****

**硕士学位论文**

**链路预测及交通恢复序列的预测**

姓 名：李星

学 号：1433384

所在院系：电子信息与工程学院

学科门类：工学

学科专业：计算机科学与技术

指导教师：关佶红 教授

二〇一七年三月

****

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Master

**Link Prediction and Traffic Recovery Sequence Prediction**

Candidate: Li Xing

Student Number: 1433384

School/Department: College of Electronic and Information Engineering

Discipline: Master of Engineering

Major: Computer Science and Technology

Supervisor: Guan Jihong

March, 2016

|  |
| --- |
| **链路预测与交通恢复序列的**  **预测**  **李星**  **同济大学** |

**学位论文版权使用授权书**

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年 月 日

**同济大学学位论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年 月 日

摘要

随着移动互联网的高速发展而产生的通过智能手机终端来进行电子交易，受到越来越多用户的青睐。交易过程中，如何保障用户账号的安全性显得尤为重要。而现阶段智能手机用户认证系统存在不少缺陷，利用移动终端用户行为特征进行身份认证成为研究的热点。本文提出一种基于手势密码触屏行为模式的认证方法，针对行为特征中的用户手势密码触屏行为习惯，分析和研究用户手势密码触屏行为的特征，进行手势密码触屏行为构建模式和认证方法的研究。

一、设计开发Android系统上手势密码触屏行为采集器。设计模拟输入手势密码环境，记录正确完成手势密码输入时的触屏行为特征，包括时间、X坐标、Y坐标、压力、接触面积等。将采集到的数据保存在json格式的数据包里，以便后续的数据分析工作。

二、提出一种手势密码触屏行为模式构建方法。分析相关领域现有行为模式构建的方法，结合移动行为新型特征，挑选出适合的触屏行为特征组成行为特征向量。通过曼哈顿计算算法和DTW算法计算行为特征向量之间的距离，组建行为特征距离向量。通过改进后的SVM算法对行为特征距离向量进行训练，获取用户手势密码触屏行为模式。

三、提出构建多用户单账号下的手势密码触屏行为模式的方法，解决现实场景下可能存在的多个用户共用一个账号的情况。针对账户下存在多个用户的情况，对混合用户数据进行k-means聚类，区分同一账号下不同用户的手势密码触屏行为数据，并针对每个用户的数据构建手势密码行为模式。当前用户输入的手势密码若能匹配任一手势密码行为模式，即可认为该用户为合法用户，否则为非法用户。

**关键词**：行为认证，手势密码，触屏特征，支持向量机，DTW，k-means算法

**ABSTRACT**

With the development of mobile internet, more and more users prefer to finish the e-Commerce by mobile phones. The security of users’ account plays an important role in the process of trading. Nowadays, the user authentication system of mobile phone has some limits and finishing the user authentication based on touch screen behavior becomes research hotspot. In this paper, presenting a method of user authentication based on password touch screen features.

Firstly, designing and developing a data collector for password touch screen behavior in Android system. Designing an environment for inputting password and recording the touch screen features, including time, X coordinate, Y coordinate, pressure, contact area when finishing the password correctly. Saving the dataset into the data package of json for the future research.

Secondly, proposing a method to build a model based on touch screen behavior. Combining analyzing the methods of building behavior models and the new mobile behavior, we can pick up some suitable features to build behavior feature vector. Building the behavior feature distance vector through Manhattan algorithm and DTW algorithm. Then we can get the model based on password touch screen behavior by the SVM classifier.

Finally, proposing a method to solve the problem of multi users with single account by building behavior models based on touch screen behavior of multi users with single account. Aiming at this situation, we take the action of cluster with the data of multi users to distinguish the difference touch screen data from one account. Meanwhile, building the behavior model for each user. If the active user’s password can match each model, we can say it’s a valid user. Besides, it’s an illegal user.

**Key Words:** behavior authentication，password，touch-screen feature，SVM，DTW，k-means

目录

[第1章 引言 4](#_Toc438380793)

[1.1 研究背景与意义 4](#_Toc438380794)

[1.2 国内外研究现状 6](#_Toc438380795)

[1.3 本文主要研究工作 7](#_Toc438380796)

[1.4 论文组织结构 8](#_Toc438380797)

[第2章 手势密码触屏行为模式构建的关键技术 10](#_Toc438380798)

[2.1 距离计算算法 10](#_Toc438380799)

[2.1.1 曼哈顿距离 10](#_Toc438380800)

[2.1.2 动态事件规整算法（Dynamic Tine Warping, DTW） 10](#_Toc438380801)

[2.2 分类算法 12](#_Toc438380802)

[第3章 数据采集和预处理 15](#_Toc438380803)

[3.1 数据采集 15](#_Toc438380804)

[3.1.1 数据采集器 15](#_Toc438380805)

[3.1.2 数据集 16](#_Toc438380806)

[3.2 数据预处理 16](#_Toc438380807)

[3.2.1 数据清理 16](#_Toc438380808)

[3.2.2 数据变换 17](#_Toc438380809)

[3.2.3 归一化 17](#_Toc438380810)

[3.3 特征提取 18](#_Toc438380811)

[第4章手势密码触屏行为模式的构建和认证 19](#_Toc438380812)

[4.1 总体架构 19](#_Toc438380813)

[4.1.1 方法总体架构 19](#_Toc438380814)

[4.1.2 概念定义 20](#_Toc438380815)

[4.2 特征距离向量集构建 21](#_Toc438380816)

[4.2.1 参照特征向量集生成 21](#_Toc438380817)

[4.2.2 特征距离向量集生成 22](#_Toc438380818)

[4.3 分类器预测和训练 22](#_Toc438380819)

[4.4 实验分析 24](#_Toc438380820)

[4.4.1 实验环境和数据 24](#_Toc438380821)

[4.4.2 实验结果分析 24](#_Toc438380822)

[第5章 多用户单账号场景下的手势密码触屏行为模式的构建 27](#_Toc438380823)

[5.1 多用户单账号场景分析 27](#_Toc438380825)

[5.2 多用户单账号手势密码触屏行为模式的构建 27](#_Toc438380826)

[5.2.1 总体架构 27](#_Toc438380827)

[5.2.2 多用户聚类 29](#_Toc438380828)

[5.2.3 手势模型构建 30](#_Toc438380829)

[5.3 实验分析 30](#_Toc438380830)

[第6章 总结与展望 34](#_Toc438380831)

[6.1 总结 34](#_Toc438380832)

[6.2 进一步工作 35](#_Toc438380833)

[致谢 36](#_Toc438380834)

[参考文献 37](#_Toc438380835)

[**个人简历、在读期间发表的学术论文与研究成果** 40](#_Toc438380836)

# 第1章 绪论

随着移动计算能力以及智能手机占有量的指数级增长，触屏智能手机已经成为每个人日常生活不可替代的重要组成部分，如何正确识别移动终端用户身份的合法性是保障用户信息和财产安全的关键。传统的安全策略存在的局限性使其难以满足移动终端用户对于安全性的需求，用户触屏行为的研究和应用逐渐成为用户身份认证领域的热点。手势密码触屏行为认证通过采集用户完成手势密码时的触屏行为数据，对该部分数据进行分析和建模，针对每个用户建立独有的手势密码触屏行为模式，作为身份认证的依据。

本章首先阐明论文的研究背景以及研究的意义与价值，然后对本文研究内容的研究现状进行阐述和分析，随后介绍本论文的研究思路与主要研究工作，最后介绍本文整体的组织结构和内容安排。

## 1.1 研究背景与意义

在真核生物中，DNA序列需要通过严格的分级组装的方式包装和高度压缩后形成染色质结构。首先，DNA序列和蛋白质相结合，形成核小体。核小体进一步压缩形成更加复杂的纤丝状的染色质结构。最后，纤丝状的染色质被进一步压缩，形成能够被包装进细胞核的染色体结构。核小体是真核生物染色质的基本组成单位，由DNA缠绕在组蛋白上构成，各核小体串联而成染色质纤维。H2A、H2B、H3和H4这四种组蛋白，每一种组蛋白各二个分子，形成一个[组蛋白八聚体](http://baike.baidu.com/view/3872893.htm)，约200bp的DNA分子盘绕在组蛋白八聚体构成的核心结构外面，形成了一个核小体。其中，缠绕在组蛋白八聚体周围的核心 DNA（core DNA）约1.65 圈，约合147个碱基对；而相邻的核小体之间的自由区域（linker DNA）约为20～50个碱基的长度。核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA复制和修复等过程中都扮演着重要角色。核小体是表观遗传机制中非常重要的组成部分。然而，核小体的形成机制仍然是一个未知的问题。核小体的形成和分布并不是随机的。要明确核小体对基因调控的影响，必须要先掌握核小体的分布规律。如何精确进行核小体定位是生物信息学一个非常热门的研究问题。近年来，对核小体定位的研究得到广泛关注，提出了很多核小体预测和定位的算法。然而如何高效准确地进行核小体定位依然是需要深入研究的问题。

目前，研究者们主要通过实验的方法来检测核小体的位置。实验方法的根本思路是用微球菌核酸酶来分解和消化两个核小体之间的连接DNA，从而得到互相分离的核小体，并且根据标识的位置进行核小体的定位。近几年，随着科技的发展，实验技术的提升，出现了很多其他的核小体定位的实验技术，例如染色质免疫共沉淀-芯片（ChIP-chip）、染色质免疫共沉淀-测序（ChIP-Seq）等。随着实验方法的发展，核小体的定位变得更加容易实现，准确度也更高。但是，这种通过实验技术来获取全基因组的核小体定位信息的方式需要大量的时间和经费。此外，随着更多不同物种的核小体位置信息的增加，如何存储通过实验方法得到的核小体的位置信息也成为研究者们面临的巨大难题。近几年来，Segal研究小组提出了基于序列特征的核小体定位理论预测模型，发现DNA序列能够为核小体定位提供重要的信息。此外，还有很多研究者发现，DNA的某些结构特征也和核小体密切相关。所以，对实验所获得的数据进行分析，使用计算方法来进行核小体定位对弥补生物实验技术的缺陷，辅助生物实验进行具有重要意义，对基因组中的核小体定位的理论预测已经成为生物信息领域热门研究课题。

核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA 复制和修复等过程中都扮演着重要角色。在染色体的任何时期，核小体的结构都是不变的。核小体是细胞内基因调控的重要环节，是表观遗传机制的重要组成部分。核小体的定位预测有望为创新药物研究和临床医学带来实际的有价值的突破，具有很好的实际应用价值。

本课题旨在找到更高效更准确的核小体定位算法，提高预测和定位精度，解释某些特定的基因调控现象，为表观遗传机制提供一定的参考意义。

## 1.2 国内外研究现状

在真核生物中，DNA序列需要通过严格的分级组装的方式包装和高度压缩后形成染色质结构。首先，DNA序列和蛋白质相结合，形成核小体。核小体进一步压缩形成更加复杂的纤丝状的染色质结构。最后，纤丝状的染色质被进一步压缩，形成能够被包装进细胞核的染色体结构。核小体是真核生物染色质的基本组成单位，由DNA缠绕在组蛋白上构成，各核小体串联而成染色质纤维。H2A, H2B, H3和H4这四种组蛋白， 每一种组蛋白各二个分子，形成一个[组蛋白八聚体](http://baike.baidu.com/view/3872893.htm" \t "_blank)，约200 bp的DNA分子盘绕在组蛋白八聚体构成的核心结构外面，形成了一个核小体。其中，缠绕在组蛋白八聚体周围的核心 DNA （core DNA） 约 1.65 圈，约合147 个碱基对；而相邻的核小体之间的自由区域（linker DNA） 约为 20～50 个碱基的长度。核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA 复制和修复等过程中都扮演着重要角色。然而，核小体定位是一个涉及 DNA、转录因子、组蛋白修饰酶和染色质重塑复合体等分子间相互作用的复杂过程。核小体的形成机制仍然是一个未知的问题。如何精确进行核小体定位是一个非常热门的研究问题。目前，研究者们主要通过实验的方法来预测核小体的位置。实验方法的根本思路是用微球菌核酸酶来分解和消化两个核小体之间的连接DNA，从而得到互相分离的核小体，并且根据标识的位置进行核小体的定位。近几年，随着科技的发展，实验技术的提升，出现了很多其他的核小体定位的实验技术，例如染色质免疫共沉淀-芯片（ChIP-chip）、染色质免疫共沉淀-测序（ChIP-Seq）等。此外，近年来，在实验数据的基础上，研究者们使用计算方法来进行核小体定位弥补生物实验技术的缺陷。在众多的影响核小体形成的因素中，包括染色质重塑子，特定的DNA结合蛋白，DNA本身的序列等已经成为近期生物信息领域的研究热点。目前已有的定位方法主要基于两大类，基于序列的核小体定位方法和基于结构的核小体定位方法。其中，基于DNA基于序列的研究方法有， Segal研究小组、Kaplan研究小组等小组提出的定位算法[1]-[7]。Trifonov等人首次将含AT和GC的二核苷酸和三核苷酸用来进行核小体的定位[8]。随后,有更多的研究学者发现呈现周期性的序列特征与核小体的形成有密切的联系。例如，G和C碱基的含量和体外核小体的分布的相关性高达50%[3]。基于序列特征的计算方法被越来越多的学者用来进行核小体的预测。但是，随后也有研究证明，DNA的序列特征并不是决定核小体组织形成的决定性因素[9]-[10]。之后，有学者就提出问题，DNA的结构化特征是否对核小体的形成有影响[11]-[13]。为了解决这个问题，一些研究学者开始对DNA序列的结构化特征和核小体的形成机制进行研究。核小体DNA数据库的一些物理化学性质，例如DNA蛋白质复合体的倾斜度和螺旋弯曲度已经被证实与核小体的绑定有显著关系[14]。基于roll-slide 模型，Tolstorukov等人发现，相邻碱基对的滑动对DNA超螺旋节距有很大的影响力，相邻碱基对的弯曲度对DNA的弯曲度起决定性作用。Miele等人发现基于二核苷酸的DNA弹性和固有的弯曲度可以用来分析核小体[13]。Morozov等人使用DNA伸缩能量函数构造了基于序列的核小体形成的生物物理学模型[15]。DNA晶体结构的二核苷酸的弯曲度也被发现和核小体DNA片段有关联[16]-[17]。对核小体晶体结构的研究分析显示碱基对的行为、核糖环的折叠和支柱扭曲共同影响着核小体DNA序列的结构变异[18]。这些研究表明，可能还存在着很多与核小体相关的结构化特征。因此，系统化的分析不同的结构化特征和区别对核小体形成有关的结构化特征变得非常重要，以便来理解核小体与基因序列的结构化特征之间确切的关系。越来越多的研究学者开始利用这些和核小体相关的结构化特征提出更多有效的新方法来进行核小体的预测，并达到了满意的实验结果。但是目前的核小体定位算法精度都还较低，大约在50%左右。主要存在三方面的问题：1、没有找到较好的特征进行核小体的定位。2、通过构造特征向量然后采用机器学习中的分类算法的过程中，采用的机器学习算法都不能达到很高的准确性。3、通过构造特征图谱然后寻找峰值的过程中，峰值寻找算法还不能很好判断峰值位置的问题。

（1）特征寻找和选择

核小体特征分为两类，序列特征和结构特征。序列特征包括K-mer、周期性、GC含量等；结构特征包括DNA变性、DNA弯曲度、Z-DNA等。目前大部分基于生物计算的核小体定位算法要么只采用序列特征，要么只采用结构特征，忽略了序列特征和结构特征的相关性。目前，已有的特征选择的方法有嵌入式、过滤式、封装式三大模型。嵌入式模型是将特征选择过程作为组成部分嵌入到学习算法理，在学习训练的过程中选择合适的特征，学习过程结束时，学习算法所用到的特征即为特征选择的结果。经典的算法是决策树分类算法，如ID3、CART以及C4.5。过滤式模型通常作为学习算法的预处理手段，独立于具体的学习算法。常用的特性评价函数如信息增益、类间距离、相关系数等等。过滤式模型在大规模数据或在线数据的处理上具有明显的优势。封装式模型直接利用学习算法的准确率作为特征或特征子集重要性程度的评价标准，相关算法有贝叶斯网络、支持向量机等。过滤式模型和封装式模型都存在各自的优势和缺点。过滤式模型效率高，但是最终的学习准确率可能不高。而封装式模型学习效果好，但是容易出现过拟合现象。

（2）相关机器学习算法

一些已有的核小体定位算法中，有一些算法是通过构造特征向量然后采用机器学习中的分类算法对DNA序列进行分类，分为有核小体的位置和没有核小体的位置。已有的机器学习方法在核小体定位上的应用有SVM[19]、HMM[24]、高斯混合模型[23]、Boosting[21]等。例如，但是目前为止，这些算法的准确率都不是很高，一般准确率都在50%左右。这些机器学习算法并不能很好的处理生物数据,并未针生物数据做一定的修改和适应。仍有很多的算法并未应用在核小体的定位上，例如J48、Decision Stump等，很多算法值得尝试。

（3）峰值寻找算法

目前已有的很多核小体定位算法大多都是通过构造特征图谱然后寻找峰值或者峰谷的位置来进行核小体的定位，例如DLaNe[24]、All- atom[28]等等。但是在峰值寻找的过程中，都不能避免局部最优问题，因而无法精确找到所需要的峰值或者峰谷。峰值寻找算法的精度大大影响核小体定位的精度。

