Y3246722

密级：_____

浙江大学

硕士学位论文

论文题目 基于层次聚类特征选择和 HF-SVM 的活动识别技术

作者姓名 付浩

指导教师 邢卫 副教授

学科(专业) 计算机科学与技术

所在学院 计算机科学与技术学院

提交日期 2017 年 3 月 8 日

万方数据

A Dissertation Submitted to Zhejiang University for the Degree of Master of Engineering

TITLE: Activity Recognition Technology Based on HF-SVM and Feature Selection using Hierarchical Clustering

Author: Hao Fu

Supervisor: Associate Professor Wei Xing

Subject: computer science and technology

College: computer science and technology

Submitted Date: March 8th, 2017

万方数据

Y3246722

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得浙江大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：付浩　　签字日期：2017年3月7日

学位论文版权使用授权书

本学位论文作者完全了解浙江大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：付浩　　导师签名：[illegible]

签字日期：2017年3月7日　　签字日期：2017年3月7日

学位论文作者毕业后去向：

工作单位：　　电话：

通讯地址：　　邮编：

万方数据

摘要

基于人体活动的智能计算是人工智能领域的一个重要研究方向，其目的是通过获取用户的状态和环境数据信息，为用户提供智能化应用服务。随着移动智能设备（如智能可穿戴设备）及其相关传感器等技术的飞速发展，基于移动智能设备的活动识别技术成为了研究的热点。由于移动智能终端在计算能力、存储空间和能量等硬件资源方面的限制，而传统机器学习模型需要巨大运算能力，基于移动智能设备的活动识别技术无法广泛应用。

针对以上问题，本文在分析传统特征选择和支持向量机的基础上，提出了一种基于层次聚类的特征选择和硬件友好型核函数的支持向量机的多类别分类方法。论文的主要算法改进和成果如下：

（1） 提出了改进的基于层次聚类算法的特征选择算法。基于层次聚类的特征选择算法使用的评价函数基于互信息和关联系数。这不能适用于活动识别领域的连续型数据。本文基于皮尔逊相关系数和共享最近邻这两种度量改进了评价函数。使用改进的基于层次聚类的特征选择算法，完成了特征提取，降低活动识别技术在模型训练过程中的复杂度。

（2） 提出了基于硬件友好型核函数的 SVM 算法。传统的 SVM 算法在模型训练和算法应用中需要大量的指数运算。本文基于高斯核函数和拉普拉斯核函数提出了硬件友好型的核函数，既保持高斯核函数抗噪声的优点，也具有较小的计算代价。

（3） 提出了基于组合分类器的活动识别技术。活动识别是多类别分类问题。传统的多类别分类算法存在分类误差大的问题。本文提出了基于 OVO 策略的 SV 算法，将每个分类器的输出结果作用于 Sigmoid 函数，然后根据每个二值分类器的投票得到最终输出。本文提出的方法避免单独使用 Sigmoid 函数取最大值时易受到噪声的干扰而预测出错，也避免了多个分类器直接投票由于分类器权重相同导致的错误。

万方数据

本文从以上三个方面对活动识别技术进行研究，展开的实验分析显示文中提出的活动识别技术具有较高准确率，具有一定的实用价值。

关键词： 用户活动识别，HF-SVM，层次聚类特征选择，核函数

万方数据

Abstract

Intelligent computing based on human activities is an important research direction in the field of artificial intelligence. Its purpose is to provide users with intelligent application services by acquiring user's state and environment data information. With the rapid development of mobile intelligent devices (such as intelligent wearable devices) and related sensors, activity recognition technology based on the intelligent equipment has become the research hotspot. Due to the limitation of hardware resources, such as computing power, storage space and energy, and the traditional machine learning model needs huge computing power, mobile intelligent device-based activity recognition technology can not be widely used.

In order to solve the above problems, in this paper, the traditional feature selection and support vector machine techniques are studied, and then proposes a multi-category classification method based on hierarchical clustering feature selection and hardware-friendly kernel function. The main algorithm improvements and achievements are as follows:

(1) An improved feature selection algorithm based on hierarchical clustering algorithm is proposed. The traditional hierarchical clustering-based feature selection algorithm uses an evaluation function based on mutual information and correlation coefficients. This does not apply to continuous data in the field of motion recognition. This paper improves the evaluation function based on Pearson correlation coefficient and shared Nearest Neighbor. Using the improved feature selection algorithm based on hierarchical clustering, the feature extraction is completed and the complexity of the activity recognition technology in the process of model training is reduced.

