

为不失一般性，设$\mathbf{s}$表示电梯群组系统状态，包括电梯的运行方向（direction）、当前位置（position）、外呼（hall calls）、内选（car calls）和新呼梯（new calls）作为群控器观测的状态信息。其中，外呼状态包括等待乘客的数量信息，即各层请求服务的人数；内选状态包括乘梯乘客的数量信息，即在各层流出的人数；新呼梯状态包括新产生的呼梯楼层和方向状态。

设u表示派梯结果，本文中电梯调度问题是将新呼梯的产生分配给群组中某台电梯，因此决策的行为就是派梯方案。为减少行为空间，定义梯号i为决策结果，即决策由哪一部电梯服务当前新产生的呼梯。需要注意的是，在任意状态下，满载电梯不作为可选的决策结果。

4.2.2.3 电梯调度系统的成本与值函数

成本与值函数的定义是与电梯调度的性能指标密切关联的。电梯调度中最常使用的性能指标是平均候梯时间，即从乘客到达至进入电梯的时间对所有乘客的平均值。本文

综合考虑等待时间、乘梯时间和电梯停靠次数作为性能评价的综合指标。

如4.2.1节假设3所述，只考虑当前系统内的乘客，而不考虑未来可能到达的乘客，这不仅是对随机性的简化，而且对于问题模型中成本的定义是有影响的。本文对成本的定义不是基于对所有乘客的等待时间、乘梯时间和停靠次数的平均值定义的，而是考虑到强化学习算法的在线执行和成本的即时性，考虑当前存在于群组系统内部的乘客（包括候梯的和乘梯的乘客），不考虑已退出系统和未进入系统的乘客。

（1）定义平均候梯时间成本

设共有 p 个乘客，乘客 i 的当前等待时间为

$$T_w(i) = t - t_i \tag{4.4}$$

式中 t——当前时刻；

t_i——乘客 i 的到达时间；

乘客 i——在上次决策时已等候而当前仍在等候的乘客。

则平均候梯时间成本为

$$R_w = \frac{1}{p}\sum_{i=1}^{p} T_w(i) \tag{4.5}$$

（2）定义平均乘梯时间成本

设梯内 p' 个乘客在上次决策时也在梯内，其中乘客 j 的当前乘梯时间为

$$T_r(j) = t - t_j \tag{4.6}$$

式中 t——当前时刻；

t_j——乘客 j 进入电梯的时间；

乘客 j——在上次决策时已等候而当前仍在等候的乘客。

则平均乘梯时间成本为

$$R_r = \frac{1}{p'}\sum_{j=1}^{p'} T_r(j) \tag{4.7}$$

（3）定义停靠次数成本

设电梯 i 的未来停靠次数为 C_i，则停靠次数成本为

$$R_c = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{4.8}$$

式中 N——电梯系统中的电梯数量。

根据式（4.5），（4.7），（4.8）定义状态和行为的综合成本为

$$R=\sqrt{R_w{}^2+R_r{}^2+R_c{}^2} \tag{4.9}$$

值函数$V^\pi(\mathbf{s})$是根据成本定义的，如3.2.1节所述可以有不同的定义形式，主要包括：有限时间区间型、无限时间区间折扣型和平均报酬型。论文使用无限时间区间折扣型定义值函数，即

$$V^\pi(\mathbf{s})=E(r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots|\mathbf{s}_t=\mathbf{s},\pi)=E(\sum_{i=0}^{\infty}\gamma^i r_{t+i+1}|\mathbf{s}_t=\mathbf{s},\pi) \tag{4.10}$$

式中 $\gamma\in(0,1)$——折扣因子。

γ将后续成本折扣到当前状态，突出了当前立即成本作为对电梯状态的评价的重要性大于后续成本的重要性，同时保证$V^\pi(\mathbf{s})$有限，使得$V^\pi(\mathbf{s})$作为策略的评价是有意义的。当策略π为随机策略时，

$$\begin{aligned}V^\pi(\mathbf{s})&=E(r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots|\mathbf{s}_t=\mathbf{s},\pi)\\&=E(\sum_{i=0}^{\infty}\gamma^i r_{t+i+1}|\mathbf{s}_t=\mathbf{s},\pi)\\&=E(r_{t+1}+\gamma\sum_{i=0}^{\infty}\gamma^i r_{t+i+2}|\mathbf{s}_t=\mathbf{s})\\&=\sum_u\pi(\mathbf{s},u)\sum_{s'}P(\mathbf{s}_{t+1}=\mathbf{s}'|\mathbf{s}_t=\mathbf{s},u_t=u)(R(\mathbf{s},u,\mathbf{s}')+\gamma E(\sum_{i=0}^{\infty}\gamma^i r_{t+i+1}|\mathbf{s}_{t+1}=\mathbf{s}'))\\&=\sum_u\pi(\mathbf{s},u)\sum_{s'}P(\mathbf{s}_{t+1}=\mathbf{s}'|\mathbf{s}_t=\mathbf{s},u_t=u)(R(\mathbf{s},u,\mathbf{s}')+\gamma V^\pi(\mathbf{s}'))\end{aligned} \tag{4.11}$$

为简化形式，方便理解，设：

$$P(\mathbf{s}_{t+1}=\mathbf{s}'|\mathbf{s}_t=\mathbf{s},u_t=u)=P^u_{\mathbf{ss}'}$$

$$R(\mathbf{s},u,\mathbf{s}')=R^u_{\mathbf{ss}'}$$

则式（4.11）最终简化为：

$$\begin{aligned}V^\pi(\mathbf{s})&=\sum_u\pi(\mathbf{s},u)\sum_{s'}P^u_{\mathbf{ss}'}(R^u_{\mathbf{ss}'}+\gamma V^\pi(\mathbf{s}'))\\&=\sum_u\pi(\mathbf{s},u)(R^u_{\mathbf{s}}+\sum_{s'}\gamma P^u_{\mathbf{ss}'}V^\pi(\mathbf{s}'))\end{aligned} \tag{4.12}$$

式（4.12）即为$V^\pi(\mathbf{s})$的Bellman方程。

在电梯群组调度问题中，最优值函数

$$V^*(\mathbf{s})=\min_\pi V^\pi(\mathbf{s})=\min_u(R^u_{\mathbf{s}}+\sum_{s'}\gamma P^u_{\mathbf{ss}'}V^*(\mathbf{s}')) \tag{4.13}$$

相应的，状态-行为值函数为

$$Q^*(\mathbf{s},u)=R_{\mathbf{s}}^{u}+\sum_{\mathbf{s}'}\gamma P_{\mathbf{ss}'}^{u}V^*(\mathbf{s}') \tag{4.14}$$

强化学习方法使用状态-行为值函数作为强化学习模型中的值函数，电梯群控调度算法就是通过强化学习算法求解式（4.14），逼近最优的状态-行为值函数，进而得到优化的派梯策略。

4.3 基于强化学习的电梯调度算法

基于强化学习的调度算法包括三部分：值迭代算法、探索性随机策略和基于前向神经网络的泛化技术。

4.3.1 值迭代算法

电梯群组调度问题以 MDP 为框架模型，值迭代算法是核心。由于群控单元作为单决策智能体以最小化长期折扣成本为目标，而且不考虑电梯群组系统的状态转移概率模型，因此可使用标准强化学习算法中的无模型折扣报酬强化学习算法 Q-learning 作为值迭代方法更新求解最优状态-行为值函数。

求解式（4.14）的值函数迭代一般规则为：

$$Q_{t+1}(\mathbf{s}_t,u_t)=Q_t(\mathbf{s}_t,u_t)+\alpha[R_{t+1}+\gamma\min_{u\in U_t}Q_t(\mathbf{s}_{t+1},u)-Q_t(\mathbf{s}_t,u_t)] \tag{4.15}$$

式中 α——强化学习的学习率，在学习过程中逐渐衰减；

γ——为成本的折扣因子，$\gamma\in(0,1)$。

群控器在每一个时间步的起始时刻观测系统状态$\mathbf{s}_t$，使用当前 Q 函数值Q_t得来的策略来选择派梯行为u_t，执行u_t后并在时间段尾观测状态行为所带来的成本值R_{t+1}和后续状态$\mathbf{s}_{t+1}$，这样就得到经验组合$<\mathbf{s}_t,u_t,R_{t+1},\mathbf{s}_{t+1}>$，根据式(4.15)更新状态行为对$<\mathbf{s}_t,u_t>$的 Q 值；在下一时间段中，开始新的迭代循环。随迭代进行，将逐渐逼近最优的状态-行为值函数$Q^*(\mathbf{s},u)$，同时隐式地获得最佳派梯策略。在 MDP 序贯决策的框架背景下，不考虑时间步的时间跨度，时间段之间也可以是不连续的，这一特点就非常适合仿真状态下的值函数学习与逼近。

上述的值函数更新算法过程是以时间步（time step）为计算基础的，但需要改造成为适应计算机运算只是发生在离散时间点上这一特点的算法过程。

改造后的值迭代算法过程为：

（1）初始化$Q_0(\mathbf{s},u)$，$\forall \mathbf{s}\in \mathrm{S},\forall u\in U_{\mathbf{s}}$

（2）在时刻 t 观测状态$\mathbf{s}_t$和R_t

（3）在前一时刻的Q_{t-1}更新为当前Q_t

$$\Delta Q_t(\mathbf{s}_{t-1},u_{t-1})=\alpha[R_t+\gamma\min_{u\in U_{\mathbf{s}}}Q_{t-1}(\mathbf{s}_t,u)-Q_{t-1}(\mathbf{s}_{t-1},u_{t-1})]$$

（4）在状态$\mathbf{s}_t$下，使用由Q_t得来的策略来选择行为u_t

（5）$\mathbf{s}_{t-1}\leftarrow\mathbf{s}_t,u_{t-1}\leftarrow u_t$

（6）下一时刻 t+1 返回 2 重新执行。

算法的图形化说明如图 4.4：

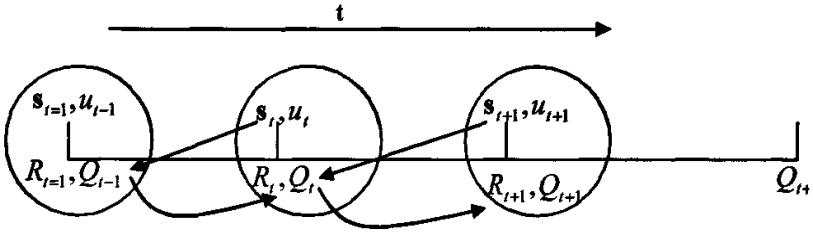

图 4.4 值迭代算法过程图解

Fig 4.4 Value iteration algorithm process

随着时间的推进，系统状态不断改变，调度单元不断的进行状态观测和决策。在观测到电梯群组系统状态的同时，得到立即的对前一时刻状态和派梯方案的成本评价，并且根据前一时刻的值函数、当前状态、立即评价和当前状态对值函数作更新。观测到的成本是对派梯方案优劣的立即评价，而值函数是对派梯方案对后续状态影响的长期综合评价。