## 1.3 本文主要研究工作

本文主要研究

本文的主要研究内容及工作分为以下几个方面：

（1）特征寻找和选择

至今为止，热点的生物属性仍然没有被完全了解，所以当前用来识别热点的相关的属性特征仍然是不够的。在众多特征中，需要根据生物特性和相应的模型进行删选。在本课题中，以序列特征为主，结合结构特征，采取序列特征和结构特征相结合的方法，并采用过滤式特征选择方法，以皮尔逊相关系数作为过滤条件，选择和寻找生物特征。

（2）机器学习算法的选择

相关机器学习方法较多，需要研究如何选择和集成相关算法以提高定位算法效率和定位精度。目前的核小体定位算法中，大多只采用了一种机器学习算法。本课题中将实验更多其他的机器学习算法，并集成相关算法以提高预测准确率。

（3）峰值寻找算法的优化

峰值寻找算法还存在不能很好判断峰值位置的问题，尚需进一步优化。本课题主要研究内容在于找到一种更加精确的峰值定位算法，解决局部最优问题，使得能够找打实际的所需的峰值（或峰谷）。峰值寻找算法的精度大大影响核小体定位的精度。

## 1.4 论文组织结构

本文的组织结构如下所列：

第一章主要介绍了研究的背景和意义，以及现有的针对手势密码触屏行为特征和认证的研究现状和基础；

第二章介绍手势密码触屏行为模式构建和身份认证过程中涉及的关键技术和算法，包括计算特征向量间距离的曼哈顿算法和DTW算法，筛选特征的主成分分析方法，用于分类的SVM及相关的参数优化算法；

第三章从数据采集、预处理、特征构建这三面入手。详细描述了Android系统上手势密码触屏数据采集器的开发，对采集来的原始数据通过数据筛选、数据格式统一、归一化等方法进行预处理，并基于预处理后数据进行手势密码触屏行为特征的构建。

第四章介绍手势密码触屏行为模式的构建和认证。首先描述了行为模式构建的总体架构，然后通过参照集特征向量生成、特征距离向量集构建以及SVM分类器等算法构建手势密码触屏行为模式并进行认证效果验证。最后，通过实验分析来获取触屏行为模式构建方案的误检率和漏检率，明确模型的认证效果。

第五章介绍的是多用户单账号场景下的手势密码触屏行为模式构建。通过k-means聚类算法区分同一账户下的不同用户，然后对每一个用户使用第四章介绍的手势密码触屏行为模式构建方法进行建模。认证时，主要匹配该账户下的任一用户的行为模式，即可认为其为合法用户。

# 第2章 链路预测方法分析

本章主要介绍了链路预测的基础概念。

## 2.1 问题描述

定义G（V,E）为一个无向无权的网络，其中V表示网络中节点的集合，E表示网络中边的集合。网络中节点的数量为N，若该网络为一个全联通网络，那么网络中边的数量应该为N\*(N-1)/2,我们表示为全集U。

现我们使用某个链路预测算法，根据网络现有的连边信息，算法会对所有未连接的边进行评分。若我们选取边e’,该边的两个端点分辨为x、y，根据链路预测算法，我们计算得到e’的分值为Sxy。将评分从高到低进行排序，分值越高表示边出现的概率越大。

链路预测中主要有两类的需求，这两类需求分别是丢失边的预测以及错误连边的预测。丢失边的预测主要是指网络中存在但是尚未被发现的边，如人类的蛋白质网络中，大量的蛋白质相互作用仍未被人类发现。错误连边主要是指实际网络中不存在，但是被错误加入其中的边，如QQ有时会向我们进行好友推荐，而实际生活中我们并不认识该人。

以下将对链路预测中的丢失边预测以及错误连边预测分别进行讨论。

2.1.1 丢失边的预测

在链路预测中，为了预测算法的准确性，我们通常将E分为两个部分，训练集ET和验证集EP。训练集和测试集的划分原则为ET EP = E,且 ET EP = ∅ 。我们会将属于U但是不属于E的边称为不存在的边，在计算算法的准确性时，我们是基于训练集中的信息，验证集仅用作数据的验证。

2.1.2 错误边的预测

与丢失边的预测略有不同，为了验证算法的准确性，我们需要模拟错误连边的情况。我们将网络中添加的边用作验证集EP，此时的训练集ET = E + EP。在堆算法的预测准确性进行计算时，我们基于的网络拓扑结构包含中网络中原本存在的边以及新加入的扰动边。

### 2.2 数据集划分

由上述关于链路预测的定义，我们发现在链路预测中我们会用到两类边，该两类边称作“训练集”和“验证集”，不同的验证集和训练的的划分会在一定的程度上影响结果的准确性。该两类边的选取主要是由验证集的选取决定的，因为一旦获得验证集EP之后，在丢失边的预测中训练集ET=E-EP ,在错误边的预测中ET=E+EP。常见的验证集的选取主要有随机抽样、逐项遍历、k折叠交叉检验，滚雪球抽样。现对该几类方法做简要的介绍，为了便于介绍，我们将以未知边的预测为例，关于错误连边中验证集的选择稍加拓展即可。

2.2.1 随机抽样

在网络G（E，V）中，共有节点N个，边M条。现在我们需要划分其中比例为p（0 < p < 1）的边作为验证集，随机抽样即从M条边中随机抽取p\*E（若非整数，向上取整）条边组成验证集。

2.2.2 逐项遍历

针对某些较小的网络，逐项遍历法是一种更为精准的数据集划分方式。在逐项遍历法中，我们每次从网络中选取一条边作为验证集，其余的边则作为训练集使用。对该条边进行预测时我们会得到一个准确率，依次遍历网络中的M条边，我们会得到M次的准确率，求M次准确率的平均值，该平均值即为我们算法的准确率。该方法相较于随机抽样法中每条边都有被选中的机会。

2.2.3 k折叠交叉检验

该方法是逐项遍历法的一种改进方法，我们将边随机分成k份，选取其中的一份作为验证集，剩余部分作为训练集，此时我们会得到一个预测准确率。遍历k份边，共得到k次的预测准确率，求该k次准确率的平均值。10折叠交叉验证是最常见的方法，此时相当于选取10%的边作为验证集，即p=0.1。当k等于M时，该方法退化为逐项交叉检验法。

2.2.4 滚雪球抽样

该方法先随机访问一些被调查者，然后再邀请这些调查者推荐调查对象。在实际应用中类似于广度优先搜索。在初始化时，我们随机选择一些节点，此后依次访问该类节点的邻居节点，并将这些点加入样本中，该过程直至满足样本的数量要求为止。

2.2.5 其他方法

其他常用的抽样方法还包括熟识者抽样、随机游走抽样、基于路径抽样。

在链路预测的相关研究中，数据集的划分多采用随机抽取的方式，为了便于比较，本文也将采取随机抽取的方式来获取验证集。此外，先前的研究当中，验证集的选取大多为固定值10%，为了验证本文提出改进算法的健壮性，我们将p值设置为[0.044,0.08,0.12.0.16,0.20]这五个不同的值。

### 2.3 经典的链路预测算法

本文选取的算法是链路预测中最为经典也是最为广泛使用的算法。该类算法属于基于节点相似性的链路预测算法，该类算法的一个前提假设是两个节点之间相似性越大，两点间存在链接的可能性也越大。

度量节点的相似性有很多种方式，比如利用节点的属性就是一种简单而又直接的方式。举例来说，若两个人具有相同的年龄，所在城市相同，职业相同，兴趣相同等，我们则认为这两个人具有很高的相似性，这两个人认识的可能性也大于属性差异较大的两个人。

基于节点属性的相似性虽然可以获得较高的链路预测准确率，但是获取节点的属性往往是十分困难的。比如我们常用的QQ，我们的很多类信息是对他人不可见的（身份证号，手机号等），即使获得了用户的一些信息，该类信息也未必真实可靠，比如说年龄、身高、性别、所在地等等。由于节点属性的难易获取，所以利用节点属性进行链路预测便遇到了很大的障碍。

与节点属性相比较，网络的拓扑结构更易于获取同时也更为可靠。基于拓扑相似性的链路预测算法的准确率高低很大程度上取决于该算法是否很好的抓住了网络的结构特征，即网络的形成机制。

基于节点的聚类系数对于链路预测算法的改进中的几种算法便是基于拓扑相似性的算法，相比于计算全局网络结构的算法，本文所选取的算法有计算复杂度较低的特点，这几种方法也常见于对改进算法的比较之中。

2.3.1 CN 指标(Common Neighbors)

对于网络中的节点vx，Γ(x)表示该节点的所有邻居，同样的Γ(y)表示节点vy的所有邻居集合。下式表示的便是节点vx和节点vy的所有共同邻居的数目，该算法认为两个节点间共同邻居的数目越多的话，该两个节点认识的可能性就越大。

Sxy = | Γ(y)|

在现实的生活中，如果两个陌生人有很多的共同好友，那么这两个人在未来认识的可能性也就越大。再例如，在科学家合作网络中，某两位科学家有着越多的共同合作者，那么这两个人在未来合作的可能性就越大。

2.3.2 Salton 指标

Salton指标又称作余弦相似性指标，kx和ky分别表示节点vx和节点vy的度的大小。与CN指标相比较，该式多出的分母部分表示在两个节点共同领居数目一定的时候，本身节点度越小，两个节点连接的可能性越大。比如在微博当中，大v之间会有着更为广阔的社交空间，所以他们会有更多的共同邻居，所以反而本身好友较少的前提下有着相同的共同邻居的两个节点在未来认识的可能性较大。

Sxy =

2.3.3 Jaccard 指标

Jaccard指标提出于一百多年前，相比于CN指标，该算法认为在共同邻居数目一定的前提下，两节点邻居数目越多，相连的可能性越小。

2.3.4 Sorensen 指标

该指标常用于生态系统网络的预测，该公式可看做是Salton指标的一个简单的变形。

Sxy =

2.3.5 大度节点有利指标(hub promoted index, HPI)

该指标主要被用来刻画新陈代谢网络中每对反应物的拓扑相似性，该公式认为在共同邻居的数目一定的前提下，度较小的节点更有利于网络的连接。

Sxy =

2.3.6 大度节点不利指标(hub depressed index, HDI)

该定义与HPI类似，知识分母部分略有不同，该公式认为在共同邻居的数目一定的前提下，度较大的节点更有利于网络的连接。

Sxy =

2.3.7 LHN 指标

该公式的形式与Sorensen类似，仅将分母部分的两个元素由相加改为了相乘。

Sxy =

2.3.8 AA 指标 (Adamic-Adar)

这是对于CN指标的一种改进，该指标的主要思想是度较小的共同邻居节点的贡献度大于度较大的共同邻居节点。

Sxy =

2.3.9 RA 指标

RA指标在形式上类似于AA指标，该算法是周涛、吕林媛等人受到网络中资源分配模型的启发而提出的。网络中不直接相连的两个节点vx和vy，vx需要借助共同邻居向vy节点发送资源。假设每个节点都有一个单位的资源并平均发送给他的共同邻居，显然vy从vy处接受的资源总量为下式中的Sxy。

Sxy =

此处缺张图

2.4 评价指标

衡量链路预测算法准确性的指标主要有AUC、准确度以及排序分三种。他们三种指标对于准确度的衡量侧重点有所不同。现分别就三种指标做简单的介绍。

2.4.1 AUC (area under the receiver operating characteristic curve)

2.4.2 准确度（Precision）

相比于auc指标，有时我们只关心排在前面的几条结果是否正确。准确度定义为在前L个预测中，预测准确的比例。假设现在的验证集大小为L，根据出现连边的可能性对于未出现的边进行排序，验证集评分在前L中的个数为m，那么准确率为：

Precision=

L大小的选择一般会对结果产生影响。

2.4.3 排序分（Ranking Score）

排序分考虑了验证集在最终排序结果中的位置。链路预测算法会对所有未知边进行排序，令H表示未知边的集合，re表示测试边e在排序中的排名。那么这条边的排序分为:

RS=

遍历所有的验证集，将各条测试边的排序分相加便得到总的排序分，显然排序分越低预测越准确。

2.5 本章小结

本章回顾了问题描述、数据集划分、评价指标以及被改进的算法，下章着重讲解算法如何改进以及实验部分。

第三章 基于聚类系数的链路预测算法

3.1 聚类系数的定义

Watts和Strogatz在1998年提出网络中聚类系数的概念，该指标为网络的一个局部性指标。针对网络中的任意一个节点vi，其聚类系数定义为该点所有邻居节点之间的连接数占领居间最大连边可能性的比例，其数学表达式为：

Ci =

式子中li表示节点vi的邻居节点之间连边的数目，ki表示节点vi度的大小，Ci即为节点vi的聚类系数大小。根据式子我们可知，在一个全网络图中，每个节点的聚类系数均为1，在一个树状结构中，每个节点的聚类系数均为0。

聚类系数表示了节点间联系程度，高的聚类系数表示节点的邻居之间相识的可能性较大。如果节点v1连接于节点v2，节点v2连接于节点v3，那么节点v3很可能与v1相连接。这种现象体现了部分节点间存在的密集连接性质。在生活中，若将人类的生活 抽象为社交网络，若某人的聚类系数较大，那么该人的朋友之间相互认识的可能 性越大，那么我们认为该人的好友之间产生连边的概率较大。基于大的聚类系数 促进网络的连接，所以本文认为可将聚类系数这一特征融入到现有算法之中， 升算法对于部分网络的预测准确度。

此处应该有图解

3.2 基于聚类系数的链路预测算法改进

对于上文中所提到的9种链路预测算法，加入聚类系数这一元素得到的改进后的算法为：

1. CCN 指标(Clustering Common Neighbors)

Sxy =

加入聚类系数后的CNN指标与CN指标的区别在于，CN指标根据两点间的共同邻居的数目进行打分，而在CNN指标看来，共同邻居节点的贡献度是不同的，聚类系数较高的共同邻居节点的贡献度较大。

1. CSalton 指标(Clustering Salton)

Sxy =

该指标与Salton指标的区别在于分母部分，同CNN指标，CSalton指标将共同邻居节点的贡献度区别看待，同样认为较大的聚类系数的共同邻居节点的贡献度较大。

1. CJaccard 指标(Clustering Jaccard)

改进原理同CCN。

1. CSorensen 指标(Clustering Sorensen)

Sxy =

改进原理同CCN。

5) CHPI 指标(Clustering  hub promoted index)

Sxy =

改进原理同CCN。

6) CHDI 指标(Clustering hub depressed index )

Sxy =

改进原理同CCN。

1. CLHN 指标(Clustering LHN)

Sxy =

改进原理同CCN。

8) CAA 指标(Clustering Adamic-Adar)

Sxy =

改进原理同CCN，CAA指标同时看重共同领居节点度大小和聚类系数的区别，算法认为较大的聚类系数较小的度大小的共同邻居节点更利于促进连边的形成。

9) CRA 指标(Clustering Resource Allocation)

Sxy =

改进原理同CAA。

3.3 数据集介绍

本章实验部分共用到五类数据集，这些数据集合是在pajek和stanford两个数据集当中手机而来。Pajek数据集发布在网站<http://vlado.fmf.uni-lj.si/pub/networks/data/上>，stanford数据集发布在网站<http://snap.stanford.edu/data/>之上。这两个网站之上搜集了大量的复杂网络相关数据供复杂网络相关的研究者自由使用。

3.3.1 数据集描述

本章共使用五类数据集，他们分别是USAir、Elegans、Football、Jazz、Karate，现对这些数据做简要的介绍。

USAir 表示的是美 国的航线信息，节点表示机场，若两个机场之间有直达航线，那么该两点之间存在一条 连接边。Elegans 表示的是线虫的神经网络，节点表示线虫的神经元，突触或间隙表示 为连边。Footbal 数据中的节点代表国家队，这些国家中的球员频繁参加国外联赛，若 国家队中有成员在另外一个国家的联赛中踢球，那么这两个国家队之间存在连边信息。 Jazz 表示 jazz 演奏家之间的合作网络。Karate 是社会网络分析领域中的经典数据集， 该网络构造了美国一所大学空手道俱乐部中 34 名成员的社会关系，若两人在现实之中 是好友关系，那么该两人之间存在一条连边。

3.3.2 网络特征的统计

表1. 网络的统计特征

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络 | |V| | |E| | D | C | <k> | <d> | H |
| Elegans | 297 | 2148 | 0.0488 | 0.2923 | 14.4646 | 2.4553 | 1.8008 |
| Football | 35 | 118 | 0.1983 | 0.3389 | 6.7428 | 2.1226 | 1.4880 |
| Jazz | 198 | 2742 | 0.1405 | 0.6174 | 27.6969 | 2.2350 | 1.3951 |
| USAir | 332 | 2126 | 0.0386 | 0.6252 | 12.8072 | 2.7381 | 3.4638 |
| Karate | 34 | 78 | 0.1390 | 0.5706 | 4.5882 | 2.4081 | 1.6932 |

|V|代表网络中节点的数目。|E|代表网络中边的数目。D表示网络的密度，计算公式为 。C 表示节点的聚类系数，上文已经给出了计算公式。<k>表 示网络中节点度的平均值。<d>表示网络的平均最短距离，该值需要计算所有节 点对之间的最短距离再求平均。H 表示网络节点度之间的异质性，若节点的度分 布越不平均，该值则越大，计算公式为，分母部分先求所有节点度的平均值再平方，分子部分是先求各个节点度的平方再求平均值

