学校代码： 10246
学　　号： 14210240026

硕 士 学 位 论 文

（学术学位）

智能家居环境下重症残疾人对家居设备操作的预测技术研究

Operation Prediction of Smart Appliances to Support People with Disabilities in Smart Homes

院　　系： 计算机科学技术学院
专　　业： 计算机软件与理论
姓　　名： 吴娥英
指 导 教 师： 顾宁　教授
完 成 日 期： 2017 年 3 月 26 日

指导小组成员名单

顾　宁　　教　授

张　亮　　教　授

卢　暾　　副教授

丁向华　　副教授

目录

目录..... I
摘要.....III
Abstract.....IV
第一章 绪论.....1
1.1 研究背景和意义.....1
1.2 国内外研究现状.....2
1.3 本文研究内容.....4
1.3.1 待解决的主要问题.....4
1.3.2 本文主要贡献.....6
1.4 论文结构.....7
第二章 相关原理与技术.....9
2.1 马尔可夫过程.....9
2.2 隐马尔可夫原理.....10
2.2.1 HMM 定义.....10
2.2.2 HMM 的基本问题.....11
2.2.3 HMM 训练算法.....13
2.3 数据填补技术.....15
2.3.1 简单填补法.....15
2.3.2 回归填补法.....15
2.3.3 多重填补法.....16
2.3.4 K 最近邻填补法（KNN）.....16
2.4 小结.....17
第三章 基于 KNN 的残疾人缺失操作填补算法.....18
3.1 操作缺失现象的分析.....18
3.1.1 操作缺失定义.....18
3.1.2 操作缺失分析.....19
3.2 K 最近邻（KNN）算法.....21
3.2.1 算法原理.....21
3.2.2 算法过程.....22
3.3 基于 KNN 的数据填补算法.....23
3.3.1 序列相似度定义.....23
3.3.2 数据填补过程.....24
3.3.3 填补结果评估.....26
3.4 小结.....28
第四章 基于 HMM 的残疾人操作预测算法.....29
4.1 环境（温度）影响用户行为.....29
4.2 基于 HMM 的残疾人操作预测.....30
4.2.1 训练 HMM.....31
4.2.2 预测操作.....36
4.3 优化状态转移函数.....38

4.3.1 HMM 中的时间信息.....38
4.3.2 三维状态转移矩阵.....41
4.4 残疾人行为操作预测模型.....41
4.4.1 数据填补模块.....42
4.4.2 操作预测模块.....43
4.4.3 通用性分析.....43
4.5 小结.....43
第五章 实验设计与分析.....45
5.1 实验数据.....45
5.1.1 数据预处理.....45
5.1.2 训练集与测试集.....46
5.2 实验设计.....46
5.2.1 实验内容.....46
5.2.2 实验评估.....47
5.3 实验结果分析.....47
5.3.1 算法参数评估.....47
5.3.2 改进模型性能评估.....49
5.4 小结.....51
第六章 总结与展望.....52
6.1 总结.....52
6.2 展望.....53
参考文献.....54
发表论文和科研情况说明.....58
致谢.....59

摘要

智能家居利用先进的计算机技术、网络通信技术，将家庭环境中的家用电器、通讯设备等家居子系统有机结合，为居住者提供高效、舒适、安全的生活环境。然而，当前的智能家居系统往往需要用户主动发起对日常设备的控制，而重症残疾人由于存在肢体、语言等方面的巨大障碍，其对智能设备的控制就会变得十分困难。

在重症残疾人智能家居环境下，传统的设备控制方式存在局限性：1）重症残疾人用户不具备正常的生活自理能力，因此，其对智能设备的操作成本将比正常人高得多；2）随着智能设备数目和种类的增加，操作指令的数量和复杂程度随之增加，使得残疾人用户面临的困难更加严峻。为了改善这些状况，有效的解决思路就是简化用户操作，因此，本文从帮助重症残疾人提高其在智能家居环境中的自理能力出发，将残疾人用户对智能设备的控制操作建模为时间序列上的有限操作集合，分析操作序列的数据特点，在一定程度上填补缺失操作，研究用户行为模式，实现了一个基于隐马尔可夫的时间操作预测模型以减小用户的操作代价，本文的主要研究工作如下：

1. 提出并实现了一个基于隐马尔可夫模型的操作序列预测算法，将温度建模为隐含状态，操作命令建模为观察状态，分析序列变化的内在属性，有效地为给定的局部操作序列预测下一步最有可能的操作，同时，改进隐马尔可夫中的状态转移函数，增加时间维度，扩展了隐马尔可夫模型的齐次性，构造出一个三维的状态转移函数。

2. 深入分析智能家居环境下重症残疾人用户对家居设备的操作行为特征，描述用户操作的缺失现象，定义缺失类型，计算序列的 Jaccard 相似度，实现了基于 K 最近邻算法的操作填补技术，参考残缺序列的 K 近邻序列集中相关操作的位置来确定待填补操作的位置，最终实现操作填补，为预测过程提供质量良好的数据集。

3. 基于真实的数据集，设计相关实验验证本文的数据填补算法和隐马尔可夫预测过程，实现对操作预测效果的评估。实验结果表明，本文提出的基于 KNN 的数据填补算法能有效对缺失操作进行填补，而基于改进的 HMM 的预测模型相比传统的 HMM 预测模型表现出了更高的预测精确率。

关键字：智能家居；重症残疾人；操作序列预测；隐马尔可夫模型；操作填补

中图分类号：TP3

Abstract

As a product of Internet of Things, Smart Homes have the abilities to combine the home appliances, communication equipments and other home subsystems together with the help of advanced computer technologies and network technologies, which aims to provide efficient, comfortable and safe living environments for the occupants. However, most current smart home systems often require their occupants to operate home appliances directly, but for people with disabilities, who usually have physical limitations to perform daily activities, it's very difficult them to perform such manual operations.

In Smart Homes, the traditional equipment control method has some limitations: 1) People with disabilities usually don't have normal self-care abilities, therefore, the cost of operating intelligent appliances will be much higher for them than normal people. 2) The operating complexity increases with the increase of numbers and types of intelligent appliances, which makes people with disabilities face more difficult problems. This paper aims to improve the disabled people's self-care abilities in Smart Homes. The operations to the intelligent appliances are modeled as a limited set of operations on time series, on the basis of this work, this paper analyzes the data characteristics, studies the user behavior patterns and implements a Hidden-Markov-based time operation prediction model to reduce the users' operation cost. The main work of this paper is as follows:

1. A Hidden-Markov model based operation sequence prediction framework is proposed and implemented, the temperature is modeled as hidden state and the operation command is modeled as observing state. Given a partial sequence, this model can predict the next most likely operation so as to recommend services to users. At the same time, it improves the state transition function in HMM, extends the Hidden Markov homogeneity by adding the time dimension, and finally constructs a three-dimensional state transition function.

2. Analyzing the behavioral characteristics of users in Smart Homes deeply, describing the lack of operation log, defining the types of incomplete sequences, implementing an operation padding technique based on K-Nearest neighbor algorithm, determining the position of missing operation in the incomplete sequence by referencing similar sequences and finally realizing the data padding process.

3. Based on the experiments on the real data set, we design relevant experiments to verify the data padding algorithm and hidden Markov prediction process in this paper, so as to evaluate the operation prediction effect. The experimental results show that the processed KNN-based data-padding algorithm can effectively fill the missing operations, while the improved HMM prediction model shows higher prediction accuracy than the traditional HMM prediction model.

Keywords: Smart Homes, People with Disabilities, Operation Sequence Prediction, Hidden Markov Model, Operation padding

Chinese Library Classification: TP3

第一章 绪论

近些年来，智能家居的发展呈方兴未艾之势，家居智能化在全球范围内得到广泛研究。智能家居系统通过物联网技术实现设备互联性并支持用户对设备的控制以达到设备可访问性的目标，最终引入高效、舒适、安全的生活方式。然而，由于智能家居操作错综复杂等因素，智能设备的控制往往需要较大的投入，对于存在肢体障碍的重症残疾人来说，人为控制设备的代价令其望而生畏。如何简化残疾人用户的设备操作方式是亟待解决的问题，而操作预测技术为我们提供了一个有效的研究思路。本章将首先介绍本文的研究背景，包括残疾人智能家居系统中面临的问题和挑战，其次介绍当前国内外研究现状，接着是本文的主要研究贡献，最后概述本文的组织结构。

1.1 研究背景和意义

智能家居概念起源甚早，但直到 1984 年美国康涅狄格州出现了首幢“智能型建筑”，智能家居生活才逐步走进大众的视野，而随着计算机技术、网络通信技术、物联网技术的迅猛发展，人们对生活环境的要求（如舒适性、效率性等）也不断提高，因此大大促进了智能家居的发展。

智能家居环境可以定义为一个能够获取其居住环境内的用户和周边因素等信息，并以此去适应居住者的生活习惯从而达到舒适、高效的居住目标的系统[1]。而这要求智能家居环境提供的服务应当具备最大化用户舒适度和最小化用户操作成本的特性。然而，当前的智能家居系统往往需要居住者主动发起对日常家居设备的控制，而其被动地提供响应服务，这必然要求居住者投入较大的人工成本去实现对智能设备的控制，进一步地，当居住者为患有肢体障碍的重症残疾人时，这样的设备控制方式存在以下两个主要问题：

1. 重症残疾人的先天障碍性与主观能动性之间的矛盾。处于特定环境下的个体往往都能对于环境的改变会做出相应的动态调整，如家居环境下，当温度不断地升高，用户为了保持生理的舒适度，便会执行开空调、开电扇等操作。然而，重症残疾人由于存在先天或后天的严重肢体障碍，导致他们在完成这样的“简单”的操作时需要比正常人付出更大的代价，甚至无法完成这些操作。
2. 重症残疾人的先天障碍性与智能家居设备复杂性之间的矛盾。随着智能家居系统的发展，智能设备的数量和种类随之增长，而操作指令的复杂程度也随

之增长，而对于自理能力不足的残疾人来说，家居设备不断复杂化，则残疾人对设备的操作控制也就变得更加困难了。以智能手机控制为例，设备越多，其涉及的指令就会越多、越复杂，那么手机控制界面的操作难度也会加大，用户在控制设备时，很有可能出现用户输入多个命令或者重复点击按钮后系统才予以响应的情况，最终引起不友好的用户体验。其次，指令长度和复杂度的增长加剧了本身就存在控制缺陷的残疾人面临的困难。

为了改善上述问题，有效的解决思路就是简化用户操作，改变请求-响应的服务模式，让智能家居系统自主学习用户行为，基于用户操作历史对其即将执行的操作进行预测，从而有效地降低用户的操作成本。

本文立足于研究智能家居环境下重症残疾人的行为习惯，统计残疾人用户的日常操作行为序列，分析序列特征并预测下一个操作。技术上，本文结合隐马尔可夫模型，将用户操作建模为隐马尔可夫过程中的观察状态，将温度（温度段）建模为隐含状态，并且考虑到温度变化的时间特性，改进了隐马尔可夫模型中的状态转移函数。同时，基于 K 近邻填补算法，填补了残疾人智能家居环境中缺失的操作数据，最终提出一个较完善的操作预测模型用于预测智能家居环境下残疾人的操作行为，实验证明，本文的预测模型相比传统的单一预测模型表现出了更好的预测效果，而合理的预测为智能家居系统自动执行设备操作提供了依据，从而实现代替用户的手动操作，减小其操作成本的目的，最终提高重症残疾人在智能家居环境下的生活自理能力。

1.2 国内外研究现状

科技的发展使得智能家居系统的研究受到广泛关注，当前国内外很多机构都致力于智能家居系统的研究，如麻省理工大学就研发出一个名为 Home of Future 的未来智能家居系统，用来分析与研究智能环境下用户的行为[2]。而美国德州大学也建立了一个 MavHome 的智能家居原型系统，并在此基础上进行多项研究[15, 3, 4]。科技巨头苹果和谷歌也分别推出了自家的家居平台 HomeKit 和 Works with Nest，用以整合第三方智能家居设备，让设备之间可以互相交流完成自主学习，最终向用户提供智能服务[14]。此外，还有 Adaptive House[5]，Gator Tech Smart House[6]等专门为残疾人、老年人等弱势群体构建的智能家居平台，致力于为弱势群体在智能家居环境下的生活提供便捷。

当前，如何构建出能满足残疾人特殊生理需求的智能家居系统是国内外研究者的热门研究问题。[12]中研究者就设计了一个具有开放的分层结构（用户层、

HMI 层、交流层）的智能家居系统去适应残疾人用户的需求，该系统基于服务发现协议去发现用户可能的需求，并且使用 Wi-Fi，蓝牙等无线技术实现用户和设备之间的交流。[11]中的 Health Smart Home 系统为有特殊需求的残疾人用户提供健康监护的功能，即安装自动化设备和一系列感应器来监督用户的健康状态以确保病人（尤其是有视觉、听力障碍的残疾人）的安全。除此之外，辅助技术的开发提高了残疾人的生活质量，如 Chen 等人就研发了名为 PR2 的辅助机器人来帮助患有严重肢体障碍的残疾人完成日常活动，包括借助 PR2 来实现刮胡子、拿杯子、挠痒等基本日常行为[13]。[16]中提出一个多模式的语言驱动系统（multi-modal Tongue Drive System），整合头部移动、语言（舌头）变化来帮助残疾人完成复杂的活动（如在 PC 上收发邮件）。

显然，上述的研究技术大多借助于网络技术、通信技术等的发展为残疾人构建更舒适的智能家居环境，比如研发新型硬件设备（辅助工具）来提高残疾人生活自理能力。事实上，近些年来，越来越多的研究者开始关注智能家居环境的“软服务”能力，也就是关注用户行为，试图从用户行为的角度出发，分析用户生活习惯，学习用户的行为模式，为其提供更“贴切”的智能服务。

智能家居环境下用户的行为分析可以大致分为三类:

一是用户活动的识别与分类（Activity Recognition），K.S.Gayathri 等人根据不同活动的时间间隔、并发以及重叠的特征，定义了简单活动 / 组合活动的识别方法（如开灯为简单活动，做早饭为组合活动）[7]。活动识别系统通常用于监控居住者和环境的变化，适当的活动模型可作为决策的依据[8]。

二是异常活动的检测（Anomalies Detection），在残疾人、老年人的智能家居环境下，异常活动的检测显得尤为重要。Wonjoon Kang 等人认为，当一个活动的持续时间大大长于或者短于正常时间，我们就可以认为其是一个异常活动而发出警告[9]，从而提高智能家居系统的安全性能。而[10]中基于对用户正常活动的学习结果（正常的规则集合）来检测异常活动的发生。

三是活动的预测（Activity Prediction），正如前面介绍，合理的预测能够使得智能家居系统具备自动化推荐服务的能力，以系统的自动执行代替用户的手工操作，从而减小用户对设备的操作成本。[1]中就介绍了 MavHome 智能家居环境下三种有效的用户行为预测算法来预测居住者的行为模式。Holder 和 Cook 也在[17]中也提出了一种预测算法，该算法计算距离某个参考活动的时间间隔来预测下一个事件。Jakkula 在[18]中提出的基于活动之间的关系来检测异常和预测活动的方法也十分有效。文[19]利用 SPEED 算法来预测智能家居中居民的活动，

SPEED 提取用户活动产生的数据（ON-OFF 的设备状态）组成序列，研究用户行为。

除此之外，机器学习在模拟与预测智能环境下用户的行为时也表现出了优越性。朴素贝叶斯、隐马尔可夫模型、动态贝叶斯网络等都是常用的用户行为建模方法。给定概率函数，观察对象和活动标签之间的依赖关系，朴素贝叶斯分类器便能预测下一步最有可能的操作[20]。而隐马尔可夫模型和动态贝叶斯也可用于建模用户活动[20,21]。而[22]中利用深度学习算法提出了一个有效的在线用户学习技术进行活动预测。Yi Yang 等人提出在智能家居环境下基于位置感知的自动化服务，即利用基于时间的马尔可夫模型（TMM）去建模居住者的地理位置模式，根据其所处地理位置来预测行为[23]。事实上，隐马尔可夫模型作为一种统计方法，极度适用于时间序列的建模，比如描述有限时间内的状态转换过程（如语音识别）。而在智能家居环境下，该模型也适用于描述时间序列下用户的行为[24]，文[28]就提出一种基于 HMM 的模拟智能环境下人类活动的数学方法。同时，也有很多研究者利用 HMM 来进行用户行为或者序列的预测[29,30,31,32]。而本文研究用户对有限设备的一系列开关操作，符合隐马尔可夫建模特征，因此，也用该模型来研究用户行为，下一节将对此具体展开介绍。

1.3 本文研究内容

1.3.1 待解决的主要问题

本文的研究基于复旦大学协同信息与系统实验室（CISL）进行的科研项目《基于智能交互控制设备的大数据残疾人自理与服务云平台》，该项目在上海市残联的支持下，实验室团队自主研发了基于智能交互控制设备的大数据残疾人自理与服务云平台，致力于改善残疾人行动不便，生活自理困难的现状，目前整个系统已经在上海市杨浦区试点运行。而本文的研究对象为杨浦区六家重症残疾人（包括渐冻人、重度残疾人）家庭，具体地，我们为这六家重症残疾人（渐冻人）家庭搭建了智能家居平台，主要利用中控器将空调、电风扇、窗户、窗帘、电灯等五种家居设备联结为一个智能家居网络，且每种设备都有开关（ON-OFF）操作，残疾人用户可以使用智能手机去控制这些设备的开关，操作方式有语音命令和操作命令两种。考虑到重症残疾人往往存在较严重的肢体障碍，为了提高其在智能家居环境下的生活自理能力，减小对家居设备的控制成本，本文认为一个有效的思路就是让智能家居系统能学习用户行为，对用户即将进行的操作进行预测。围绕学习用户行为习惯，预测给定操作序列的下一步最有可能的操作这一研究内

容，需要解决以下问题：

- **如何考虑残疾人用户在智能家居环境下的特殊性：**相比正常用户，残疾人用户在智能家居环境下的行为表现出一定的特殊性。一个显著特征是存在操作记录缺失现象，也就是说，用户的一些操作记录并没有被录入到数据库中，因此数据库中记录的操作序列就是不完整的。究其原因，在本文的研究项目下，大多数残疾人用户对智能控制方式的依赖性较强，如日常生活时残疾人用户由于存在身体缺陷无法完成设备的手动控制，因此需要借助智能系统实现对家居的控制。而当残疾人家属在家时，这些用户对智能控制方式的依赖会大大减弱，因为这时候很多操作是家属代替残疾人用户去手动执行的，比如夜晚家属为残疾人用户的房间关灯，而这样一个手动操作不会被记录。尽管普通用户也会存在操作缺失现象，但普通用户的操作缺失具有一般规律，即何时何地都有可能去手动执行某些操作，而残疾人用户的操作缺失现象更倾向于当在家属在家时发生，因此，如何去填补残疾人用户下这些缺失的数据具有重要的研究意义。