Q-learning 值函数迭代的算法过程在应用于一些状态空间较小的问题时，状态-行为值函数的存储是通过表格实现的。而在应用于电梯群组动态调度问题的研究时，还需考虑：行为选择的探索-利用问题和值函数的泛化问题。

4.3.2 行为选择的探索性随机策略

由 4.3.1 节中的算法过程看出，在调度过程中，值函数的修改过程与调度决策的执行过程是同时进行的，而策略的选择过程就是策略的修改过程。因此，强化学习运算不断逼近最优值函数的过程也就是不断修改派梯策略的过程。当派梯策略被优化后，就面临一个问题，行为选择的探索-利用问题，也就是说：是利用已获得的、可以产生高回

报的派梯方案，还是选择搜索未知的派梯方案。

搜索新方案能够带来长期的性能改善，可以帮助收敛到最优策略，利用旧方案可以帮助系统短期性能改善，但可能得到次优解。目前有几种探索性的行为选择策略，包括贪婪搜索、随机策略和基于间隔的技术等。其中，贪婪策略最为简单，根据当前最大的状态-行为值来选择行为。随机策略是以概率随机选择行为，一些随机策略为鼓励初始探索而以大概率值起始。无论贪婪策略还是随机策略，都需要当前的 Q 函数的值来决定选取的决策。

使用 Gibbs 分布构造探索性的随机策略，以概率选择行为。派梯行为的产生概率的计算式为：

$$prob(u_t)=\frac{e^{-Q(s_t,u_t)/T}}{\sum_{u} e^{-Q(s_t,u)/T}} \qquad u_t \in \{i \mid i=1,2,3,4\} \tag{4.16}$$

其中，$T>0$ 是温度参数，在学习过程中逐渐衰减。由式（4.16）可以看出，在当前状态 s_t 下，最终选择哪部电梯服务是由当前的 Q 函数决定的。Q 值越大，则该派梯方案越不合理，相应的被选择的概率就越低，这就是与直观的感觉相符的。参数 T 是较为重要的参数，控制着派梯选择的随机性，影响向未知行为空间探索的强度。在计算初期，Q*的估计值很不精确，则使用大的 T 值，各行为的选择概率相同，鼓励不同派梯方案的探索；在计算后期，Q*的估计值逐渐精确，使用小的 T 值，则评价值（成本）低的方案有更高的被选择的概率，同时仍然允许探索其他方案。

4.3.3 值函数的泛化方法

泛化（generalization）是指在整个状态空间的有限子集中的经历被一般化，进而产生在更大状态空间子集上的逼近值。运用强化学习方法研究电梯群组调度面临值函数的泛化问题，原因有二：

一是在值函数迭代过程中，需要存储 Q 函数。一般的强化学习在应用中，值函数均使用表格存储，表格的入口是状态和行为，出口是状态行为值，表格为每一个状态-行为对建立一个条目来存储值函数。然而，这对于电梯群组动态调度这种状态空间庞大的问题来说是极不现实的，使用表格存储值函数将耗费大量内存和对表格搜索的计算时间。

二是在调度的实际运行中，可能有很多已经遇到过的状态不会原原本本的再现。在这种情况下，学习的唯一途径就是由以前经历过的状态泛化（推广）到从来没有遇到过

的状态。

这类泛化常被称为函数逼近，这是因为它从目标函数中抽取样本，并试图由这些样本一般化地构造整个函数的一个逼近。函数逼近是有监督学习的一个实例，是在机器学习、人工神经网络、模式识别和统计曲线拟和中的主要研究课题。原则上，在这些领域中所研究的任何方法均可用于强化学习中。将函数逼近方法作为泛化技术与强化学习算法结合，也是强化学习理论与监督学习方法的结合。

在电梯群组调度问题中，尝试使用三层前馈神经网络作为状态-行为值函数的逼近器，选择的理由是因为前馈神经网络可以逼近任意的非线性函数。

设神经网络的输入为 **X**，网络权值为 **W**，输出为 **Y**，则前馈神经网络建立起的输入和输出的函数关系为 $\mathbf{Y}=\psi(\mathbf{X},\mathbf{W})$。使用三层前馈神经网络逼近强化学习中状态-行为值函数，只需将电梯状态 **s** 和派梯行为 u 映射到网络的输入 **X**，将评价函数 Q 映射到网络的输出 **Y**，则 $Q=\psi(\mathbf{X},\mathbf{W})$。可见，神经网络的函数逼近方法将定义在系统状态和行为的值函数参数化表达了。因此，以 MDP 为背景的电梯群组调度问题的值函数逼近问题就转变成为三层前馈神经网络权值的训练问题。

训练的样本数据是群控器在线观测状态并执行调度的过程中获得的。强化学习迭代更新值函数的方法与神经网络将值函数泛化到整个状态空间的结合点就是：强化学习的值更新算法提供神经网络的训练数据。即：前一时刻的状态-行为对与现在时刻更新后的状态-行为值$<(\mathbf{s}_{t-1},u_{t-1}),Q_t>$作为训练数据去调整网络权值，其中 Q_t 为理想输出，可由式（4.15）计算得出。用于值函数存储的神经网络结构如图 4.5 所示。

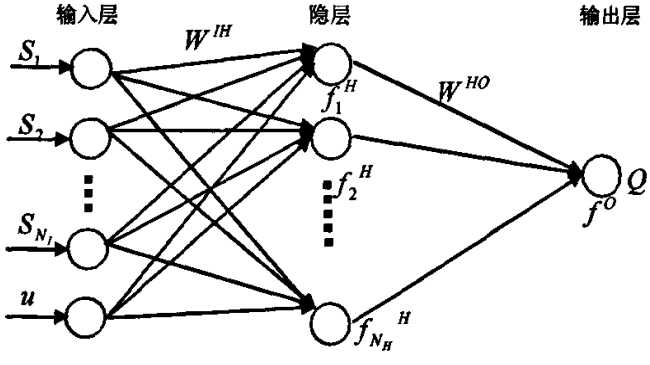

图 4.5 前馈神经网络结构

Fig 4.5 Feed-forward neural network structure

在图 4.5 中，网络输入分为两类，一类是电梯状态 **S**，**S** 为 N_I 维向量；一类是派梯

方案u。N_H为隐层神经元个数，输出层为单神经元。W^{IH}为输入层到隐层的权值矩阵，w_{ji}^{IH}为输入层第i个神经元到隐层第j个神经元的权值；$\mathbf{W}^{HO}$为隐层到输出层的权值矩阵，w_j^{HO}为隐层到输出层的权值。$f_j^H, j=1,2\cdots N_H$为隐层第j个神经元的激活函数，f^O为输出层神经元的激活函数。

设网络输入向量为$X=(s^T \quad u)^T=\{x_i \mid i=1,2,...,N_I+1\}$，隐层神经元输出向量为$P_H=\{p_j \mid j=1,2,....,N_H\}$。

则在t时刻，各变量的实例为

$X_t=\{x_{it}\}$，$P_{Ht}=\{p_{jt}\}$，$W^{IH}{}_t=\{w_{ji}^{IH}{}_t\}$，$\boldsymbol{\theta}^H{}_t=\{\theta_j^H{}_t\}$，$W^{HO}{}_t=\{w_j^{HO}{}_t\}$，$\theta^O{}_t$

网络的前向计算过程中，

$$v_{jt}=\sum_{i=1}^{N_I+2} w_{ji}^{IH}{}_t x_{it} \tag{4.17}$$

$$p_{jt}=f_j^H\left(\sum_{i=1}^{N_I+2} w_{ji}^{IH}{}_t x_{it}-\theta_j^H{}_t\right) \tag{4.18}$$

$$Q_t=f^O\left(\sum_{j=1}^{N_H+1} w_j^{HO}{}_t p_{jt}-\theta^O{}_t\right) \tag{4.19}$$

在t时刻使用$<(\mathbf{s}_{t-1},u_{t-1}),Q_t>$作为网络的训练样本，网络权值更新算法遵循 Delta 学习规则，以最小化输出误差目标。

定义t时刻的输出误差为

$$E_t=\frac{1}{2}e_t^2=\frac{1}{2}(Q_t-Q_{t-1})^2=\frac{1}{2}\Delta Q_t^2 \tag{4.20}$$

使用梯度下降法（gradient descent）更新权值，权值更新与输出误差的负梯度成正比，输出层与输入层权值，隐层和输出层的阈值更新方程分别为：

$$\Delta w_j^{HO}{}_t=-\eta\frac{\partial E}{\partial e}\frac{\partial e}{\partial Q}\frac{\partial Q}{\partial w_j^{HO}}\Big|_t=\eta e_t f^{O\prime}|_t\ p_{jt} \tag{4.21}$$

$$\Delta w_{ji}^{IH}{}_t=-\eta\frac{\partial E}{\partial w_{ji}^{IH}}\Big|_t=-\frac{\partial E}{\partial e}\frac{\partial e}{\partial p_j}\frac{\partial p_j}{\partial w_{ji}^{IH}}\Big|_t=\eta e_t f^{O\prime}|_t\ w_j^{HO}{}_t f_j^{H\prime}|_t\ x_{it} \tag{4.22}$$

$$\Delta\theta_j^H{}_t=\eta e_t f_j^{H\prime}|_t\ w_j^{HO}{}_t \tag{4.23}$$

$$\Delta\theta^O{}_t=\eta e_t f^{O\prime}|_t \tag{4.24}$$

其中，$\eta>0$为常数。

隐层神经元的激活函数为 Sigmoid 函数，输出层神经元的激活函数为线性函数。则

$$f_j^H = \frac{1}{1+e^{-av_j}} \tag{4.25}$$

$$f_j^H{}'|_t = ap_{jt}(1-p_{jt}) \tag{4.26}$$

$$f^O{}'|_t = 1 \tag{4.27}$$

式（4.25）中，参数$a>0$，控制着隐层激活函数的倾斜程度和进入输出饱和的快慢。

将式（4.26）和式（4.27）代入式（4.21）、（4.22）、（4.23）和（4.24），则权值和阈值更新方程为

$$\Delta w_j^{HO}{}_t = \eta \Delta Q_t p_{jt} \tag{4.28}$$

$$\Delta w_{ji}^{IH}{}_t = a\eta \Delta Q_t w_j^{HO}{}_t p_{jt}(1-p_{jt})x_{it} \tag{4.29}$$

$$\Delta \theta_j^H{}_t = \eta \Delta Q_t w_j^{HO}{}_t p_{jt}(1-p_{jt}) \tag{4.30}$$

$$\Delta \theta^O{}_t = \eta \Delta Q_t \tag{4.31}$$

4.3.4 电梯调度系统的调度算法

电梯调度系统是一个非线性、随机离散动态系统，存在大量不确定性，难以建立精确的数学模型。本文将基于人工免疫交通模式识别的群控系统控制器引入学习过程中，使系统在不同交通模式下选择相应的神经网络参数，从而获得更合理的派梯结果，其中派梯控制部分用强化学习方式实现。