3.4 实验结果分析

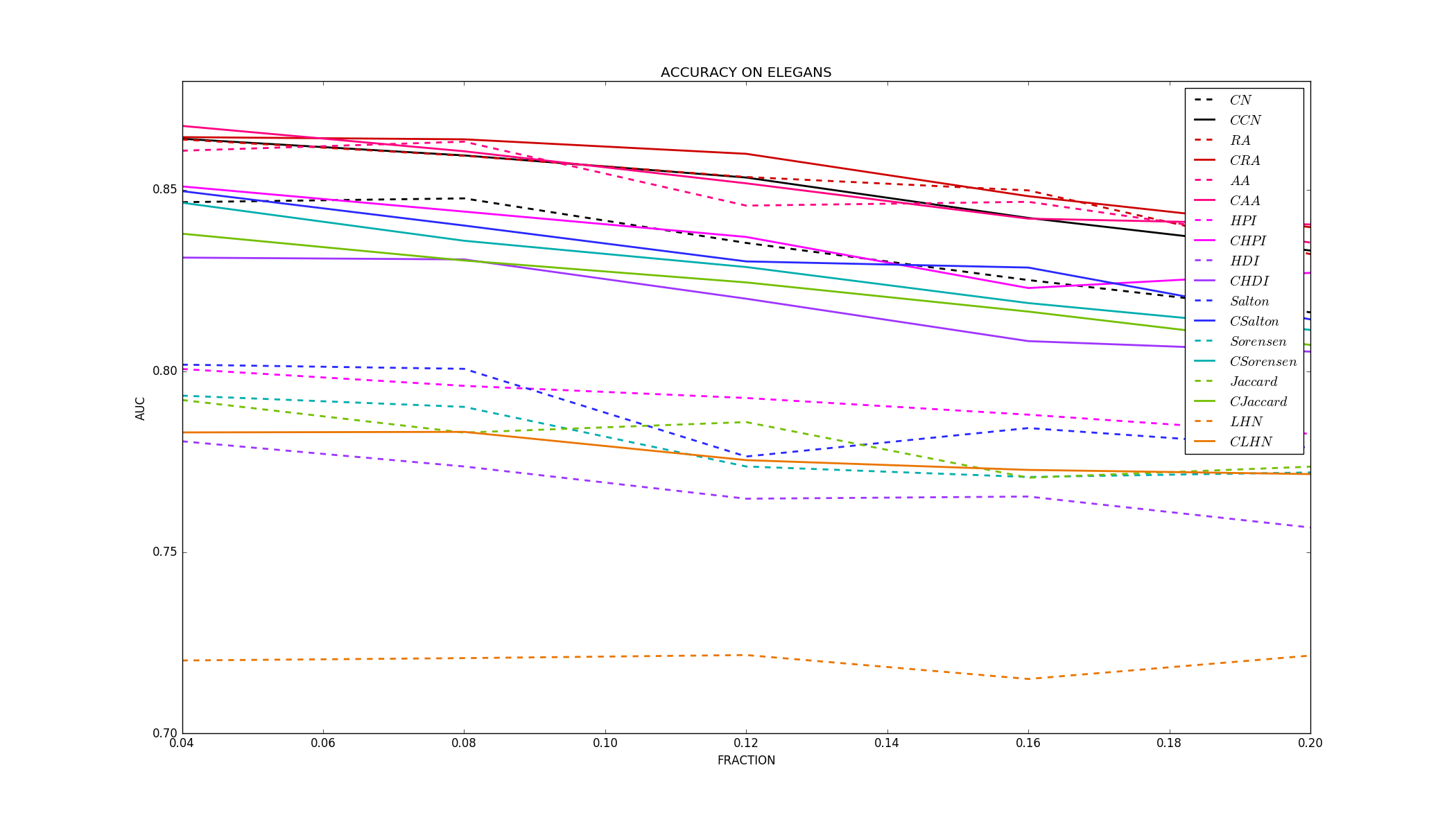
3.4.1 参数设定

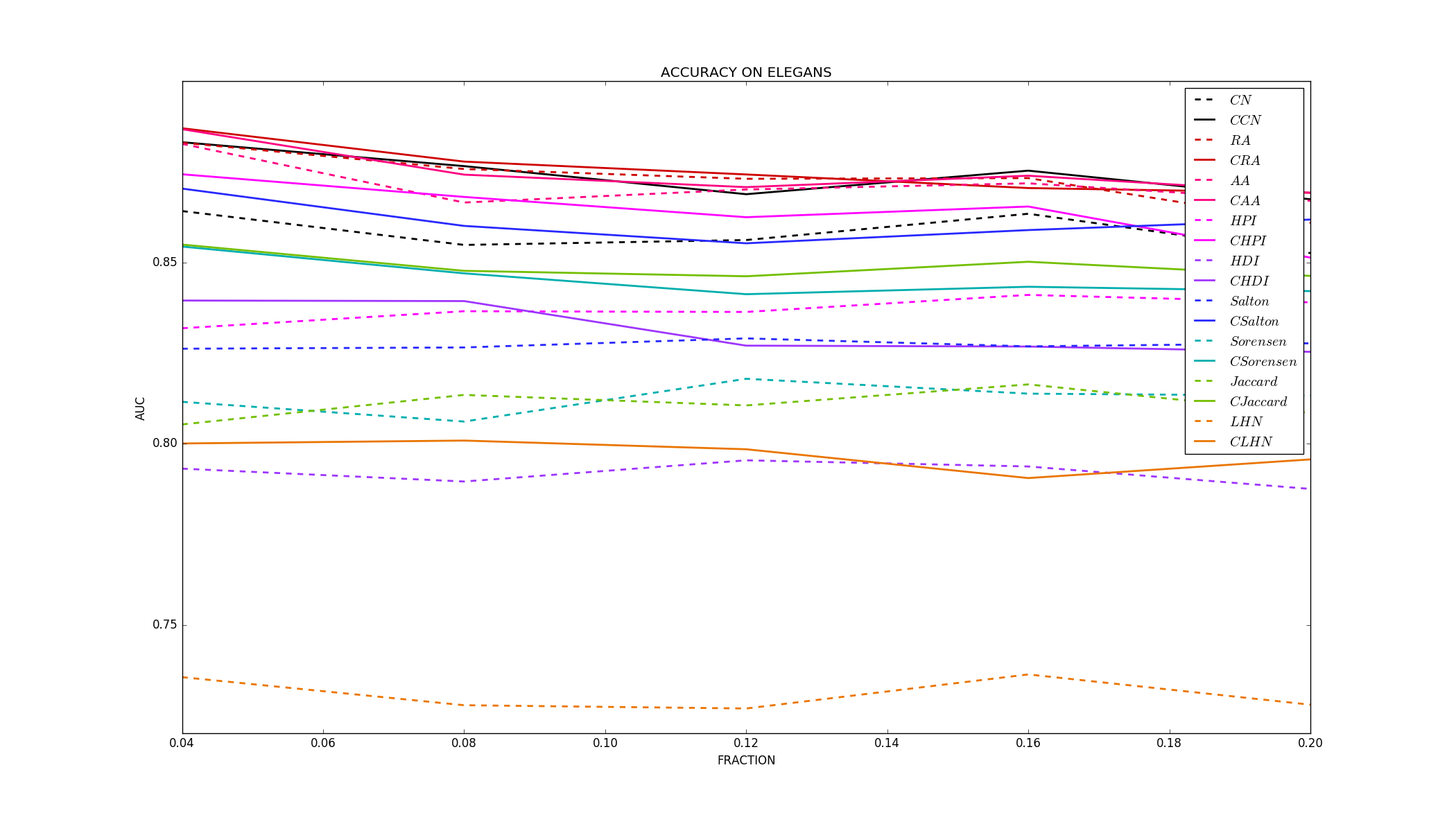
为了保证试验的准确定，当我们确定某一个 p 值之后，我们会对测试集进行 100 次划分，划分之间彼此不关联，这主要是防止测试集的随机性对试验结果造 成的误差。测试集划分好之后，我们将内层循环次数确定为 1000 次，也就是说 总循环比较的次数为 100000 次，该次数对应于上述 AUC 指标中的 n。

3.4.2 实验结果以及分析

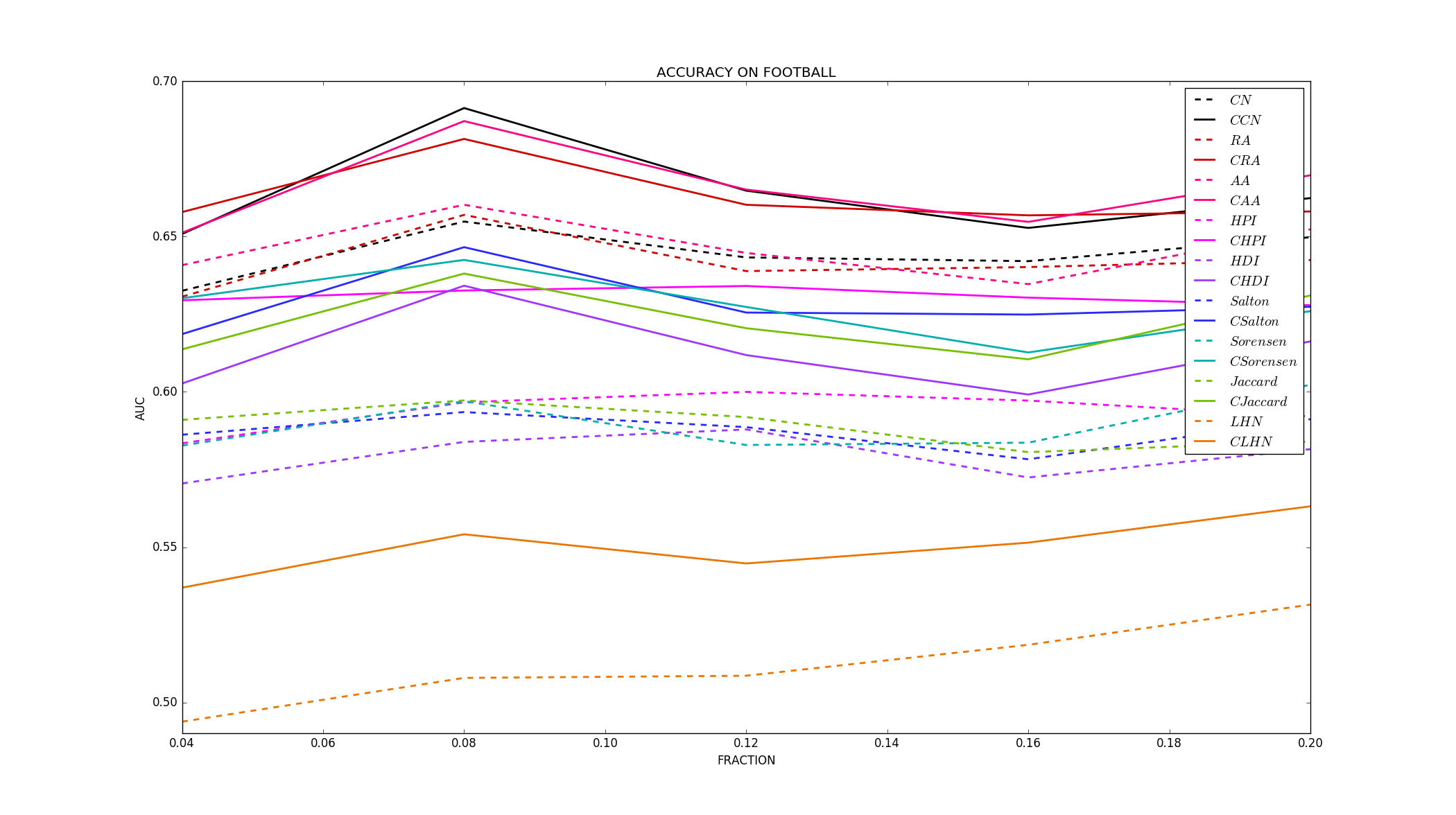
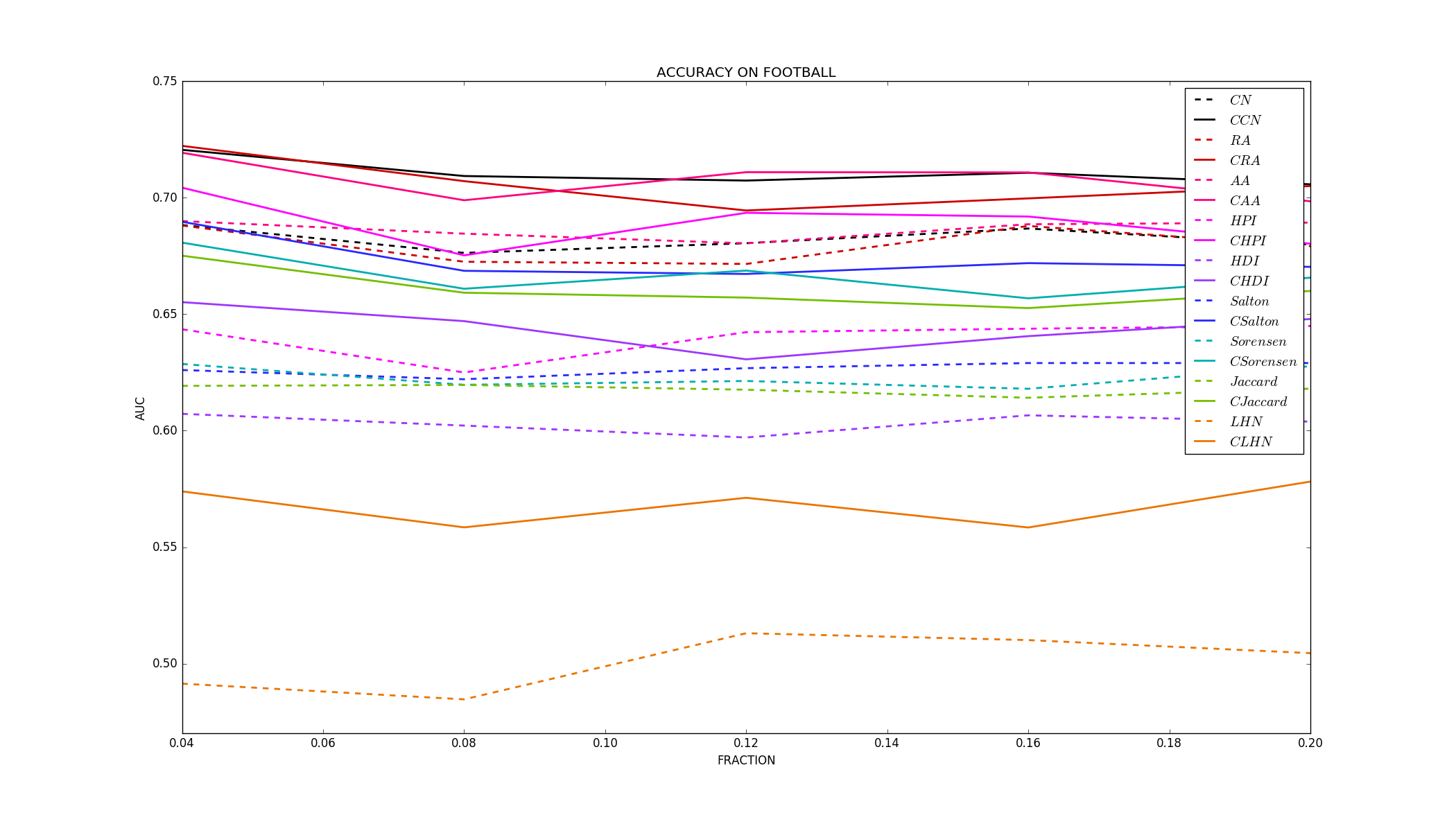
(1)Elegans 网络中缺失边的发现与错误边的纠正，左侧为缺失边的发现，右 侧为错误边的纠正。下图中共 18 种算法进行比较，其中原算法 9 种，加入聚类 系数改进后的算法 9 种。为了便于比较，我们将某一算法和改进后的算法一一对 应，这主要体现在绘图时线段的颜色，如下图所示 CN 和加入聚类系数后的 CCN 算法均由黑色线段表示，RA 和 CRA 均由红色表示。此外，原算法用虚线表示， 改进后的算法用实线表示。结果图的横坐标表示所选取的训练集的比例 p，纵坐 标表示算法的预测准确度。

从 Elegans 的实验结果图可以看出改进后的算法相比较原算法在准确率上有所上升，该效果的 升不仅在于发现丢失边的情况同时适用于错误连边的纠正。随 着测试集比例的 升，改进后的方法同样优于原算法。