- **如何建模残疾人用户的操作行为：**在本文的智能家居环境下，残疾人用户通过手机控制家居设备，执行设备的开关操作，假设有两种设备对应的四个开关操作，分别记为 A，a，B，b，那么一定时间周期内（比如一天），用户执行的操作可记录为 ABbaAaABab 这样的操作序列，而研究问题可以转化为给定这样的一个操作序列，如何预测下一步最有可能的操作。另外，调查发现残疾人用户的行为受到环境的影响，比如温度升高，用户开空调的可能性也会增大，也就是，用户行为是对环境（温度）变化敏感的，或者说用户能够对环境变化做出自适应调整，因此，是否能够在建模残疾人用户的行为时将此环境因素也考虑在内也值得研究。

在上述问题中，前者的解决思路是为残疾人用户产生的缺失操作进行填补，而后者的解决方案应是构造合适的模型去学习用户的操作行为，从而进一步去预测用户操作。因此，本文从探讨这两个问题出发，整理用户需求，利用基于 KNN 的数据填补技术来填补不完整序列中的缺失操作，即给定一个不完整的操作序列，得到数据集中与其相似的其他序列组成参考序列集，根据参考序列集中待填补操作（已知）所处的位置索引的前后关系来确定目标序列中该缺失操作应处的位置。而另一方面，在进行深入研究之后，本文运用隐马尔可夫模型模拟用户操作过程，将残疾人对智能设备的操作表示为时间序列下各个操作的集合，即时间轴上的操作序列。同时，整合操作过程中的两个不确定性；一是环境（温度）变化的不确定性，二是用户行为受到温度变化影响的不确定性，本文利用隐马尔可夫模型[27]

来描述这样的过程，即将温度建模为隐含状态，将用户操作建模为观察状态，隐含状态的变化指导观察状态的变化。而研究问题即可转化为训练一个 HMM，利用该模型去预测给定操作序列的下一步最可能的操作。

1.3.2 本文主要贡献

围绕 1.3.1 中提出的问题，本文展开了一系列的后续工作，明确了以智能家居环境下重症残疾人的操作序列预测方法作为研究对象的基础上，查询文献资料，了解国内外研究现状，完成残疾人用户的操作行为分析，最终提出一个操作序列预测模型，该模型整合数据填补模块与操作预测模块，并对相关技术进行改进，最后完成模型的实现并通过实验验证。图 1.1 即为本文的研究过程。

本文的主要贡献如下：

1. 深入分析智能家居环境下残疾人用户的操作行为的特殊性，即存在操作缺失现象和对环境变化的敏感性。在此基础上了解当前国内外研究现状，结合残疾人用户的操作需求，探究时间序列上用户操作行为的变化特征和规律，分析智能家居环境下残疾人的行为习惯。

2. 提出了一个智能家居环境下重症残疾人对设备操作的预测模型，该模型包含两部分：基于 KNN 的操作填补模块和基于 HMM 的操作预测模块。操作填补模块将数据集中的数据进行预处理，填补不完整序列；而操作预测模块用来预测给定序列的下一步最有可能的操作。

3. 实现并改进了一个基于隐马尔可夫模型的操作预测算法，将温度建模为 HMM 中的隐含状态，将操作建模为 HMM 中的观察状态，分析用户产生的操作序列的变化特征，学习用户行为习惯。同时，改进 HMM 中的状态转移矩阵，扩展了齐次性原理，引入时间信息，构造出三维状态转移函数，最终得到一个改进的非静态的 HMM。

4. 实现并改进了一个基于 KNN 的数据填补算法，该填补算法用于填补不完整序列，通过改进 Jaccard 相似度计算方法，获取不完整序列的相似序列集，根据相似序列集中缺失操作所处的位置来确定该缺失操作在目标序列中应处的位置，最终完成数据的填补，为操作预测模块提供质量良好的数据集。

5. 本文采集了真实的数据，设计相关实验来验证本文的填补算法与预测算法的有效性。实验表明，基于 KNN 的操作填补技术能够较有效地为重症残疾人用户产生的缺失操作进行填补，而本文改进的 HMM 预测算法相比传统的

HMM 算法在预测残疾人用户的下一步操作时表现出了更高的预测准确率，从而表明了本文的预测模型是有效的，有助于减小残疾人用户对家居设备的操作成本。

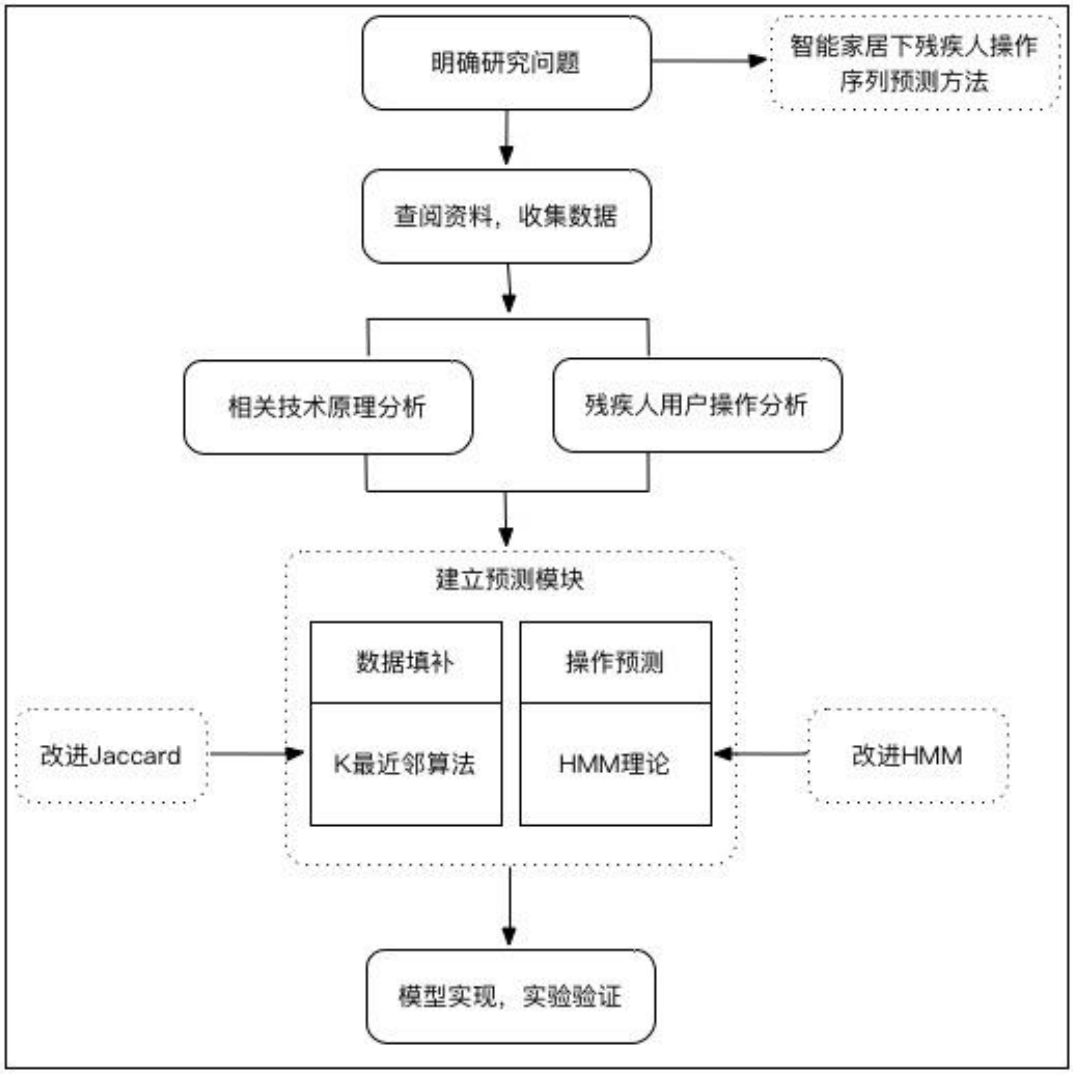

图 1.1 本文研究过程

1.4 论文结构

本文主要分为六个部分：

第一章：绪论。介绍智能家居环境下重症残疾人的行为研究的背景与意义，介绍当前国内外研究现状，探讨残疾人用户在智能家居环境下的操作行为特殊性。最后，总结了本文的主要贡献。

第二章：相关原理与技术。详细介绍隐马尔可夫模型的原理，紧接着介绍隐马尔可夫模型的几个基本问题：评价问题、解码问题以及学习问题（即 HMM 训练算法）。其次，介绍相关的缺失数据的填补算法，包括简单填补法、回归填补法和 KNN 填补算法。

第三章：基于 KNN 的残疾人缺失操作填补算法。首先对残疾人用户的操作日志进行分析，定义操作缺失类型，紧接着介绍 K 近邻填补算法的原理，包括改进 Jaccard 相似度计算方法，参考集的选取，填补过程等，最后通过实验对填补结果进行评估。

第四章：基于 HMM 的残疾人操作预测算法。本章详细介绍了 HMM 预测算法的工作原理，包括模型训练算法，预测过程的实现，优化状态转移函数等。最后介绍了一个完整的预测模型，包含数据填补模块和操作预测模块。

第五章：实验设计与分析。采用残疾人家庭中真实的用户操作数据与中国气象网站的天气数据对本文的工作进行实验验证，并对实验结果进行分析与可视化展示。

第六章：总结与展望。总结本文的主要工作，提出下一步的工作计划。

第二章 相关原理与技术

为了描述智能家居环境下残疾人的操作行为，本文引入隐马尔可夫模型(Hidden Markov Model, HMM)，本章节将首先介绍隐马尔可夫模型的原理，包括马尔可夫过程，隐马尔可夫模型的定义与几个基本问题。其次，本章将详细介绍模型的 Baum-Welch 训练算法。最后介绍几种常见的数据缺失填补方法。

2.1 马尔可夫过程

马尔可夫过程[33,34]最早由俄国数学家 A.A 马尔可夫提出，它是一个具有马尔可夫性质的随机过程。马尔可夫性质（也为无后效性）可以表示为在一个随机过程中，已知当前时刻 T 的状态为 S，则下一时刻 T+1 的状态只与 T 时刻的状态相关，而与 T 时刻之前的状态无关。即给定当前的知识的情况下，过去（当前以前的情况）对于预测将来是无关的。因此，马尔可夫性质可以形式化地表示为：$P(S_{n+1}|S_0,S_1,\dots,S_i,\dots,S_n)=P(S_{n+1}|S_n)$，其中，$S_i$表示时刻$T=i$时的状态，而整个公式表示$T=n+1$时刻的状态概率仅与$T=n$时刻的状态相关。

一个典型的马尔可夫随机过程例子：我们假设天气只有晴天，阴天，和雨天三种，分别用E_1,E_2,E_3表示。假设某 10 天的天气状况为$E_1E_2E_2E_3E_1E_3E_3E_2E_1E_2$，根据概率统计，我们可得出当某一天的天气为$E_1$，则接下来一天的天气为$E_1$的概率$P(E_1|E_1)=0/3$，为$E_2$的概率$P(E_2|E_1)=2/3$，为$E_3$的概率$P(E_3|E_1)=1/3$。也就是说，我们得到了当前状态为$E_1$时，下一时刻的状态转移概率。同理，我们也可以得出当前天气为E_2，E_3时接下来一天的天气概率。

一般的，假设一个马尔可夫随机过程中有N个状态，则每一时刻就可能发生N^2种状态转移，而这每一种转移都有一个状态转移概率，即从某一个状态转移到另一个状态的概率。这所有的N^2个状态转移概率就可以用一个状态转移矩阵表示，如图 2.1，其中，任意状态转移概率a_{ij}表示当前状态为i，下一时刻状态为j的概率。

$$\begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{N1} & \cdots & a_{Nj} & \cdots & a_{NN} \end{bmatrix}$$

图 2.1 马尔可夫状态转移矩阵

2.2 隐马尔可夫原理

隐马尔可夫模型[36]是一个含有隐含的未知参数的马尔可夫过程，区别于马尔可夫过程中可见的状态（如上一节中的天气），隐马尔可夫模型中的状态是隐含的，不可观察的，但是这些隐含的状态指导着观察状态的改变，因此，观察者可以通过可观察的参数来确定该过程中的隐含状态。换言之，隐马尔可夫模型是一个双重的随机过程[35]，一是状态转移是随机变化的，二是可观察状态是随着状态转移随机变化的。

扩充 2.1 节中的天气例子，假设天气状况（晴天、阴天、雨天）是未知的，而天气影响着植被的叶面潮湿度（干燥、微干、微湿、潮湿），正常情况下，天气越晴朗，植被叶面也越干燥。假设天气状态分别为S_1，S_2，S_3，叶面潮湿度分别为O_1，O_2，O_3，O_4，假设某段时间观察到的叶面潮湿情况为$O_1O_1O_2O_3O_2O_4O_2O_1O_2O_1$，而实际的天气情况为$S_1S_1S_3S_3S_2S_3S_2S_1S_2S_2$，则我们可以叶面潮湿度建模为观察状态，天气状态建模为隐含状态，从而构成一个隐马尔可夫模型。如图 2.2 所示，上层的观察值序列是可见的，而下层的模型是不可见的，但是天气变化影响着潮湿度的变化，即隐含状态的变化指导观察状态的变化。

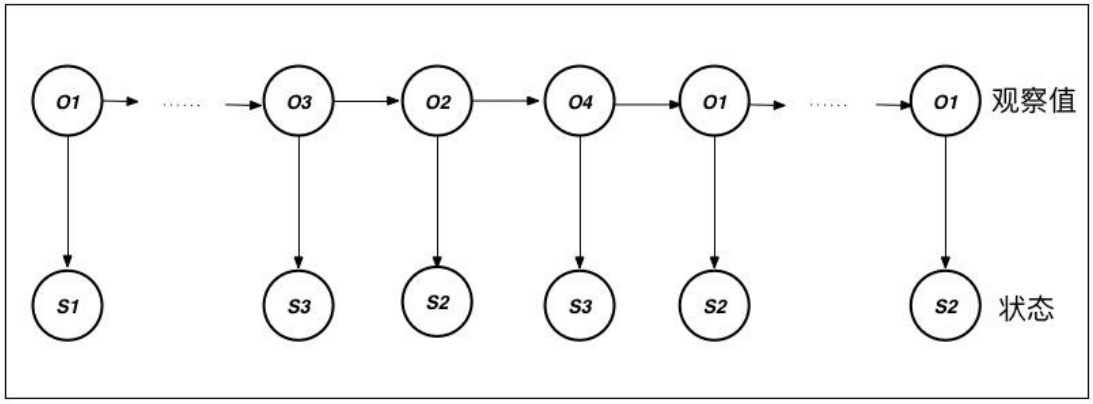

图 2.2 隐马尔可夫模型示意图

2.2.1 HMM 定义

一般地，隐马尔可夫模型包含以下五个基本要素：

隐含状态的有限集合S：$S = \{S_1, S_2, \dots, S_N\}$，其中 N 表示隐含状态的个数。

观察状态的有限集合O：$O = \{O_1, O_2, \dots, O_M\}$，其中 M 表示观察状态的个数。

状态转移概率矩阵A：$A = \{a_{ij}\}$，其中A为一个$N \times N$的矩阵，而a_{ij}表示当前

状态为S_i，下一时刻状态转移到S_j的概率，可形式化为$a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$，其中$q_t$表示当前时刻的状态。

发射概率矩阵B：$B = \{b_{jk}\}$，其中B是一个$N \times M$的矩阵，而b_{jk}表示某一时刻的状态为S_j，而观察状态为O_k的概率，可表示为$b_{jk} = P(v_t = O_k | q_t = S_j)$，其中$v_t$和$q_t$分别表示当前时刻的隐含状态和观察状态。

初始概率π：$\pi = \{\pi_i\}$，表示在初始时刻即$t = 1$时各个隐含状态出现的概率，且所有状态出现的概率之和为 1。

通常，隐马尔可夫模型表示为$\lambda = (A, B, \pi)$。图 2.3 为隐马尔可夫模型示意图，S表示隐含状态，O表示观察状态，其中实线表示隐含状态之间存在转移过程，而虚线表示发射概率，即每个隐含状态都有相应的概率指导各个观察状态的产生。

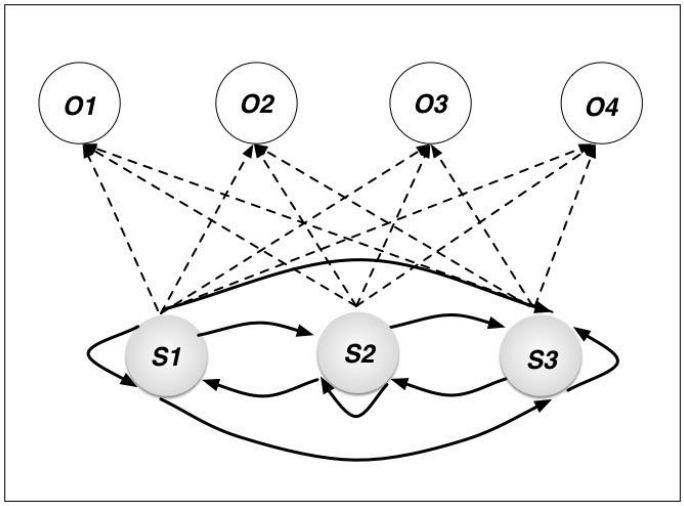

图 2.3 HMM 示意图

2.2.2 HMM 的基本问题

已知隐马尔可夫模型的基本原理和定义之后，可用于指导解决以下问题：

（一）评价问题[57]，即假设 HMM 的参数已知，求某个观察状态序列在此 HMM 中出现的概率。

假设给定一个隐马尔可夫模型λ，可观察状态序列$O = o_1 o_2 \ldots o_T$，求在所有可能的隐藏状态序列下，此观察序列的概率即$P(O|\lambda)$。我们可以通过前向算法(Forward Algorithm)[37]来解决此问题：考察序列的第一个时刻$T = 1$，用$\alpha_1(j)$表示在第一个时刻下，隐含状态为S_j的情况下观察状态o_1出现的概率，由 HMM 的定义我们可以推导出：

$$t = 1, \quad \alpha_1(j) = \pi_j b_{jo_1} \tag{2.1}$$

而隐含状态S_j有N种可能，因此$T = 1$时可观察序列在所有可能的隐含序列下的概率为$P = \sum_{j=1}^{N} \alpha_1(j)$。递归，假设已经计算前$t$个时刻，考察时刻$T = t + 1$，注意，该时刻的每一个可能的隐含状态都可以由$t$时刻的$N$个隐含状态转移生成，如图 2.3 所示，当$N = 3$，$t + 1$时刻的每个状态都有三条路径可达，分别为前一时刻的三个状态转移生成。因此，在某个状态S_j下到观察状态序列$O' = o_1 o_2 \dots o_{t+1}$的概率为：

$$t > 1, \quad \alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij}\, b_{jo_{t+1}} \tag{2.2}$$