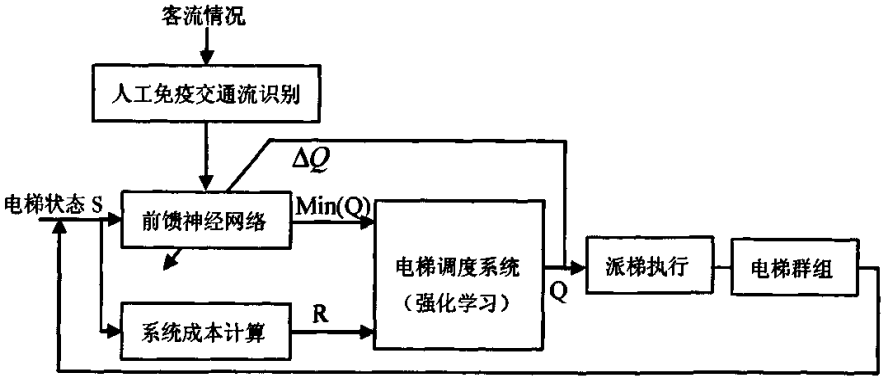

图 4.6　电梯群组调度系统的结构

Fig 4.6 Structure of elevator group scheduling system

电梯调度系统的结构如图 4.6 所示。电梯调度系统由人工免疫交通模式识别模块和

强化学习两部分组成。第二章对电梯交通流的聚类分析共分为七种交通模式，其中上行高峰、下行高峰模式、午饭后上行高峰模式、午饭前下行高峰模式和随机层间模式是调度中比较复杂的问题，是本文重点解决的问题，在这五种模式下的试验结果也可以作为评价系统性能的依据。不同时间段交通流的情况有所不同，为了使派梯策略能够适应当前的交通状况，令系统每隔一段时间（一般为 5 分钟）对交通流进行一次辨识，在各种交通模式下，前馈神经网络的参数会有相应的变化。这种基于交通模式识别的电梯调度可以提高电梯系统的运行效率，降低能耗。确定前馈神经网络的参数后，经过泛化，得到系统的状态-行为值函数的最小值 $\min(Q)$，再根据当前电梯群组的状态可以计算出系统的成本 R，由这两个参数确定当前状态-行为值函数 Q，得到派梯方案 u，更新电梯群组的状态。而派梯前后的 Q 值差作为神经网络权值和阈值的修正依据，使神经网络逐渐逼近最优派梯。

概括来说，使用 Q-learning 值函数更新算法、基于 Gibbs 分布的随机行为选择策略和三层前馈神经网络的泛化技术联合解决电梯群组调度策略。

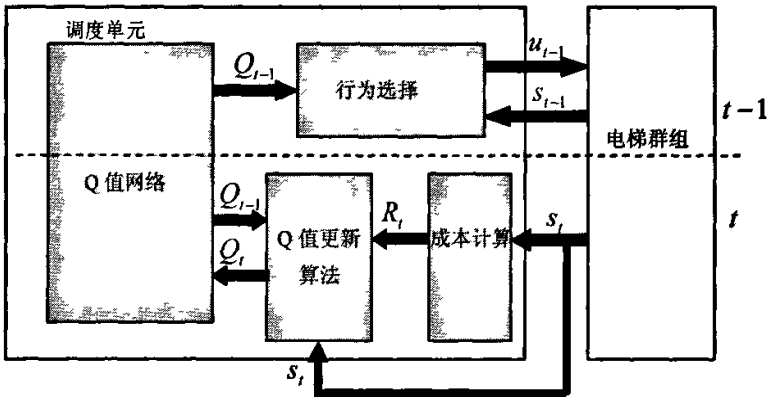

图 4.7 电梯群组调度的算法过程

Fig 4.7 Algorithm process of Elevator group scheduling system

图 4.7 中展示了群组调度系统由 $t-1$ 时刻过渡到 t 时刻所经历的计算过程。可以看出，调度单元内部成四部分：行为选择模块、成本计算模块、Q 值更新算法模块和 Q 值网络。其中前三者是计算过程的功能模块，Q 值网络作为电梯群组调度问题值函数的存储空间。Q-learning 的值迭代算法由 Q 值更新算法模块实现，行为选择模块实现随机行为选择策略，而 Q 值函数的泛化在 Q 值更新算法模块与 Q 值网络的交互中完成。

在结合了随机行为选择策略和神经网络泛化方法后，调度算法的简要计算过程为：

初始化神经网络所有的权值，初始化当前状态$\mathbf{s}_t$；

Repeat (每一步)

（1）时刻 t 观测状态$\mathbf{s}_t$；

（2）$\mathbf{s}_t$代入神经网络，计算每个行为 u 的$Q(\mathbf{s}_t,u)$，求出$\min_{u\in U_s} Q(\mathbf{s}_t,u)$；

（3）在状态$\mathbf{s}_t$下，使用由$Q_t(\mathbf{s}_t,u)$得来的 Gibbs 分布，以概率选择派梯方案u_t；

（4）$Q^{output} \leftarrow Q_t(\mathbf{s}_t,u_t)$，并执行派梯方案$u_t$；

（5）观察经过派梯后系统的新状态$\mathbf{s}_{t+1}$，并得到即时报酬R_{t+1}；

（6）在新状态$\mathbf{s}_{t+1}$下，对所有派梯方案用神经网络计算$Q_t(\mathbf{s}_{t+1},u)$，求出$\min_{u\in U_s} Q(\mathbf{s}_{t+1},u)$；

（7）利用$\Delta Q_t(\mathbf{s}_t,u_t)=\alpha[R_{t+1}+\gamma\min_{u\in U_s} Q_t(\mathbf{s}_{t+1},u)-Q_t(\mathbf{s}_t,u_t)]$计算出$\hat{Q}_t(\mathbf{s}_t,u_t)$，令$Q^{tagart} \leftarrow \hat{Q}_t(\mathbf{s}_t,u_t)$；

（8）根据Q^{output}、Q^{target}、$\mathbf{s}_t$和u_t，使用 Delta 规则更新神经网络权值；

（9）$\mathbf{s}_t \leftarrow \mathbf{s}_{t+1}$；

Until 仿真时间结束。

4.4 小结

本文将调度问题置于 MDP 模型下，定义系统状态、行为、成本和值函数等基本要素，构成问题模型。研究基于强化学习的调度问题求解算法。强化学习用于实际问题求解时，必然面对值函数更新算法问题、利用与探索问题和值函数的泛化问题。本文运用 Q-learning 作为强化学习模型中值函数的基本迭代算法，异步地逼近最优的状态-行为值函数；为保证在学习过程中既能保证利用已有的优化策略，又能保证探索未知的策略来寻求长期的性能改善，运用 Gibbs 分布构造随机性的行为选择策略；使用前馈型神经网络作为值函数的泛化方法，将值函数的更新过程转变为网络权值的训练过程，在观测状态和派梯决策的同时，使用更新的状态-行为值训练神经网络，将值函数的逼近过程转变为神将网络权值的修改过程。将 Q-learning 算法、随机行为选择策略和神经网络的泛化技术综合起来，构成解决电梯群组调度的强化学习算法。

第五章 仿真实验与结果分析

在电梯群组调度算法的设计过程中需要测试群控算法的性能，从安全可靠性和运行成本考虑，采用计算机建立仿真模型来验证算法的可行性。前面几章研究了电梯交通流的聚类分析和基于强化学习的电梯群组调度问题，并提出了综合考虑平均等候时间、平均乘梯时间和电梯停靠次数的综合成本，利用强化学习算法作为在与环境的交互过程中学习最优策略的方法。在这一章主要设计电梯虚拟仿真环境，构建基于强化学习算法的电梯调度组件单元的软件结构，定义强化学习算法框架中各部分功能函数的接口，实现调度单元并执行仿真，与其它调度方法进行比较，并对仿真结果进行分析，验证了算法的可行性和有效性。

5.1 电梯群组调度的仿真环境

电梯群控调度算法的实现与仿真包括电梯群控虚拟仿真环境和调度算法两大部分，都是借助 MATLAB 的运算能力实现并运行的。

5.1.1 电梯群组调度的虚拟仿真环境

电梯群控虚拟仿真环境是设计、测试、评价电梯群组调度方法、模拟电梯交通流、仿真电梯运行过程的重要工具。虚拟仿真环境的基本构造如图 5.1 所示。

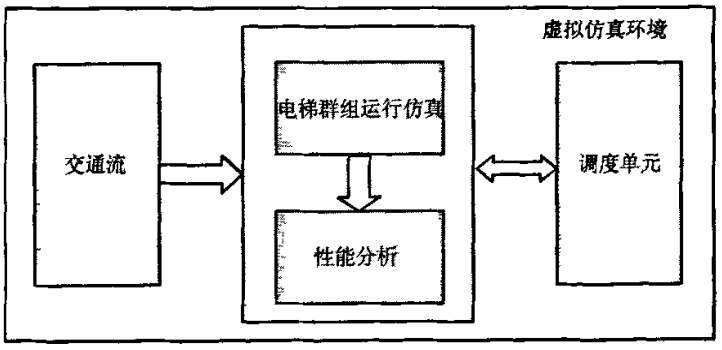

图 5.1 电梯群控虚拟仿真环境结构图

Fig 5.1The structure of virtual environment of elevator group control system

由图 5.1 看出，电梯群控虚拟仿真环境将交通流发生器和调度单元与群组运行仿真和性能分析组件分离。这种结构的开放性，使得调度单元作为分立的模块单独开发成为可能。基于强化学习的电梯群组调度方法就是以调度单元的形式实现，通过与虚拟环境的交互完成仿真，同时依靠虚拟环境的性能评价功能统计必要数据。

5.1.2 基于强化学习的电梯调度算法的运行环境

基于强化学习的电梯群组调度的运行需要观测群组系统状态并实时更新值函数，需要大量的矩阵运算。考虑到 MATLAB 强大的矩阵运算能力，本文将值函数更新中涉及的大量矩阵运算功能交与 MATLAB 运行。调度算法单元的运行环境如图 5.2 所示。

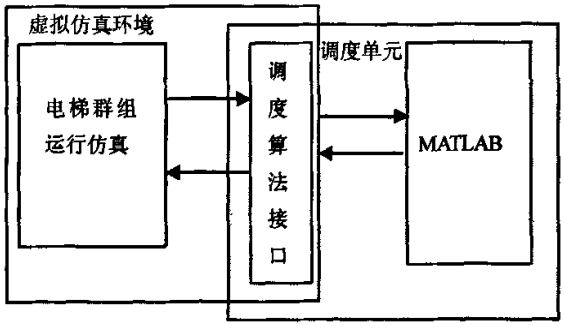

图 5.2 基于强化学习电梯调度算法的运行环境

Fig 5.2 Running environment of the elevator group scheduling algorithm based on reinforcement learning