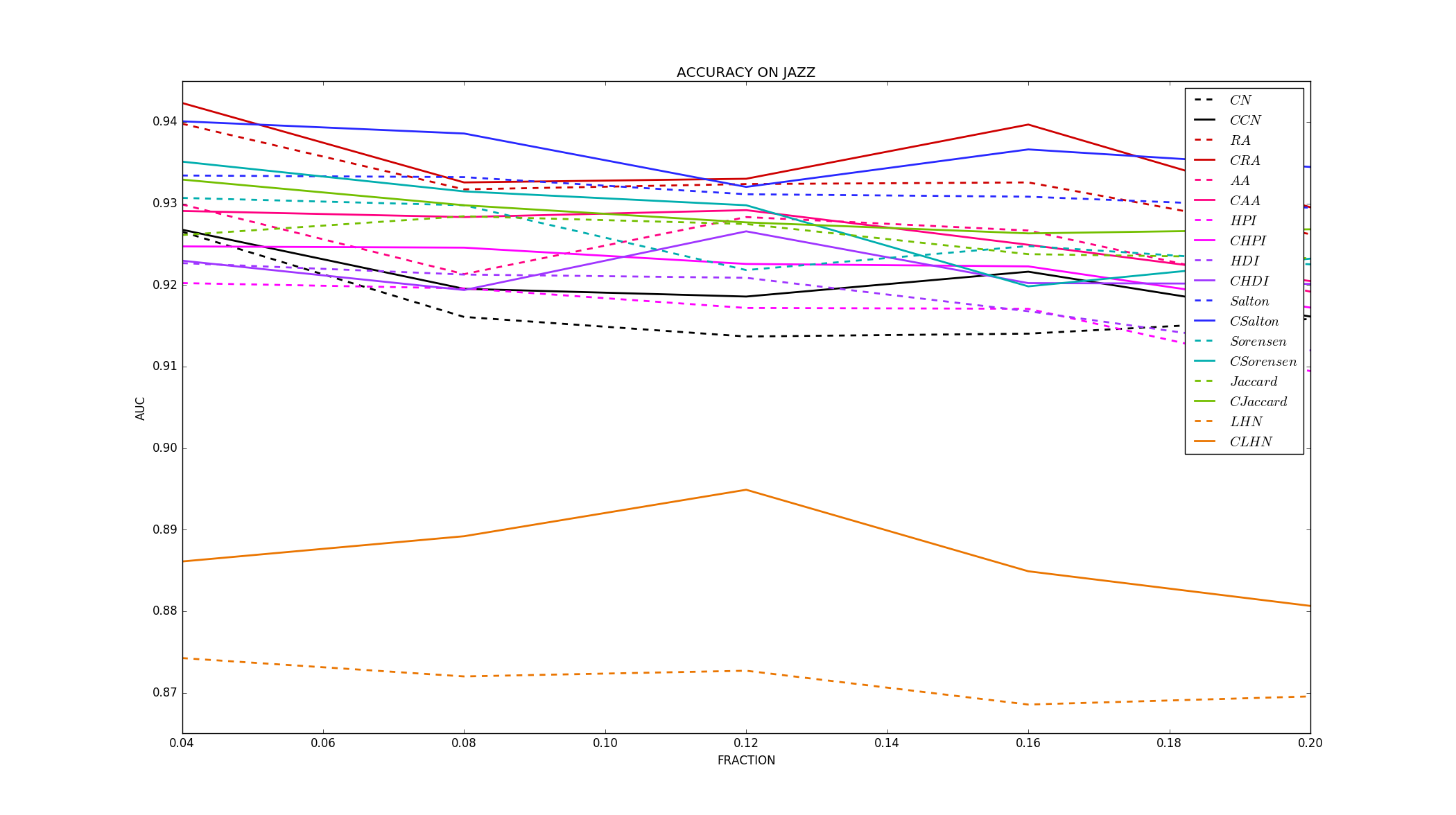


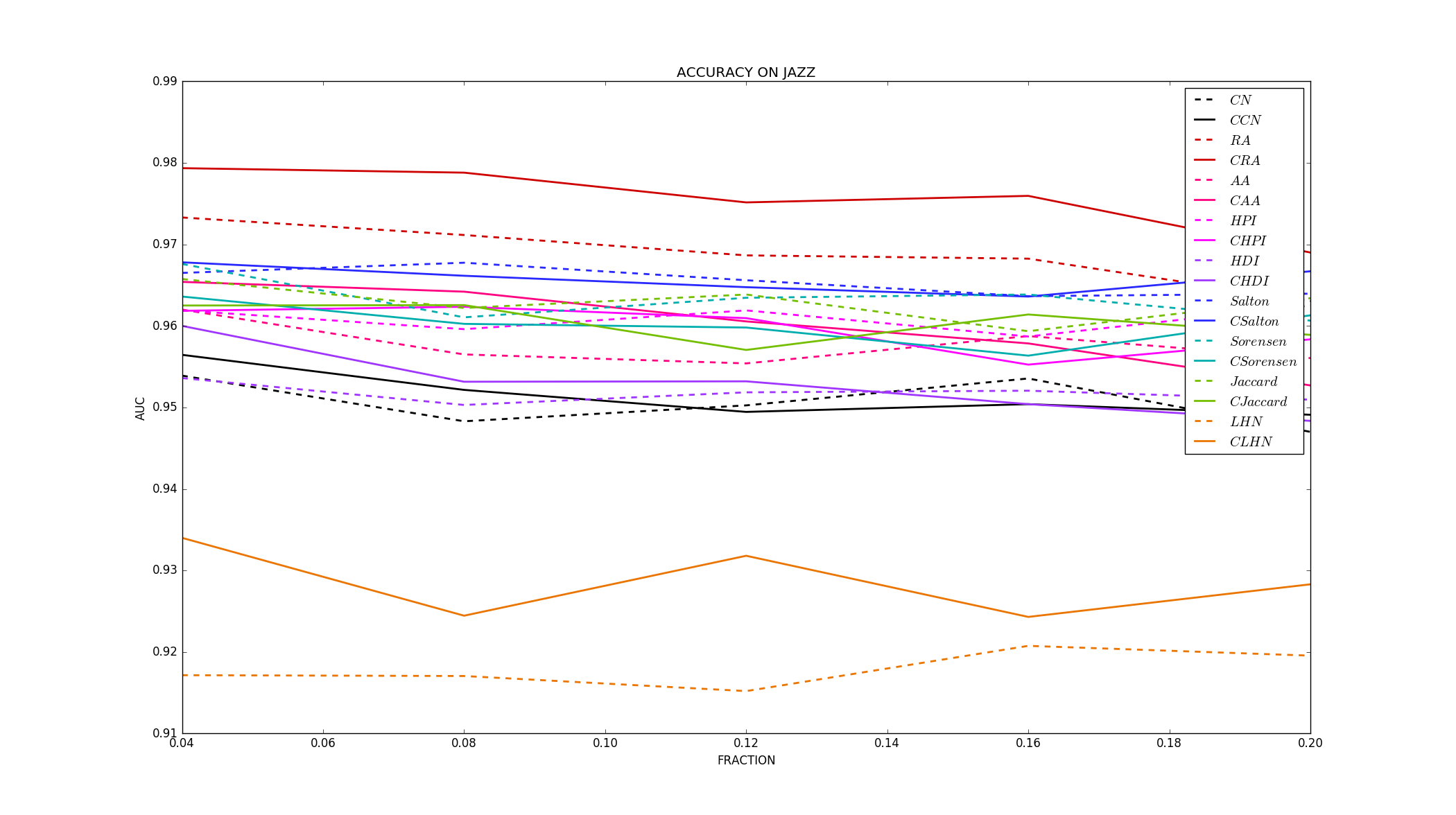


(2)Football 网络的实验结果如下图所示。实验结果表明改进后的算法在整体上明显 优于原先的算法，此外在链路预测的过程中我们发现当 p=0.08 时，预测的准确率达到 最高值。

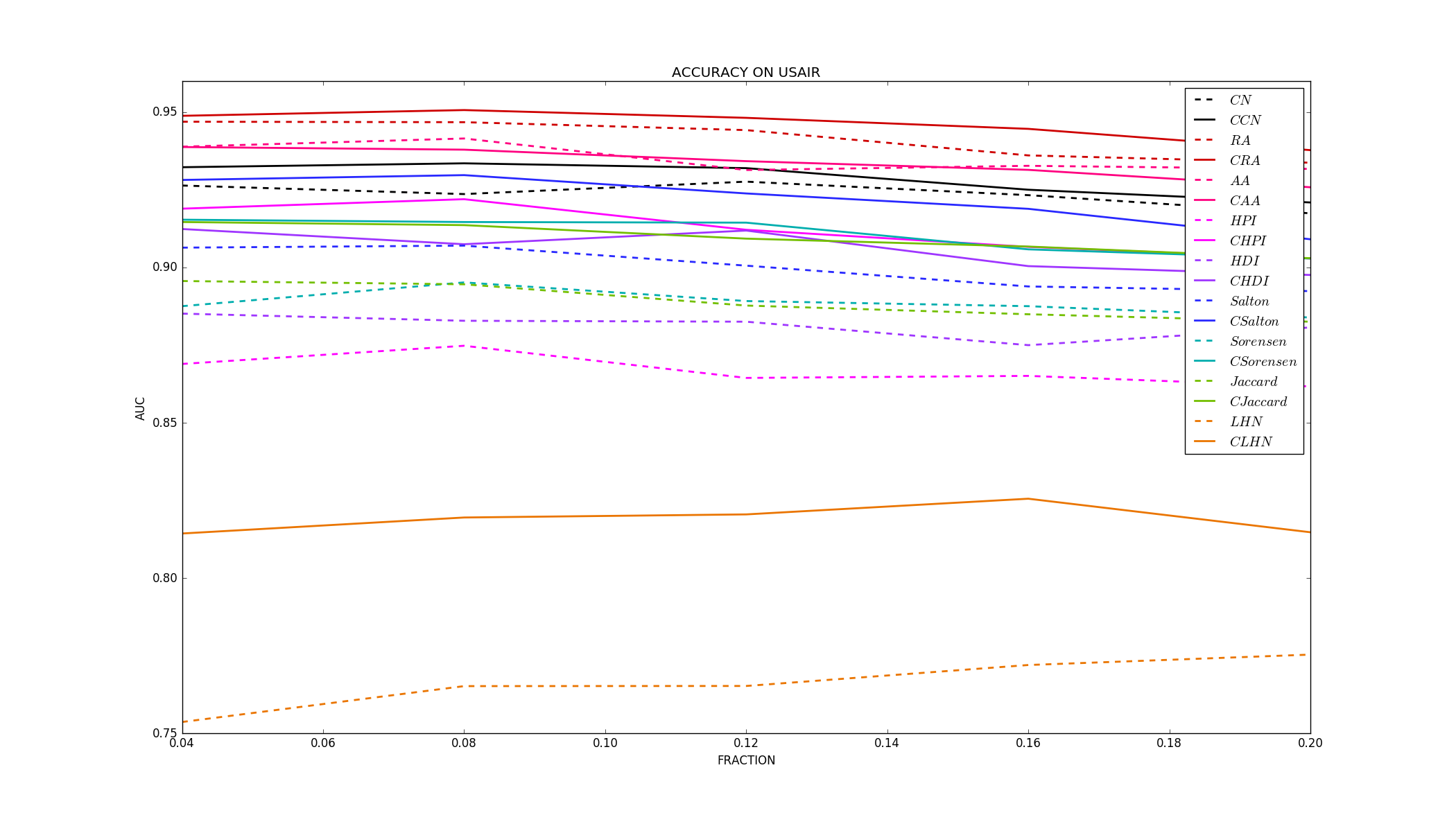
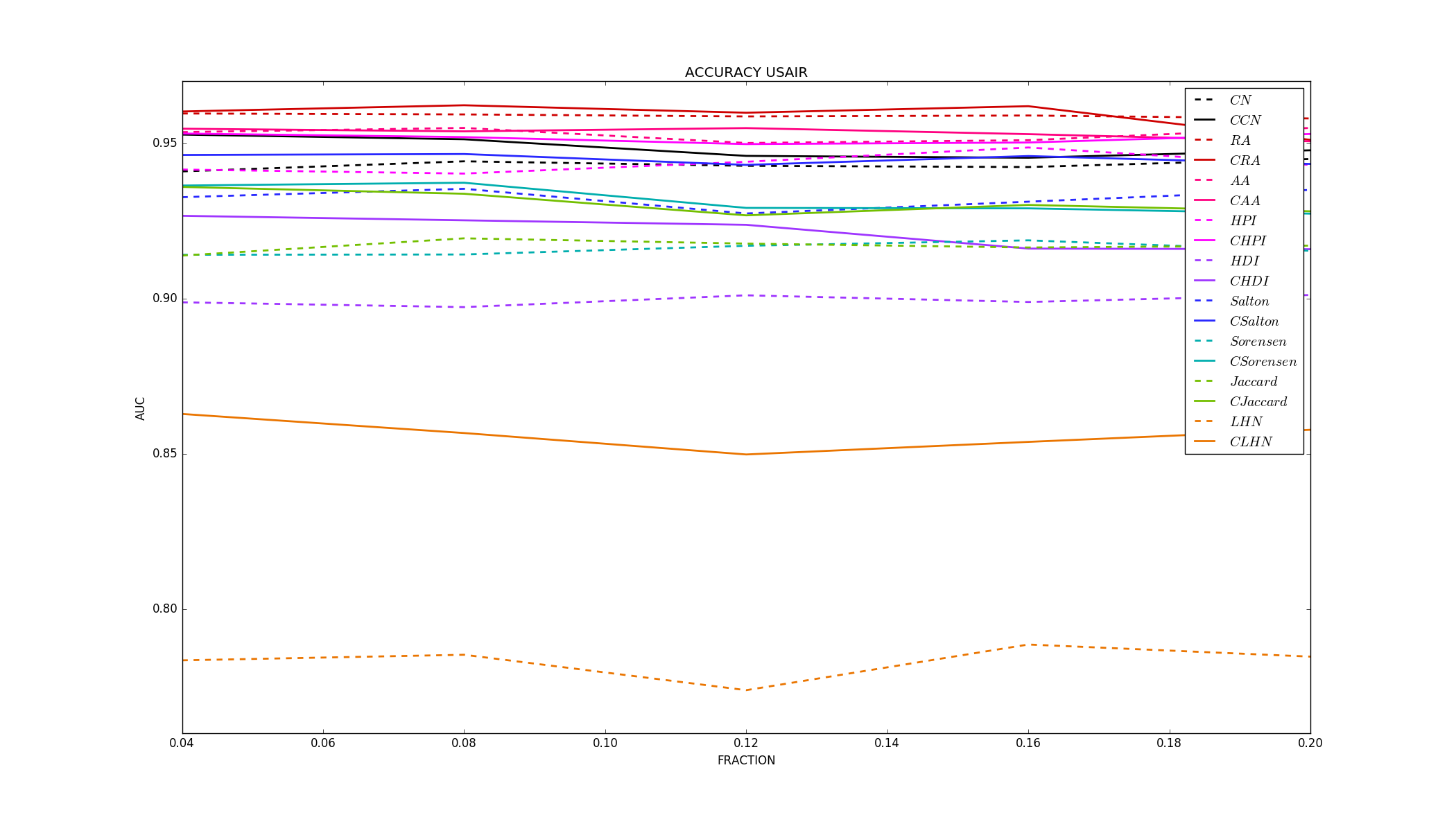