也就是通过到达$t + 1$时刻的所有路径的概率之和来计算到达该状态的概率，最后，完整序列的概率$P(O|\lambda) = \sum_{j=1}^{N} \alpha_T(j)$。显然，前向算法采用来递归的思想，在给定的 HMM 下计算某个可观察状态序列的概率时，通过计算到达终点的路径之和的概率与相应发射概率的联合概率，最终实现给定观察状态序列在所有可能的隐藏状态序列下的前向概率。

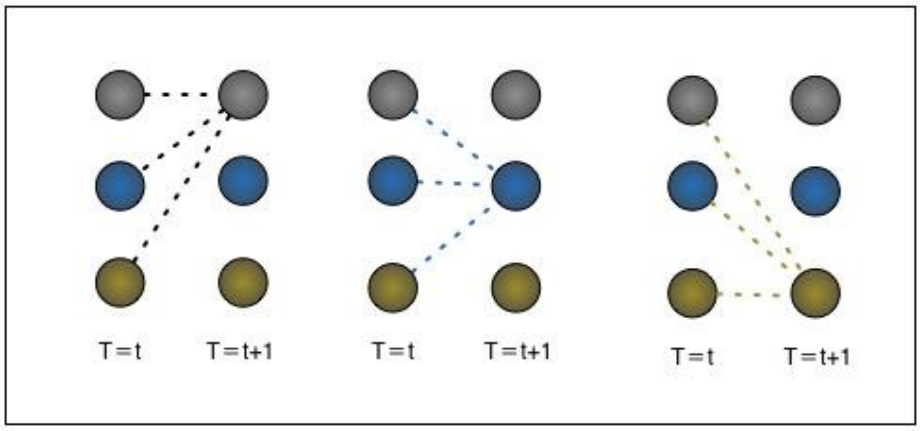

图 2.3 HMM 某一时刻状态转移示意图

（二）解码问题[57]，即根据可观察的状态序列找出潜在的最有可能的隐含状态序列。

假设给定一个隐马尔可夫模型λ，可观察状态序列$O = o_1 o_2 \dots o_T$，计算最有可能（概率最大）的隐藏状态序列。如图 2.4，在所有可能的路径中，我们需要找到一条最优的路径，即图中的红色实线，也就是要求的最有可能的隐藏状态序列，而黑色虚线为所有路径（序列）。

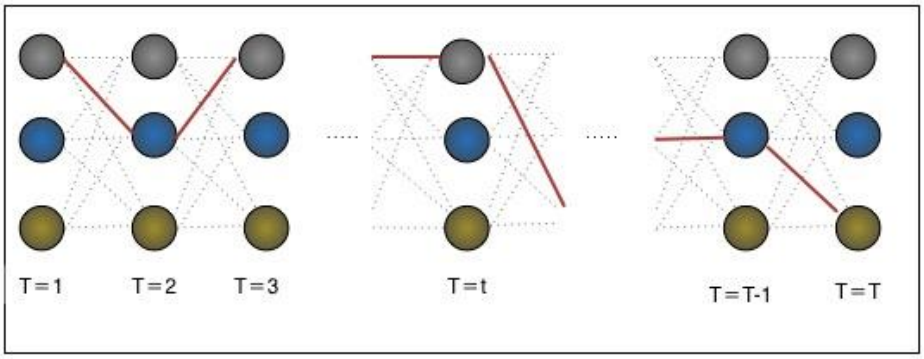

图 2.4 HMM 最优路径示意图

维特比算法(Viterbi)[37]可解决此问题：我们用$\delta(t,i)$表示t时刻，到达状态i的所有可能路径（序列）中概率最大的路径的概率。因此，我们的目标就是求得$t=T$时刻每个状态的最大概率，得到此刻概率最大的状态所对应的路径，即为对应的全局最优路径（序列）。

当$t=1$时，到达该状态的最可能路径还不存在，即没有$t: 0 \to 1$，因此：

$$\delta(1,i) = \pi_i b_{io_1} \tag{2.3}$$

当$t>1$时，考虑计算t时刻的部分概率，需要知道$t-1$时刻的部分概率以及转移到t时刻的状态转移概率，取其中最大的联合概率作为t时刻的部分概率。我们用一个指针φ来表示到达t时刻状态的最大局部状态的前一个状态，即：

$$\varphi_t(i) = \underset{j}{^{argmax}} \left(\delta(t-1,j) \cdot a_{ji}\right) \tag{2.4}$$

其中，$\underset{j}{^{argmax}}$表示能最大化后面表达式的j的值，也就是说，状态j所对应的路径是到达$\mathrm{T}=t-1$时刻的最优路径，下一时刻的状态转移一定从状态j发出，而$\mathrm{T}=t$时刻的$\delta(t,i)$为：

$$\delta(t,i) = \underset{j}{^{max}} \left(\delta(t-1,j) \cdot a_{ji} \cdot b_{io_t}\right) \tag{2.5}$$

换言之，某一时刻的部分最优概率等于前一时刻的部分最优概率乘以状态转移概率以及发射概率。最后，通过回溯法，得到最后时刻T的最优概率以及对应的状态，依次往前，得到整个全局的最优路径（序列）。

2.2.3 HMM 训练算法

本文中，为了预测智能家居环境下残疾人用户的操作行为，我们需要训练出

一个符合用户行为特征的 HMM。隐马尔可夫的训练过程可以这样表示：给定可观察状态序列$O = o_1 o_2 \dots o_T$，找到一个最优的参数模型$\lambda(A, B, \pi)$，使得$P(O|\lambda)$最大。

一般地，Baum-Welch 算法[37,58]，也称前向后向算法是目前使用最广泛的 HMM 训练算法。它的基本思想就是首先初始化一个 HMM，该初始值可以是一个错误的猜测，然后利用梯度下降的思想，通过已知的训练集不断地去减小该初始值的误差，使之更精确地描述训练样本，最后得到一个稳定且收敛的 HMM，注意该过程最后得到的 HMM 的最大似然估计是一个局部最优解。该算法存在以下几个基本步骤：

1) 初始化：原则上，我们可以设定模型的初始参数$\lambda(A, B, \pi)$为任意值，但由于该算法得到的是一个局部最优解，因此，初始参数的选择至关重要。

2) 计算前向变量和后向变量[58]：上一节已经介绍前向变量的原理，即$\alpha_t(i) = P(o_1 o_2 \dots o_t, q_t = S_i|\lambda)$，表示在给定模型$\lambda$且$t$时刻序列的隐含状态为$S_i$的情况下，局部观察序列$o_1 o_2 \dots o_t$(从$t = 1$时刻到$t$时刻)出现的概率，计算过程参考 2.2.2。同理，我们可以引入后向变量$\beta_t(i) = P(o_{t+1} \dots o_T, q_t = S_i|\lambda)$，它表示在给定模型$\lambda$且$t$时刻序列的隐含状态为$S_i$的情况下，局部观察序列$o_{t+1} \dots o_T$(从$t + 1$时刻到终止时刻)出现的概率：

$$t = T, \quad \beta_T(i) = 1 \tag{2.6}$$

$$t < T, \quad \beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij}\, b_{jo_{t+1}} \tag{2.7}$$

3) 更新参数：基于前向变量α和后向变量β，我们便可以引入两个新的变量：

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_{jo_{t+1}} \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_{jo_{t+1}} \beta_{t+1}(j)} \tag{2.8}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{2.9}$$

在上述公式中，$\xi_t(i,j)$表示在t时刻隐含状态为S_i，且下一时刻转移到S_j的概率。$\gamma_t(i)$表示在t时刻隐含状态为S_i，则下一时刻从S_i转移到所有状态的概率之和。基于$\xi_t(i,j)$和$\gamma_t(i)$，便可更新 HMM 模型：

$$\pi_i^* = \gamma_1(i) \tag{2.10}$$

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2.11}$$

$$b_{jk}^{*} = \frac{\sum_{t=1,o_t=k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.12}$$

至此，一个新的 HMM 参数模型$\lambda^{*} = (\pi^{*}, A^{*}, B^{*})$生成，重复此更新过程直至收敛。最后获得的模型即是最符合给定训练序列的模型，可用于残疾人行为预测。

2.3 数据填补技术

在数据统计过程中，数据缺失现象十分常见，在本文研究过程中，我们发现在智能家居环境下残疾人用户的操作行为也存在着数据缺失现象，最典型的例子为系统会缺失用户对设备的手动控制操作，使得数据库中的操作序列不完整。事实上，在整个数据挖掘领域，数据缺失都常常发生且不可避免，因此，如何填补缺失数据显得尤为重要。

2.3.1 简单填补法

均值填补法是一种简单的数据填补算法，它的原理十分简单：当数据类型为连续型数据时，计算数据集中各个完整数据的平均值，将此作为数据缺失填充值即可。当数据类型为离散型数据时，选取数据集中的中位数作为填充值。

均值填补法的优点是简单方便，处理数据的成本较低。但是，简单的均值填补法会导致数据集产生有偏估计，一定程度上丢失数据的原本个性，使得整个数据集的趋势变得更加平缓，即数据集方差和标准差都会变小。

2.3.2 回归填补法

回归填补法[59]也是一种常见的数据填补算法，它首先选取缺失变量的若干相关属性作为回归过程的自变量，然后针对缺失变量建立回归方程，利用该回归方程计算缺失值的期望，期望值即为填充值。

然而，回归填补法也会扭曲数据集，因为它仅考虑目标变量的若干相关变量，却没有考虑其他变量的影响。更重要的是，并非所有数据点都能拟合到回归方程中，这就需要建立一个误差估计来评判回归方程的准确性，与此同时，回归填充要求自变量和因变量之间存在强烈的相关关系，而在实际数据集中往往不存在。

2.3.3 多重填补法

多重填补法[38,39]是一种较精确的数据填充算法，它的基本原理为使用包含 M 个插补值的向量去填补缺失数据。首先，依次使用向量中的每一个元素去填补缺失数据，每次得到一个完全的数据集，共产生 M 个完全数据集。其次，用标准的数据分析方法分析每个完全数据集。最后，对所有的完全数据集进行综合，得到最终的估计值。

多重填补法能够产生有效的数据填补值，因为它不使用单一值来填充数据，而是考虑 M 个插补值，M 一般大于等于 20，而这样一个向量样本能够充分考虑数据集的不确定性，产生较好的效果。

2.3.4 K 最近邻填补法（KNN）

K 最近邻算法[60]最早在 1967 年被提出来，它是目前在数据挖掘领域应用广泛的分类、聚类的算法。K 最近邻算法认为数据集合中的项的距离越近，则项的相似度越高。利用该性质可计算缺失数据项的 K 最近邻，得到相似集合，并根据相似集合来填补缺失数据。

K 最近邻填补法的总体思路如下：

第一步，设定 K 值，计算缺失数据项与数据集中其他数据项的相似性。对于任意两个数据项X与Y，假设数据项$X=(x_1,x_2)$，数据项$Y=(y_1,y_2)$，则常见的相似性计算方法有：

1) 欧氏距离：即计算m维空间中两个点的距离。则X与Y的相似度为$sim(X,Y)=\sqrt{(x_1-y_1)^2+(x_1-y_1)^2}$。
2) 余弦相似度[40, 41, 42]：通过两个m维向量的夹角来衡量相似性，夹角越小，相似性越大。则X与Y的相似度为$sim(X,Y)=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}}$。
3) Pearson 系数[43]：Pearson 系数是一种线性相关系数，用于衡量两个向量的线性相关程度。则X与Y的相似度为$sim(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$，其中$Cov(X,Y)$表示两者的协方差，$Var(X)$和$Var(Y)$表示两者的方差。

第二步，根据与目标项的相似度，将数据集中的数据项按照相似度从大到小排序，选取前 K 个数据项作为目标项的参考集合。

第三步，根据选出的 K 个数据项填充缺失数据，一般可以选取这 K 个数据

项中相应属性的均值、众数等进行填充。

K 最近邻算法具有较好的适应性，既能对连续型数据进行填充，也能对离散型数据进行填充。但其缺点是需要事先指定 K 值的大小，而 K 值的大小会影响填充的效果，另外，当数据量较大时，该算法开销较大。

2.4 小结

本章主要介绍两方面的相关技术，一是隐马尔可夫原理，该部分是后续预测算法的理论基础。二是数据填补算法，主要介绍了当前主流的几种数据填补算法的基本原理以及对比分析其优缺点。本章的工作为后续章节的工作奠定了理论基础，具有重要的意义。

第三章 基于 KNN 的残疾人缺失操作填补算法

残疾人智能家居系统是以残疾人为中心，致力于提高残疾人用户的生活自理能力的家居平台。为了深入了解残疾人用户在智能家居环境下的行为特征，本章分析了在残疾人用户对设备的操作控制过程中存在的数据缺失现象，然后提出基于 K 最近邻的数据填补算法，接着分别介绍 KNN 算法的原理以及基于 KNN 的数据填补技术，并对该填补技术进行简单的评估。

3.1 操作缺失现象的分析

3.1.1 操作缺失定义

本文的研究对象为杨浦区六家重症残疾人（包括渐冻人、重度残疾人）家庭，具体地，这些家庭中安装的智能家居主要为电灯、窗户、窗帘、空调、电风扇五种设备，而对这些设备的操作分为开、关两种。对应关系如表 3.1 所示。

设备	操作类型	
	开	关
电灯	A	a
窗户	B	b
空调	C	c
电扇	D	d
窗帘	E	e

表 3.1 设备与操作类型对应表

为了方便研究，我们以其中的一位重症残疾人用户“用户 1”作为案例进行分析，“用户 1”是一个患有肌肉萎缩性侧索硬化症（即渐冻人）的中年女性，平时与其儿子生活，儿子有稳定工作，节假日在家照顾母亲。我们采集了“用户 1”在 2015 年 7 月 14 日到 10 月 14 日以及 2016 年 7 月 14 日到 10 月 14 日历时六个月（186 天）期间产生的操作数据。同时，以天为单位，对用户一天内产生的操作数据进行整理与划分，形成一个操作序列集合，可表示为$dataset_{user1} = \{seq_i | 0 \leq i \leq 185$ *且*i*为整数*$\}$。

分析序列集合$dataset_{user1}$，发现其中的序列存在一定“残缺”问题，如$seq_j = \{ABbCcADd\}$为某一天产生的操作序列，分析可知该序列在第 0 位和第 5 位的操作均为A（开灯），而在这中间没有对应的a操作（关灯），正常情况下，假设每一个操作都是有效的，则可以推出在子序列$subSeq_j = \{ABbCcA\}$中丢失了a操作，且a操作应在第 1 位到第 5 位之间的某一位置上，而可能的填补结果如图 3.1 所示，即a操作在两个连续的A操作之间的任意位置都可能出现。为了更好的表示残缺序列，我们将其定义为：

定义 3.1：如果某个操作序列中接连（可不连续）出现两个或以上的相同操作X，则相应子序列段必定缺失对应操作x，那么，整个序列为不完整序列。

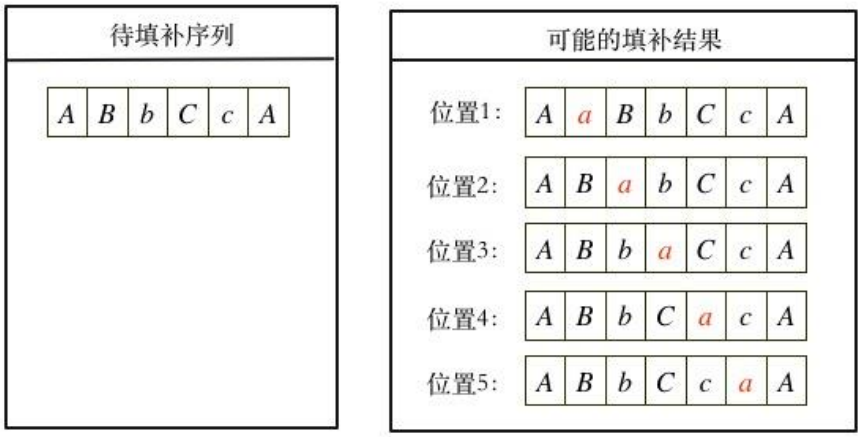

图 3.1 操作序列缺失示意图

3.1.2 操作缺失分析

进一步的分析之后，本文发现残疾人用户在进行设备操作控制时，产生的序列中缺失操作的现象十分普遍，我们统计了研究对象中六个用户的所有操作序列，统计操作序列中完整序列与不完整序列的比例，得到结果如图 3.2 所示，其中上方红色柱状为用户不完整序列的数目，而下方青蓝色柱状为用户完整序列的数目，由图可见，这六个用户产生的操作序列中都存在普遍的操作缺失现象，而不完整序列所占的比例在 35%到 55%之间，这直接反映了操作缺失的普遍性。

在此基础上，我们以用户 1 为例，统计了其在某一段时间内每天的操作序列变化趋势，得到结果如图 3.3，该图横坐标为日期（星期几），红色柱状表示当天生成的操作序列是存在操作缺失的，而青蓝色表示完整序列。该图可以得到的信息为：一是用户 1 产生的操作序列中存在缺失的现象；二是用户 1 对智能家居设备的操作控制存在一定的周期性，在工作日，该用户产生的平均操作次数要远大