5.2 软件设计与实现

5.2.1 电梯调度算法结构

算法软件不太复杂，因而运用传统的模块化设计与开发的思想进行强化学习调度算法的设计与开发工作。

算法的功能结构共分为五大部分：调度算法接口模块、状态观测模块、成本计算模块、Q 函数值更新模块和行为选择模块。其间关系如图 5.3 所示。

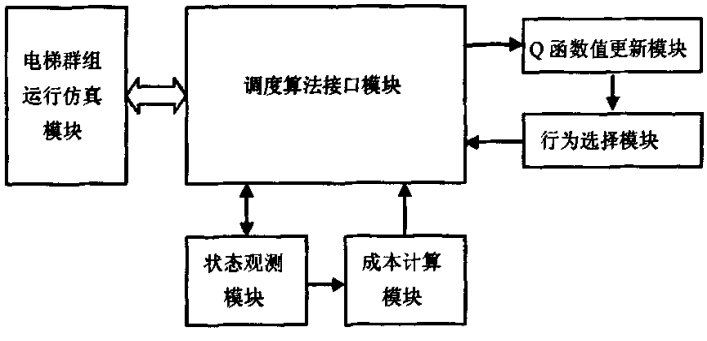

图 5.3 电梯群组调度算法软件结构

Fig 5.3 The structure of the elevator group scheduling algorithm

调度算法接口模块负责与虚拟环境中的电梯群组运行仿真模块通信，接受电梯状态和乘客信息，送出派梯结果。状态观测模块将来自虚拟环境的电梯状态和乘客信息的结构重组为强化学习模型的一般形式，方便模块的重构。成本计算模块根据当前状态计算

对前一时刻状态和所作决策的成本评价，提供给值迭代算法。Q 函数值更新模块根据历史经历、当前状态和成本评价，更新神经网络权值，而行为选择模块根据当前更新后的 Q 值函数，利用 Gibbs 分布依概率选择派梯行为，并将结果通过接口模块回送给虚拟环境。

5.2.2 电梯群组调度系统的程序流程

根据 4.3.4 节中的电梯群组调度算法描述和图 5.3，可以获得基于强化学习算法的电梯群组动态调度算法的运行步骤。对于电梯群组动态调度的具体领域背景以及算法中要注意的细节，已经有详细的阐述。基于强化学习的电梯群组调度算法的总体流程如图 5.4 所示。

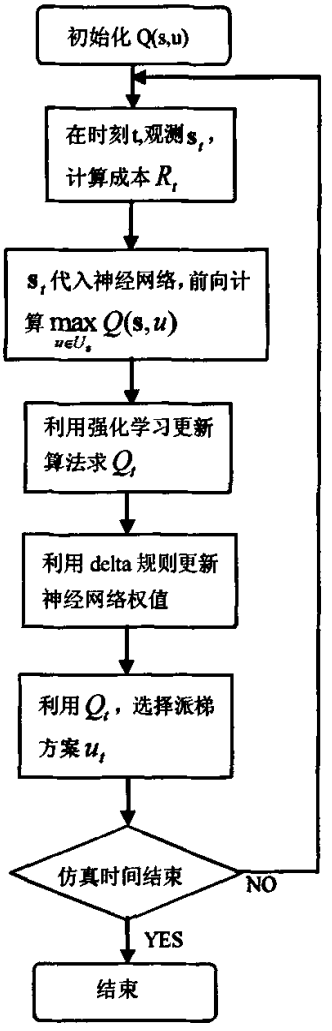

图 5.4 基于强化学习的电梯群组调度算法的总体流程

Fig 5.4 The whole process of elevator group scheduling algorithm based on reinforcement learning

在执行总体流程时，各关键模块发挥各自的作用，实现电梯群组调度算法，其程序流程图如图 5.5 到图 5.8 所示。

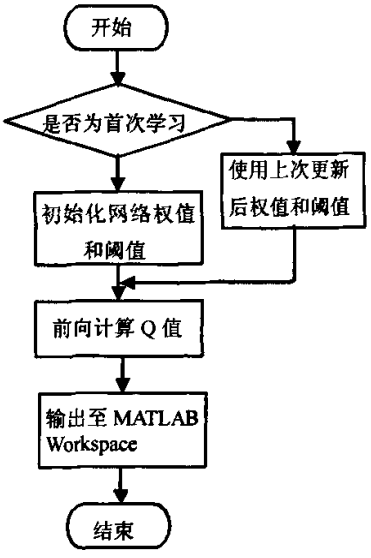

图 5.5 初始化过程

Fig 5.5 Initialization

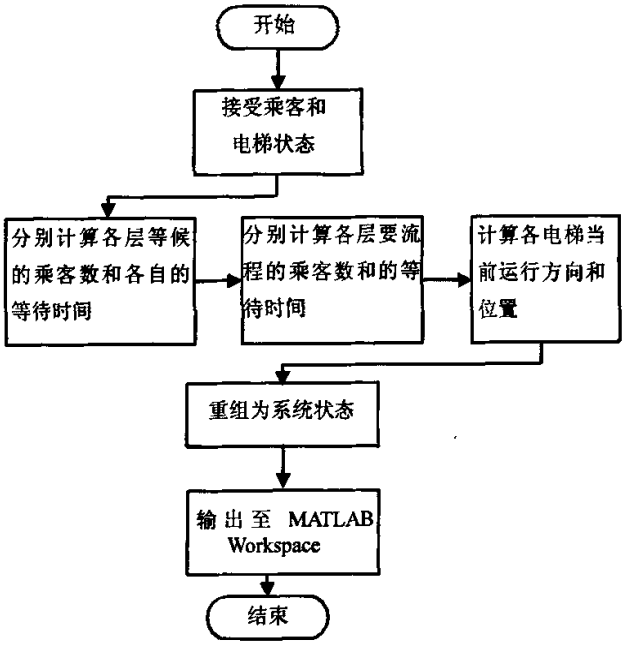

图 5.6 状态观测过程

Fig 5.6 Status observation

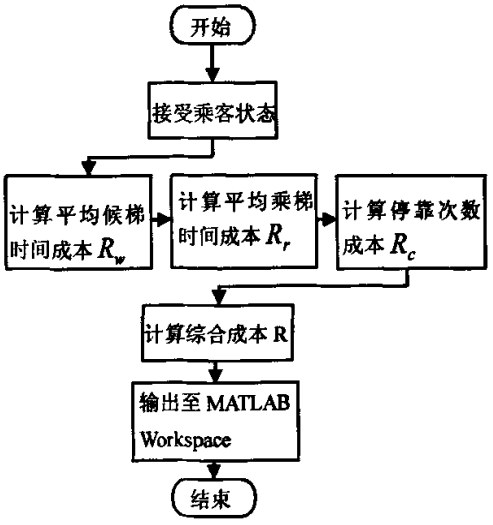

图 5.7 成本计算过程

Fig 5.7 The calculation of the rewards

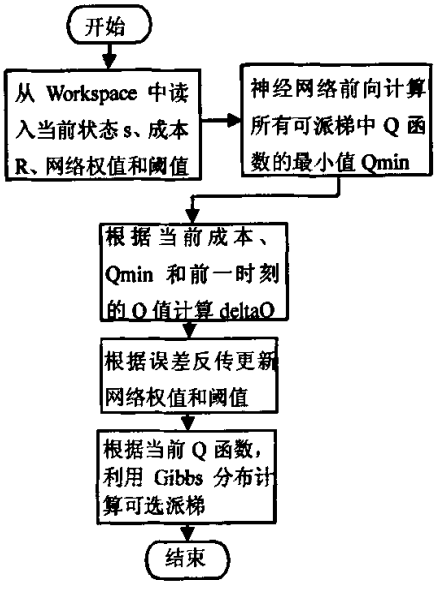

图 5.8 Q 函数更新和派梯决策过程

Fig 5.8 The renewing of the Q function and dispatching decision process

5.3 电梯动态调度仿真实验与结果分析

5.3.1 电梯群组仿真环境设定

5.3.1.1 电梯调度系统参数设定

调度算法的仿真平台是 4 部电梯 20 层大楼。电梯参数和大楼参数如表 5.1 和表 5.2

所示。

表 5.1 电梯参数

Table 5.1 The parameters of the elevator group

电梯参数	具体数值
额定速度(m/s)	1.75
加速度(m/s^2)	1
加加速度(m/s^3)	1
额定容量(人)	15
开关门时间(s)	1

表 5.2 大楼参数

Table 5.2 The parameters of the building

大楼参数	具体数值
楼层数	20
楼层高度(m)	4

5.3.1.2 强化学习算法参数的设定

基于强化学习的电梯调度算法中的参数包括：折扣因子γ、强化学习率α、Delta学习规则中的学习常数η和网络隐层神经元的激活函数a和Gibbs分布中的温度参数T。

强化学习的学习率α随时间逐渐衰减，衰减律为

$$\alpha_t = b\alpha_{t-1} = b^t\alpha_0 \tag{5.1}$$

其中，α_0为初始学习率，b为衰减因子，控制学习率的衰减速度，仿真时设$b=0.999$，α_0=0.99。

Gibbs分布中的参数T也是随仿真时间递减的，递减规律为

$$T_k = d^k T_0 \tag{5.2}$$

其中，T_0是参数T的初值，k是仿真消逝的时间，在仿真试验中，每500秒(计算机时间)k做加一运算，仿真时设$T=1000, d=0.98$。

其他参数均为常数。设定式(4.15)中折扣因子γ=0.8，式(4.25)中$a=0.3$，式(4.28)中η=1。

5.3.2 电梯调度系统仿真运行与分析

5.3.2.1 Q函数的学习过程

强化学习方法是一种从经验中通过学习值函数来寻找优化策略的方法。因此，产生交通流，由虚拟环境调用群组调度算法产生派梯的仿真过程就是强化学习方法的学习过程。

仿真时利用电梯虚拟仿真环境的交通流发生器产生上高峰的交通流，用于算法训练。上高峰交通流曲线图如图 5.9 所示。

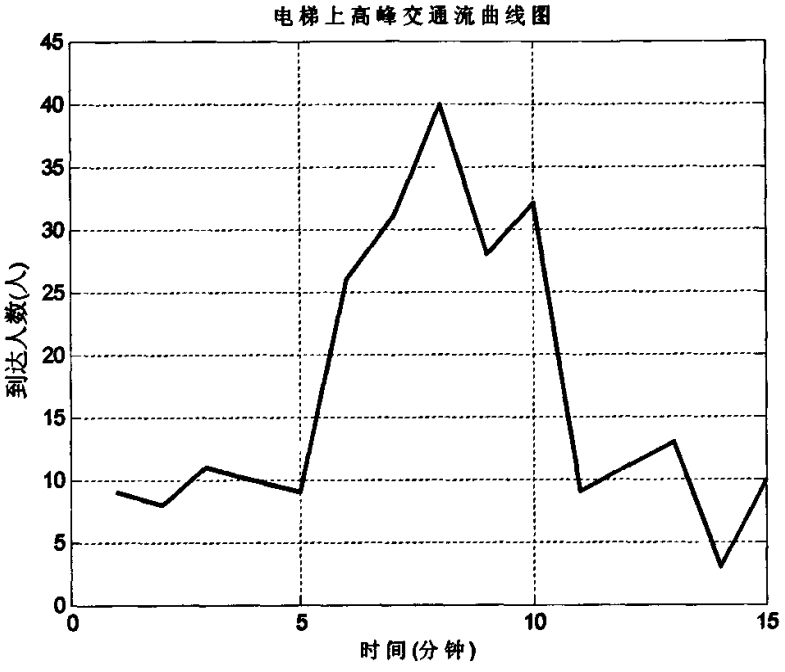

图 5.9 训练交通流曲线图

Fig 5.9 The graph of elevator traffic flow for training

交通流延续时间为 15 分钟，共有 413 人到达门厅请求上行。使用该交通流曲线重复执行，训练强化学习调度算法，共反复执行 16 次。之所以反复使用统计意义上相同的交通流，是为了研究强化学习的算法过程对人为定义的 Q 值函数的逼近作用。