(3)Jazz 网络的实验结果如下图所示。改进后的算法只在少数几个点之上不能超越原 算法，整体表现仍然十分优异。

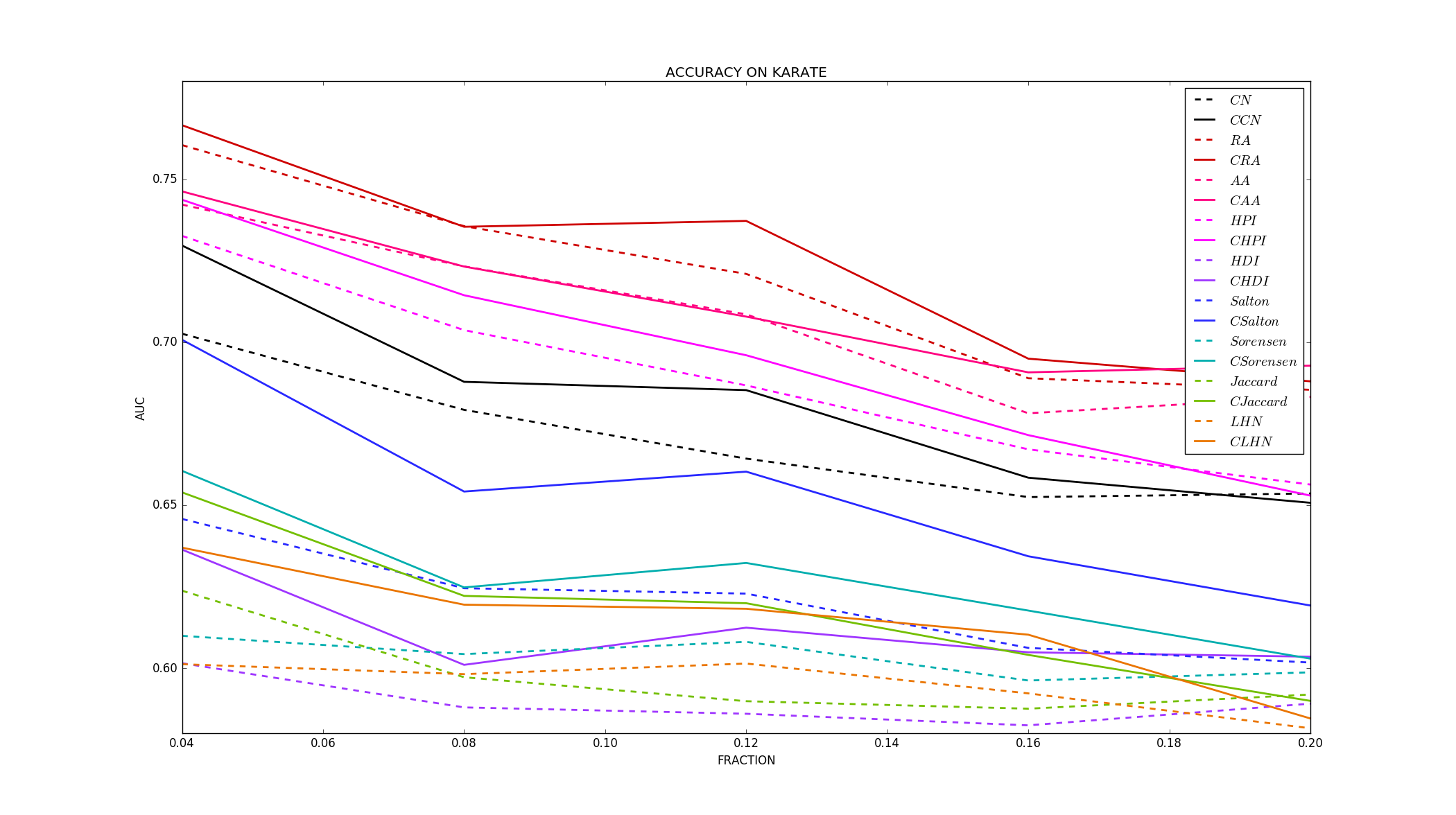


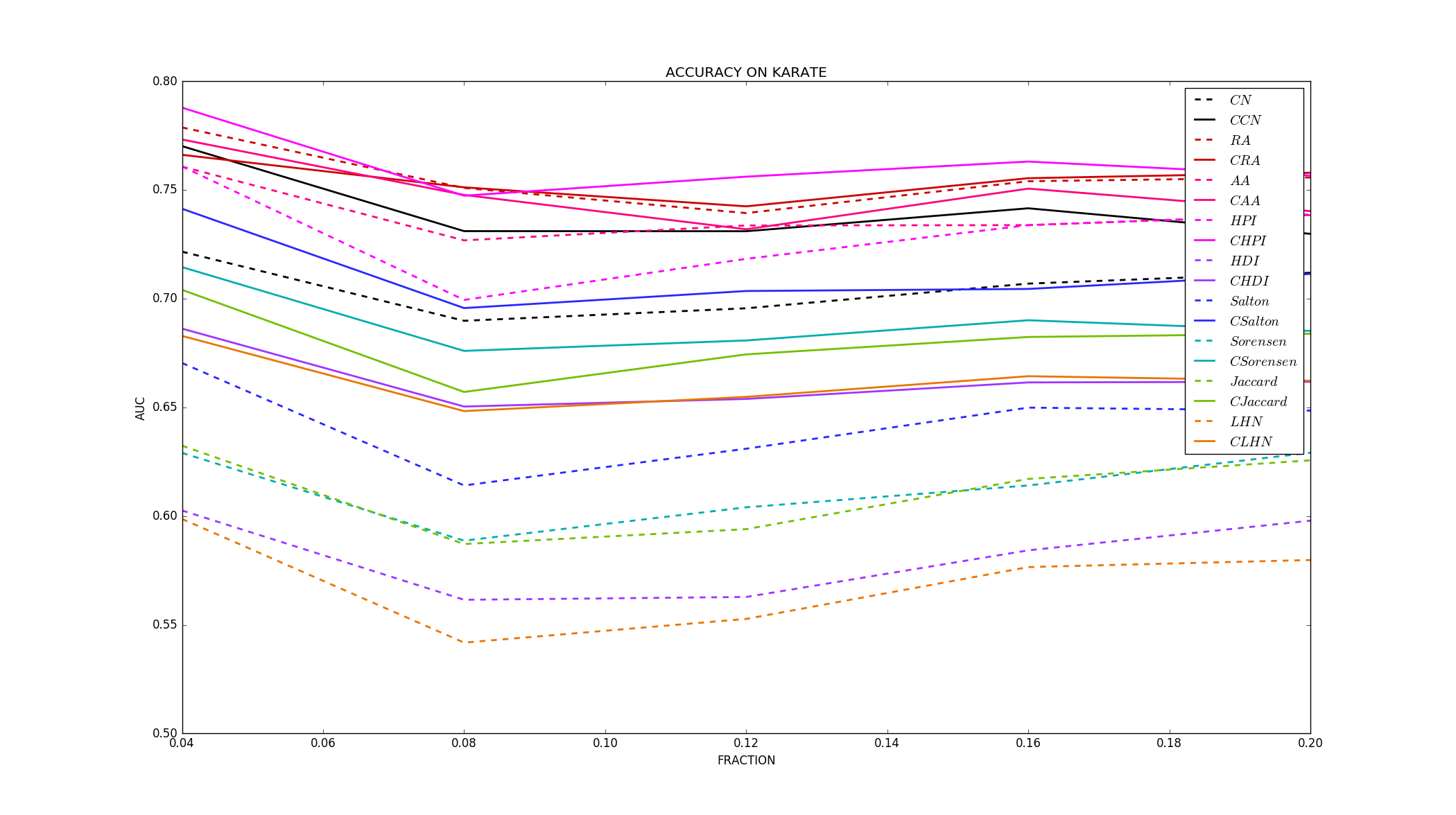


(4)USAir网络的实验结果如下图所示。该实验结果分层现象十分明显，改进后的算法 曲线明显在改进前算法的曲线之上，这表明加入聚类系数这一节点的局部信息之后，该 9 种算法在美国航空网络的预测上效果明显 升。



(5)Karate 网络的实验结果如下图所示。与上述四种网络结果共同验证了算法改进的 有效性，此外我们发现这些链路预测算法对错误连边的发现要略优于预测不存在的边。





3.5 本章小结

聚类系数表示节点邻居之间联系的紧密程度，较高的聚类系数表明邻居间较 高的连接概率。受该指标的影响，我们将其特性添加到现有的 9 种链路预测算法 之中得到了我们改进后的 9 种算法。通过在 5 种真实网络上的实验，我们发现了 改进后的算法在这 5 种网络上有着更好的表现。除此之外，针对不同大小的训练 集划分比例，我们进行了大量的比较实验，过往研究很少会探讨训练集 p 的大小 对于实验结果的影响。我们发现不同大小的训练集对于实验结果会产生影响，但 改进后的算法与原算法的相对大小不变。最后，我们的链路预测实验还包含了算 法在错误连边上的发掘能力，实验表明，若我们有算法 A、B，算法 A 相比 B 在 丢失边的发掘能力上可能会更好，但若换做是错误边的发现，那么 B 算法可能会 优于A算法。

本研究只简单的考虑了节点的聚类系数对于链路预测算法的影响，除此之 外，在未来的研究当中我们可将更多的网络特征融入到算法之中来进行算法的 升。本次实验的网络只有 5 种，更多数据的实验有待补充，我们不排除特定网络 之上，改进后算法的准确率不如原算法。本次实验网络的节点只有几百个，更大 规模网络的链路预测研究有待进一步深入。

第四章 交通拥堵问题的研究

时间序列是一种常见的数据表现形式，对于时间序列处理而言，一个普遍的任务就是比较两个序列的相似性。在时间序列中，需要比较相似性的两段时间序列的长度通常不相等，因此使用传统的距离计算算法无法有效求得两个时间序列之间的距离（或相似性）。

DTW（动态时间规整）算法是专门用来衡量两个长度不同的时间序列之间相似度的的方法。它是一个典型的动态优化问题，用满足一定条件的时间规整函数描述测试模板和参考模板的时间对应关系，求解两模板匹配时累计距离最小所对应的规整函数。

假设有两个序列Q和C，长度分别为n和m：和。为了对齐这两个序列，先构造一个的矩阵网格，矩阵元素表示和两个点的距离。每一个矩阵元素表示点和的对齐。动态规整算法可以归结为寻找一条通过此网格中若干格点的路径，路径king通过的格点即为两个序列进行计算的对齐的点。

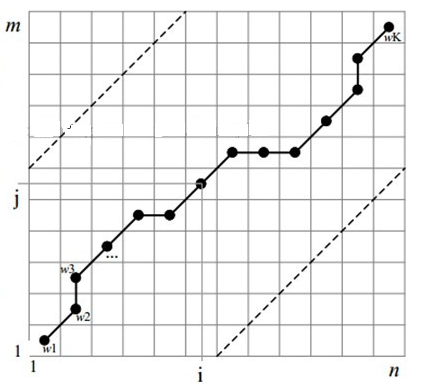


图2.1 触屏行为连续型特征距离计算模型

将该路径定位为warping path规整路径，并用W来表示。W的第k个元素定义为，定义了序列Q和C的映射。即：且。

对于这条路径，需要满足如下几个约束：

1. 边界条件：。即选定的路径必定从左下角出发，在右上角结束。
2. 连续性：如果，那么对于路径的下一个点需要满足和。也就是不可能跨过点去匹配，只能和自己相邻的点对齐。这样保存Q和C中的每个坐标都在W中出现。
3. 单调性：如果，那么对于路径的下一个点需要满足0。这限制W上面的点必须是随着时间单调进行的。

对于满足上述约束的路径，选择总的距离和最小的那条。因此DTW算法可表示为：。

手势密码触屏行为数据中的X坐标、Y坐标、压力、接触面积、速度、接触面积长轴和接触面积短轴等均为时间序列数据（连续型数据）。在不同次的手势密码行为中采集而来时间序列数据的维度通常是不一致的。为了方便后期分类算法和聚类算法的使用，需要通过DTW算法将不同维度的时间序列数据（连续型数据）转换成同维度的特征距离向量。

手势密码触屏行为连续型特征向量表示完成一次手势密码行为所包含离散特征的向量。其中n表示特征有n维，即。

手势密码触屏行为连续特征距离向量为该行为连续特征向量通过DTW算法计算对应特征获得的向量，表示为。且。

## 2.4 评价指标

分类和认证的目的在于将未知身份的用户数据进行分类处理，判定用户属于合法用户还是非法用户，从而实现用户身份认证的过程。本小节主要介绍手势密码行为认证中使用的SVM分类算法以及文章中使用的SVM参数优化算法。

2.5 本章小结

支持向量机（Support Vector Machine，SVM）可用于模式分类和非线性回归。它的主要思想是建立一个分类超平面作为决策曲面，使得正例和反例之间的隔离边缘被最大化；支持向量机的理论基础是统计学习理论，更精确地说，支持向量机是结构风险最小化的近似实现。

这个原理基于这样的事实：学习机器在测试数据上的误差率以训练误差率和一个依赖于VC维数的项的和为界，在可分模式情况下，支持向量机对于前一项的值为零，并且使第二项最小化。

手势密码触屏用户行为认证模式是典型的而二分类问题，，可将用户数据分为正常用户数据和异常用户数据，而SVM是经典的二分类分类器。同时，SVM算法要求样本数相对较少，适合数据量不大的分类认证。故而本文选择SVM作为模式构建的分类器算法。

SVM通过寻找一个超平面将不同的数据集分类成二部分。在训练阶段，SVM对训练集使用RBF核函数。在测试阶段，会计算测试集和经SVM训练而得到的超平面之间距离作为分类的分数，来预测该用户是正常用户数据还是异常用户数据。

2.2.2 SVM参数优化算法

在SVM做分类预测时需要调节惩罚参数c和核函数参数g，才能得到比较理想的预测分类准确率。

关于SVM参数的优化选取，国际上并没有公认统一的最好的办法，本文使用的方法是让c和g在一定的范围取值，对于取定的c和g，把训练集作为原始数据集并利用K-CV方法得到再此组c和g下训练集验证分类准确率，最终取使得训练集验证分类准确率最高的那组c和g作为最佳的参数。并选择能够达到最高验证分类准确率中参数c最小的那组c和g作为最佳的参数。

因为过高的c会导致过学习状态发生，即训练集分类准确率很高而测试集分类准确率很低，所以在能够达到最高验证分类准确率中的所有的成对c和g中认为较小的惩罚参数c是更佳的选择对象。

增加了参数优化算法后的SVM分类流程如下。



图2.2 SVM整体流程

2.3 k-means聚类算法

k-means算法是一种基于形心的分类方法，簇的形心就是这个簇的中心点，这类方法采用簇的形心来代表这个簇。形心的定义方法有很多种，例如用该簇内所有点的均值来定义形心，或者用簇内所有点的中心值来定义。k-means算法把形心定义簇内所有点的平均值[33,34,35]。

对于聚类效果的评价，可以采用簇内变差来度量。假定簇内一个对象为x，该簇的形心为，用来表示这个对象到簇的形心的欧式距离。簇的簇内变差为这个簇内所有对象到形心的欧式距离的平方和：。 其中，E表示为簇内变差，x为簇内的某个数据对象，为簇的形心。聚类的最终目标是使得每个簇尽可能紧凑，而簇与簇之间独立。

k-means算法的聚类流程如下：首先，在包含了n个数据集的集合Data中，随机选择k个对象，每个对象表示一个簇的最初的均值，也就是簇的形心。对其余的每个对象，分别计算它们到这k个对象的欧式距离，将它们分配到距离最近的那个形心代表的簇中。然后，重新计算每个簇内对象的均值，产生簇新的形心。利用新生成的簇的形心，重新分配Data中的对象。重复这个过程，直到不再发生变化，即新一轮的簇的形心和前一轮的相同为止。

k-means的聚类效果如下图所示



图2.3 k-mean的聚类效果图

本文中主要用k-means来解决手势密码触屏行为多用户单账号场景下多个用户的数据混在一起的现实问题。通过聚类来区分各个数据的来源，作为后期进一步研究工作开展的基石。

# 第3章 数据采集和预处理

本章主要介绍手势密码触屏行为数据的采集和预处理。首先介绍了手势密码触屏行为数据采集器的设计和开发，及后续实验分析中使用到的数据集。之后介绍了针对手势密码触屏行为原始数据进行数据预处理，包括数据清理、数据变换和归一化。最后，介绍了手势密码行为特征的选取及相应行为特征向量的构建。

## 3.1 数据采集

在本节中先介绍了手势密码触屏行为数据采集器的设计；继而讲解了在Android平台下开发的采集手势密码触屏行为数据的APP的实现，并展示了保存在json格式数据包下数据解析到PC端时的样式；最后介绍了应用在后续试验分析的两个数据集。

### 3.1.1 数据采集器

为了采集用户在完成手势密码过程中手指在屏幕上滑动时，X坐标、Y坐标、压力、接触面积、速度等特征数据的变化，用java语言开发了一个Android平台上的手势密码触屏行为数据采集器。

该采集器模拟一个手势密码输入环境，每个用户输入用户名后，即可开始数据采集工作。采集器记录了完成手势密码时，手指与屏幕接触时的时间点、X坐标、Y坐标、压力、接触面积、接触面积长轴和短轴等行为数据。每完成一次手势密码输入，即将该次行为数据记录在Android终端上以该用户名命名的json格式的数据包中。主流的数据分析语言，如matlab、python、R等，均支持对json数据包的直接读取使用。后期可将采集到的各个用户的数据包导入PC端，进行相应的数据分析处理。

数据采集器的开发平台为Eclipse，开发语言为Java，开发以后的APP应用在Android系统终端上。

手势密码触屏数据的采集的实现过程中主要使用了MotionEvent对象。MotionEvent包含了关于发生触摸的位置和时间的信息，以及触摸事件的其他细节。

通过getAction()方法，可以获得触摸事件，如ACTION\_DOWN、ACTION\_UP、ACTION\_MOVE等。此处我们选择ACTION\_MOVE，即表示用户的手指在触摸屏上移动时的状态。

为了进一步获取ACTION\_MOVE动作下的触屏数据，需要调用event.getDowntime, event.getX(), event.getY(), event.getPressure(), event.getSize()等方法能分别记录触屏时的时间、X坐标、Y坐标、压力以及接触面积等数据。

开发实现的采集器如下图所示。

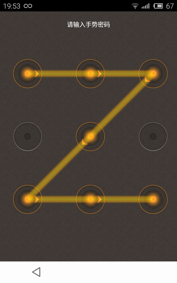


图3.1 手势密码模拟器

3.1.2 解析JSON格式数据包

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式。JSON完全独立于语言的文本格式，这些特性使JSON成为理想的数据交换语言。易于人阅读和编写，同时也易于机器解析和生成。简单地说，JSON可以将JavaScript对象中表示的一组数据转换为字符串，然后就可以在函数之间轻松地传递这个字符串，或者在异步应用程序中将字符串从Web客户机传递个服务器端程序。

JSON建构于两种结构：（1）“名称/值”对的集合（A collection of name/value pairs）。不同的语言中，它被理解为对象，记录，结构，字典，哈希表，有键列表，或者关联数组；（2）值的有序列表，在大部分语言中，它被理解为数组。

后续的数据分析，主要应用的软件和语言是matlab。在用matlab解析JSON格式时运用的工具包是jsonlab-1.1。我们的目的是讲保存在Android手机上的JSON格式的数据包里的数据解析出来，主要运用的函数是 dat=loadjson(['examples' filesep 'example1.json'])，其中examples表示文件所在的文件夹名，example1.json表示JSON格式的文件名。

对于每一次完成的手势密码，其采集到的数据如下图所示。其中有三个数组，accelerationPoints、flingPoints以及orientationPoints。第一个数组的数据表示完成手势密码时加速度传感器采集到的相关数据，第二个数组的数据表示触屏行为相关数据，第三个数组的数据表示方向传感器采集到的相关数据。本文实验选用的是触屏行为相关数据，其他类的数据可供后续研究使用。数组后面的数字204表示完成该次手势密码过程中，采集了204个数据点。