于在周末产生的操作次数，并且用户的操作次数以星期为单位呈现规律变化。在进行调研之后发现，该用户的家属在工作日外出上班，因此导致该用户在此时间段对智能家居系统的依赖性较高，会更频繁地使用该系统来对设备进行操作以图方便省力。而到了非工作日，往往残疾人家属在家，很多设备的操作都由残疾人家属完成（比如开关灯），这样的手工操作不被记录，因此，系统采集到的数据也就变少了。

总结图 3.2 和图 3.3，我们可以得出结论：智能家居环境下残疾人用户产生的操作中存在普遍的操作缺失现象。那些由残疾人家属代替残疾人用户手工完成的操作会被丢失，因此，为了充分挖掘残疾人用户的行为习惯，采用合适的技术填补缺失数据是有意义的。

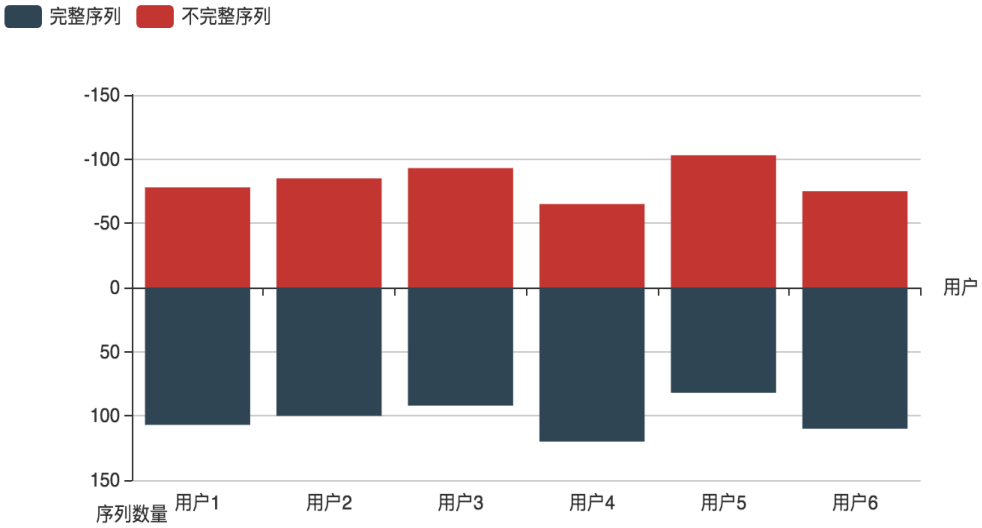

图 3.2 用户完整序列与不完整序列

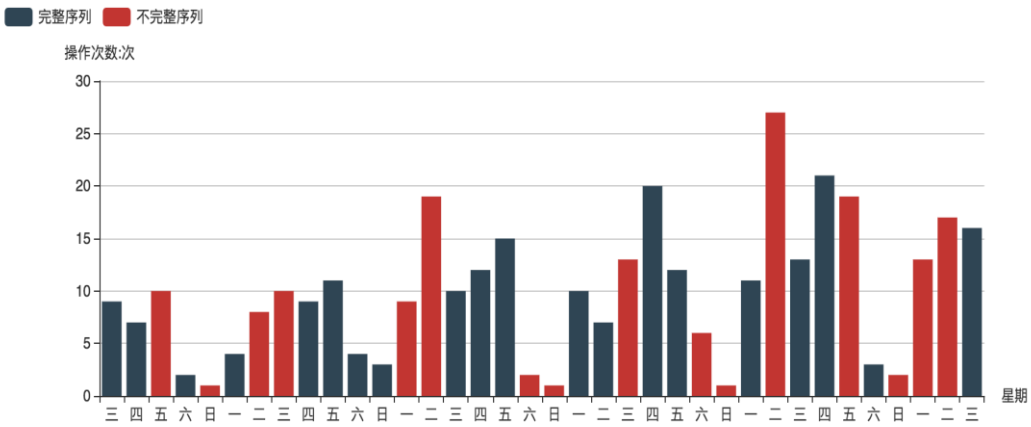

图 3.3 某段时间内用户 1 的操作次数趋势图

3.2 K 最近邻（KNN）算法

3.2.1 算法原理

K 最近邻算法（KNN）是机器学习中最简单常见的分类算法，该算法的原理是在一个特征空间中，如果一个样本的 K 个最相似的样本属于同一类，那么该样本也属于该类。也就是说，决策某一样本属于什么分类时，只关注与其最近的有限个样本（假设都已正确分类）所属的类别。如图 3.4 就是 KNN 分类算法的一个原理图：假设图中红色圆圈表示正类，蓝色圆圈表示负类，则如何判断原点样本的类别？根据 K 最近邻算法，待确定样本的最近邻居为哪一类别，则该样本也属于那一类别，因此，假设 K＝2，则原点应属于负类（概率为2/3），假设 K＝3，则原点应属于正类（概率为3/5），即在一定程度上，K 的取值决定了原点样本所属的类别。

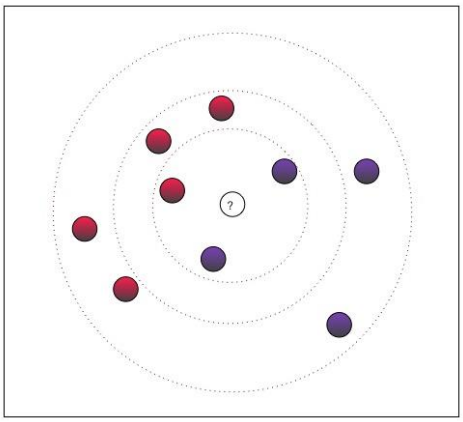

图 3.4 K 最近邻原理图

至此，我们可以推出 KNN 算法的几个关键要素[47]：K 值的选择，距离的度量方法，还有分类规则。首先是 K 值的选择，图 3.4 反映出 K 值的选择直接影响着样本的分类结果。事实上，如果 K 值过小，则由于参考样本过少而容易发生过度拟合现象，因为只有和样本距离最近的小部分“邻居”才影响着样本的分类结果。相反，如果 K 值过大，则意味着与样本距离较远的训练实例也会影响样本的分类结果，这就会导致错误分类。其次是距离的度量方法，距离的度量也就是样本之间相似性的度量，KNN 认为两个样本之间的距离越近，则越为近邻。一般的距离度量方法有 2.3.4 小节介绍的欧氏距离、余弦相似度、皮尔森系数等。最后一个关键要素为分类规则，一般采用多数表决法，即根据样本的 K 近邻的多数类决定该样本所属的类别。

3.2.2 算法过程

K 最近邻算法的实现过程大致可以表示为：输入训练样本集（包含各个样本及其所属的类别），待确定的样本，通过一定的算法规则，输出目标样本所属的类别，算法 3.1 即为 K 最近邻算法的实现过程：

算法 3.1：K 最近邻算法（KNN）

输入：

训练集：$T = \{element_i = (X, Y)\}$，其中$X$和$Y$表示第$i$个训练样本和它所属的类；

待分类样本：x；

最近邻参数：K；

距离函数：$sim(arg1, arg2)$；该方法计算两个参数之间的距离（相似度）；

输出：

待分类样本x所属的类y

算法过程：

1. *simSet* ← *{}, simKSet* ← *{}*
2. *for each* $element \in T$ *do*
3. *calculate* $sim(x, element.X)$
4. *simSet.push(*$sim(x, element.X)$*,element)*
5. *end*
6. *sort(simSet)*//根据sim(x, element.X)的大小降序排列
7. *simKSet* ← *simSet.sub(0,K-1)*//选取 simSet 中前 K 个元素作为最近邻
8. *for each unit* ∈ *simKSet*
9. *count(*$unit.element.Y$*)*//统计 simKSet 中每个类别出现的次数
10. *end*
11. *y* ← *biggest(Y)*//出现次数最多的类即为输出结果
12. *return y*

从上面的算法实现过程可以得知，K 最近邻算法的优点是原理简单，计算方便，且算法的复杂度较低，当训练样本为 N，样本的特征空间维度为 D 时，该算法的复杂度为$O(DN)$。而 K 最近邻算法的缺点是计算量较大，因为对于每一个未分类样本都要计算其与训练集中所有样本的距离。除此之外，当样本不平衡

时，如果某一类样本数量很大，则当输入待预测样本时，预测结果会受到大容量样本的影响。

3.3 基于 KNN 的数据填补算法

3.1 节分析了智能家居环境下残疾人用户存在数据缺失现象，即数据库无法保存一些用户（家属）对设备的手动操作记录。而本节将介绍如何利用 K 最近邻算法填补这些缺失的操作。K 最近邻算法是一种典型的分类算法，但是也可用于回归，也就是说，在特征空间中，假设某一向量的某一属性值缺失，则可以将此缺失值填补为该向量的 K 近邻中这一属性的平均值，以此填补缺失数据。本文就将基于 KNN 算法进行数据填补，大致思路就是首先得到待填补序列的 K 个最近邻（完整序列），然后根据最近邻中目标操作的位置索引的前后关系来确定缺失操作在待填补序列中的位置，最终实现数据填补。

3.3.1 序列相似度定义

K 最近邻算法的基本原理是数据集合中项的距离越近，则相似度越高，在本文，即为序列的距离越近，则相似度越高。由于本文的数据项可表示形如$Seq = \{D, d, E, B, b, E, e\}$的一组离散的操作符号的集合，偏好数值型计算的余弦相似性等相似性计算方法并不适合，因此本文采用 Jaccard 系数[44, 45]计算序列相似性。

Jaccard 系数是一种测量样本集合之间的相似性（差异性）的统计方法，表示为两个集合的交集与并集的比，形如公式(3.1)[46]：

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \tag{3.1}$$

其中，$|A \cap B|$表示集合 A 和集合 B 的交集，$|A \cup B|$表示两者的并集。$Jaccard(A,B)$的取值范围为$0 \leq Jaccard(A,B) \leq 1$，当 A 和 B 均为空集，则$Jaccard(A,B) = 1$。

假设存在序列$Seq1 = \{A, D, d, a, C, E, c, C\}$和序列$Seq2 = \{D, d, A, a, D, d, B, C\}$，根据公式(3.1)，则$Seq1$与$Seq2$的相似性为：

$$sim(Sqe1, Seq2) = Jaccard(Seq1, Seq2) = \frac{|Seq1 \cap Seq2|}{|Seq1 \cup Seq2|} \tag{3.2}$$

其中，$|Seq1 \cap Seq2|$表示两个序列中的共同操作数，$|Seq1 \cup Seq2|$表示两个

序列中操作数的总和。但是，在序列$Seq1$和$Seq2$中都存在重复的属性，如$Seq1$中出现两个C，$Seq2$中出现两个D，两个d，即同一个操作在一个序列中多次出现，为了更好地表示这种操作数叠加的现象，本文对 Jaccard 相似度进行改进，得到公式(3.3)：

$$sim(Sqe1, Seq2) = \frac{\sum_{k=1}^{N} \min(count(Seq1(O_k)), count(seq2(O_k)))}{\sum_{k=1}^{N} \max(count(Seq1(O_k)), count(Seq2(O_k)))} \quad (3.3)$$

其中，$count(Seq1(O_k))$表示序列$Seq1$中操作符O_k出现的次数，k取值为 1 到 N，N 为操作种类（N＝10）。而min()和max()函数分别取最小值和最大值。至此，我们得到了一个改进的基于 Jaccard 系数的操作序列相似性计算方法。

3.3.2 数据填补过程

我们已知在智能家居环境下重症残疾人用户对家居设备的操作过程中存在不同程度的操作缺失现象，本文受到 K 最近邻算法可用于回归填补的启发，提出并实现了一个基于 KNN 的数据填补算法。

图 3.5 为本文的填补过程示意图：首先将每一个用户产生序列分为不完整序列集合和完整序列集合，对于不完整序列集合中的任意序列，我们假设其缺失操作为两个或者两个以上，则我们需要从该不完整序列中提取多个不完整子序列，分别对应每一个缺失操作。分析可知，当缺失操作为多个时，缺失情况分为两种：一为嵌套缺失，二为并列缺失。

定义 3.2：嵌套缺失：即一个序列中存在缺失操作，而在其对应的不完整子序列中还存在一个或多个其他的不完整子序列。形如$sequence = \{\dots, X, \dots, Y, \dots, Y \dots, X, \dots\}$。则从此序列可以分步提取两个对应的不完整子序列，首先提取$subSeq1 = \{Y, \dots, Y\}$，填补之后再提取第二个子序列$subSeq2 = \{, X, \dots Y, \dots, y, \dots, Y, \dots X\}$。

定义 3.3: 并列缺失：即一个序列中存在多个并列的缺失操作。形如$sequence = \{\dots, X, \dots, X, \dots, Y \dots, Y, \dots\}$。则从此序列可以同时提取两个并列的不完整子序列，分别为$subSeq1 = \{X, \dots, X\}$和$subSeq2 = \{Y, \dots, Y\}$。

因此，对于图 3.5 中不完整序列集合中的任一不完整序列，都可以将其拆分为若干个形如$seq = \{X, \dots, X\}$ 的单一不完整子序列。而针对该子序列的填补过程包含四个基本步骤：

第一步：计算不完整子序列$seq = \{X, \dots, X\}$与完整序列集中对应的各个完整

子序列（形如$\{X,\dots,x,\dots,X\}$）的序列相似度（Jaccard 系数）。

第二步：将上述的完整子序列按照相似度进行降序排序，选取相似度最高的前 N 个子序列作为seq的前 N 个最近邻。

第三步：对最近邻集合中各个序列中的x的前一个操作进行计数。

第四步：对第三步的结果进行整理，得到一个填补候选列表，该列表列出了各个操作出现在x前的次数，对其降序排序，选取出现次数最高的操作（称为目标操作），将待填补操作x插入到序列seq中第一个目标操作出现位置的后一个位置中即可。

在该过程中，如果seq中不存在目标操作，则选取出现次数次之的操作为目标操作，以此类推。最后将该不完整子序列填补为完整子序列，并得到一个相对完整序列，判断原序列是否还有操作缺失，若是，则循环上述填补过程，反之，其便成为一个已填补的完整序列。

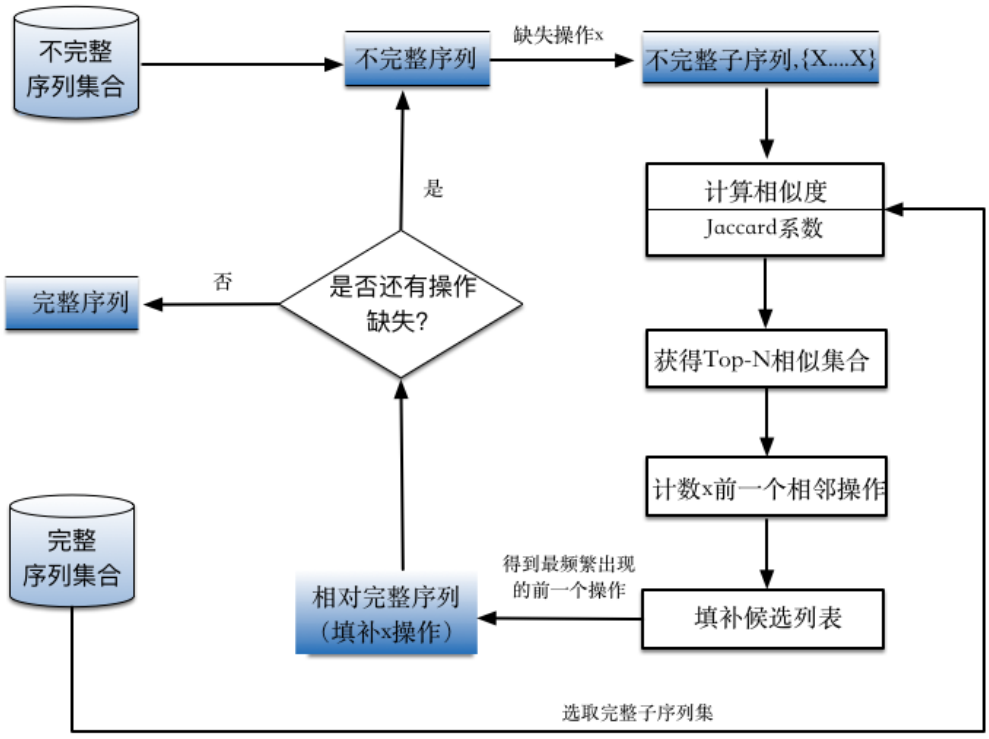

图 3.5 数据填补过程示意图

算法 3.2 描述了对单一不完整子序列$seq=\{X,\dots,X\}$的数据填补过程，其中算法的输入为不完整子序列$seq=\{X,\dots,X\}$，对应的完整子序列集合（形如$\{X,\dots x,\dots,X\}$的序列的集合），最近邻个数 N，通过基于 KNN 的数据填补技术，输出完整子序列$seq'=\{X,\dots,x,\dots X\}$。

算法 3.2：基于 KNN 的残疾人缺失操作填补算法

输入：

单一缺失子序列：$seq = \{X, \dots, X\}$，缺失操作x；

完整子序列集合：$seqs$；

最近邻参数：N；

输出：

完整子序列：$seq' = \{X, \dots, x, \dots, X\}$

算法过程：

```
1.  sims ← {}, simsN ← {}, list ← {}
2.  for each oneSeq ∈ seqs
3.      oneSim ← sim(seq, oneSeq)
4.      sims.put(oneSeq.index, oneSim)
5.  end
6.  sims.sort() //降序排列
7.  simsN ← sims.sub(0, N-1)
8.  for each element ∈ simsN
9.      oneSeq' ← seqs.get(element.getKey())
10.     prev ← before(x) //找到操作 x 前一个操作
11.     list.put(prev, count++) //计数，填入待填补列表
12. end
13. target ← list.first() //第一个元素为目标操作
14. seq.insert(x, target) //将 x 插入到 target 的后面
15. end
```

3.3.3 填补结果评估

针对基于 KNN 的数据填补算法，本文设计了相关实验来验证填补的效果。具体思路为：将每个用户在 186 天中产生的完整序列根据操作类型分解成形如$\{X, \dots x, \dots, X\}$的十种完整子序列（对应本文的十种操作）作为数据集，对于每一种完整子序列集合，进行 10 折交叉验证，将数据集分为 1 份测试集和 9 份训练集，对于测试集中的子序列，形如$seq = \{X, \dots, x, \dots, X\}$，删除操作$x$，根据填补

算法将删除了的操作x填补到相应的位置中，并与原子序列进行比较以评价填补效果。表 3.2 表示实验期间采集的六个用户的操作序列数目，并提取出相应的不同种类的完整子序列的个数。