(2) The SVM algorithm based on hardware-friendly kernel function is proposed. Traditional SVM algorithm needs a lot of exponential operations in model training and algorithm application. Based on the Gauss kernel function and the Laplacian kernel function, this paper proposes a hardware-friendly kernel

万方数据

function, which not only has the advantage of preserving the anti-noise of Gaussian kernel function, but also has a small computational cost.

(3) Propose the activity recognition technology based on combinatorial classifier. Activity identification is a multi-class classification problem. The traditional multi-class classification algorithm has the problem of large classification error. In this paper, the SV algorithm based on OVO strategy is proposed. The output of multiple classifiers is applied to the Sigmoid function, and the final output is obtained according to the vote of each binary classifier. The method proposed in this paper avoids using Sigmoid function to get the maximum value, and it is easy to be affected by the noise. It also avoids the errors caused by the same classifier weight.

In this paper, the activity recognition technology is studied from the above three aspects, and the experimental analysis shows that the proposed activity recognition technology has high accuracy and practical value.

Keywords: Human activity recognition, Hardware-Friendly Support Vector, Feature Selection Based on Hierarchical Clustering Machine, Kernel

万方数据

目录

万方数据

万方数据

万方数据

图目录

万方数据

表目录

万方数据

第1章 绪论

1.1 研究背景及意义

随着各类穿戴智能产品的大量普及和移动互联网技术的深度发展，传统健康养老产业迈入了智能时代，衍生出大量的互联网创新产品与运营模式[1]。可穿戴智能产品带来的是一种新的健康生活方式，通过不断地量化分析从用户的生活中获取到的数据，帮助人们更好地进行保持健康生活状态[29]。

自从20世纪90年代末期，研究人员对获取和识别人体活动过程中的动作和行为展开了深入的研究。如果从识别动作的目的和持续时间两方面进行划分，可将相关的研究分为两大类：以主动交互为目的的手势类活动识别和用户被动的日常活动识别。前者研究的问题是识别短暂和规范的交互动作，而后者针对的是长期且多样的日常行为[1]。本文研究的是用户日常行为识别。

目前，用户行为识别领域出现了大量的基于各种方式实现的识别技术。其中较为普遍的方法是基于计算机视觉和图像视频处理的技术，依据照相机或者摄像机采集的图像资料，从静止的图像或视频中提取并识别用户的活动行为[2]。如文献[30]基于图像视觉领域的深度学习技术研究了人的交互式别问题。基于图像的技术虽已取得了较为丰富的研究成果，但是并没有在日常生活中得到较为广泛的应用，其根本原因在于基于图像的识别方法存在如下问题：(1)，该类识别方法依赖于外部的图像采集设备，适用范围被限制在已经部署了相关设备的环境中，而部署图像采集设备需要花费很大的代价，推广起来并不容易。(2)，要求用户处在能够被图像采集设备所观察到的区域中，如果图像采集设备是固定的地点部署的，则设备观察到的区域将非常受限，即使对于可移动的采集设备也会受到联网能力和部署条件的限制。(3)，图像能够传达或者蕴含的信息非常丰富，可能存在着除了用户活动之外其他隐私活动，这些活动信息有被泄漏的风险。(4)，图像采集设备容易受到外在环境条件的影响。(5)，图像数据识别和处理，尤其是视频流中的图像数据识别和处理，需要很高的网络带宽，强大的计算和存储能力，因此很难

万方数据

在现有的设备中做到实时的数据处理和反馈。

过去的十年，移动智能装备尤其是智能手机、智能手环等得到飞速的发展和迅速的普及推广。目前，智能手机的销量是个人电脑的4倍，有大约一半的成年人拥有一部智能手机，这一比例将在2020年达到80%[3]。在迅速普及的同时，智能手机的处理能力也得到迅速的提升。比如高通骁龙新款821处理器，它是一款4核64位处理器，单核速度可达2.4GHz，下载速度最高可以达到600Mbps[4]。目前智能手机搭载了数十种传感器，如陀螺仪、加速度计、气压传感器、重力感应器等。智能手机强大的计算存储能力、联网能力及其内置的丰富的传感器，为利用智能手机及其内置传感器进行用户行的识别奠定了基础[9]。