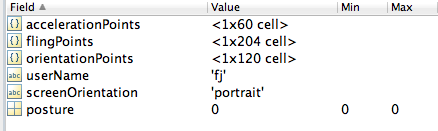


图3.2 每次手势密码采集到的数据

对于每个数据点上采集到的数据，如下图所示。eventTime表示手指在屏幕上移动到该点的时间，positionX表示手指在屏幕上的X坐标，positionY表示手指在屏幕上的Y坐标，pressure表示手指在屏幕上的压力，size表示手指在屏幕上的接触面积，touchMajor表示手指与屏幕接触面积的长轴，touchMinor表示手指与屏幕接触面积的短轴。

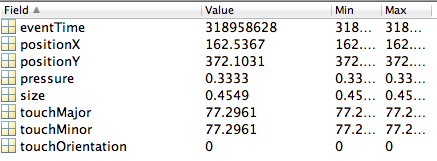


图3.3 每个数据点上采集到的数据

将完成一次手势密码所采集的所有数据点的数据，读取下来，保存至一个数组，如下图所示。以上述数据为例，该次手势密码触屏行为数据将以一个204\*7的数组形式供后续的实验分析。其中，204表示数据采集器在完成该次手势密码触屏行为的过程中采集了204个数据点；7表示总归有tStamp、Xpos、Ypos、Pressure、Size、touchMajor和touchMinor等7个触屏行为特征。



图3.4 数据采集器采集原始数据

### 3.1.3 数据集

本文实验使用了两个数据集。

数据集一是由德国慕尼黑大学的Alexander De Luca教授提供的公开数据集，该数据集是由Android移动手机（Android2.1或者更高的操作系统）上的模拟手势密码解锁的APP采集而来。该数据集是由33名采集者耗时21天在各自的手机上采集而来，对于每个用户，都有30次正样本手势密码行为数据以及99次负样本手势密码行为数据。总计有近4000次手势密码行为数据。

使用该数据集的意义在于能够更好的同手势密码触屏行为模式构建其他研究者的研究方案进行类比。

数据集二是通过上一小节描述的手势密码触屏行为数据采集器采集而来的。数据采集基于同一部Android手机，总归有15名采集者，每5名采集者完成同一个手势密码图案，每个采集者需要完成相同的手势密码120次，因此累计完成手势密码共有1800次。

完成的手势密码采集图案如下图所示。

图3.5 三个手势密码采集图案

## 3.2 数据预处理

通过数据采集器采集而来的数据往往掺杂着脏数据，无法直接进行数据挖掘，或挖掘结果差强人意。为了提高数据挖掘的质量，因此有必要在数据分析前，对原始数据进行预处理。本文实验使用的数据预处理方法主要包括数据清理、数据变换以及归一化。

### 3.2.1 数据清理

由于手势密码数据采集器设计实现过程中的问题或者采集者完成手势密码时产生的问题，会出现比该用户正常手势密码行为序列明显更长或者更短的序列。图3.3所示为同一用户完成同一手势密码行为的数据中的正常序列、过长序列、过短序列。显然，算法需要将数据集中的过长序列和过短序列进行清洗。

数据清理采用通用的基于规则的方法。此处的过滤规则为：将手势密码序列长度超过平均长度一倍的或者序列长度不足平均长度一半的序列数据视为无效数据，从原始数据集中剔除。

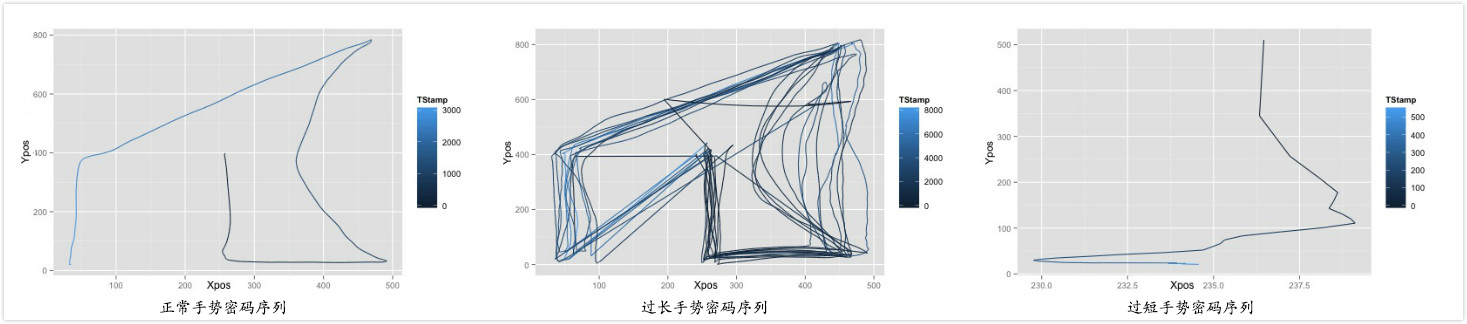


图3.6 正常异常手势密码序列对比

### 3.2.2 数据变换

如图3.1所示，用户完成一次手势密码行为时，数据采集器采集到的时间特征是该时刻的时间点，为了方便后续的数据分析处理，需要将改时间转换成以手指触屏为0时刻点的相对时间。

通过各个点的绝对时间减去第一个点的绝对时间，即可获得相对第一个点的相对时间变化。

### 3.2.3 归一化

如图3.1所示，用户完成一次手势密码行为的数据包括时间、X坐标、Y坐标、压力、接触面积、接触面积的长轴和短轴等基本特征。由于不同的特征有不同的量纲，而不同量纲的特征对应的数值差距特别明显。因此，需要通过归一化的算法，将有量纲的行为特征数据经过变换化为无量纲的行为特征数据。

对于每次手势密码行为数据，若对数据集整体进行归一化，由于量纲数量级差别悬殊，会使较低数据集的属性变为0，会使原始信息过多丧失。因此，须针对每一个特征维度进行归一化。归一化的方法采用min-max标准化（Min-Max Normalization），是对原始数据的线性变化，使结果映射到[0,1]之间。其中max为样本数据的最大值，min为样本数据的最小值。转换函数如下：x’=(x-min)/(max-min)。

## 3.3 特征提取

如图3.1所示，通过手势密码触屏行为数据采集器采集到的数据，主要包括随时间变化的X坐标、Y坐标、压力、接触面积、接触面积的长轴和接触面积的短轴等。

通过分析这些原始数据中的特征，可以提取更多地反应用户行为的特征。可以通过X坐标、Y坐标的变化获取随时间变化的速度和加速度特征；通过最后采集点的时间减去第一个采集点的时间，可以得到完成手势密码的总时间，同类型的还可以提取对应手势密码行为采集点的个数；对于随时间变化的X坐标、Y坐标、压力、接触面积、接触面积的长轴和短轴等特征，可以分别计算其标准差、均值、最大值、最小值。

根据是否为时间序列数据，将上述特征分为连续型特征和离散型特征。连续型特征即为随时间动态变化的特征，包括原始数据中的X坐标、Y坐标、压力、接触面积、接触面积的长轴和短轴，以及经过二次挖掘的速度、加速度。离散型特征则表示反应手势密码触屏行为全局性能的特征，比如总时间、数据采集点的个数、各个连续型特征的均值、标准差、最大值、最小值等。

如所示即为本文中所涉及的手势密码触屏行为特征。

表3.1 手势密码触屏行为特征

|  |  |  |
| --- | --- | --- |
|  | **触屏特征类别** | **手势触屏特征** |
| 连续型特征 | 原始数据类 | X坐标、Y坐标、压力、接触面积随时间变化 |
| 二次挖掘类 | 速度、加速度随时间变化 |
| 离散型特征 | 整体类 | 总时间、采集点的个数 |
| 方差类 | 连续型特征的方差 |
| 均值类 | 连续型特征的均值 |
| 特殊类 | 连续型特征的最大值和最小值 |

# 第4章手势密码触屏行为模式的构建和认证

本章主要介绍了手势密码触屏行为模式的构建和认证。首先介绍了行为模式构建方法的总体架构及后文所涉及的概念介绍。然后介绍了手势密码行为特征距离向量计算的方法，包括参照集特征向量的计算，以及利用参照集特征向量计算特征距离向量的方法。接着介绍了如何利用SVM分类器，针对特征距离向量，进行训练和预测。最后介绍了基于本文提出的手势密码触屏行为模式构建方法进行实验的实验结果以及同相关领域研究实验结果的比较。

## 4.1 总体架构

### 4.1.1 方法总体架构

在完成手势密码的过程中，不同用户往往有不同的完成手势密码的行为习惯，而移动终端上压力传感器采集的各类行为数据就是挖掘用户行为习惯的主要依据。面向手势密码触屏行为模式的用户认证方法就是基于移动终端上的行为数据，是在较短时间内以较高精确度完成用户认证的一种方法。

该方法的总体架构如图4.1所示。它包括三个主要模块：（1）预处理模块；（2）特征距离向量集构建模块；（3）训练/预测模块。

预处理模块包括对原始数据中的异常部分的清洗、数据变化和数据归一化，在上文已介绍。

特征距离向量集构建模块包括基于获取特征向量参照集，及基于特征向量集计算特征距离向量集，其中特征向量分量间的距离计算用到曼哈顿算法和DTW算法。

训练/预测模块则是将数据以特征距离向量集形式构建训练集和测试集，通过SVM分类器对训练集训练获取分类模型，继而用测试集进行预测，获取代表分类效果的漏检率和误检率。

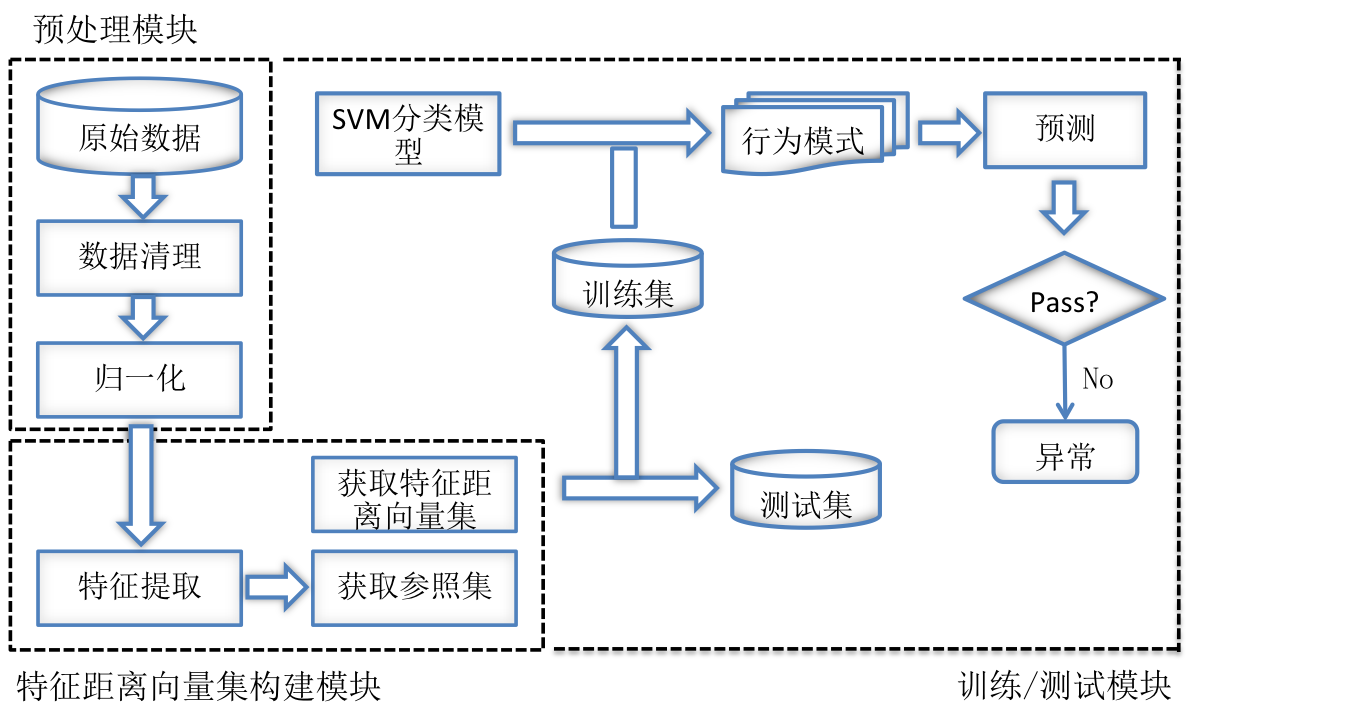


图4.1 用户认证方法总体架构

### 4.1.2 概念定义

本小节对后续章节中所涉及的概念定义进行统一阐述。依次给出行为特征向量、参照特征向量、特征距离向量以及特征距离向量集等概念的定义。

定义**1**: 行为特征向量表示完成一次手势密码行为所包含特征的向量。其中m表示特征有m维，即

定义**2**：参照特征向量表示最能代表每个用户合法行为的特征向量，用来作为计算特征距离向量的基准。表示为

定义**3**：特征距离向量为每个用户参照特征向量与其余行为特征向量通过距离算法计算对应特征获得的向量。表示为

定义**4**：特征距离向量集由一系列的特征距离向量与特征距离向量标签对组成，可作为训练集和测试集。其中di表示第i个特征距离向量，yi标识该特征距离向量是来自合法用户还是异常用户，n表示该集合包括n个特征距离向量对。特征距离向量集表示为

## 4.2 特征距离向量集构建

### 4.2.1 参照特征向量集生成

由于连续型特征是随时间变化的高维度行为特征，不能将行为特征向量直接应用到各类分类器中。因此，需要先使用距离计算算法将行为特征向量转换为行为特征距离向量。而在将所有行为特征向量转换为行为特征距离向量前，需要先求能一定程度上代表该用户的参照行为特征向量。

在获取参照行为特征向量的过程中，同样应用到了距离计算算法。

对于行为特征向量中的连续型特征，使用DTW（Dynamic Time Warping）算法来计算。这是由于来自两个样本数据的连续型特征，包含的数据序列维度很可能是不一样的，传统的距离计算算法只能计算维度相同的向量间的距离，而DTW是专门用来计算两个不同维度的数据序列之间的距离的算法。

对于行为特征向量中的离散型特征，则使用曼哈顿距离算法来计算。这是由于对于不同样本数据，离散型特征维度相同。可以直接用传统的距离算法进行计算，并且曼哈顿距离算法在键盘鼠标行为分析的前人研究工作中已经被证明能获得更好的实验效果。

算法1给出了参照行为特征向量的生成过程。对于每个用户取前五次的行为特征向量，通过distance\_calulate方法依次计算每个行为特征向量到其余行为特征向量距离和。用min\_distance方法获取所求距离和最小的向量，该向量即为擦抹自己参照行为特征向量。

|  |
| --- |
| 算法1：行为特征向量参照集生成过程 |
| **Input:**  datacell：计算参照行为特征向量的用户行为特征向量数组  **Output:**  ref\_set：参照行为特征向量  1.**for** i=1 to 5||j=1 to 5 **do**  2. ref\_distance[i]=ref\_distance[i]+ sum(distance  3. \_calculate(datacell[i],datacell[j]));  4.**end**  5. min\_distance=min(ref\_distance);  6. k=find(ref\_distance==min\_distance);  7. ref\_set=datacell[k]; |

### 4.2.2 特征距离向量集生成

每一个用户的数据，都需要生成一个参照行为特征向量。在生成参照行为特征向量以后，就可以通过计算对应用户新的行为特征向量和参照行为特征向量间的距离来获取该用户的特征距离向量。

对于每个用户都需要计算行为特征距离向量训练集和测试集。本文中，取10次用户的正常数据和10次用户的异常数据，作为该用户行为模型训练的训练集合。将数据集中的其余数据作为测试集。当然，训练集和测试集都需要由行为特征向量转换为行为特征距离向量。

算法2给出了获取行为特征距离向量的训练集和测试集的方法。先计算行为特征向量距离集的长度，然后对于其中的每个行为特征向量，通过distance\_calculate算法计算其中每个特征与参照行为特征向量对应特征的距离，从而获得特征距离向量集。

|  |
| --- |
| 算法2：获取行为特征距离向量集合 |
| **Input:**  feature\_datacell：行为特征向量集  ref\_set：参照行为特征向量  **Output:**  disvec\_datacell：特征距离向量集   1. len=length(feature\_datacell); 2. **for** i=1 to len **do** 3. discev\_datacell[i] =distance\_calculate(ref\_set, feature\_datacell [i]); 4. **end** |

## 4.3 分类器预测和训练

在本文中，选择SVM算法作为行为模式构建的分类器算法。选择该分类器的原因如下：（1）SVM是经典的二分类分类器，而用户行为认证问题，可将用户数据分为正常用户数据和异常用户数据，是典型的二分类问题；（2）SVM算法要求的样本数相对较少，适合数据量不大的实验数据集。

SVM是通过寻找一个超平面从而将不同的数据集分成两部分的分类器。在训练阶段，SVM对训练集使用RBF 核函数。在测试阶段，会计算测试集和经SVM训练而得的超平面之间的距离作为分类的分数，来预测该测试数据为正常用户数据还是异常用户数据。

在上文中，经处理后已获得被SVM分类器处理的特征距离向量集，而在分类模块中还需要对用户的训练结果和测试数据打上相应的标签，标注相应的行为特征距离向量来源于正常用户还是异常用户。通常正常用户标注为1，异常用户标注为-1。