	不完整序列	完整序列	A-A 完整子序列	a-a 完整子序列	B-B 完整子序列	b-b 完整子序列	C-C 完整子序列	c-c 完整子序列	D-D 完整子序列	d-d 完整子序列	E-E 完整子序列	e-e 完整子序列
用户 1	79	107	220	213	183	177	164	166	232	238	180	192
用户 2	86	100	206	211	190	198	178	172	220	214	188	194
用户 3	93	93	180	184	202	210	223	220	170	170	176	181
用户 4	66	120	240	252	108	117	94	78	260	258	100	109
用户 5	103	83	149	134	78	85	60	65	128	124	107	102
用户 6	75	111	201	206	225	218	230	198	124	102	208	210

表 3.2 数据采集期间各个用户的完整子序列，单位：个

在实验过程中，假设当前的数据集为某用户的$A-A$完整子序列集合，大小为T,则将T个完整子序列随机分为T_C个（一份）测试子序列和T_X个（九份）训练子序列，删除测试集合中的操作a，对于每一条缺失的测试子序列，利用 KNN 填补过程将缺失的操作a填补到合适的位置，并与原来的测试子序列进行比较，若两者相同，则表示填补结果正确，反之，填补结果错误。本实验设置最近邻个数为 20，对于每一组数据集合都进行 10 次交叉实验。此外，我们定义 KNN 填补算法的召回率$Recall$来评估算法的性能，公式如下：

$$Recall = \frac{|T_D|}{|T_C|} \tag{3.4}$$

其中的$|T_C|$即为测试集的大小，$|T_D|$即为测试集中填补正确的序列集合的大小。而$Recall$的取值范围为[0，1]，理想情况下，$Recall$为 1 表示所有的填补结果都正确，而$Recall$为 0 表示所有的填补结果都错误。显然，$Recall$越接近 1，

表示本文的填补算法效果越好。表 3.3 为填补结果，从表中可知，在本文的实验数据上，该算法的填补召回率在[0.66，0.76]之间，整体的召回率约为 0.71，说明该填补算法能够有效地对缺失操作进行填补，因此，我们利用本文的基于 KNN 的数据填补技术对各个残疾人用户产生的不完整序列进行填补，最后得到所有用户的所有的完整序列集合，作为下一章节的实验数据。

用户 ID	召回率										平均召回率 Recall
	A-A 子序列	a-a 子序列	B-B 子序列	b-b 子序列	C-C 子序列	c-c 子序列	D-D 子序列	d-d 子序列	E-E 子序列	e-e 子序列	
用户 1	0.70	0.74	0.77	0.79	0.69	0.82	0.72	0.73	0.76	0.88	0.76
用户 2	0.78	0.76	0.64	0.61	0.72	0.63	0.69	0.69	0.66	0.62	0.68
用户 3	0.71	0.82	0.77	0.76	0.79	0.66	0.69	0.72	0.71	0.67	0.73
用户 4	0.68	0.73	0.74	0.64	0.67	0.70	0.71	0.75	0.68	0.70	0.70
用户 5	0.64	0.70	0.64	0.67	0.58	0.71	0.60	0.67	0.69	0.70	0.66
用户 6	0.80	0.78	0.69	0.62	0.79	0.76	0.70	0.75	0.78	0.73	0.74

表 3.3 基于 KNN 的数据填补算法评估结果，最近邻：20

3.4 小结

本章主要介绍了智能家居环境下残疾人用户普遍存在操作缺失现象，分析了该现象出现的原因。针对操作缺失的发生，本文提出并且实现了一种基于 KNN 的数据填补技术，本章详细描述了该填补算法的过程，最后对填补结果进行实验评估。本章完成了数据填补过程，为下一章节的操作预测部分提供质量良好的数据集。

第四章 基于 HMM 的残疾人操作预测算法

本章扩展了之前章节所讨论的隐马尔可夫模型的原理和应用，对智能家居环境下重症残疾人用户对家居设备的操作序列进行预测。本章节分析了环境（温度）对用户行为的影响，将温度建模为 HMM 的隐含状态，预测随着温度变化用户操作行为的变化。同时，根据温度变化的性质，引入时间因子，改进 HMM 中的状态转移函数。最后本章节描述了一个完整的基于 HMM 的残疾人用户操作预测模型。

4.1 环境（温度）影响用户行为

研究表明，不同温度环境下用户的热感受是不同的。图 4.1 展示了温度变化与用户舒适度的关系[48]，其中横坐标表示温度，纵坐标表示热度感觉。图中的红色区域表示用户感知为“太暖和”，绿色区域表示用户感知为“舒适”，蓝色区域表示用户感知为“太冷”。从纵向看，当温度为 20℃左右时，用户感到“舒适”的概率最高，而感到太冷或者太热的概率较小。而随着温度的升高，用户感到“太暖和”的概率渐渐增大，而感到“太冷”的概率在不断减小。

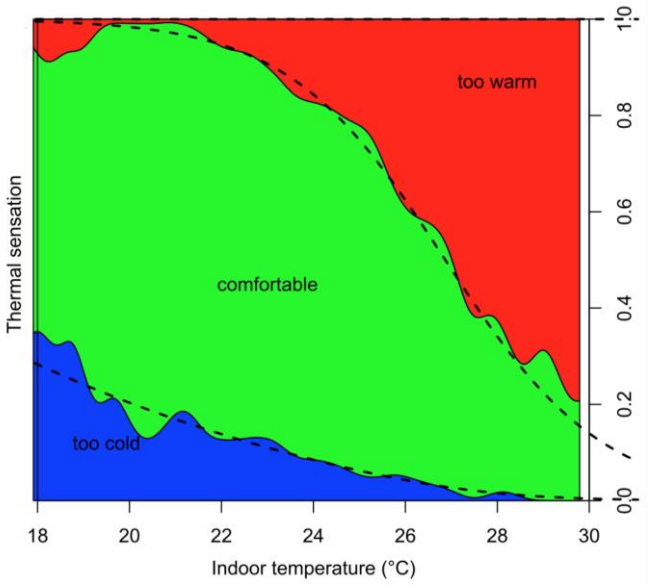

图 4.1 温度与用户舒适度关系图[48]

事实上，当室内环境发生变化时，处于室内环境中的用户会对这样的变化做出应激反应以保持自身的舒适度[49]。也就是说，居住者并不是室内环境变化的被动接受者，相反，当室内环境改变（如温度升高，光照变强）导致居住者感到

不舒服，他们会做出一些行为来保持自身的舒适感。这些行为涉及到与窗户、照明设备、恒温设备等的交互来调整室内环境[50]。如 Gunay 等人[50]就发现根据光照强度的改变，居住者会通过调整百叶窗来保持室内光照和温度的舒适度，因此提出一个自适应光照强度的百叶窗自动控制算法。

而在智能家居环境下，重症残疾人用户对家居设备的控制也会随着温度的变化而做出相应的改变。如当清晨温度较低时，用户的操作主要为开窗帘、开窗户等，而随着温度渐渐升高，我们发现残疾人用户打开电扇，空调等的次数就会明显增多，并且这些操作会频繁反复地发生。也就是说，随着温度的改变，残疾人用户对设备的操作也会作出相应的改变，以保持自己在智能家居环境下的舒适度。因此，本文在研究智能家居环境下残疾人用户的行为特征时，将温度建模为隐马尔可夫模型中的隐含状态，将用户对家居设备的操作建模为隐马尔可夫模型中的观察状态，从而研究环境（温度）变化下残疾人用户的操作序列的变化。

4.2 基于 HMM 的残疾人操作预测

HMM 已经被广泛应用于时间序列的分析和预测，典型的应用场景包括语音识别[51]，心电图分析[52]等。本节将采用原始的 HMM 方法去预测智能家居环境下残疾人用户对家居设备的操作。该方法的分为以下步骤：第一步，建立初始的 HMM，并使用训练集数据重估参数。第二步，预测给定局部序列的下一步最有可能的操作，该过程分为两小步：一是根据下一步操作有 N 种可能，得到 N 个局部序列（原局部序列加上下一步操作），并计算这些序列在此 HMM 中出现的概率，二是得到概率最大的序列所对应的下一步操作，即为预测值。图 4.2 即为具体步骤。

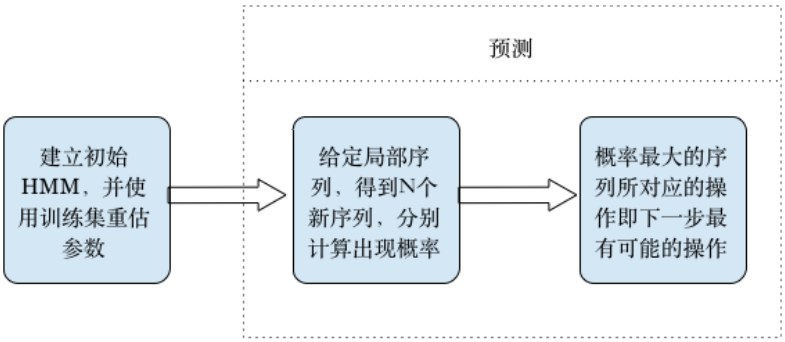

图 4.2 HMM 预测用户操作的步骤

4.2.1 训练 HMM

由于用户行为特征的多样性，本文分别为每个用户训练个性化的 HMM。该过程描述为：随机初始化一个原始的 HMM，此时假设训练集中有L条完整操作序列，记为$Seqs = \{O(1), O(2), \dots, O(L)\}$，则利用这一组训练序列去不断重估原始 HMM 的各个参数，每次得到一个新的 HMM，使得新的 HMM 越来越能准确地描述训练集。形式化的表示方式为：找到一个 HMM，记为λ，使得$P(Seqs|\lambda)$最大化。一般的，我们采用 Baum-Welch 算法来训练 HMM，算法的具体原理已经在第二章的 2.2.3 节给出。

下面着重 HMM 训练算法的实现过程：我们已知 HMM 包含一组隐含状态S，一组观察值O，状态转移概率矩阵$A = \{a_{ij}\}$，发射概率矩阵$B = \{b_{jk}\}$以及初始概率分布$\pi = \{\pi_i\}$。图 4.3 就是隐马尔可夫链，包含一条隐含链和一条观测链，整个过程中观察值随着隐含状态的改变而生成。

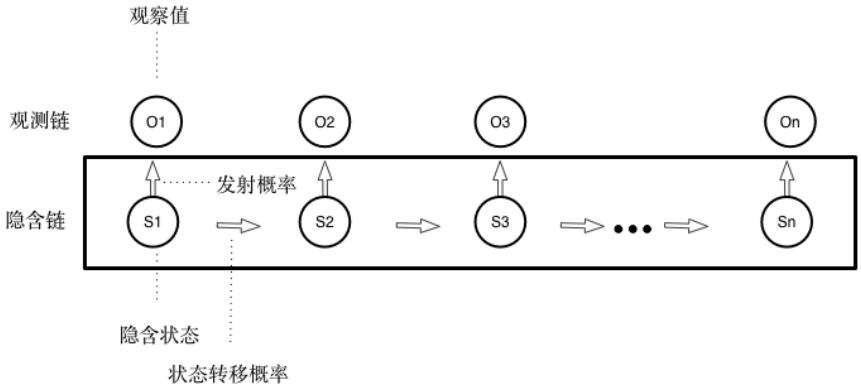

图 4.3 HMM 观测链与隐含链

HMM 的训练过程即为确定上图中各个参数的过程：

- 隐含状态S：本文选用温度作为隐含状态，将采集到的温度（如温度范围为[T_s℃，T_e℃]）以相同的间隔离散化为 N（如 N＝5）个状态，则分别用S_0，S_1，S_2，S_3，S_4表示。
- 观察状态O：表 3.1 已定义本文的观察值，即对五个特定家居设备的开关操作，表示范围从$\{A-a\}$到$\{E-e\}$，对应观察状态$O_0, O_1, \dots, O_8, O_9$。
- 初始概率分布π：一般地，可以随机选取π_i的数值使之符合$\sum_{i=1}^{N}\pi_i = 1$。本文统计了数据集中各个温度状态出现的频率，将此作为初始的概率以减小初始误差。

- 状态转移概率A：已知状态转移函数为$A=\{a_{ij}\}$。其中a_{ij}表示这一时刻状态为S_i，下一时刻状态为S_j的概率，这一概率可以表示为$P=\frac{\text{状态转移}S_i\to S_j\text{的次数（遍历整条链）}}{\text{状态转移从}S_i\text{出发的次数}}$。
- 发射概率B：已知状态转移函数为$B=\{b_{jk}\}$。其中b_{jk}表示当前的隐含状态为S_j，观察值为O_k的概率，这一概率可以表示为$P=\frac{\text{隐含状态}S_j\text{，观察值}O_k\text{的次数}}{\text{隐含状态为}S_j\text{的次数}}$。

明确各个参数的物理意义之后，算法 4.1 给出了 Baum-Welch 的具体过程，其中输入为一组观察序列$Seqs$，输出为隐马尔可夫模型。算法初始化一个 HMM 为λ，每次更新得到一个新的λ^*，且$P(Seqs|\lambda^*)>P(Seqs|\lambda)$，直至收敛得到一个满意的 HMM。

算法 4.1：Baum-Welch 训练 HMM 算法

输入：

一组观察序列：$Seqs=\{O(1),O(2),\dots,O(L)\}$；

迭代次数：$nIterations$；

输出：

隐马尔可夫模型：λ^*

算法过程：

1. 初始化模型$\lambda(\pi,A,B)$，随机给定参数π_i，a_{ij}，b_{jk}，使其满足条件$\{\sum_{i=1}^{N}\pi_i=1,\sum_{j=1}^{N}a_{ij}=1,\sum_{k=1}^{M}b_{jk}=1\}$，令训练次数$n=0$;
2. *do*
3. n ← n + 1;
4. 根据公式（*2.8*），（*2.9*），计算$\xi_t(i,j)$和$\gamma_t(i)$;
5. 将上述计算的$\xi_t(i,j)$和$\gamma_t(i)$代入公式（*2.10*），（*2.11*）和（*2.12*），得到π^*,a_{ij}^*,b_{jk}^*，从而得到一个新的 HMM，表示为$\lambda^*=(\pi^*,A^*,B^*)$;*//具体计算过程可见算法 4.2 与算法 4.3*
6. *while*$(n<nIterations)$;
7. *return* λ^*;

上述算法的关键步骤即为每次更新参数a_{ij}^*和b_{jk}^*，回顾公式（2.11）：$a_{ij}^*=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$和公式（2.12）：$b_{jk}^*=\frac{\sum_{t=1,o_t=k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$。可以发现两者都涉及变量$\xi_t(i,j)$和变量$\gamma_t(i)$，其中，$\xi_t(i,j)$的物理意义为：给定 HMM，序列在$t$时刻状态

为S_i，而在$t+1$时刻的状态为S_j的概率。$\gamma_t(i)$的物理意义表示为t时刻状态为S_i的概率。a_{ij}^*可表示为：纵观整个序列，在所有时刻上从状态S_i转移到状态S_j的期望次数除以在所有时刻上从状态S_i转移出去的期望次数。b_{jk}^*可表示为纵观整个序列，从状态S_j下观察到操作为k的期望次数除以从其他状态转移到状态S_j的期望次数。算法 4.2 和算法 4.3 分别给出了$\xi_t(i,j)$和$\gamma_t(i)$的计算过程，而算法 4.4 给出了更新参数a_{ij}^*和b_{jk}^*的过程。

算法 4.2 对参数ξ进行估计，其中的已知参数$alpha$和$beta$分别表示前向变量（见算法 4.5）和后向变量（与算法 4.5 类似，可自行推导），该算法以算法 4.5 为基础。

算法 4.2：参数估计ξ(estimateXi 算法)

输入：

一个观察序列：$seq = \{O(1), O(2), \dots, O(T)\}$;

隐马尔可夫模型：hmm，（其中 A[][]表示状态转移矩阵，B[][]表示发射矩阵）；

输出：

参数ξ的矩阵：*xi*[][][];

算法过程：

```
1.  init xi[][][]
2.  for(t = 1 to seq.size()-1)
3.      subSeq ← seq[1,t]
4.      prob ← forward(subSeq)//算法 4.5
5.      for(i = 1 to hmm.N)
6.          for(j = 1 to hmm.N)
7.              xi[t][i][j]←alpha[t][i]*A[i][j]*B[j][O_{t+1}]*beta[t+1][j]/prob
8.          end
9.      end
10. end
11. return xi[][][]
```

算法 4.3 对参数γ进行估计，其中的已知参数xi为算法 4.2 的结果。

算法 4.3：参数估计γ(estimateGamma 算法)

输入：

参数ξ的矩阵：xi[][][];

隐马尔科夫模型：*hmm*;

输出：

参数γ的矩阵：*gamma*[][];

算法过程：

```
1. init gamma[][]
2. for(t = 1 to xi.length)
3.     for(i = 1 to hmm.N)
4.         for(j = 1 to hmm.N)
5.             gamma[t][i] ← gamma[t][i]+xi[t][i][j]
6.         end
7.     end
8. end
9. return gamma[][]
```

算法 4.4 给出了根据上述两个参数ξ和γ，对 HMM 模型中的π^*, a_{ij}^*和b_{jk}^*进行重估的过程。其中参数xi为算法 4.2 的结果，参数$gamma$为算法 4.3 的结果。

算法 4.4：更新 HMM 模型

输入：

一组训练序列：$seqs = \{seq1, seq2, \ldots\}$;

隐马尔科夫模型：*hmm*;

输出：

新隐马尔可夫模型：*nhmm*;