在每次算法执行的过程中，记录实际的 Q 值曲线，对前后两次的 Q 值曲线作差，得到误差曲线。为较明显的说明问题，从结果中抽取了第 2 次仿真与第 1 次仿真、第 6 次与第 5 次、第 12 次与第 11 次以及第 16 次与第 15 次仿真之间 Q 值的误差曲线。误差曲线如图 5.10 所示。

由图 5.10 可以看出，在重复的仿真训练过程中，前后两次 Q 值曲线间的误差是逐渐减小的，说明了论文使用的强化学习方法对 Q 值函数的逐渐逼近作用。

仅仅使用一种交通流对算法训练是不全面的，也是不符合实际的，这是由于在实际电梯群组运行过程中，交通模式是变化的，交通流的强度也在不断的改变。一天中的不同时段和不同天的相同时段，电梯群组系统所面对的交通状态均是不同的。因此，论文还生成其他交通流用于算法的训练，这样，保留不同交通流训练后的神经网络权值和阈值，当进行在线派梯时，根据第二章中采用的人工免疫聚类方法，辨识当前交通流模式，

根据不同的模式采用不同的网络权值和阈值进行派梯。

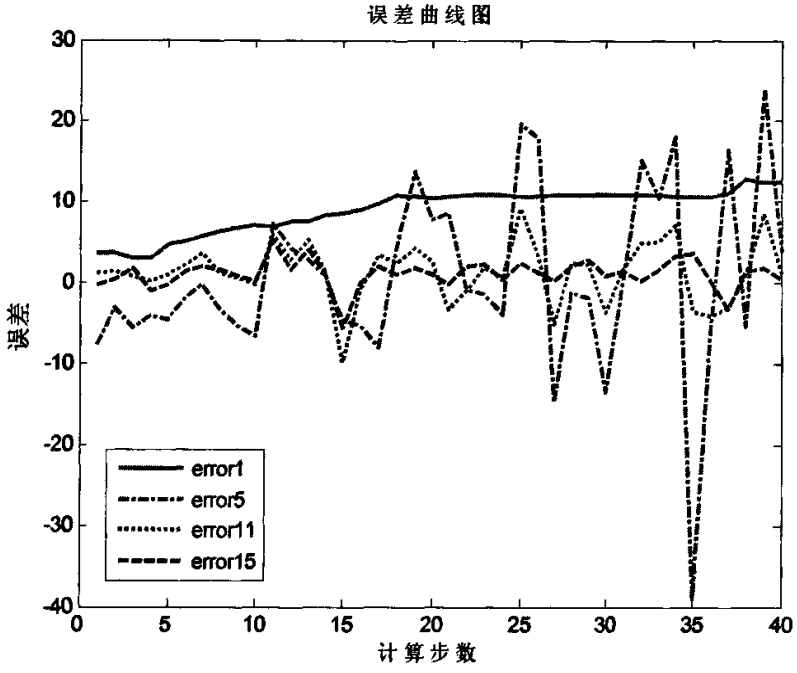

图 5.10 Q 函数误差曲线图

Fig 5.10 Error graph of the Q function

在强化学习算法训练过程中所使用的其他交通流包括：

交通流 1：上高峰交通流模式，在 15 分钟内达到 413 人，交通流统计如表 5.3。

交通流 2：午饭上高峰交通流模式，在 15 分钟内达到 215 人，交通流统计如表 5.4。

交通流 3：随机层间交通模式，15 分钟内到达 220 人，交通流统计如表 5.5。

交通流 4：午饭下高峰交通流模式，在 15 分钟内到达 215 人，交通流统计如表 5.6。

交通流 5：下高峰交通流模式，在 15 分钟内到达 355 人，交通流统计如表 5.7。

表 5.3 交通流 1 的乘客到达统计表

Table 5.3 Statistics of passengers arriving in the traffic flow 1

时间(分)	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15
上行人数	37	41	32	36	32	36	40	28	40	35	31	26	24	28	35
下行人数	5	1	2	4	1	4	3	1	2	4	1	0	4	3	1
层间人数	4	5	4	1	2	1	5	1	4	2	2	4	1	3	1

表 5.4 交通流 2 的乘客到达统计表

Table 5.4 Statistics of passengers arriving in the traffic flow 2

时间(分)	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15
上行人数	18	12	14	15	10	16	18	21	24	18	10	16	14	17	19
下行人数	1	2	5	1	3	2	3	1	2	1	1	1	0	1	1
层间人数	2	1	3	4	1	4	1	5	2	4	1	1	5	3	2

表 5.5 交通流 3 的乘客到达统计表

Table 5.5 Statistics of passengers arriving in the traffic flow 3

时间(分)	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15
上行人数	5	6	2	1	8	0	2	1	4	3	8	1	3	4	2
下行人数	4	2	1	4	1	7	1	3	5	1	0	4	7	2	1
层间人数	12	18	15	12	10	15	13	15	25	24	26	20	10	11	9

表 5.6 交通流 4 的乘客到达统计表

Table 5.6 Statistics of passengers arriving in the traffic flow 4

时间(分)	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15
上行人数	1	5	3	2	0	1	2	4	1	2	1	5	3	2	0
下行人数	14	16	13	10	27	26	14	25	11	14	16	24	18	19	15
层间人数	0	2	4	3	1	5	3	1	0	5	1	0	7	1	2

表 5.7 交通流 5 的乘客到达统计表

Table 5.7 Statistics of passengers arriving in the traffic flow 5

时间(分)	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15
上行人数	1	0	1	2	3	4	1	2	5	1	3	1	2	3	1
下行人数	25	16	24	34	35	29	34	36	37	41	43	34	36	41	45
层间人数	2	1	4	1	3	4	1	2	4	2	1	1	1	1	4

在仿真中，选择不同的交通流用于调度，基于强化学习方法的调度算法的学习过程是与调度的运行同步的，每一种交通流模式的仿真训练都是对同一个初始神经网络进行调整，但在不断的学习中，不同的神经网络权值和阈值将适应不同的交通流状态。并在不断地学习中，由于强化学习算法的学习率在学习过程中是不断衰减的，这使得学习强度逐渐减缓，并且随机行为选择策略中参数T的逐渐降低，调度单元逐渐地不再探索，而是利用已经学到的优化派梯策略。

5.3.2.2 不同调度方法的比较

本文提出的基于强化学习的电梯调度方法是建立在对电梯交通流人工免疫聚类分析基础之上的，并综合考虑了平均候梯时间、平均乘梯时间和电梯停靠次数，计算出综合成本作为电梯动态调度的重要依据。在进行在线派梯之前，对聚类分析出的各种交通流模式进行训练，得到不同交通流模式下对应的电梯调度系统的神经网络参数，以便进行模式识别后采用不同的网络参数。为了验证系统性能，在五种交通流条件下，将基于强化学习的调度方法与其他方法在对应相同的交通流条件下的调度结果加以比较。比较结果如表 5.8 所示。参与比较的另外两种算法为：一种是基于遗传算法的电梯群控方法[53]，使用 GA 表示。另一种是传统的静态分区方法，静态分区将 20 层大楼平均分为 4 个区

域，一部电梯负责一个区域，静态分区法使用 SZ 表示。基于强化学习的电梯调度方法使用 RL 表示。

表 5.8 三种调度方法的性能比较

Table 5.8 The comparison of three kinds of elevator group scheduling methods

交通流模式	方法	平均候梯时间	最长候梯时间	平均乘梯时间	最长乘梯时间	平均停靠次数	综合成本
上行高峰模式	RL	52.326	150	35.265	61	6.5	63.544
	GA	75.165	242	53.591	106	8.9	-
	SZ	31.784	125	38.267	86	6.8	-
午饭后上行模式	RL	32.146	65	26.254	56	5.3	35.979
	GA	64.815	221	52.851	95	7.2	-
	SZ	25.562	63	39.950	92	5.9	-
随机层间模式	RL	12.356	49	15.265	35	5	20.266
	GA	17.354	50	17.533	43	10.6	-
	SZ	20.204	59	21.814	48	5.9	-
午饭前下行模式	RL	45.263	157	22.418	61	5.8	50.510
	GA	28.662	135	21.091	48	6.9	-
	SZ	20.232	75	20.972	45	5.7	-
下行高峰模式	RL	60.254	195	58.458	162	6.9	84.234
	GA	85.124	252	65.872	265	5.6	-
	SZ	98.716	259	68.456	145	6.5	-

从表中可以看出，在不同的交通流条件下，不同调度算法的性能表现是不同的。在强度较大的上高峰期间，静态分区展现出更好的调度效果，这是由于该方法通过分区将高到达率的人群分流，自然形成多个候梯队列。同时这也是对乘客目的楼层信息的间接获得方法，因为只有那些目的楼层在电梯服务区内的乘客才会走向这部电梯，对乘客目的楼层信息的获知必然减少乘客的候梯时间。但是在随机层间模式下和下行高峰模式下，由于乘客的分布楼层不均，静态分区方法使得电梯的利用不均衡，进而乘客的平均候梯时间和平均乘梯时间都较长。

基于遗传算法的电梯调度方法在强度不大的随机层间模式下表现是较为优秀的，乘客的平均候梯时间较短，但这是以平均乘梯时间和电梯停靠次数为代价的。也就是说，用装载更多的乘客来换取乘客平均候梯时间的缩短。而且，在强度很大的上行高峰期间，基于遗传算法的调度方法出现了极强的不适用性，极高的到达率使得遗传算法的候梯时间过长，而且在其他性能指标方面也不乐观。

基于强化学习的电梯调度算法在五种交通流下的表现比较平均，在上行高峰模式下，优于基于遗传算法的派梯方法，能够应付强度较大的上行高峰交通，并且停靠次数

不多。上行高峰模式和午饭后上行高峰模式，同属于上行模式，但高峰强度不同，两者的平均候梯时间指标变化较大，其他指标变化不大。在随机层间模式下，方法的性能较优，能够具有很好的平均候梯时间，同时在平均乘梯时间和电梯停靠次数方面优于另外两种调度方法。但在下行高峰模式和午饭前下行高峰模式下，该方法将电梯派向高层，使得底层乘客的平均候梯时间加长，系统性能变差。