算法3给出了分类阶段训练和预测的过程。首先通过SVMcgForClass函数针对训练集进行交叉验证调参，获取能获得最优训练效果的c和g。进而通过svmtrain函数对训练集进行训练，获取用户行为模式model。最后通过svmpredict函数对测试集在训练的model中验证效果，accuracy为测试集在模型上的验证准确度。

|  |
| --- |
| 算法3：分类器训练和预测 |
| **Input:**  train\_set：特征距离向量训练集  train\_label：训练集数据对应标签  test\_set：特征距离向量测试集  test\_label：测试集数据对应标签  **Output:**  accuracy：预测效果准确度  1. [bestc,bestg]=SVMcgForClass(train\_set,train\_label);  2. model=svmtrain(train\_set,train\_label,bestc,bestg);  3. accuracy=svmpredict(test\_set,test\_label,model); |

## 4.4 实验分析

### 4.4.1 实验环境和数据

本文实验在mac操作系统下的matlab2012b平台上进行相关模拟实验。实验数据使用的是上文提到的数据集一。

通过每个用户的前五条数据生成特征向量参照集，取30%的合法用户数据和等量的异常用户数据作为训练集。剩余的数据作为测试集。

每一条完成触屏手势密码行为的数据，由以下几项构成：TStamp，Xpos，Ypos，Pressure，Size和Speed。其中TStamp表示完成手势密码时采集器记录下来的时间变化，Xpos和Ypos表示在各个时间点上记录下来的X坐标和Y坐标，Pressure和Size表示在各个时间点上记录下来的手指对触摸屏的压力和手指跟触摸屏的接触面积，Speed表示各个时间点上手指的移动速度。

### 4.4.2 实验结果分析

上文预处理部分介绍了如何去除原始数据中的脏数据以及如何通过归一化来消除不同特征间的量纲影响。算法1描述了在完成特征提取后的行为特征向量集中生成参照集。算法2阐述了如何通过参照集向量和行为特征向量集获取行为特征距离向量集。算法3则给出了SVM分类器对行为特征距离向量集进行训练和预测的过程。该部分则是展示实验结果和分析[320。

根据分析各个用户的手势密码触屏行为数据可以发现，在所有数据中均能采集到X坐标、Y坐标以及由时间和坐标衍生的速度等特征，而压力、接触面积等特征，则在部分数据中由于硬件设备问题存在数据缺失、数据变化不明显等缺陷。因此，首先根据这两类特征分别对用户的手势密码触屏行为数据进行建模认证。如下表所示，对坐标速度类特征，当FRR为18.3%时，FAR为10.8%；对压力接触面积类特征，当FRR为29.45%时，FAR为35.31%。结果表明，即便缺失压力接触面积类特征，仅凭坐标速度类特征，依然能获得较好的认证效果；而存在数据缺失的压力接触面积类特征，同样也能起到不错的用户认证效果。

表4.1 两类移动特征FRR和FAR对比

|  |  |  |
| --- | --- | --- |
|  | FRR | FAR |
| Xpos、Ypos、Speed类 | 18.3% | 10.8% |
| Pressure、Size类 | 29.45% | 35.31% |

下图则直观反应Xpos、Ypos、Press、Size等四个特征对用户行为认证效果的影响。对于Xpos和Ypos随时间变化的曲线，正常用户数据normal1、normal2和异常用户数据attack1存在明显的区别。对于完整的Press数据随时间变化的曲线，同样能区分正常用户normal1、normal2与异常用户attack1。但变化趋势相较于Xpos和Ypos则略显不明显，因此认证效果比之前两者也稍差。对于只能采集到固定值而不能采集到趋势的Size特征，则正常用户normal1、normal2与异常用户attack1则没有任何区别，无法作为用户行为认证的依据。因此，倘若Press和Size存在数据缺失或数据变化不明显等问题，则基于这两类特征的认证效果会明显下降。

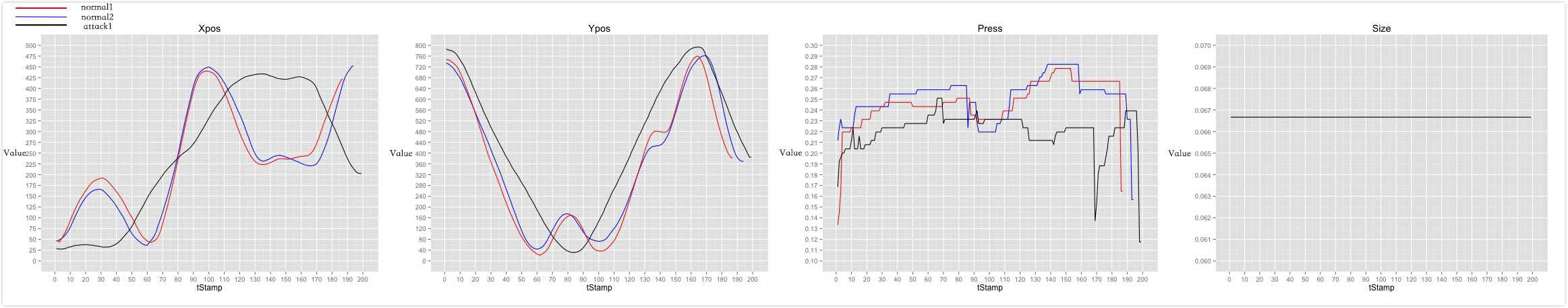


图4.2 正常用户与异常用户Xpos、Ypos、Press、Size等特征表现出的区别

将上诉两类特征结合，进一步的实验结果显示：经过特征提取后的动态特征和静态特征中存在部分特征对认证效果不起作用或者起反作用。经多次实验和调试，在动态特征为X坐标、Y坐标、压力、接触面积、速度随时间变化，静态特征为总时间、第一点的压力、第一点的接触面积、速度的方差时，认证效果为最优。

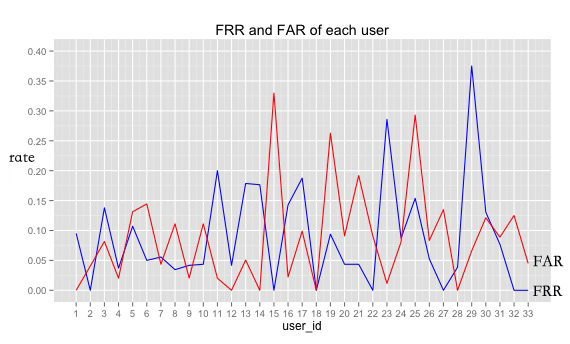
如下图所示，为每个用户在上述模型构建后的认证效果。90%以上用户的FRR和FAR值都在小于20%，70%左右用户的FRR和FAR值在小于10%。而对于认证效果极佳的2，4，5，7，9，12，18，28，33等用户，均能获得小于5%的FRR和FAR。 

图4.3 各用户FRR和FRR对照图

如下表所示，本文实验结果显示，在FRR为8.76%时，FAR为8.85%。相比参考文献[19]在相同数据集的实验中获得19%的FRR和21%的FAR的实验结果，在认证效果上得到了较大幅度的提升。

表4.2 本文实验结果与数据来源文章实验结果对比

|  |  |  |
| --- | --- | --- |
|  | FRR | FAR |
| 本文实验结果 | 8.76% | 8.85% |
| 参考文献19实验结果 | 19% | 21% |

# 第5章 多用户单账号场景下的手势密码触屏行为模式的构建

在本章中主要讲述了多用户单账户场景下的手势密码触屏行为模式的构建。首先，对多用户单账号场景进行刻画，描述该场景下的特征以及单用户账号的手势触屏行为模式针对该场景的局限性。接着，介绍了多用户单账号手势密码触屏行为模式的构建方法，包括方法的总体架构、基于多用户混合数据的聚类方法的应用、以及相应的实验分析对比。

## 5.1 多用户单账号场景分析

前期工作中，已完成单用户单账户的手势密码触屏行为模式构建工作。通过数据预处理、特征距离向量生成以及SVM分类器分类等模块完成手势密码触屏行为模式的构建。进一步的实现分析结果页表明单用户单账户的手势密码触屏行为认证有较低的误检率和漏检率，分别为8.85%和8.76%。

当然，单用户单账户较为理想的手势密码触屏行为模式构建的场景。在现实的应用中，存在情侣、家庭成员共用一个手势密码的情形，即多用户单账号场景。对于来自同一个“家庭”的用户来完成同一手势密码，可通过认证；对于不属于这个“家庭”的用户来完成这一手势，则不能通过认证，被认定为非法用户。

对于多用户单账号的手势密码触屏行为模式场景，若不区分数据源，将其视为一个用户进行模式构建，将会严重影响模型构建的准确性，从而影响认证效果。假设有A、B两个用户共用一个账号，如果建模的数据来自于A，那么该模型自然会拒绝所有来自B的数据；如果建模的数据来自于A和B，那么所得的模型对A和B的认证效果都不会太好。毋庸置疑，直接把单用户单账户的模式构建方法应用在多用户单账号的场景下，认证效果必然极差。

所以对于多用户单账号场景，模型构建的关键就在于区分数据源，将同一个账号下的不同用户的手势密码触屏行为数据进行区分。

## 5.2 多用户单账号手势密码触屏行为模式的构建

### 5.2.1 总体架构

相比单用户单账号的手势密码触屏行为模式构建，多用户单账号模式构建的关键在于对数据来源的区分，将同一账号下的多个用户的手势密码触屏行为数据进行区分。即增加对同一账号下的数据进行聚类处理的模块，然后对聚类后的每个用户的数据手势密码触屏行为模式的建模，依次进行认证。

如下图所示，即为多用户单账号手势密码触屏行为模式的总体架构。分为预处理模块、多用户聚类模块及手势密码触屏行为模式构建模块。

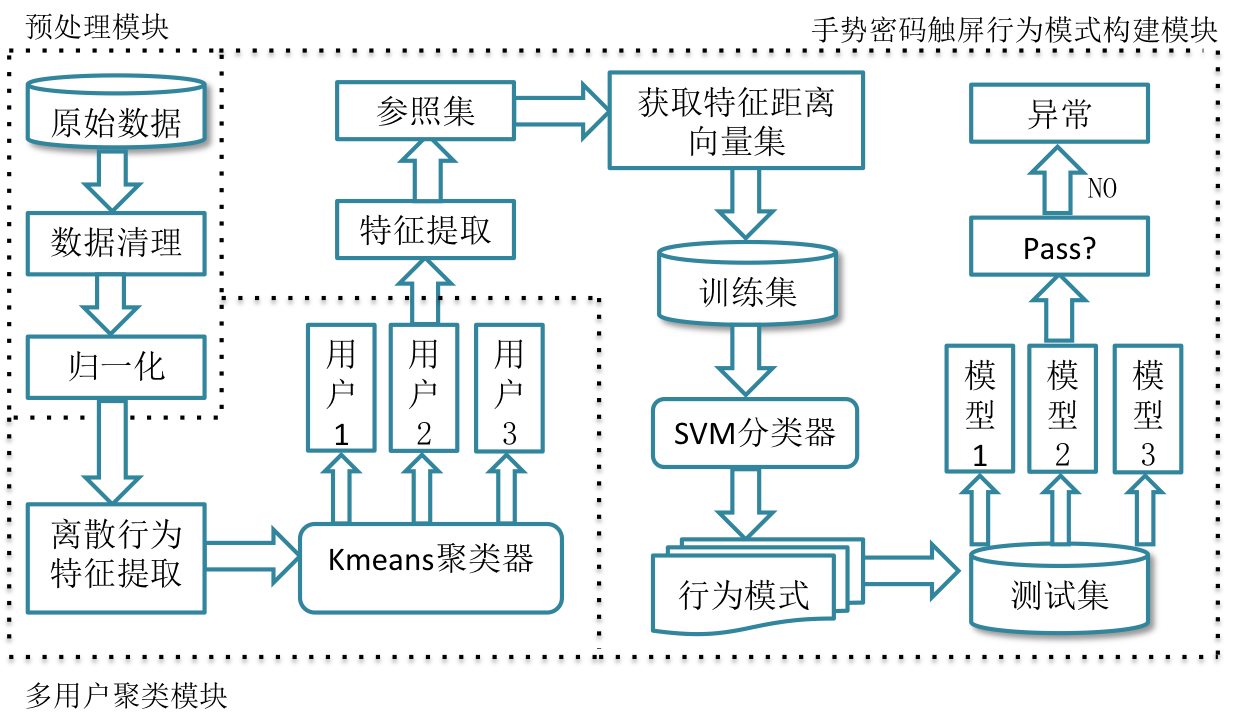


图5.1 多用户单账号手势密码触屏行为模式总体架构

预处理模块同3.2节预处理所述一致，分为数据清理、数据转换以及归一化。该模块将手势密码触屏行为数据采集器采集而来的原始数据，处理成能被应用于后续数据分析处理的格式。既提高了聚类分类模型的准确性，也提高了相应分类聚类的效率。

多用户聚类模块的目的在于将同一个账号多个用户混杂在一起的数据，分清数据来源。在这一模块，先提取手势密码触屏行为的离散型行为特征，再运用K-means聚类器基于各次手势密码触屏行为的离散行为特征进行聚类处理，将数据聚成若干个簇，每个簇对应一个用户的手势密码触屏行为数据。

手势密码触屏行为模式构建模块类似第四章所阐述的手势密码触屏行为模式的构建。对聚类后的每个用户的数据，提取相应的连续型行为特征和离散型行为特征。再基于各个用户的行为特征向量，计算获取参照集特征向量。通过距离计算算法计算各个行为特征向量与参照集特征向量的距离，获取特征距离向量集。最后通过SVM分类器进行分类训练和认证。

对于多用户单账号的手势密码触屏行为模式，如果新来的手势密码触屏行为数据，匹配了该“家庭”内任一用户的手势密码触屏行为数据，即可认为该次手势密码行为属于该“家庭”的成员；如果全部不匹配，则认为该次手势密码行为数据属于异常行为。

### 5.2.2 多用户聚类

本文中选用的聚类算法是k-means算法[37]。k-means算法是基于欧式距离，算法过程中将对象划分到欧式距离最近的形心代表的簇中，因此k-mean数据类算法的最终结果为彼此不交叉的球状簇。对于存在多用户单账号的手势密码触屏行为数据，每个球状簇代表一个用户的手势密码行为数据。簇内同一用户数据的欧式距离接近，而距离其它簇的欧式距离较远。下图所示为二维空间中两个用户手势密码触屏行为数据的示意图[36]。



图5.2 二维空间中两个用户的手势密码触屏行为数据示意图

多用户单账号场景下，手势密码触屏行为数据采集器采集而来的是由多个用户数据混杂的数据。在计算连续型特征间的距离时，首先通过算法找到能相对最好代表该用户的参照集特征向量。由于多个用户数据混杂，因此无法找到能代表每一个用户的参照集特征向量。

因此，在对多个用户聚类划分成多个用户数据时，构建的行为特征向量的特征选择离散型特征，如总时间，采集点的个数，X坐标、Y坐标、压力、接触面积、接触面积的长轴和短轴等特征的均值、标准差、最大值、最小值。相比连续型特征，离散型特征提供的信息量上有所减少，但是鉴于大多数行为认证数据取的都是离散型特征。显然，其仍具有不错的认证效果。

对于聚类的效果可以用purity方法评估。Purity方法是极为简单的一种聚类评价方法，只需计算正确聚类的手势密码触屏行为数据数占总手势密码触屏行为数据数的比例：。其中

是聚类的集合。表示第k个聚类的集合。表示的是手势密码触屏行为特征数据的集合。 表示第j次手势密码触屏行为特征数据。N表示总的手势密码触屏行为特征数据的数目。

### 5.2.3 手势模型构建

对于多用户单账号场景下的手势密码触屏行为模式构建，通过上一小节描述的聚类算法。可将多个用户混杂在一起的数据分清数据来源，聚类成若干个“家庭”用户。对于每个家庭用户均需要构建一个该用户对应的手势密码触屏行为模式。

此时的数据已经聚类成不同的用户，被打上该用户的标签。因此在手势密码触屏行为模式构建时，可以应用连续型手势密码触屏行为特征数据。本文选取的连续型手势密码触屏行为特征包括随时间变化的X坐标、Y坐标、压力、接触面积、速度、接触面积的长轴和接触面积的短轴等。当然行为特征向量的组成中也包括离散型的手势密码触屏行为特征，如总时间、采集点的个数、各项连续型行为特征的均值、标准差、最大值，最小值。此外，还包括一些特殊特征，比如第一点的压力、接触面积，最后一点的压力、接触面积等。

后续的模式构建方案类似于单用户单账号的手势密码触屏行为模式的构建。根据行为特征向量，计算获取参照集行为特征向量，然后根据距离计算算法形成行为特征距离向量集。通过SVM分类器对行为特征距离向量集进行分类训练。当然，在多用户单账号场景下，对新的用户行为数据进行测试时，新数据匹配任一用户的手势密码触屏行为模式，即可认为该数据为合法用户的操作；若对所有用户的手势密码触屏行为模式，均不匹配，则认为该数据为非法用户的操作。

## 5.3 实验分析

本文实验在mac操作系统下的matlab2012b平台上进行相应的模型构建和认证。实验数据使用的是上文提到的数据集二。该数据集总共采集了三种类型的手势密码，每个手势密码由5个人来完成，总共15人，每人完成120次手势密码输入操作，总计1800次手势密码输入操作。

聚类的数据为混合用户的所有数据，根据模拟情侣和家庭环境两种情景，分成两人的数据混合以及三人的数据混合。

该模型聚类效果的评估使用Purity指标，模型认证效果的评估使用FRR（漏检率）和FAR（误检率）。

**实验1：对多用户单账号混合数据直接使用单用户单账号模型构建方法处理（仅作了两个用户混合数据结果分析，三个用户混合效果只会更差）**

该实验中对两个用户各选5条数据总计10条数据，训练参照集行为特征向量；训练集为取两个用户各10条数据做正样本，取20条同手势的其他用户数据作为负样本；测试集为取两个用户各40条数据做正样本，取80条同手势的其他用户数据作为负样本。