算法过程：

```
1. init aijNum[],aijDen[],allGamma[][][]; // aijNum 为更新a_ij^*的分子，
```

aijDen 为更新a_{ij}^{*}的分母。allGamma 为gamma 数组

```
2.  for(each seq in seqs)
3.      update xi[][][] according to  算法4.2
4.      update gamma[][] according to  算法4.3
5.      allGamma[i++]←gamma[][]
6.      for(i=1 to hmm.N)
7.          for(t=1 to seq.size()-1)
8.              aijDen[i] ←aijDen[i]+gamma[t][i]
9.              for(j=1 to hmm.N)
10.                 aijNum[i][j]←aijNum[i][j]+xi[t][i][j]
11.             end
12.         end
13.     end
14. end
15. //更新a*ij
16. for(i=1 to hmm.N)
17.     for(j=1 to hmm.N)
18.         nhmm.setAij(i, j, aijNum[i][j]/aijDen[i]);
19.     end
20. end
21. //更新π*
22. for(o=1 to seqs.size)
23.     for(i=1 to hmm.N)
24.         nhmm.setPi(i, nhmm.getPi(i)+ allGamma[o][1][i]/seqs.size)
25.     end
26. end
27. //更新b*jk
28. for(o=1 to seqs.size)
29.     for(i=1 to hmm.N)
30.         for(k=1 to hmm.M)
```

```
31.                 bjkNum ←0
32.                 for(t=1 to seq.size)
33.                     if(seq[t] == k)
34.                         bjkNum ← bjkNum+allGamma[o][t][i]
35.                     end
36.                     sum←sum+ allGamma[o][t][i]
37.                 end
38.                 nhmm.setBjk(j, k, bjkNum/sum)
39.             end
40.         end
41. end
42. return nhmm
```

至此，我们详细介绍了 HMM 训练算法，值得注意的是，Baum-Welch 算法执行过程中，会不断重估 HMM 的参数，直到两次 HMM 之间的距离（常见的 KL 距离）小于某个阈值（如$criterion = 0.05$）或者迭代次数达到某一阈值（如$iteration = 60$）。设定的距离过大或者迭代次数过少会导致得到的 HMM 不能很好地描述训练集，而设定的距离太小或者迭代次数过多又会产生过拟合现象，因此，阈值的选择也十分重要。实验表明在本文的数据集上当迭代次数为 60 时效果最佳。

4.2.2 预测操作

基于上一小节得到的最能描述给定序列的 HMM，本小节将介绍如何利用该训练好的HMM预测局部序列的下一步最有可能的操作。假设当前的HMM为λ，局部序列为$seq = \{ABDdEbeC\}$，如何预测下一步最有可能的操作？该预测过程主要分为两步，一是计算所有可能的新序列在该 HMM 中出现的概率，二是得到概率最大的序列所对应的操作，即为下一步最有可能的操作。

具体过程如下：

当前的局部序列为$seq = \{ABDdEbeC\}$，由于操作种类有十个，因此理论上下一步操作有可能为十种里面的任一种，如图 4.4 所示。因此可以得到十个候选序列（假设分别为seq_1到seq_{10}），接下来就分别计算这十个候选序列在λ中出现的概率即可，也就是 2.2.2 节中介绍的评价问题：已知 HMM 的模型参数，计算

给定可观察序列的概率。

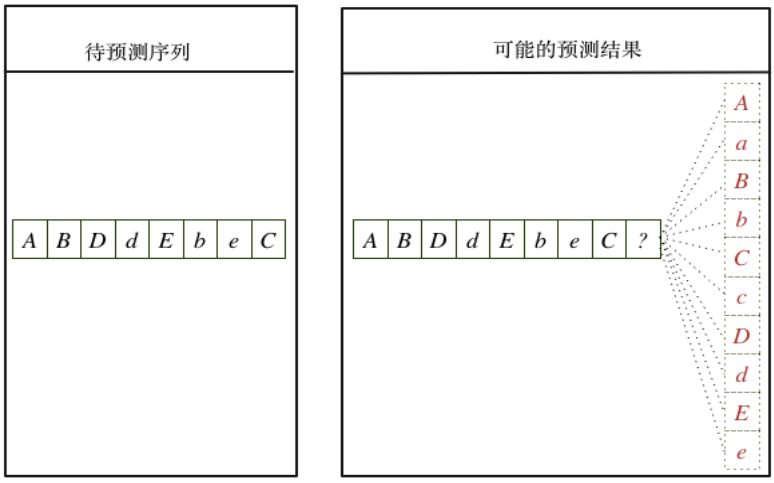

图 4.4 操作序列预测示意图

前向算法用于计算上述概率值：已知seq的长度为 8，即总时刻$T=8$，目标为计算$t=9$时刻的操作。由于十个候选序列的前八个操作是相同，所以计算过程一致：首先，计算$t=1$时刻的局部序列$seq=\{A\}$的概率，由公式（2.1）计算可得此时该序列在λ中出现的前向概率α_1。接着计算$t=2$时刻的局部序列$seq=\{AB\}$的概率，由于$t=2$时刻的状态由$t=1$时刻转变而来，且每一时刻有$N(N=5)$种可能的隐含状态，则该时刻的概率可表示为前一时刻的前向概率α_1与包含这一时刻所有状态的概率的联合概率，公式（2.2）可计算出此时该序列在λ中出现的前向概率α_2，同理，我们可得$t=3$时刻的前向概率α_3，直至$t=8$时刻的前向概率α_8，α_8表示了序列$seq=\{ABDdEbeC\}$在λ中出现的概率。接着计算$t=9$ 时刻seq_1，seq_2，…，seq_{10}的概率，则同理利用$t=8$时刻的前向概率α_8来分别计算各个可能的候选序列的前向概率，得到α_9的十个可能的值，分别记为α_9^1，α_9^2，…，α_9^{10}，从中得到最大值，假设最大值为α_9^2，则对应的操作a为$t=9$时刻最有可能的操作，至此，预测过程完毕。

算法 4.5：前向算法计算序列概率(forward 算法)

输入：

一个观察序列：$seq=\{O(1),O(2),\dots,O(T)\}$；

隐马尔可夫模型：hmm，（其中 A[][]表示状态转移矩阵，B[][]表示发射矩阵）；

输出：

seq出现在hmm中的概率：$prob$

算法过程：

```
1.  //初始化，t=1
2.  for(i = 1 to hmm.N)
3.      alpha[1][i]←hmm.pi[i]*hmm.B[i][O1]   //t=1 时刻的前向变量
4.  end
5.  //迭代，根据 t 时刻求 t+1 时刻的前向变量
6.  for(t = 1 to T-1)
7.      for(j = 1 to N)
8.          sum←0.0
9.          for(i = 1 to N)
10.             sum←sum+alpha[t][i]*hmm.A[i][j]
11.         end
12.     alpha[t+1][j] ← sum*hmm.B[j][Ot+1]
13.     end
14. end
15. //计算序列的概率
16. prob←0.0
17. for(j=1 to N)
18.     prob←prob+alpha[T][j]
19. end
```

4.3 优化状态转移函数

4.3.1 HMM 中的时间信息

隐马尔可夫模型中有两个重要的假设，分别是有限历史性假设和齐次性假设[53]。有限历史性假设即为 2.1 节中介绍的马尔科夫性质，即某一时刻系统状态的概率分布只与它前一时刻的状态有关。而齐次性假设表示任意时刻的状态转移不会随着时间的改变而改变，即隐马尔可夫模型中的状态转移矩阵是不变的。近些年来，为了提高模型的应用效果，研究者往往会对上述假设提出改进。高阶隐马尔可夫模型可以认为是对有限历史性假设的扩充，即我们在考虑系统处于某一时刻状态的概率分布时，不再是仅仅考虑它的前一时刻的状态，而是考虑前N个时刻的状态，即N阶隐马尔可夫模型。当然高阶隐马尔可夫模型也存在一定的问

题：如随着阶数增大，系统的开销会大大增加。而另一方面，也可以对齐次性假设进行扩展，如[54]介绍了一种非齐次的隐马尔可夫模型用于语言建模过程并取得了较好的效果。

考虑本文的隐马尔可夫模型中，隐含状态表示为温度，则齐次性假设为隐含状态的转移过程与时间无关，即隐含状态的转移概率是不变的。但事实上，一个直观的例子就是在一天中，清晨温度从 20℃升高到 22℃的概率要比下午温度从20℃升高到22℃的概率高得多，也就是说，相同的状态转移发生在不同时刻的概率是不一样的。因此，本文试图将这种变化的特征引入隐马尔可夫模型中，添加时间信息，扩展齐次性假设，构造出一个非齐次的隐马尔可夫模型。实验证明本文的非齐次的隐马尔可夫模型在预测残疾人用户的操作时比传统的隐马尔可夫模型具有更高的准确率。

时间信息表示了状态出现的一种先后顺序关系，从隐马尔可夫模型的隐含链来看，状态转移的过程本身就是一种时间序列，为了形象化地表示这样的时间关系，我们引入了状态的位置索引，因为对于一条具体的隐含链，状态之间的相对位置关系就显示了它们出现的先后顺序。那么如何将位置索引与状态转移矩阵联系起来？当然，我们不能盲目地认为位置越靠前则状态转移概率越高或者位置越靠后则状态转移概率越高。事实上，我们引入状态位置的平均值和标准差来表示位置与状态转移的关系。如下面的例子：

假设存在隐含序列$seqHidden = \{S_1, S_2, S_3, S_2, S_1, S_2\}$；则该隐含序列对应的位置序列为$seqPos = \{0,1,2,3,4,5\}$，我们发现状态转移$S_1 \to S_2$发生在位置 0 和位置 4，则这个状态转移的平均位置为$\frac{(0+1)/2+(4+5)/2}{2} = 2.5$，标准差为$\sqrt[2]{\frac{1}{2}((0.5-2.5)^2 + (4.5-2.5)^2)} = 2$，也就是说，状态转移$S_1 \to S_2$发生的平均位置索引为 2.5，而这个状态转移发生的离散程度为 2。为了进一步挖掘各个状态转移的位置关系，我们从中国气象网站[55]采集了用户 1 在一段时间内每次操作设备时对应的上海市天气数据，并将采集到的温度[18℃,38℃]均匀划分为五个温度段，从低到高表示为$\{S_0, \dots, S_4\}$，表 4.1 和表 4.2 分别为用户 1 的状态（温度）转移发生的位置的平均值和标准差，而表中的值为-1.0 时表示不存在这样的状态转移过程。

结束状态 / 起始状态	S_0	S_1	S_2	S_3	S_4
S_0	6.50	4.00	-1.0	-1.0	-1.0
S_1	9.50	8.19	1.83	2.50	-1.0
S_2	-1.0	7.10	6.07	3.00	-1.0
S_3	-1.0	4.50	7.75	6.75	2.50
S_4	-1.0	-1.0	2.50	6.50	3.10

表 4.1 状态转移发生的位置的平均值（user=用户 1，隐含状态数 N=5）

结束状态 / 起始状态	S_0	S_1	S_2	S_3	S_4
S_0	4.75	3.84	-1.0	-1.0	-1.0
S_1	2.45	5.01	0.47	0.82	-1.0
S_2	-1.0	2.94	2.96	0.50	-1.0
S_3	-1.0	0.32	2.95	1.96	0.32
S_4	-1.0	-1.0	0.32	0.32	1.85

表 4.2 状态转移发生的位置的标准差（user=用户 1，隐含状态数 N=5）

表 4.1 是状态转移过程中两个相邻状态的位置的平均值，它反映了这对状态在整个马尔科夫过程中的均值位置。其中，均值越小表示这对状态的转移越有可能发生在序列的靠前位置，反之，均值越大表示这对状态的转移越有可能发生在序列的靠后位置。从表 4.1 还可以得出的结论是当我们在预测隐马尔可夫过程时，如果一对状态的位置越接近训练集中统计的关于这对状态的位置均值，那么表示此时这对状态的转移是越合理的。但是，当一对状态的位置离训练集中的位置均值较远时，并不代表这样的转移就是不合理的，因为有些状态转移会频繁地发生在不同的位置。因此，我们引入标准差来辅助判断状态转移的合理性，表 4.2 即为状态转移过程中两个相邻状态的位置的标准差，它反映了这对状态在整个马尔科夫过程中的活跃度，其中，标准差越小，表示这对状态越稳定，反之，表示这对状态越活跃。也就是说，随着标准差的增大，位置信息对状态的约束就会越小，

状态转移发生在不同位置上的合理性也就越高。

4.3.2 三维状态转移矩阵

回顾 2.2.1 中 HMM 的定义，状态转移函数为$f = a_{ij}$，其中只有两个变量i和j分别表示两个状态，为了加入状态间的位置信息，我们引入了第三个变量t来构造一个非静态的状态转移函数f'：

$$f' = a_{ijt} = a_{ij} \times \omega \tag{4.1}$$

其中t表示状态转移$S_i \to S_j$发生时S_i的位置索引，而ω为表示状态活跃度的缩放系数，本文定义ω如下：

$$\omega = \begin{cases} 1.0, & SD(i,j) = -1.0 \\ e^{\frac{SD(i,j)}{|p-\bar{p}|}}, & 0 \le \frac{SD(i,j)}{|p-\bar{p}|} \le ln1.5 \\ 1.5, & \frac{SD(i,j)}{|p-\bar{p}|} > ln1.5 \end{cases} \tag{4.2}$$

其中，$SD(i,j)$表示状态转移$S_i \to S_j$发生在训练集中的位置的标准差。p表示这对状态的当前位置，可表示为$p = (t + t + 1)/2$，$\bar{p}$表示这对状态在训练集中的位置的平均值。原则上，$\frac{SD(i,j)}{|p-\bar{p}|}$的取值范围为 0 到正无穷，这里，我们设置它的取值范围为 0 到$ln1.5$，即当它的值大于$ln1.5$时取$ln1.5$。那么，ω的最终取值范围为 1 到 1.5，也就是说，HMM 中的静态的状态转移矩阵中每个元素都有 1 到 1.5 倍的缩放倍数。

至此，我们添加了时间信息t，将状态转移函数从二维的a_{ij} 转变为三维的a_{ijt}，也就是说状态转移发生的位置影响了状态转移的概率，扩展了齐次性假设，得到了一个非静态的隐马尔可夫模型，实验证明本文的非静态的隐马尔可夫模型在预测智能家居环境下残疾人用户的操作序列时表现出了更好的预测效果。

4.4 残疾人行为操作预测模型

本文从智能家居环境下重症残疾人对家居设备的操作行为的需求入手，设计了一个基于隐马尔可夫模型的残疾人用户操作预测模型，解决了残疾人用户对设备控制过程中用户的肢体障碍性与设备控制复杂性之间的矛盾。具体工作原理为：利用用户的以往操作记录来预测将来最有可能发生的操作。该模型如图 4.5 所示，主要由两部分组成：基于 KNN 的数据填补模块和基于 HMM 操作预测模块。数据填补模块包含计算相似度，获得相似序列集合，填补缺失数据三个步骤，而操作预测模块包括训练 HMM，计算序列出现的概率，预测下一步操作三个步骤。

具体过程为：将用户产生的操作数据进行初步的整理与清洗，然后进入基于KNN 的数据填补模块，填补不完整序列中的缺失操作，得到完整的数据集，并将之分为训练集与测试集，训练集进入基于 HMM 的操作预测模块进行 HMM 的训练，并用测试集的数据进行实验验证，最后将预测结果返回给残疾人用户。

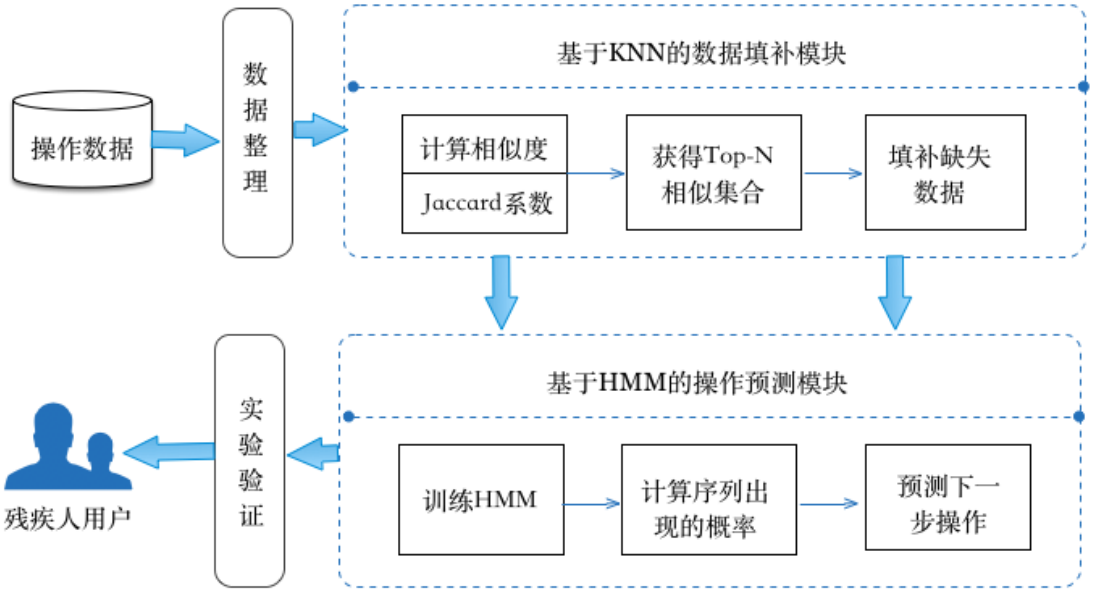

图 4.5 残疾人操作预测模型示意图

4.4.1 数据填补模块

数据填补模块的主要作用是填补残疾人操作设备过程中产生的缺失数据，由于本文的数据是上海市若干残疾人家庭在智能家居环境下产生的真实原始的数据，存在一定的质量问题（如数据杂乱冗余、操作缺失等），因此本文将数据进行初步的整理清洗之后，对缺失操作进行了填补。

由图 4.5 可知该模块有三个主要步骤：

计算相似度：该步骤主要用于计算序列之间的相似度，相似度计算方法为 Jaccard 系数，根据序列之间共有的操作来评判序列的相关度。

获得 Top-N 相似集合：该步骤主要将上一步骤中的结果进行排序，得到待填补序列的 N 个最相似的序列，组成数据填补的参考集。

填补缺失数据：该步骤将为缺失序列进行填补，根据上一步骤得到的相似集合，得到缺失操作在参考集的序列中出现的位置，以此作为依据将缺失操作填补到缺失序列的相应位置。

该填补模块的具体工作原理与工作过程已在第三章详细介绍，最后，该模块得到完整的序列集合，构成本

文章最终的实验数据，我们将其划分为训练集和测试集用于后续的实验。

4.4.2 操作预测模块

操作预测模块主要用来预测给定序列的接下一步最有可能的操作。该模块的数据来源于上一模块得到的完整数据集，其中的训练集用于算法的训练，得到一个最符合每个用户的行为习惯的 HMM，而测试集用于验证本文算法的准确率。