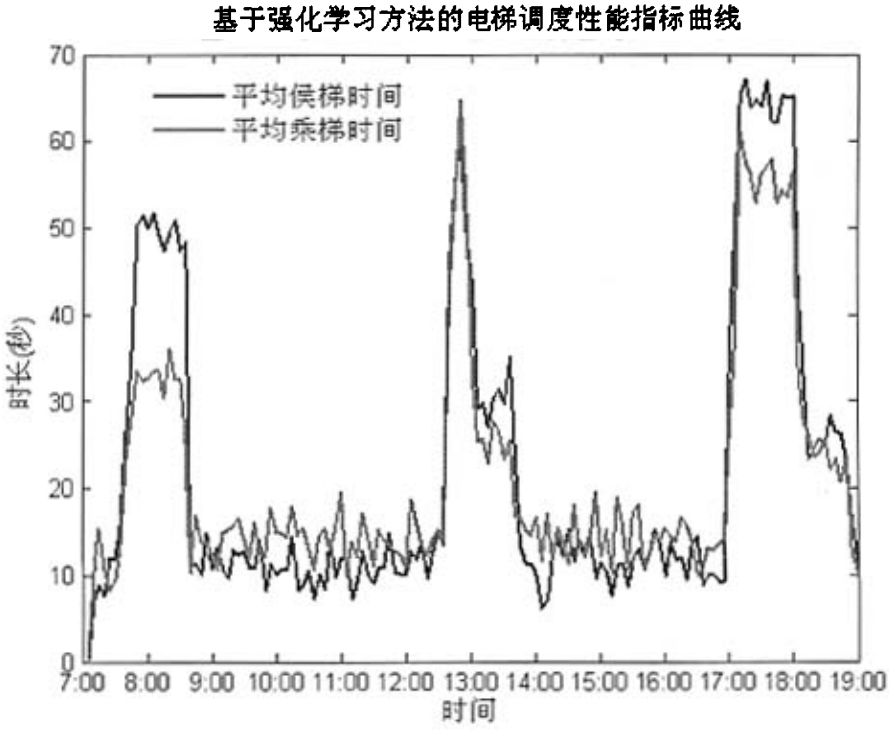

图 5.11 基于强化学习方法的电梯调度性能指标图

Fig 5.11 Performance of the elevator group scheduling system based on reinforcement learning algorithm

经过这五种交通流的训练，将训练后的网络参数作为在线派梯的参数，利用交通流发生器随机生成一天的大楼交通流，每隔 5 分钟检测当前的交通流特征数据，计算与各聚类中心的欧式距离大小，确定当前的交通流模式，再选择不同的参数进行派梯，并不断进行在线学习，完成电梯群组的动态调度。图 5.11 描述了基于强化学习的电梯调度系统的性能指标。

由表 5.8 和图 5.11 可以看出，不同的交通流条件下，调度算法的性能表现是不同的。在强度较大的上高峰期间和随机层间模式下，基于强化学习方法的动态调度算法表现出了良好的调度效果，而在下高峰交通模式下，方法将电梯派向高层，使得底层乘客的候梯时间加长。不过总体看来，基于强化学习的电梯调度算法对不同交通流有普遍的适应

性而且表现较为平均，在随机层间模式和上高峰交通模式下性能较优，当交通流强度下降时性能会有更大的改善。而且算法自身的学习能力在不断变化的环境中始终存在，这使得调度算法能够更好的适应不稳定的交通状况。

5.4 小结

基于强化学习的电梯群组调度算法是在电梯群控虚拟仿真环境的基础上实现的。本章阐述了算法软件的功能结构，将软件在功能上分为调度算法接口模块、状态观测模块、成本计算模块、Q 函数值更新模块和行为选择模块；通过物理配置图说明算法软件的模块分布与交互机制。对算法软件及其关键模块的过程使用了图形化说明。

仿真实验中，在设定的电梯和大楼参数的基础上，将算法模块用于电梯群组的仿真调度中。在一个上高峰交通流条件下，对算法更新的 Q 值函数做了跟踪记录，对相邻两次的 Q 值曲线做求差处理，通过对前后 4 次误差曲线的比较，说明了强化学习方法对 Q 值函数的逼近过程。

本文共使用 5 种不同的交通流作为算法的执行条件，将基于强化学习的调度算法同基于遗传算法的调度方法和静态分区方法进行比较，在上高峰模式和随机层间模式下呈现出一定的优越性，充分说明了基于强化学习的电梯群组动态调度算法的有效性和可行性。

第六章 结 论

本文运用人工智能领域的新理论和新方法，对电梯交通客流和电梯群组调度算法进行了分析和研究。对电梯交通流进行人工免疫聚类分析，根据分析结果对电梯交通流进行在线模式识别，综合考虑平均候梯时间、平均乘梯时间和电梯停靠次数的前提下，提出了一种基于强化学习的电梯群组调度方法，作为在与环境的交互过程中学习最优策略的方法。

本文首先将聚类思想引入到人工免疫系统理论及其抗体的免疫激励、免疫抑制和免疫记忆机制中，提出了一种新型的人工免疫聚类算法，对电梯交通的客流数据实施特征提取，获取电梯交通流人工免疫记忆数据集，然后采用人工免疫聚类方法对电梯交通流进行分类和模式识别。聚类结果清晰地体现了电梯客流的特征，符合交通流的实际。

本文的重点是研究强化学习方法在电梯群组调度中的应用，强化学习是在随机序贯决策问题中学习最优控制策略的理论和技术的整体，是在与环境的交互中学习的计算性方法。本文通过运用 MDP 框架模型化电梯调度问题，定义了模型中的各个要素，并且综合考虑平均等待时间、平均乘梯时间和电梯停靠次数这三个因素，计算出的综合成本作为性能评价的综合指标。面对在应用中存在的利用与探索问题和值函数的泛化问题，使用随机行为选择策略和前馈神经网络分别作为解决的方案，并与 Q-learning 的值迭代算法结合，共同构成电梯群组调度算法。仿真结果证明，在对交通流进行反复训练时，Q 值曲线间的误差是逐渐减小的，说明了本文使用的强化学习方法对 Q 值函数的逐渐逼近作用。最后，在交通流模式识别的基础上，设计了基于强化学习算法的电梯动态调度系统。使用 5 种不同交通流作为算法的执行条件，将基于强化学习的调度算法与其它方法作比较，呈现一定的优越性，尤其在上高峰模式和随机层间模式下有更好的适用性，充分说明了基于强化学习的电梯群组动态调度算法的有效性和可行性。

但是，由于作者水平有限和时间紧张，论文研究是存在不足与遗憾的：

（1）在强化学习应用于电梯群组调度研究中，存在一些假设条件和理想状态，还需研究在条件放宽时的调度方法。

（2）强化学习方法本身的快速增长使论文中的研究难免存在片面性。

（3）论文中强化学习使用 MDP 作为框架，而没有纳入电梯群组运行的环境模型，这使得所建立的问题模型和算法本身还有很大的提升空间。

存在的不足意味着未来还有工作要做。未来的工作可以分为两方面，一方面是继续深入分析电梯交通流，另一方面是研究强化学习算法本身。强化学习是多学科交叉的研究领域，目前的难点主要集中在分布式条件下，状态部分可观的环境中，智能体优化的行为（控制、调度）策略的学习问题。这种环境的特点都是电梯群组系统具备的，因此这种条件下的强化学习算法都是可以在电梯群组调度问题中检验的。

参考文献

1. Y. Sakai, K. Kurosawa. Development of elevator supervisory group control system with artificial intelligence [J], Hitachi Review, 1984, 33(1):25-30
2. 宗群，尚晓光. 电梯群控系统虚拟仿真环境设计[J]，制造业自动化，1999，10（4）：24-26
3. 黄静，郭勇. 对目前国内外电梯控制技术的研究[J]，电脑与信息技术，2000，5（17）：64-66
4. Kenji Sasaki. Elevator group supervisory control system using neural networks [J], Elevator World, 1996, 44(2): 81-86
5. 宗群，尚晓光，岳有军等. 电梯群控系统的交通模式识别[J]，控制与决策，2001，9（8）：163-166
6. 黄敏，崔宝同，顾树生. 基于小波神经网络的电梯交通流预测[J]，控制与决策，2006，5（21）：589-592
7. K. Igarashi, S. Take, T. Ishikawa. Supervisory control for elevator group with fuzzy expert system [C], IEEE International Conference of Industrial Technology, 1994, (12), 133-137
8. Ming Ho, Brent Robertson. Elevator group supervisory control using fuzzy logic [C], Canadian Conference on Electrical and Computer Engineer, 1994, (2): 825-828
9. Chang Bum Kim, Kyoung A. Seong, Hyung Lee-Kwang, Je. A fuzzy approach to elevator group control system [J], IEEE Trans. on System, Man and Cybernetics, 1995,25 (6):985-990
10. Chin-teng Lin, George Lee. Neural-network-based fuzzy Logic Control and decision System [J], IEEE Trans on Computers, 1991, 40(12): 1320-1336
11. Markon S.H, Kita, Y. Nishkawa. Adaptive optimum elevator group control by use of neural network [J], Transaction of the Institute of System, Control and Information Engineers, 1994, 7(12): 487-497
12. 万健如，刘春江，刘洪池. 基于前向神经网络最佳派梯智能控制[J]，系统工程与电子技术，2003，4（25）：466-468
13. 顾晨，赵国军，刘铮. 基于人工神经网络的自适应模糊电梯群控系统[J]，计算机测量与控制，2003，11（12）：947-949
14. Atsuya Fujino, Toshimitsu Tobita et al. An elevator group control system with floor

attribute control method and system optimization using genetic algorithm[C], IEEE 21nd International Conference on Industrial Electronics, 1995:1502-1507

15. 万健如，李保海. 群控电梯改进型遗传算法最佳派梯方法[J]， 起重运输机械， 2005（4）：51-53
16. Andrew Watkins, Lois Boggess. A new classifier based on resource limited artificial immune systems[C], Proceedings of the 2002 Congress on Evolutionary Computation, 2002, (2): 1546-1551
17. 刘静，钟伟才，刘芳等. 免疫进化免疫算法[M]， 电子学报，2001，29(12A)：1868-1872
18. P.K. Harmer, P.D. Williams, G.H. Gunsch etc. An artificial immune system architecture for computer security applications [J], IEEE Transactions on evolutionary computation, 2002, 6(3):252-280
19. P. J. Costa Branco, J.A. Dente, R. Vilela Mendes. Using immunology principles for fault detection [J], IEEE Transactions on industrial electronics, 2003, 50(2): 362-373
20. Y.S. Ding. A nonlinear PID controller based on fuzzy-tuned immune feedback law [C], Proceedings of the 3rd World Congress on Intelligent Control and Automation, 2000, (3):1576-1580
21. 李春华，毛宗源. 基于人工免疫算法的多目标函数优化[J]，计算机测量与控制，2005，13（3）：278-280
22. Soa. T. P, Kuok H S, Liu S K. New developments in elevator traffic analysis [J], Elevator Technology, IEEE 1993, (5):212-221
23. 郑延军，张惠侨，叶庆泰. 电梯群控系统客流分析与仿真[J]，计算机工程与应用，2001，(22)：139-141
24. 汤效琴，戴汝源，徐琪. 数据挖掘中变量聚类方法的应用研究[J]，计算机工程与应用，2004，(24)：171-173
25. 焦李成，杜海峰.人工免疫系统进展与展望[J]，电子学报，2003，31（10）：1540-1547
26. DASGUPTA D，FORREST S. Artificial immune systems in industrial application [C], Proc of Second Int Conf on Intelligent Processing and Manufacturing of Materials, Honolulu, USA:IEEEPress,1999:257-267
27. 肖人彬，王磊. 人工免疫系统：原理、模型、分析及展望[J]，计算机学报，2002，25（12）：1282-1291
28. 林学颜，张玲.现代细胞与分子免疫学[M]，北京：科学出版社，1999：46-62
29. L.N.de Castro, F.J.Von Zuben. An evolutionary immune system network for data

clustering [C], Proceedings of Sixth Brazilian Symposium on Neural Network, Rio de Janerio, 2000: 84-89