下表为多用户单账号混合数据直接使用单用户单账号模式构建方法实验效果与单用户单账号账号使用该模式构建方法实验效果的对比。

表5.1 两种方案实验效果对比

|  |  |  |
| --- | --- | --- |
|  | FRR | FAR |
| 本实验 | 58.75% | 21.25% |
| 单用户单账号实验 | 9.65% | 8.76% |

实验结果表明直接将多用户单账号混合数据通过单用户单账户模型进行认证，效果极差。理论上，随机判断某个用户是或者不是合法用户的概率为50%，而本实验的FRR也达到58.75%，比随机判断的效果都要差。因此，改进原有模型算法以适应多用户单账号场景是有必要的。

**实验2：对多用户单账号混合数据使用聚类算法，将数据来源分类，对每条数据打上类型标签**

本实验使用的聚类算法为k-means算法。选取的离散型特征如下：总时间、采集点的个数、连续型特征（X坐标、Y坐标、压力、接触面积、接触面积长轴、接触面积短轴）的均值、标准、最大值、最小值。

实验分别了两个用户数据和三个用户数据的混合，两个用户混合的单个实验数据为240条（2\*120条），三个用户混合的单个实验数据为360（3\*120条）。任一种情况有30种分类，所有总共有60种情形的单个实验。

如下表所示为手势密码1在未归一化和归一化时进行聚类效果的比较，实验表明在做聚类实验时，将不同量纲的特征数据统一为同一量纲的数值是能较大幅度提升聚类效果的。因此，在数据预处理时，需要应用相应的归一化算法统一不同量纲的特征数据。

表5.2 手势密码1未归一化和归一化条件下，聚类效果比较

|  |  |  |
| --- | --- | --- |
|  | Purity | Purity |
| 未归一化 | 87.92% | 74.58% |
| 归一化 | 98.96% | 98.40% |

如下表所示为手势密码1，手势密码2，手势密码3在两个用户和三个用户的数据混合状况下的聚类效果，以及三个手势总体的聚类效果。从实验结果，我们可以发现在两个用户的数据混合时，聚类的总体效果可以达到98.17%；在三个用户的数据混合时，聚类的总体效果可以达到97.41%。该结果表明，通过k-means算法对多用户混合的手势密码触屏行为数据有非常好的聚类效果，该算法能实现多个用户数据混合在一起时，对数据来源加上标签的功能。

表5.3 三个手势密码在两个用户混合和三个用户混合时的聚类效果

|  |  |  |
| --- | --- | --- |
| 手势密码 | Purity（双用户） | Purity（三用户） |
| Password1 | 98．96% | 98．40% |
| Passowrd2 | 98.33% | 91.72% |
| Password3 | 97.21% | 96.92% |
| 总体 | 98.17% | 97.41% |

**实验3：对多用户单账号混合数据先使用聚类算法，再使用分类算法**

本实验使用的聚类算法为k-means算法，使用的分类算法为SVM。

实验分别了两个用户和三个用户的混合，对于聚类实验，两个用户混合的单个实验数据为240条（2\*120条），三个用户混合的单个实验数据为360（3\*120条）。任一种情况有30种分类，所有总共有60种情形的单个实验。

聚类以后，对混合用户数据打上标签。对其中的任一用户，选择10条该用户数据，训练参照集行为特征向量；选取20条该用户数据为正样本，20条同手势的非该家庭用户数据作为负样本，作为实验的训练集；选取该用户的其余数据为正样本，取等量同手势的非该家庭用户数据作为负样本，作为实验的测试集。

如下表所示，改进后的手势密码触屏行为模型认证效果明显提升。整体效果而言，FRR达到了13.86%，FAR达到了8.08%。对于单个手势，最优的FRR达到了10.51%，最优的FAR达到了7.51%。相比相同的多用户单账号数据使用上文的单用户单账号手势密码触屏行为模型的实验结果，即58.75%的FRR和21.25%的FAR，改进后的实验效果显著提升。

因此，从本节中描述的一系列实验结果可以表明，改进后的手势密码触屏行为模型，即通过先聚类再分类的认证模式，能够解决多用户单账号场景的实际问题。

表5.4 各个认证方案效果对比

|  |  |  |
| --- | --- | --- |
|  | FRR | FAR |
| Password1 | 13.54% | 7.51% |
| Password2 | 17.53% | 7.54% |
| Password3 | 10.51% | 9.01% |
| 多用户单账号（聚类） | 13.86% | 8.08% |
| 多用户单账号（未聚类） | 58.75% | 21.25% |
| 单用户单账号 | 9.65% | 8.76% |

# 第6章 总结与展望

## 6.1 总结

随着移动计算能力以及智能手机占有量的指数级增长，触屏智能手机已经成为每个人日常生活不可替代的重要组成部分，如何正确识别移动终端用户身份的合法性是保障用户信息和财产安全的关键。除了传统的用户名密码认证、辅助设备认证、生物特征认证以外，利用用户习惯来认证用户身份的研究和应用正在成为相关领域的热点。

本文针对用户行为习惯中的手势密码触屏行为习惯进行深入的研究和探讨，首先介绍了相关领域的国内外研究现状，并详细介绍了用户手势密码触屏行为认证的主要技术。第三章主要描述了手势密码触屏行为数据采集器的设计实现，以及对采集而来的原始数据进行数据预处理、特征提取等方面的工作。在第四章，提出手势密码触屏行为模式的构建方法，包括参照集行为特征向量的构建，行为特征距离向量的计算以及通过SVM分类器进行训练建模，以及相应的实验结果与分析。第五章中提出针对现实场景中可能存在的多用户共用一个账号情况的问题，通过使用k-means算法聚类对混合数据进行标签，给出多用户手势密码触屏行为模式构建的改进方案，并给出相应的实验结果与分析。

本文主要做的工作总结如下：

1. 手势密码触屏行为数据采集器的设计与实现。研究传统用户行为的特征选取，结合调研的Android移动终端上的传感器及相应可采集的数据，选取要采集的特征数据。并学习相关数据包和类的应用，在Android系统上开发了模拟完成手势密码的APP，在完成手势密码的过程中，记录相应特征数据到json格式的数据包中，为后续的数据分析和模式建模打好基础。
2. 提出了一种手势密码触屏行为模式的构建方案。研究学习产统用户行为建模认证方法，结合Android移动端新增的用户行为特征数据，提出了手势密码触屏行为模式构建方法。相比同一数据集使用其他模式构建方法的实验效果，FAR和FRR都有不小幅度的改良，达到预期的实验效果。
3. 现实场景多用户单账户手势密码触屏行为问题的解决。通过分析多用户单账号手势密码触屏行为数据习惯，研究和分析聚类算法，通过k-means聚类算法，将多用户的手势密码触屏行为数据聚类处理后建模，实现多个用户共用同一账号下的手势密码触屏行为模式构建。实验结果分析得出预期的效果。观测聚类前后的对比认证效果，FRR和FRR相比聚类前，均由显著提升，保持较好的认证水平。

## 6.2 进一步工作

手势密码触屏行为认证在传统的认证基础上，无需特殊硬件支持，在用户登录时无打扰的进行了手势密码触屏行为数据的采集和认证，使得用户账号的安全性能得以提升。当然此认证方法也存在诸多问题，例如效率不够高，算法复杂，认证为非实时认证。本文对手势密码触屏行为认证的研究不够完善，存在许多不足之处，包括以下几点需要进一步探讨：

1. 本文研究的手势密码触屏行为模式仅仅针对于用户登录系统时输入手势密码的触屏行为认证，为一次性认证过程。对于登录系统成功的用户，没有持续性的认证过程。对于登陆后的用户，此时无法保证在更替用户后，仍然对用户有身份认证的功能。如何在用户持续使用的过程中实现手势密码触屏行为认证为后续的研究点。
2. 文中中提出的针对连续型特征向量使用DTW算法计算特征距离向量的方法，算法较为复杂，在实际应用的场景下，会带来认证延迟，影响用户的使用体验。后续研究将针对提高算法效率，简化算法过程，研究如何快速计算连续型特征向量间的距离，并达到较好的认证效果。
3. 在本研究的实验分析中，发现了完成手势密码时的姿势对认证效果同样存在影响。现实情况下，完成手势密码时可能存在站姿、坐姿、躺姿等各种姿势，如果能分清楚完成手势时的姿势，理论上应该能提升手势密码认证效果。

致谢

一转眼，三年研究生生涯就快落下帷幕。细细回忆这三年来的点点滴滴，满满的感动萦绕心头。

回顾我的研究生生涯，最要感谢是我的导师关佶红教授。关老师严谨求实的治学态度、乐观正面的生活态度都时时刻刻影响着我。在生活中，关老师更是给予了无微不至的关心。遇到这样一位好导师，是我的幸运。感谢恩师，今后的工作和生活，会带着你的教诲，孜孜不倦的努力和前行。

同时，特别感谢复旦大学的周水庚教授。周老师对学术严谨、一丝不苟的态度深深的影响着我，为我们树立了一个好标杆。周老师牺牲了自己所有的业余时间，在学术上给予我们全面专业的指导和帮助。感觉周老师！

感谢甘杨兰师姐，在我的研究生期间，甘师姐在学术和生活上都给予了我非常大的帮助。记得刚入学的时候，对研究方向一无所知，是甘师姐不厌其烦的指导我帮助我，对我的学术研究和论文发表都给予了最大的帮助。感谢甘师姐！

感谢实验室师兄、师姐们的指导和同门们的鼎力相助。DMB实验室是个和睦温馨的大家庭，生活上互相帮助，学术上互相切磋。这份同门情谊我永难忘。

感谢挚友卜宏达、朱海泉、王佳晟一路的支持和帮助。能结交几位知己挚友，乃人生一大幸事。

感谢我的父亲母亲，这么多年来支持我，理解我，做我永远坚强的后盾。你们是我前进的动力。希望自己早日事业有成，回报你们多年的养育之恩。

在快要毕业之际，感谢这三年遇到的所有老师和同学。祝所有同学都有一个锦绣前程。

感谢参加论文评审和答辩的各位专家和学者对本文的评阅和指导。

参考文献

[1] 艾瑞咨询.2014年Q1中国第三移动支付市场交易规模报告[R/OL]. <http://www.199it.com/archives/241923.html>

[2] 中国电子银行网. 2012中国手机银行安全性调研报告[R/OL].http://ec.iresearch.cn/shopping/20130731/206796.shtml

[3] 中国信息产业网人民邮电报.三大移动支付安全问题凸显[R/OL]. <http://www.mpaypass.com.cn/news/201404/25115627.html>

[4] 360安全中心.2014年第一期中国移动安全支付报告[R/OL]. <http://policy.catr.cn/zcyj/201405/t20140504_1016393.htm>

[5] Meng Y, Wong D S, Schlegel R. Touch gestures based biometric authentication scheme for touchscreen mobile phones[C]//Information Security and Cryptology. Springer Berlin Heidelberg, 2013: 331-350

[6] Clarke N L, Furnell S M. Authentication of users on mobile telephones–A survey of attitudes and practices[J]. Computers & Security, 2005, 24(7): 519-527.

[7] Jermyn I, Mayer A J, Monrose F, et al. The Design and Analysis of Graphical Passwords[C]//Usenix Security. 1999.

[8] Tari F, Ozok A, Holden S H. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords[C]//Proceedings of the second symposium on Usable privacy and security. ACM, 2006: 56-66.

[9] Riha Z. Toward reliable user authentication through biometrics[J]. IEEE Security & Privacy, 2003, 1(3): 45-49.

[10] Francis L, Mayes K, Hancke G, et al. A location based security framework for authenticating mobile phones[C]//Proceedings of the 2nd International Workshop on Middleware for Pervasive Mobile and Embedded Computing. ACM, 2010: 5.

[11] Monrose F, Rubin A. Authentication via keystroke dynamics[C]//Proceedings of the 4th ACM conference on Computer and communications security. ACM, 1997: 48-56.

[12] Pusara M, Brodley C E. User re-authentication via mouse movements[C]//Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security. ACM, 2004: 1-8.

[13] Mantyjarvi J, Lindholm M, Vildjiounaite E, et al. Identifying users of portable devices from gait pattern with accelerometers[C]//Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on. IEEE, 2005, 2: ii/973-ii/976 Vol. 2.

[14] Saevanee H, Bhatarakosol P. User authentication using combination of behavioral biometrics over the touchpad acting like touch screen of mobile device[C]//Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on. IEEE, 2008: 82-86.

[15] Cai Z, Shen C, Wang M, et al. Mobile authentication through touch-behavior features[M]//Biometric Recognition. Springer International Publishing, 2013: 386-393.

[16] Shen C, Cai Z, Guan X, et al. User authentication through mouse dynamics[J]. Information Forensics and Security, IEEE Transactions on, 2013, 8(1): 16-30.

[17] Jakobsson M, Shi E, Golle P, et al. Implicit authentication for mobile devices[C]//Proceedings of the 4th USENIX conference on Hot topics in security. USENIX Association, 2009: 9-9.

[18] Mayrhofer R, Gellersen H. Shake well before use: Intuitive and secure pairing of mobile devices[J]. Mobile Computing, IEEE Transactions on, 2009, 8(6): 792-806.

[19] De Luca A, Hang A, Brudy F, et al. Touch me once and i know it's you!: implicit authentication based on touch screen patterns[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012: 987-996.

[20] Frank M, Biedert R, Ma E, et al. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication[J]. Information Forensics and Security, IEEE Transactions on, 2013, 8(1): 136-148.

[21] Li L, Zhao X, Xue G. Unobservable Re-authentication for Smartphones[C]//NDSS. 2013

[22] Coventry, L., De Angeli, A., Johnson, G. Usability and biometric verification at the ATM interface. In Proceedings CHI 2003. ACM Press (2003), 153-160

[23]Wood, H.M. The use of passwords for controlled access to remote computer systems and services. In Proceedings AFIPS 1977. ACM Press (1977), 27-33

[24] Sonkamble, S., Thool, R., Sonkamble, B. Survey of bio- metric recognition systems and their applications. Journal of Theoretical and Applied Information Technology, 11(1). (2010), 45-51

[25] Rokita, J. Krzyzak, A., Suen, C.Y. Cell phones personal authentication systems using multimodal biometrics. In Proceedings ICIAR 2008. Springer (2008), 1013-1022

[26] Marcel, S., Cool, C., Atanasoaei, C., Tarsetti, F., Pesán, J., Matejka, P., Cernocky, J., Helistekangas, M., Turtinen, M. MOBIO: mobile biometric face and speaker authentication, In Proceedings CVPR 2010

[27] Weiss, R., De Luca, A. PassShapes: utilizing stroke based authentication to increase password memorability. In Proceedings NordiCHI 2008. ACM Press (2008), 383- 392

[28] C. Castelluccia and P. Mutaf. Shake them up!: a movement-based pairing protocol for cpu-constrained devices. In MobiSys, pages 51–64, 2005

[29] F. Okumura, A. Kubota, Y. Hatori, K. Matsuo, M. Hashimoto, and A. Koike. A study on biometric authentication based on arm sweep action with accel- eration sensor. In Proceedings of ISPACS’06, pages 219–222. IEEE, 2006

[30] M. Pusara and C. Brodley. User re-authentication via mouse movements. In Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security, pages 1–8. ACM, 2004

[31] Sae-Bae, N., Ahmed, K., Isbister, K., Memon, N.: Biometric-Rich Gestures: a Novel Ap- proach to Authentication on Multi-Touch Devices. In: Proc. ACM Annu. Conf. Human Factors in Computing Systems, pp. 977–986 (2012)

[32] Yuhao Wu, Zhijun Ding. A Mobile Authentication Method Based on Touch Screen Behavior for Password Pattern⋆. Journal of Computational Information Systems 11:22(2015) 1-12

[33]Tran T N, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometrics and Intelligent Laboratory Systems, 2013, 120: 92-96

[34]Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 2013, 40(1): 200-210.

[35] Na S, Xumin L, Yong G. Research on k-means clustering algorithm: An improved k-means clustering algorithm//Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on. IEEE, 2010, pp. 63-67

[36]Xiaomeng Zhang, Peihai Zhao, Mimi Wang. Keystroke Dynamics in Password Authentication for Multi-user Account⋆. Journal of Computational Information Systems, 2015, 11(1): 321-331

[37]家炜. 数据挖掘: 概念与技术. 机械工业出版社. 2012. p293-308

[38]Jiang C H, Shieh S, Liu J C. Keystroke statistical learning model for web authentication//Proceedings of the 2nd ACM symposium on Information, computer and communications security. ACM, 2007, pp. 359-361

**个人简历、在读期间发表的学术论文与研究成果**

**个人简历：**

卢赟，女，1990年10月生。

2013年6月毕业于浙江工业大学计算机科学与技术学院，计算机科学与技术专业，获工学学士学位。

2013年9月入同济大学就读硕士研究生。

**已发表论文：**

1. Yun Lu, YanglanGan, Jihong Guan\*, and Shuigeng Zhou. An integrative analysis of nucleosome occupancy and positioning using diverse sequence dependent properties. Accepted by Neurocomputing, 2015.

**参与科研项目：**

1. 国家自然基金项目“基于生物信息计算的核小体定位动态机制研究”，2013.09-2016.03.