由图 4.5 可知该过程也有三个步骤：

训练 HMM：4.2 节给出了实现训练过程的具体细节，Baum-Welch 算法会根据给定序列训练出一个最佳的 HMM 模型，算法的核心在于各个参数的迭代更新。

计算给定序列出现的概率：该步骤基于上一步骤训练出的 HMM，计算给定序列在该 HMM 中出现的概率。

预测下一步操作：将上一步骤中得到的各种可能的操作所对应的序列的概率进行排序，得到最有可能的操作，从而完成预测。

本章节的前面三小节也具体给出了整个预测模块的工作原理与具体细节。该模块结束后，进入实验验证环节，即通过测试集去验证算法的准确率，最终将该预测结果返回给残疾人用户。

4.4.3 通用性分析

本文提出的用户操作预测模型虽然是基于智能家居环境下残疾人的用户行为特征的，但仍具有一定的通用性和借鉴意义。首先，从应用场景看，本文预测了智能家居环境中残疾人用户的操作，但是依然适用于预测普通用户的操作行为。其次，从研究方法上看，基于 KNN 的数据填补技术可以指导其他研究过程中的数据预处理环节。而隐马尔可夫模型在时间序列的分析方面一直是一个重要的研究方法，因此，在其他用户行为序列的分析时也可借鉴本文的预测方法。

4.5 小结

本章节核心介绍了智能家居环境中残疾人用户对家居设备的操作预测过程，

从输入特征的选择（温度）到模型的训练过程（Baum-Welch 算法的具体过程），然后引入时间信息，改进 HMM 中的状态转移矩阵，得到三维的状态转移函数。最后提出了一个初步完整的残疾人行为预测模型并进行了简单的分析，而下一章节的将对本章节的工作进行实验检验。

第五章 实验设计与分析

针对本文提出的智能家居环境下重症残疾人对家居设备的操作预测模型，本章节设计了相关实验来验证该预测模型的有效性。实验数据来源于上海市若干残疾人家庭中真实的设备操作数据。实验内容主要为评估本文的预测模型是否能准确地预测出给定操作序列的下一步最有可能的操作，通过预测结果的精确率来评估预测的质量。最后分析了本文的预测模型的实验结果。

5.1 实验数据

为了更好地评估本文的预测模型的有效性，我们采集了杨浦区若干个残疾人用户产生的真实的设备操作数据来设计相关实验。值得注意的是，本文提出的智能家居环境下残疾人操作的预测模型是个性化的，即预测模型会学习每个用户的操作习惯后对其后续的操作行为进行预测，而基于每个用户产生的操作的实验结果显示了该预测模型对该用户模拟学习的效果，这暗示了本文的实验中用户之间是相互独立的，且实验与用户数目的多少无关。因此，本文采集了六个残疾人家庭在 2015 年 7 月 14 号至 10 月 14 号和 2016 年 7 月 14 号至 10 月 14 号两年之间共 186 天产生的对家居设备的操作数据，其中，设备操作类型已在表 3.1 给出，一共有五种设备对应的十个操作。另外，本文从中国气象网站[55]采集了上述时间段内的上海市白天气温，得到温度范围为[18℃,38℃]。最终，所有用户对设备的操作数据和对应的天气数据构成了本文的数据集。

5.1.1 数据预处理

数据预处理过程分为两个阶段：

第一阶段为初步的数据整理：首先，将采集到的六个用户的操作数据进行整理，按照时间关系，将每天产生的操作形成一条操作序列。如某天用户 1 产生的操作为<开窗帘，开窗，开灯，关灯，开电扇，关电扇，开空调，关窗>，则可生成对应的序列为$seq = \{EBAaDdCb\}$。最终，生成对应每个用户每天的操作序列，形成操作序列集。其次，将采集到的温度范围[18℃,38℃]进行均匀地离散化，根据 HMM 训练过程中的具体情况，可分成若干个温度段，假设训练过程定义隐含状态个数为 5，则可按照每隔 4℃的温度间隔将上述温度划分为 5 个温度段，分别用S_0, S_1, S_2, S_3, S_4表示。

第二阶段为缺失操作的填补：该阶段主要对存在缺失操作的序列进行数据填补，第三章详细介绍了操作缺失的情况和填补算法，简而言之为利用 KNN 算法将缺失操作填补到序列的相应位置中去，最终形成完整序列。需要说明，基于 KNN 的填补算法的评估实验详见 3.3.3，本章节主要设计验证预测算法有效性的实验，而实验数据来源于填补算法产生的结果数据集。

5.1.2 训练集与测试集

通过数据预处理（数据填补）之后得到的完整的操作序列集合作为本章实验的数据集，我们将数据集划分为训练集和测试集。在数据划分过程中，我们依照 K（本文 K＝10）折交叉验证原理，将数据集合等分为 10 个数据子集合，每次实验时，取其中的一份为临时的测试集，另外九份为临时的训练集，实验时用测试集去验证训练集训练得到的预测模型的效果。整个实验过程重复进行 10 次。

5.2 实验设计

5.2.1 实验内容

本章的实验内容是针对每一个用户，利用本文的基于改进的隐马尔可夫预测模型对训练集进行训练，得到最能描述该用户行为特征的模型，然后利用测试集评估训练模型的精确率。一次具体的实验过程分为训练过程和测试过程，训练过程的输入为训练集，输出为一个 HMM。而测试过程中，针对每一条测试序列，假设长度为 N，则可以进行 N－1 次测试。图 5.1 显示具体的测试次数，假设当前测试序列为$seq = \{EBAaDdCb\}$，长度为 8，则可以进行 7 次测试。每次测试时都基于前面的有限个操作组成的局部序列去预测下一步的操作，如果预测的操作与实际的后一个操作相同，则表示预测成功，否则预测失败。

需要说明，如果预测结果是正确的，则在后续的研究中可考虑智能家居系统直接代替残疾人用户去执行这个操作，而残疾人用户只需被动接受服务；或者智能家居系统将预测结果推荐给残疾人用户，用户去决定是否执行这个操作。而如果预测结果是错误的，则在后续的研究中可考虑当非期望操作发生时，残疾人用户通过控制器（手机）发出语音或者指令命令去重新执行正确的操作。

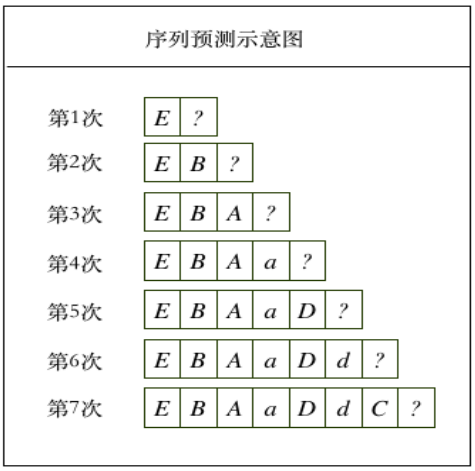

图 5.1 序列预测示意图

5.2.2 实验评估

本文采用精确率(precision)来评估预测效果。本文的精确率可表示为预测的总次数中正确预测数所占的比例。我们令$|P_{correct}|$表示预测正确的次数，而$|P_{incorrect}|$表示预测错误的次数，则精确率可表示为：

$$precision = \frac{|P_{correct}|}{|P_{correct}|+|P_{incorrect}|} \tag{5.1}$$

由公式定义可知，$precision$的取值范围为[0，1]，其值越接近 1 表示精确率越高，预测的效果越好，反之，表示精确率越低，预测的效果越不好。

5.3 实验结果分析

本文的实验结果分析主要分为两部分，一是对训练算法本身的评估，包括 HMM 训练过程中隐含状态的个数选择，算法执行过程中的收敛条件等。二是评估基于三维状态转移函数的改进模型的预测性能，比较不同参数条件下的模型效果。

5.3.1 算法参数评估

本文 4.2 节详细介绍了 HMM 训练算法（Baum-Welch 算法）的具体过程，Baum-Welch 算法的基本原理就是根据一组给定的操作序列找出最能描述该序列特征的隐马尔可夫模型，确定该隐马尔可夫模型的各个组成元素。

结合本文的实验数据与隐马尔可夫模型的定义，我们可知：1.HMM 中的隐含状态（温度）是确定的，但是隐含状态的个数是不确定的；2.观察状态（开关

操作）是确定的，观察状态的种类也是确定的，即五种设备的对应十个操作；3. 操作序列已知，则一旦确定了隐含状态和观察状态的各个属性，即可根据算法 4.4 确定 HMM 的状态转移矩阵和发射矩阵，得到最终的隐马尔可夫模型。

因此，本小节的第一个实验就是确定 HMM 训练过程中隐含状态的个数，确定隐含状态的个数是一个复杂的过程，数学上常用的方法是利用狄利克雷过程（Dirichlet Process）[56]，应用到这里，即使用 HDP-HMM，但是该方法模型复杂，难度较大，且效果也未必好，考虑到 HDP－HMM 的学习成本较大且应用性不强，本文从实际问题出发，着眼于隐含状态的物理意义，根据温度变化范围[18℃，38℃]和实验结果估计出合理的隐含状态个数。实验过程中，根据每个用户产生的操作序列分别训练 HMM，评估模型的标准为预测的精确率（Precision）。实验设定 Baum-Welch 算法的迭代次数ϕ＝60，

图 5.2 显示了根据各个用户产生的操作序列得到的不同隐含状态个数下的模型的预测效果。从图中可以看出，模型预测效果会根据隐含状态数目的变化而变化，且各个用户上的曲线变化是具有相似性的，这也体现了模型的通用型。进一步分析，可以发现当隐含状态个数少于 5 时，隐马尔可夫模型的预测效果会随着隐含状态数的增多而有一定的提高，其中，隐含状态个数为 1 表示只有一个隐含存在，即不存在状态转移，只需考虑该状态下各个观察值的发射概率。而当隐含状态个数大于 5 之后，预测效果的提高变得不明显。因此，综合考虑下本文选取隐马尔可夫的隐含状态个数$N = 5$。

本小节的第二个实验关于 HMM 训练过程的终止条件，已知 Baum-Welch 算法一个重要问题是确定训练过程中何时的模型是最优的模型，即何时终止参数迭代过程。根据不同用户产生的操作序列，本文研究了不同迭代次数下获得的 HMM 的预测效果，结果如图 5.3 所示，其中统一实验过程中的隐含状态个数$N = 5$。从图 5.3 可知，当迭代次数较小（小于 30）时，所获得的 HMM 的预测精确率较低（20%至 30%左右），说明此阶段的 HMM 还未训练成熟。而随着迭代次数的增加，模型参数不断更新，则得到的 HMM 越来越能描述训练集的数据特征，因此，模型对操作序列的预测精确率快速提高。本文的实验表明，当迭代次数约为 60 次时，对应的模型的预测精确率最高，为 70%左右。而随着迭代次数的进一步增加，模型的预测精确率趋于缓和，当迭代次数大于 80 次时，过度拟合导致了预测精确率出现了一定程度的下降。

总结图 5.2 和图 5.3，我们探究了隐马尔可夫模型训练过程中各个参数的选择，最终根据实验结果确定在本文的研究中，当隐马尔可夫模型的隐含状态个数为 5，实验迭代次数为 60 时（ϕ＝60, $N = 5$），最能描述本文的用户的操作特征。

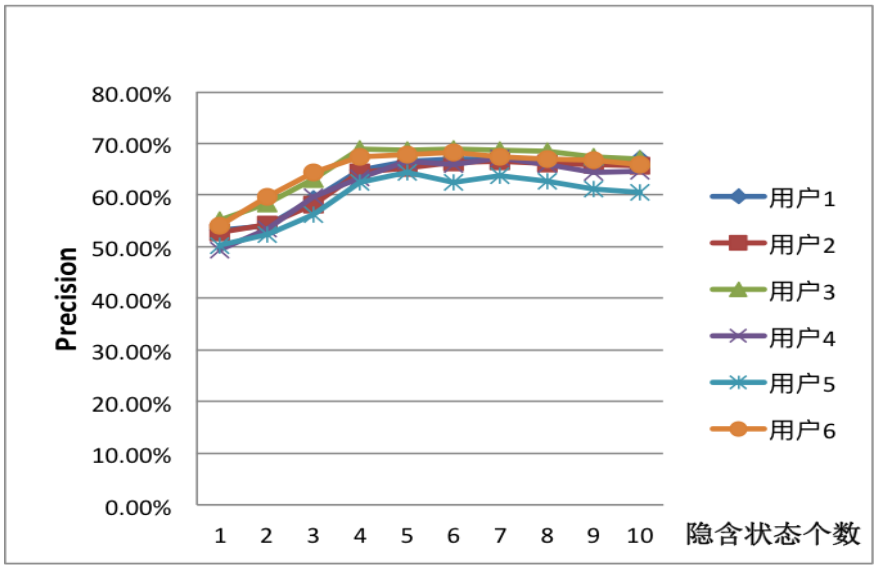

图 5.2 不同隐含状态的 HMM 的预测精确率（迭代次数ϕ=60）

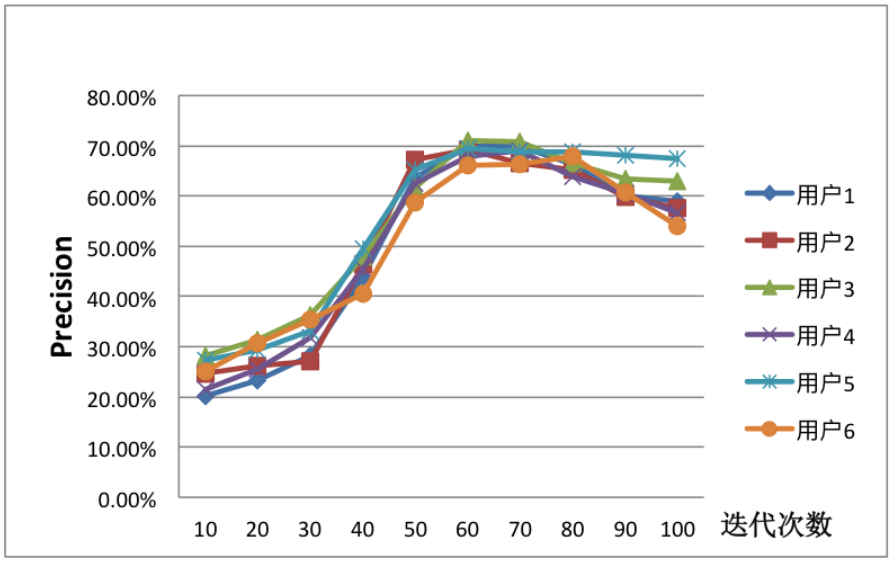

图 5.3 不同迭代次数的 HMM 的预测精确率（隐含状态数N=5）

5.3.2 改进模型性能评估

本文 4.3 节扩展了隐马尔可夫模型的齐次性假设，认为状态转移矩阵不再是固定不变的，引入时间信息（在这里用状态之间的位置顺序来表示），将二维的静态状态转移矩阵变为三维的非静态状态转移矩阵，从而得到了一个改进的基于三维状态转移矩阵的隐马尔可夫模型，为了验证本文的改进的模型（Improved Hidden Markov Model, IHMM）的预测效果，我们将其与传统的模型（Hidden Markov Model, HMM）进行了对比实验。

在实验过程中，我们依然设定隐含状态数 N＝5，迭代次数ϕ＝60,研究不同用户生成的模型的预测效果。其中，考虑到训练算法只获得局部最优解，所以初始参数的选择也具有重要意义。因此，本实验同时观察了四种模型的预测效果，分别表示为 HMM-Random(即随机初始参数的 HMM),HMM-Optimal(即优化初始参数的 HMM)，IHMM-Random(即随机初始参数的 IHMM)，IHMM-Optimal(即优化初始参数的 IHMM)。优化初始参数表示对模型$\lambda(A,B,\pi)$中的三个参数进行优化，其中，数据集上不同状态转移的统计概率作为初始状态转移矩阵A，不同状态下的不同操作的统计概率作为初始发射矩阵B,而将各个状态出现的次数的统计概率作为初始概率分布π。

以用户 1 为例子，我们研究了上述四种隐马尔可夫模型在不同大小的数据集上表现出的不同的预测性能。图 5.4 为实验结果，从图中可知如下信息：一，与随机初始参数的模型（HMM-Random, IHMM-Random）相比，相对应的优化初始参数的模型（HMM-Optimal,IHMM-Optimal）的预测精确率更高，其中，HMM-Optimal 的精确率比 HMM-Random 的精确率高 2.5%左右，而 IHMM-Optimal 的精确率比 IHMM-Optimal 的精确率高 4%左右，尽管精确率的提高有限，但足以说明初始参数选择的重要性。二，由图中曲线可以看出，本文改进的隐马尔可夫模型（紫色）能将预测精确率从传统的隐马尔可夫模型（红色）的 68%提高到 77%－78%左右，说明本文改进的基于非静态的状态转移矩阵的隐马尔可夫模型在预测残疾人操作行为时表现出了更好的性能。三，随着实验数据集的增加，模型的预测效果越来越好，这说明足够多的实验数据也是十分重要的。

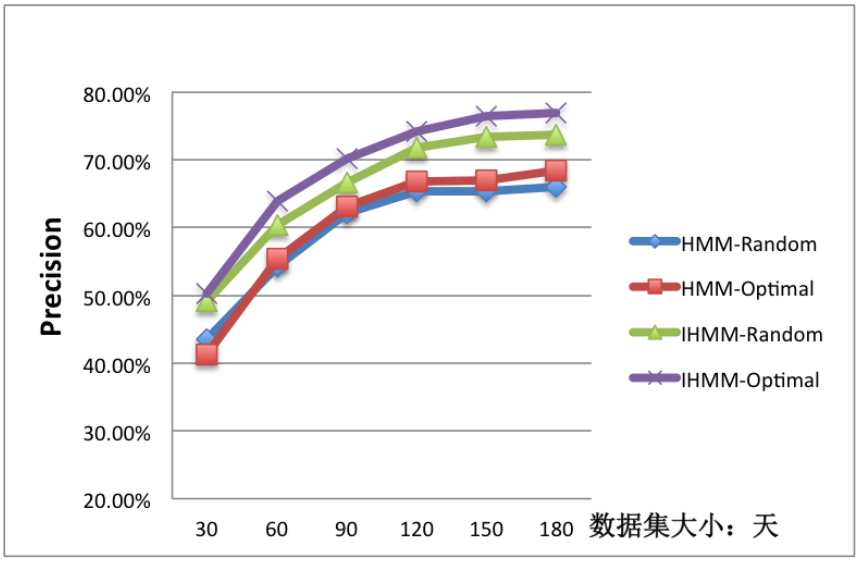

图 5.4 四种不同模型(HMM-Random, HMM-Optimal, IHMM-Random, IHMM-Optimal)的在不同大小数据集上的预测精确度（N＝5, ϕ＝60,user=用户 1）