30. L.N. de Castro and J.Timmis. An artificial immune network for multimodal function optimization [C], Proceedings of the 2002 International Conference on Evolutionary computation, Honolulu, 2002: 674-699
31. Sathyanath S, Sabin F. An AIS approach to a color image classification problem in a real time industrial application [C], Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Arizona, 2001, 2285-2290
32. 高阳，陈世福，路鑫. 强化学习研究综述[M]，自动化学报，2004，11(30)：86-100
33. Kaelbling L P. A Situated Automata approach to the design of embedded agents [J], SIGART Bulletin, 1991, 2(4):85-88
34. Minsky M L. Theory of neural analog reinforcement systems and its application to the brain model problem [D], New Jersey, USA: Princeton University, 1954
35. Bush R R., Mosteller F. Stochastic models for learning [M], New York: Wiley, 1955
36. Waltz M D, Fu K S. A heuristic approach to reinforcement learning control systems [J], IEEE Trans. Automatic Control, 1965, 10 (3): 390-398
37. Samuel A L. Some studies in machine learning using the game of checkers [J], IBM Journal on Research and Development, 1967, (11): 601–617
38. Widrow B, Gupta N K, Maitra S. Punish/reward: Learning with a critic in adaptive threshold system [J], IEEE Trans. on Systems, Man, and Cybernetics, 1973, 3(5): 455-465
39. Saridis G N. Self-organizing control of stochastic system [M], New York: Marcel Dekker, 1977, 319-332
40. N.J. van Eck, van Wezel MM. Application of reinforcement learning to the game of Othello [J], Computer and Operations Research, In Press, Corrected Proof, Available online 5 December 2006: 1-19
41. Richard S. Sutton, Doina Precup, Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning [J], Artificial Intelligence, 1999, (122): 181-211
42. Andrew G. Barto, Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning [J], Discrete Event Dynamic Systems: Theory and Applications, 2003, 13(1-2): 41-77
43. 赵玮，王荫清. 随机运筹学[M]，北京：高等教育出版社，1993
44. 张飞舟，范跃祖等. 智能交通系统中的公交车辆动态调度研究[J]，公路交通科技，

2002，19（3）：123-126

45. S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results [J], Machine Learning, 1996, (22): 159-196
46. 林联明，王浩，王一雄. 基于神经网络的 Sarsa 强化学习算法[J]，计算机技术与发展，2006，16(1)：30-32
47. 陆鑫，高阳，李宁等. 基于神经网络的强化学习算法研究[J]，计算机研究与发展，2002，39(8)：981-985
48. Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J], Journal of Artificial Intelligence Research, 1996, 4(2): 237-285
49. R.S. Sutton. Dyna, an integrated architecture for learning, planning and reacting [J], SIGART Bulletin2, 1991: 160-163
50. R.S. Sutton, A. G. Barto, R. J. Williams. Reinforcement learning is direct adaptive optimal control [J], Control Systems Magazine, IEEE, 1992, 12(2): 19-22
51. Tadepalli, P. and Ok, D. Model-based average reward reinforcement learning. Artificial Intelligence [J], 1998, (100): 177-224
52. A.G. Barto, S.J. Bradtke, S.P. Singh. Learning to act using real-time dynamic programming [J], Artificial Intelligence, 1995, (73): 81-138
53. 薛丽华. 基于遗传算法的模糊神经网络电梯群控[D]，天津：天津大学，2002

致 谢

三年的学习生活将以本论文结束，其间的欢乐、艰辛、老师的教诲、家人及朋友的关怀和支持、同学的帮助，一言难尽，让我毕生难忘！

衷心感谢我的导师刘建昌教授！入学三年来，无论在生活上还是在学习中，刘老师都给予我无微不至的关心和极大的帮助，回想自己所取得的每一个进步，无不凝结着导师的心血。从专业知识的学习、到项目的实践以及毕业论文的完成的整个过程中，始终得到导师热情、耐心、细致的指导和关怀。同时刘老师严谨的治学风格、渊博的知识、忘我工作态度、无私的奉献精神深深地影响我，使我深受启迪。在此我要向我的导师致以最衷心的感谢和深深的敬意！

衷心感谢我的老师杨卫国老师、李洪儒老师、周玮老师和钱晓龙老师，在我研究生学习期间，他们为我的学习和研究工作做了大量的无私的工作，自己的成长进步也包含他们大量的心血，在此真挚的向他们说一声——谢谢！

在东北大学罗克韦尔实验室工作和论文撰写期间，得到了于尊龙、王晗、郭丹宁、郭金来、赵万庆、刘贵龙、胡国梁、贺生明等同学的热情帮助和支持，他们的思维和知识开拓了我的眼界，在此一并表示衷心的感谢！

感谢与我同宿舍的王欢、王红霞和刘阳，在平时的讨论中她们曾给我灵感和触发。感谢我的朋友和家人对我生活的关照和精神的鼓励，是他们无私的奉献使我顺利的完成学业和论文的写作。

基于强化学习算法的电梯调度系统的研究

作者： 马黎华
学位授予单位： 东北大学

相似文献(10条)

1.学位论文 王珺 基于人工免疫模型的数据挖掘及其应用 2007

生物免疫系统是一个高度进化的生物系统,它旨在区分外部有害抗原和自身组织,从而清除抗原,保持有机体的稳定.从计算的角度来看,生物免疫系统是一个高度并行、分布、自适应和自组织的系统,具有很强的学习、识别、记忆和特征提取的能力.人们很希望能从生物免疫系统的运行机制中获取灵感,由此开发出了面向应用的免疫系统计算模型--人工免疫系统.

人工免疫系统是模仿自然免疫系统功能的一种智能方法,它实现一种受生物免疫系统启发、通过学习外界物质的自然防御机理的学习技术,提供噪声忍耐、无教师学习、自组织、记忆等进化学习机理,结合了分类器、神经网络和机器推理等系统的一些优点,因此具有提供新颖的解决问题方法的潜力.

近年来,随着基础医学研究的深入,人们对免疫系统的机理也有了越来越清楚的认识.鉴于此,人们开始将免疫系统的特性运用到工程领域,通过改进原有的方法或者新型的方法,以期获得性能同样优良的各种方法和系统.目前人工免疫系统已被自动控制、计算机安全和病毒检测、异常和故障诊断、知识发掘、优化等多个研究领域所认可并接受.

虽然人工免疫系统具有很大的发展潜力,但其研究尚处于起步阶段.受自身发展的限制,相对于遗传算法在数据挖掘中的广泛应用,人工免疫系统在数据挖掘中的应用屈指可数.人工免疫系统现在更多的是被用于函数优化、工程控制及入侵检测方面.目前对于基于人工免疫系统的数据挖掘技术的研究,多使用的是基本的免疫算法,而忽略了免疫网络模型等其它免疫机理所具有的独特优势.而且这类算法多是处于实验性阶段,即多用于小型数据库以证明免疫算法的可行性,而很少有将其运用到大型数据库中的案例.

本文致力于研究人工免疫系统及其在数据挖掘中的应用问题,目的是通过人工免疫系统网络模型所具有的独特的免疫机理,发挥其在数据挖掘中的优势,并进一步实现其在大型数据库中的应用.主要工作为:1. 深入探讨并研究了人工免疫系统模型根据人工免疫系统的生物学原理基础,系统地研究了人工免疫系统中几种有代表性的免疫网络模型:aiNet模型、骨髓模型及有限资源人工免疫模型等,并根据本文的需要重点研究了aiNet模型.2.提出了两种新型的基于人工免疫网络模型 aiNet 模型的聚类算法研究了基于人工免疫网络模型的聚类挖掘技术,提出了两种新型的基于人工免疫网络模型 aiNet 模型的数据挖掘聚类算法--AiFCM算法和 aiNHA 算法,并通过仿真实验分别证明了两种算法的有效性,从而将人工免疫网络模型运用到了知识发现中.3.将基于人工免疫模型的聚类技术应用到酒店前台管理数据库系统中通过SQl Serve 2000构建出了面向酒店前台管理的数据库系统,并将基于人工免疫模型的聚类挖掘技术应用到该系统中,得到了令人满意的效果.从而实现了免疫算法面对大型数据库时高效率并行搜索、学习记忆、较强的全局寻优能力等优势.

2.期刊论文 李海芳.王莉.LI Hai-fang.WANG Li 基于人工免疫的图像聚类算法的研究 -计算机工程与应用 2007,43(20)

随着图像检索系统的发展,合理组织和管理图像数据库已渐渐成为用户检索的关键所在.因此提出了一种基于人工免疫的图像聚类算法,通过模拟抗体捕获抗原的机制,对由Internet上的60幅图像组成的图像库进行了聚类分析.聚类之前提取了图像的平均颜色特征,获得了每幅图像的mean值,同时,还比较了传统K均值聚类算法与本文算法的性能.实验结果表明,算法计算时间少、聚类误差低、聚类能力强,能有效地提高了检索效率.