基于移动智能装备的活动识别技术通过智能装备（如智能手机）内置的传感器采集与用户活动相关的数据，用于对活动进行识别，因此避免了对外部设备的依赖，使得技术适用场景范围扩大；另外，由于这些传感器（如加速度计、心率传感器等）提供的数据包含的敏感信息较少且不经专业解析不具备直接的可读性，因此减少了用户隐私被泄漏的风险，而种类丰富的传感器类型也为识别用户多样的活动提供了较为充分的信息。相关的研究中使用较多的传感器类型是（加速度计和陀螺仪等）动作传感器。动作传感器可以测量智能设备跟随人体所做的物理运动的幅度和方向，因此非常适合用于推断和预测用户的行为活动和物理状态。

1.2 研究现状

利用传感器检测到的活动相关数据信号，可以使用不同的技术来识别移动对象的活动行为。近几年，研究人员提出利用隐马尔科夫（Hidden Markov Models，HMM）和动态贝叶斯网（Dynamic Bayesian Networks，DBN）等概率学习模型来进行行为的识别与分类[6]。Lafferty 等人提出了条件随机域（Conditional Random Fields，CRF）的概率模型，该算法模型可以用于序列分析和词性标注，可以用于获取到的传感器数据序列识别和分类活动[6] 。孙泽浩等使用半马尔科夫模型对活动持续的时间序列信号进行统计分析来识别用户活动识别[1]。李文洋等利用基于谱聚类分析和隐马尔可夫模型（Spectral clustering and Hidden Markov Models，SC-

万方数据

HMM）智能手机采集到的信号进行分析从而实现对日常行为进行识别[6]。

姚毓凯等在多级网格搜索（MGS）算法解决支持向量机相关参数寻优问题，并提出对少数类样本数据实施基于 SMOTE 算法的过抽样插值的算法来求解不均衡类问题[2]。具有整形参数的 SVM 模型特别适合传感器网络等硬件资源受限的应用场景，文献[7]提出并讨论分析了具有整数参数的 SVM 模型，并讨论了整数参数支持向量机的参数寻优问题。文中用求解混合整型参数的二次规划（MIPQ）问题时用到的混合整形分支定界算法去求解整型参数的 SVM 模型，此外文中讨论了在使用分支定界的方法的时使用单变量的 SMO 算法来快速的求解上下界。文献[8]探讨了使用定点小数为参数的 SVM 模型。文中讨论了硬件友好型的核函数。在训练模型时，文中基于调和旋转数字计算机算法（Coordinate Rotation Digital Computer，CORDIC）提出了一种改进的迭代的方法，最后论文讨论了迭代算法的收敛速度和参数寻优过程中的上下界。文献[9]提出了基于一种硬件友好型支持向量机（hardware-friendly SVM，HF-SVM）模型的多类别分类的活动识别模型，对步行，上楼梯，下楼梯，站起，坐下，躺下这 6 种行为进行识别，并给出了一种类似于调和旋转数字计算机（CORDIC）那样的迭代算法去求解支持向量。文献[10] 提出了基于层次聚类的特征选择算法，并且针对现有的特征选择算法所使用的信息熵在整个特征选择过程中保持不变的缺点，提出了动态互信息和条件动态互信息的类内类间距离度量标准。范昕炜提出了推广能力较好的超球面支持向量机模型并且提出了加权支持向量机来解决类别的差异对分类精度的影响问题[31]。

1.3 本文的研究内容和创新点

传统的人体活动识别技术主要依赖与图像视频的处理技术，这类技术的应用推广受设备部署和天气状况等因素的的影响很大。这几年随着智能手机和智能穿戴设备等技术的兴起，基于移动智能设备内置的传感器的人体活动识别技术显现出了极大地应用价值。

传统的机器学习领域的一些算法在模型的学习和训练过程中需要强大的计算能力和大规模的存储能力。虽然模型的学习和训练过程可以通过离线计算得到，

万方数据

学习到的模型参数在用于活动识别时仍然消耗大量的资源和能量，这在应用较多的智能手机上并不现实。另一方面，研究人员对于在嵌入式设备和传感器网络中应用支持向量机算法的模型进行了研究并提出了一些方法，这类算法虽然具有较低的计算代价和能耗，但是其计算精度并不高。