	HMM-Random	HMM-Optimal	IHMM-Random	IHMM-Optimal
用户 1	66.1%	68.5%	73.8%	77.2%
用户 2	65.7%	68.4%	72.3%	76.8%
用户 3	68.1%	70.2%	73.5%	75.1%
用户 4	66.2%	67.4%	71.8%	76.5%
用户 5	64.9%	66.8%	70.5%	75.4%
用户 6	67.7%	69.2%	72.9%	73.3%

表 5.1 各个用户下不同模型的预测精确度对比（N=5，ϕ=60，数据集为 186 天）

表 5.1 则列出了所有用户下各个不同的隐马尔可夫模型的预测精确率，其中设定隐含状态数N=5，训练迭代次数ϕ=60，数据集选取 186 天的数据。可以看出，同一种模型对各个用户的预测精确率相似，而不同模型之间的差异性较大。而总体来看，本文的改进的隐马尔可夫模型的预测精确率比传统的模型要高，集中在 73%到 77%之间，具有实际应用意义。

总结表 5.1 和图 5.4，我们对比了本文改进的隐马尔可夫模型与传统的隐马尔可夫模型在预测智能家居环境下重症残疾人的操作时的精确率，分析了不同参数模型的差异，得出的结论是本文的改进的隐马尔可夫模型具有更优异的预测性能。

5.4 小结

本章节详细介绍了本文的实验环节，首先介绍数据来源和预处理过程，其次介绍实验原理和实验评估标准，最后详细分析了两个实验的过程与结果，分别为训练算法的参数估计和改进算法性能评估，确定了最适合用户的模型参数，对比分析了改进算法与传统算法的预测性能，最终验证了本文的基于非静态的状态转移矩阵的隐马尔可夫模型在预测智能家居环境下重症残疾人的操作时具有比传统模型更高的精确率，具备更高的应用价值。

第六章 总结与展望

6.1 总结

随着信息时代的到来，智能家居系统走进了普通百姓的生活，它通过先进的计算机网络技术，实现了智能家居设备的互联与用户对设备的控制访问，为居住者提供了高效、舒适的生活环境。然而，当前的智能家居系统往往需要居住者主动发起对家居设备的控制，而系统被动地提供响应，这必然要求居住者投入较大的成本去实现对设备的控制。进一步地，当居住者为重症残疾人时，这样的设备控制方式会带来一定问题：1）重症残疾人通常存在先天或后天的严重肢体障碍，导致他们无法完成很多简单的日常行为，因此，智能家居环境下的设备控制方式必然会给丧失一定自理能力的残疾人用户带来巨大的困难；2）随着智能家居系统的发展，智能设备的数量和种类随之增长，而操作指令的复杂程度也随之增长，这无疑进一步加重了残疾人用户对设备的操作负担。因此，如何简化智能家居环境中重症残疾人用户对设备的操作控制方式是一个亟待解决的问题，一个有效的思路为预测用户对设备的操作。近年来，智能家居环境下的用户行为预测分析技术也成为国内外学者研究的热点，其基本原理为分析用户行为，对用户产生的活动进行识别分类，或对异常活动进行检测，或预测用户即将发生的活动等。然而，当前的用户行为分析技术大多关注正常用户的正常行为，并没有考虑到残疾人用户的特殊性，因此，这些用户分析技术并不能很好地适用于智能家居环境下残疾人用户的操作行为分析。

基于智能家居环境下残疾人用户对设备控制过程中存在的问题，本文围绕学习用户行为习惯，预测用户将来最有可能的操作这一研究内容，分析了残疾人用户存在的特殊性：1）残疾人用户对设备的操作存在比普通用户更频繁的缺失现象；2）残疾人用户的操作行为受到环境因素的影响，那么该如何建模这样的行为模式。针对这些特殊性，本文制定了相应的技术路线，扩展了基于 KNN 的数据填补技术，改进了基于 HMM 的操作预测技术，最终实现了预测智能家居环境下残疾人用户的操作。本文的主要贡献如下：

1. 提出一个基于改进的隐马尔可夫模型的操作预测模型，探究了隐马尔可夫模型中的几个基本问题，并训练出一个符合残疾人用户行为特征的预测模型，在此基础上，扩展隐马尔可夫的齐次性原理，引入隐含变量间的位置顺序作为时间信息，将二维的状态转移函数扩展为三维的非静态状态转移矩阵，实现了一个改进的预测模型，该模型通过数据填补和操作预测两大模块，实现

对用户操作的预测，最终降低了残疾人用户对智能家居环境下设备的操作成本，提高了残疾人用户的生活自理能力。

2. 实现了基于 K 最近邻的缺失操作填补算法，该算法通过改进的 Jaccard 相似度来计算序列之间的相似性，对于不完整序列，获取数据集中与其最相似的 K 个序列组成参考集，根据参考集中缺失操作所在的位置来确定其在不完整序列中应处的位置，从而完成缺失操作的填补，为后续的用户操作预测模型提供了良好的数据集。
3. 收集真实数据集作为本文的实验数据，填补缺失数据，建立相应的训练集和测试集，设计相关实验验证本文的数据填补算法和隐马尔可夫预测过程。实验结果表明，本文的基于 KNN 的数据填补算法能有效地对缺失数据进行填补，而本文改进的 HMM 的操作预测算法相比传统的 HMM 表现出了更高的预测精确率，最终验证了本文的操作预测模型的有效性。

6.2 展望

本文致力于降低智能家居环境下残疾人用户对家居设备的操作控制成本，基于用户历史记录实现了操作预测过程。针对本文已完成的研究工作和研究过程中发现的问题，后续可展开如下工作：1）在预测算法方面：本文采用一阶隐马尔可夫模型实现算法的训练与预测，即下一步操作只与上一步操作相关，而与之前的操作无关，后续工作可以研究高阶（N 阶）算法的性能，如 N＝2 则表示下一步操作与前两步操作相关，亦可引入动态滑动窗口，比较各种模型的性能。2）在用户行为建模过程方面：本文只考虑了环境温度对用户行为的影响，后续可以考虑其他因素如室内湿度，光照等对用户操作行为的影响。3）在实验验证方面：目前仅收集了六个残疾人用户在 6 个月内产生的操作数据进行实验，为了更好地追踪用户行为特征，后续可以收集更多用户在更长时间内产生的数据，不断扩充数据集，进行深入研究。

参考文献

[1] S. K. Das, D. J. Cook, A. Battacharya., E. O. I. Heierman, and T. Lin, The role of prediction algorithms in the MavHome smart home architecture [J]. IEEE Wireless Communications, 2003, 9(6): 77-84.

[2] Stephen S. Intille, Designing a Home of the Future [J], IEEE Pervasive Computing, 2002, 1:80-86.

[3] Karthik Gopalratnam, Diane J. Cook, Online sequential prediction via incremental parsing: The Active LeZi Algorithm [J], IEEE Intellgient Systems, Vol.22, 52-58, 2007.

[4] Abhishek Roy, Sajal K. Das, Kalyan Basu, A Predictive Framework for Location-Aware Resource Management in Smart Homes [J], IEEE Transaction on Mobile Computing, 2007, Vol.6, 1270-1283.

[5] U. of Colorado, "The adaptive house," http://www.cs.colorado.edu/%20mozer/house/, last visited: 1/09/2014.

[6] U. of Florida, "Gator tech smart house" [EB/OL]. http://www.icta.ufl.edu/gt.htm, last visited: 1/09/2014

[7] K.S.Gayathri, Susan Elias, and S.Shivashankar, Composite activity recognition in smart homes using Markov Logic Network[C], UIC/ATC/ScalCom 2014:880-887.

[8] Muhammad Raisul Alam, Mamun Bin Ibne Reaz, and Mohd. Alauddin Mohd. Ali. A review of smart homes – past, present, and future [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 42(6): 1190-1203,2012.

[9] W. Kang, and D. Shin, Detecting and predicting of abnormal behavior using hierarchical Markov Model in smart home network[C]. Proceedings of 2010 IEEE the 17th International Conference on Industrial Engineering and Engineering Management (Volume 1). 2010:410 - 414.

[10] An Cong Tran. Application of description logic learning in abnormal behavior detection in Smart Homes[C]. IEEE RIVF International Conference. 2015: 7-12.

[11] Diulie J. Freitas, Tiago B. Marcondes, Luis H. V. Nakamura, Rodolfo I. Meneguette, A health smart home system to report incidents for disabled people [C]. Proceedings of the 2015 International Conference on Distributed Computing in Sensor Systems (DCOSS). 2015: 210-211.

[12] Ghorbel, Mahmoud, Maria-Teresa Segarra, Jérome Kerdreux, Ronan Keryell, Andre Thepaut, and Mounir Mokhtari. Networking and communication in smart home for people with disabilities [C]. International Conference on Computers for Handicapped Persons. Springer Berlin Heidelberg, 2004: 937-944.

[13] Chen Chen T L, Ciocarlie M, Cousins S, et al. Robots for humanity: using assistive robotics to empower people with disabilities [J]. IEEE Robotics & Automation Magazine, 2013, 20(1): 30-39.

[14] 残疾用户盘点智能家居设备[EB/OL]. http://techcrunch.cn/2014/09/10/smart-accessibility/

[15] Cook D J, Youngblood M, Heierman E O I, et al. MavHome: An agent-based smart home [M]. 2003.

[16] Sahadat M N, Alreja A, Srikrishnan P, et al. A multimodal human computer interface combining head

movement, speech and tongue motion for people with severe disabilities[C]. Biomedical Circuits and Systems Conference. IEEE, 2015.

[17] Holder L B, Cook D J. Automated activity-aware prompting for activity initiation [J]. Gerontechnology, 2013, 11(4): 534-544.

[18] Jakkula V, Cook D J. Mining Sensor Data in Smart Environment for Temporal Activity Prediction [J]. Poster Session at the Acm Sigkdd, 2007.

[19] Alam M R, Reaz M B I, Mohd Ali M A. SPEED: An Inhabitant Activity Prediction Algorithm for Smart Homes [J]. IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, 2012, 42(4): 985-990.

[20] Chen L, Hoey J, Nugent C D, et al. Sensor-Based Activity Recognition [J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2012, 42(6): 790-808.

[21] Atallah L, Yang G Z. The use of pervasive sensing for behaviour profiling — a survey [J]. Pervasive & Mobile Computing, 2009, 5(5): 447-464.

[22] Choi S, Kim E, Oh S. Human behavior prediction for smart homes using deep learning [J]. 2013:173-179.

[23] Y Yang, Z Wang, Q Zhang, Y Yang. A time based marov model for automatic position-dependent services in smart home [C]. Chinese Control and Decision Conference, 2010.

[24] Pentland A, Liu A. Modeling and prediction of human behavior [J]. Neural Computation, 1999, 11(1): 229-42.

[25] Sarawagi S, Kirpal A. Efficient set joins on similarity predicates [C]. ACM SIGMOD International Conference on Management of Data, Paris, France, June. 2004:743-754.

[26] Medo M, Zhang Y C, Zhou T. Adaptive model for recommendation of news [J]. Epl, 2009, 88(3): 38005-38010.

[27] L. E. Baum, An inequality and associated maximization technique in statistical estimation for probablistic functions of Markov processes [J]. Inequalities, 1972, 3:1-8.

[28] Noury N, Hadidi T. Simulation of human activity in a Health Smart Home with HMM [C]. IEEE, International Conference on E-Health Networking, Applications and Services. 2013:125-129.

[29] Hamlet A J, Crane C D. Robotic Behavior Prediction Using Hidden Markov Models [J]. Eprint Arxiv, 2014.

[30] Mathew W, Raposo R, Martins B. Predicting future locations with hidden Markov models [C]. Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012: 911-918.

[31] Gonz ález A M, Roque A M S, Garc á-Gonz ález J. Modeling and forecasting electricity prices with input/output hidden Markov models [J]. IEEE Transactions on Power Systems, 2005, 20(1): 13-24.

[32] Sultana A, Hamou-Lhadj A, Couture M. An improved Hidden Markov Model for anomaly detection using frequent common patterns [J]. Am Omng Rvy, 2012:1113-1117.

[33] A.A. Markov. "Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga". Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, tom 15, pp. 135–156, 1906.

[34] Meyn S, Tweedie R L. Markov chains and stochastic stability [M]. Springer-Verlag, 1993.

[35] Kohlmorgen J, Lemm S, Muller K R. Fast Change Point Detection in Switching Dynamics Using a Hidden Markov Model of Prediction Experts [C]. Proceedings of Artificial Neural Networks, 1999: 204-209.

[36] Schuster‐Böckler B, Bateman A. An Introduction to Hidden Markov Models [J]. 2007, Appendix 3(Appendix 3): 4 - 16.

[37] Ieee L R R F. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[38] Rubin D B. Inference and missing data. Biometrika, 1976, 63(3): 581—592.

[39] Rubin D B. Multiple Imputation after 18+ Years [J]. Journal of the American Statistical Association, 1996, 91(434): 473-489.

[40] Sahon G. Automatic Text Processing. Addison-Wesley, 1989.

[41] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley, Wesley Press, 1999.

[42] Cosine similarity [EB/OL]. http://en.wikipedia.org/wiki/Cosine_similarity.

[43] Pearson product-moment correlation coefficient [EB/OL]. http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

[44] Li D, Lv Q, Xie X, et al. Interest-based real-time content recommendation in online social communities [J]. Knowledge-Based Systems, 2012, 28: 1-12.

[45] Das A S, Datar M, Garg A, et al. Google news personalization: scalable online collaborative filtering [C]. Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 271-280.

[46] Jaccard Index [EB/OL]. https://en.wikipedia.org/wiki/Jaccard_index.

[47] KNN algorithm [EB/OL]. http://baike.baidu.com/view/8349300.htm

[48] Daum D, Haldi F, Morel N. A personalized measure of thermal comfort for building controls [J]. Building & Environment, 2011, 46(1): 3-11.

[49] J.F. Nicol, M.A. Humphreys, A stochastic approach to thermal com-forteoccupant behavior and energy use in buildings, Ashrae Trans. 110 (2004).

[50] Gunay H B, O'Brien W, Beausoleil-Morrison I, et al. Development and implementation of an adaptive lighting and blinds control algorithm [J]. Building & Environment, 2016.

[51] Xie H, Anreae P, Zhang M. Learning Models for English Speech Recognition[C]. Proceedings of the 27th Conference on Australasian Computer Science, 2004:323-329.

[52] Coast D A, Stern R M, Cano G G. An Approach to Cardiac Arhythmia Analysis Using Hidden Markov Models [J]. IEEE Transactions on Biomedical Engineering, 1990, 37(9): 826-836.

[53] Christopher D. Manning and Hinrich Schutze. Foundation of Statistic Natural Language Processing. The

MIT Press.

[54] J. H. Xiao, and B. Q. Liu, “A non-stationary hidden Markov Model and its application on pinyin-to-character conversion,” in JSCL, 2005.

[55] NWS Website [EB/OL]. http://www.cma.gov.cn/2011qxfw/2011qtqyb/

[56] Dirichlet Process [EB/OL]. http://www.datalearner.com/blog/1051459673766843

[57] Hidden Markov Model [EB/OL]. http://baike.baidu.com/link?url=f6jl9JhkzYNj55UHr1y0X1BqbtBTdfdXcfP7o8bejtKG8QGOW6vt529xOOz8Rg6fcTQj8-fSzk8zfAYq8qkvPw8YdUESWjHEpoM4gsFN6qUvh6bjXG58ClM5S-U1GEb41km6PkQtWYmRu-4B3LRv8CX5lFg-JJ26dGtqh0MJlEq

[58] Baum L E, Pitrie T, Souls G. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains [J]. Annual Mathematical Statistics, 1970, 41:164-171.

[59] 岳勇，田考聪. 数据缺失及其填补方法综述 [J]. 预防医学情报杂志, 2005, 21(6): 683-685.

[60] 闭小梅，闭瑞华. KNN 算法综述[J]. 科技创新导报, 2009(14): 31-31.

发表论文和科研情况说明

[1] Wu, E., Zhang, P., Lu, T., Gu, H., & Gu, N. Behavior prediction using an improved Hidden Markov Model to support people with disabilities in smart homes [C]//IEEE, International Conference on Computer Supported Cooperative Work in Design, IEEE, 2016:560-565.

[2] 2014年9月到2015年12月参与实验室项目《基于智能家居交互控制设备的大数据残疾人自理与服务云平台》。

[3] 2014年9月至2016年5月参与实验室申请的国家自然科学基金重点项目，“面向社会秩序的社会计算理论和方法研究”，项目编号为61332008。

致谢

三年前，我怀着一颗憧憬的心踏进了复旦大学协同信息与系统实验室，如今，我即将敲响毕业的钟声。一路走来，我收获了太多人的帮助与关心，在这里我想向你们道一声感谢。

首先，十分感谢我的研究生导师顾宁教授，在这三年的研究生生涯，顾老师在各个方面都给予了我极大的帮助。科研上，他是一个治学严谨的导师，悉心指导我的科研工作，帮助我确立研究方向，鼓励我探究科研问题，全程指导我论文的完成；生活中，他是一个和蔼的长辈，热心帮助包括我在内的实验室同学们在生活中遇到的困难，使得实验室充满了温暖。再次感谢顾老师，您对我的谆谆教诲，我将铭记于心。

其次，十分感谢实验室的卢暾老师和丁向华老师，你们是我的良师益友，谢谢你们在我科研过程中给予我的巨大帮助，你们知识渊博，兢兢业业，是我学习的榜样。感谢刘铁江老师在生活工作中给予我的帮助，您对学生关怀备至，照顾着每一个同学。特别感谢张鹏师兄，在我论文书写阶段对我的悉心指导与帮助。最后，感谢实验室大家庭里所有的兄弟姐妹，谢谢你们给予我的帮助，祝福你们学业有成，前程似锦。

我还要特别感谢我的家人，你们永远是我最温暖的依靠。谢谢父母的养育之恩，你们给了我这世上最无私的爱，我一定会好好报答你们。谢谢姐姐，给我毫无保留的信任与支持。感谢所有的亲人朋友，谢谢你们！

复旦大学
学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：_____ 日期：_____

复旦大学
学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____