3.学位论文 钟将 基于人工免疫的入侵分析技术研究 2005

人工免疫系统研究旨在抽取生物免疫系统中独特的信息处理机制，研究和设计相应的模型和算法，进而应用于解决各种复杂问题。人工免疫系统是继人工神经网络、进化计算之后新的智能计算研究领域，是生命科学和计算机科学相交叉而形成的交叉学科研究热点。生物免疫系统的基本功能是识别自我和非我，并将非我分类清除，具有免疫识别、免疫记忆、免疫调节、免疫宽容和免疫监视等功能特征，是一个自适应、自学习、自组织、并行处理和分布协调的复杂系统。

入侵检测是当前网络安全研究的重要内容之一。由于生物免疫系统是一个具有很强自我保护功能的系统，它能够有效识别己知和未知抗原的特性，给当前入侵检测系统的研究带来了很大的启发，因而基于生物免疫系统的入侵检测机制的研究显得更加引人注目，对促进当前网络安全的研究具有十分重要的意义。

本文从生物免疫系统突出的自适应识别能力的角度出发来开展基于生物免疫原理的数据分析和模式识别方面的研究，并进而实现入侵检测模式的自动提取。这种思想与当前广泛采用的基于机器学习或数据挖掘的入侵检测是一致的。当前基于数据挖掘的入侵检测主要采用聚类分析、分类分析以及异常检测来建立检测模式，并识别网络中的入侵行为。本文基于人工免疫聚类算法为线索研究了动态聚类算法，异常检测算法以及自适应分类算法，并在实际网络特征连接数据集上进行了一定的仿真实验，其结果验证了这些方法的可行性和有效性。本论文的主要研究内容如下：

1. 介绍了作为全文理论基础的自然免疫学方面的基础知识，内容包括免疫记忆、免疫识别、自适应免疫应答等。探讨了人工免疫系统常用的编码方式，免疫操作以及常见的人工免疫模型。

2. 设计了聚类可行解的算法。由于网络特征数据集中可能包含未知的入侵模式，因此不能预先设定聚类簇的数量，为了在保持聚类分析精度的前提下提高动态聚类算法的效率，本文借鉴了可行解的思想，提出可划分聚类数据集的聚类可行解的概念，设计了一种基于人工免疫网络的聚类可行解的获取算法，并对算法获得聚类可行解的条件和概率进行了一定的理论分析。

3. 结合GA，Tabu搜索等智能优化算法在聚类可行解的基础上实现了动态聚类分析，并在此基础上提出了基于“约简-优化”原理的两阶段动态聚类算法的框架。本文提出了一种基于人工免疫网络的聚类簇数量的估计算法，以解决约简算法中参数选择的问题。

4. 本文提出基于人工免疫聚类算法的异常检测算法，算法采用了一种基于距离的异常度量因子，可以方便地筛选数据集中最为突出的异常数据，因此该算法能够依据不同的安全策略来调节异常容忍因子以平衡检测率和漏报率之间的矛盾。该方法采用无标记的训练数据集，因此能自动适应于不同的网络和应用环境。

5. 本文中分析了基于均匀粒度的聚类方法构造分类器存在着与先验知识之间不协调的问题。提出了根据多粒度原理、基于人工免疫聚类来获取代表点集来构造分类器的方法，在一定程度上克服聚类结果与先验知识之间的矛盾，并提高了分类器的分类准确度和推广性。实验结果表明基于此分类器的入侵检测的平均检测率和误报率都保持了较高的性能。

4.期刊论文 郭晨.梁家荣.梁美莲.GUO Chen.LIANG Jia-rong.LIANG Mei-lian 基于人工免疫的进化性数据聚类分析的研究 -微电子学与计算机2006,23(11)

文章提出了一种新型的基于人工免疫系统的进化性聚类分析方法,通过这种方法不仅可以实现对数据的聚类,过滤冗余数据,并且可以根据已知数据的趋势生成未知进化数据,以达到对数据的进化简约处理的功效.

5.学位论文 童玲 电梯交通流多模式预测方法的研究 2005

论文主要研究了电梯交通流的多模式预测方法。

良好的电梯交通流分析能够为电梯群组的调度单元提供有益的、具有预见性的指导。论文在认真分析了电梯交通流的特点和规律的基础上，提出了一种混合型的多模式预测方法：该方法首先利用人工免疫聚类算法对电梯交通流进行离线的模式识别和分类，然后在此基础上利用高斯混合模型(GMM)对具有多种模式的电梯交通流进行数学建模，从而实现对电梯交通流的在线预测。

在处理电梯交通流预测问题过程，深入讨论了现有各种预测方法的特点和应用背景，提出了对电梯交通流进行离线的定性模式分类和在线的定量预测相结合的设计方案。由于电梯交通流随时间变化呈现出规律各异的多种交通模式，因此，采用新兴的人工免疫聚类算法对电梯交通流进行交通模式的识别和聚类分析，突破了传统的四种交通流模式：空闲、随机层间、上高峰和下高峰模式的局限，将交通流细分为8种交通流模式，便于理解和提高后续数学建模和预测的准确度，同时计算出各模式对应的浓度，初步完成对电梯交通流的定性分析。利用高斯混合模型对电梯交通流的概率密度函数进行数

学逼近，令高斯混合模型中高斯分量个数及各分量的加权系数的初始值分别等于交通流聚类分析得到的交通流模式种类个数和各交通流模式的浓度，然后通过EM算法在线优化估计电梯交通流高斯混合模型的参数，解决了一般的极大似然估计法中似然函数表达式难以构造，或者似然函数解析困难的问题，同时确保了参数估计过程的收敛性。在交通流的数学模型基础上实现电梯交通流预测。由于人工免疫聚类算法和高斯混合模型的结合，提高了电梯交通流预测的精度，同时为EM算法学习提供了良好的开始，大大地减少在线学习的迭代次数，缩短了预测时间。成功地将定性分析的智能方法和定量分析的数学建模方法有效地结合起来，完成了离线分析和在线学习的紧密衔接，这是论文最具创新之处。

最后，在Matlab6.5环境下实现了电梯交通流多模式预测方法的仿真，通过各种实验测试和仿真结果的对比分析，得出本文研究的多模式预测方法能够实现对电梯交通流进行预测的结论。

6.期刊论文 王磊.吉欢.徐庆征.WANG Lei.JI Huan.XU Qing-zheng 基于人工免疫粒子群优化算法的动态聚类分析-西安理工大学学报2008,24(4)

模糊C-均值聚类算法受初始化影响较大,在迭代时容易陷入局部极小值.将粒子群优化算法与模糊C-均值聚类算法相结合,提出一种新颖的动态聚类算法.该算法利用人工免疫思想改进粒子群优化过程,在很大程度上避免了粒子群算法和聚类算法早熟现象的发生,全局搜索能力和局部搜索能力优于同类算法.利用聚类理论中的经验规则kmax≤n 来确定聚类数k的搜索范围,在最优粒子基础上进化新一级种群,该方案可有效提高算法的收敛速度.两组数据的仿真实验表明,新算法优于传统模糊C-均值聚类算法,具有收敛速度快和解的精度高的特点.

7.学位论文 徐春鸽 人工免疫系统研究及其在数据聚类中的应用 2007

生物免疫系统是一种高度并行的自适应信息学习系统，它能自适应地识别和排除侵入机体的抗原性异物，并且具有学习、记忆和自适应调节能力，能够维护机体内环境的稳定。近年来，人们逐渐意识到生物免疫机制对开发新的计算智能的启示意义，人工免疫系统(Artificial Imune System，AIS)即是受生物免疫系统启示而设计的新型算法和模型的统称，已经用于机器学习、异常和故障诊断、机器人行为仿真、机器人控制、网络入侵检测等领域，表现出较卓越的性能和效率，它已经成为计算智能大家庭的卓有成效的新成员。

但是由于人工免疫系统只是近些年发展起来的新的研究领域，虽然关于它的研究已被越来越多的人所重视，并逐渐成为关于新的进化计算和模式识别方法研究的一个热点，但是，整个研究工作仍然显得比较零乱，缺乏系统性，有些方法名为人工免疫方法，实质是对原有方法的改造，失去了应用免疫系统的意义；另外，对于人工免疫系统的模型和应用缺乏深入的探讨，许多模型和方法并未真正体现出人工免疫系统的应用特点。

并且近年来在人工免疫系统应用研究中，关于人工免疫系统应用于数据聚类方面的研究逐渐多起来，但是目前的人工免疫系统应用于数据聚类的过程中存在有许多不足，例如：聚类方法中的很多参数需要人们根据经验判断手工输入，而不能根据聚类过程的进行自适应地变化；参数的设置不合理而造成参数对于网络的变化过于敏感；对网络中的参数分析不充分等等。

本文正是基于这样的认识，希望在免疫系统模型构造，算法实现以及应用于数据聚类方法的参数设置问题作深入的探讨。

首先有针对性的介绍了生物免疫系统的一些基本概念、系统组成、功能原理等；概述了人工免疫系统的一些主要原理及应用，在此基础上，提出了人工免疫系统的一般模型框架。然后，对现有的典型的人工免疫系统应用于数据聚类的算法进行详细分析，并在现有算法的基础上，针对其不足，提出了改进算法，通过仿真实验进行了验证。总的来说，本文的主要研究集中于：

1. 在认真分析和总结现有人工免疫系统的各种形式的基础上，提出了人工免疫系统的一般模型框架及其应用方向；

2. 以人工免疫系统的一般模型为指导，对现有的聚类分析方法进行研究，针对现有的人工免疫系统在解决聚类问题时的缺陷，提出了一种新型的自适应聚类人工免疫算法，在该算法中网络抗体间的抑制阈值、抗体的克隆数目、抗体的选择和再选择数目、抗体的变异的大小都随网络进化而自适应变化，使最终网络结构更加符合原始数据的内在结构，同时，对算法的参数进行了细致深入的分析和优化，减弱了算法对问题的依赖性并且重要参数可随运算过程自适应确定，减轻了用户需自行预定重要参数的负担，最后进行了仿真实验进行验证。

8.期刊论文 吴启明.陈昊.Wu Qi-ming.CHEN Hao 核聚类人工免疫网络模型的设计 -电脑知识与技术2009,5(26)

针对聚类分析问题和人工免疫系统的应用特性,构造出一种新型的人工免疫系统的应用形式-核聚类人工免疫网络.

9.学位论文 王育才 基于人工免疫网络的电子商务个性化推荐算法研究 2006

随着互联网的不断发展，电子商务网站越来越受到重视。作为电子商务企业对外的一个门户，如何通过为用户提供更加个性化的服务，提高其商品的吸引力，进而为企业带来更大的收益，就成为了网站所面临的核心问题。

电子商务个性化推荐系统是解决这一问题的有效手段，通过准确推荐与指导可以提高用户在网站体验的舒适度，建立稳固的客户关系。在为数众多的推荐技术中协同过滤技术由于其算法机理十分符合推荐活动的实际需求，因而受到了许多学者的青睐，同时也被广泛地应用于商业推荐系统中。

本文首先简要介绍了在电子商务中采用个性化推荐的必要性。接下来分析了协同过滤技术的基本原理，并对几种比较常用的协同过滤算法进行了详细的说明。然后文章通过对自然免疫学和人工免疫学理论的研究，着重讨论了将人工免疫网络技术应用于电子商务个性化推荐的思想，提出了使用形态空间模型对推荐技术及其存在问题的解释方法。本文还在深入研究Steve算法的基础上，提出了应用聚类分析技术的改进算法——聚类免疫推荐算法(CINR)；另外通过分析原算法(Steve算法)的可并行性，又进一步提出了针对原算法的并行化改进算法——并行免疫推荐算法(PINR)。最后，通过对算法的时间复杂度分析和仿真实验都证明了这两个算法在推荐实时响应速度方面对原算法有比较大的改进。

10.期刊论文 李彬.田联房.毛宗源.LI Bin.TIAN Lian-fang.MAO Zong-yuan 基于人工免疫的灰度图像多阈值自动分割 -计算机工程与设计2007,28(1)

为了实现灰度图像的自动分类以及自动分割,提出了一种基于人工免疫及最优分类数的灰度图像多阈值自动分割方法.定义了灰度图像最优分类数目标函数;接着运用人工免疫算法,结合最优分类数函数对灰度图像进行自动分类,并产生最优的多阈值,从而使得图像的全自动分割成为可能.该人工免疫算法中,抗原是指最优分类数目标函数,而抗体是指最优的多阈值.通过实验证明,分类清晰,效果良好.