结合前人的研究技术和当今智能手机和智能穿戴等技术的发展现状，本文改进了基于层次聚类的特征选择和改进 HF-SVM 的人体活动识别技术，主要创新工作有以下内容：

1) 使用改进的基于层次聚类算法的特征选择算法。基于层次聚类的特征选择算法是近两年提出的算法，在该算法活动识别领域尚未使用。然而基于层次聚类算法利用了互信息和关联系数两种评价函数来作为类内距离和类间距离的度量。这样的距离度量决定了算法只能应用于离散型和标称型数据类型的数据。在活动识别领域，预处理和特征提取得到的可用数据是连续型的。虽然现在有很多方法可以将连续性数据离散化，但是连续数据的离散化过程将丢失很多信息。根据提取到的数据的特点，本文改进了算法的评价函数。文中使用皮尔逊相关系数来计算候选特征与分类类别之间的距离，使用共享最近邻（Shared Nearest Neighborhood，SNN）来计算候选特征与特征子集的距离。

2) 使用基于硬件友好型核函数的 SVM 算法。传统的 SVM 算法模型的训练需要强大的硬件运算能力和存储能力，在模型用于识别活动的过程中，利用决策函数（分割超平面）做分类需要进行大量的指数运算。现在智能手机虽然可以进行e^x 这样的指数运算，但是相比2^n 只需要一些移位和加减法运算，进行e^x 这样的运算需要的计算能力要求更高。而文献[8]提出的基于硬件友好型核函数的方法，受到时硬件资源限制，提出了一种迭代运算的算法来求解这个二次规划问题。这个算法为了减少计算量牺牲了精度。因此，本文中改进了硬件友好型核函数，同时使用 SMO 算法来求解支持向量机问题，使之既保持高斯核函数抗噪声的优点，也具有较小的计算代价。

万方数据

3) 基于组合分类器的活动识别技术。多类别分类问题通常有 OVO 和 OVA 这两种策略来转化成多个二分类问题。OVA 策略往往由于正负样本不平衡数量差距很大而导致分类误差很大。OVO 策略训练除的分类器个数比较多，需要组合多个分类器得到最终结果。本文提出了基于 OVO 策略的 SV 算法，将多个分类器的输出结果作用于 Sigmoid 函数，然后根据每个二值分类器的投票得到最终输出。这个方法可以避免单独使用 Sigmoid 函数取最大值时易受到噪声的干扰而预测出错，也避免了多个分类器直接投票由于分类器权重相同导致的错误。

1.4 论文组织结构

在智能手机配置日新月异、移动互联网迅速发展、“互联网+健康”大力倡导的今天，研究利用智能手机内置的传感器采集的数据建模分析识别用户的活动识具有深厚的现实意义。

本文基于前人利用智能手机内置的传感器进行活动识别的研究，提出了基于层次聚类的特征选择和硬件友好型 SVM 的用户活动识别技术。具体章节如下：

第一章：绪论。首先介绍活动识别问题的研究背景和研究意义，之后分析了基于智能手机及其内置传感器进行活动识别领域的研究现状，最后简单介绍本文的研究内容和创新点。

第二章：相关理论知识技术。本章首先介绍了智能手机机器内置的传感器技术。其次介绍了统计学习领域的基础理论知识，这些理论知识是本文分析依赖的基础。最后介绍活动识别相关的技术，主要是采集到的信号数据所需要的数据预处理和活动识别的过程。

第三章：基于层次聚类的特征选择算法。本章简述了特征选择的理论基础包括特征选择的过程、方法分类、生成特征子集的策略和评价函数。然后在前人研究的基础上提出了改进的基于层次聚类的特征选择算法。本章结合活动识别的数据特点改进了特征选择过程中的“类内”、“类间”距离度量，即，以皮尔逊相关系数来度量候选特征和分类类别之间的距离，以共享最近邻来度量已选特征子集

万方数据

之间的类内距离。基于这两种距离度量提出了新的评价函数，用于算法在最大相关性和最小类内冗余特征信息之间的平衡。最后本文给出了基于层次聚类的特征选择算法的算法思想和流程，并做实验评估算法的性能。

第四章：改进的 HF-SVM。本章首先介绍了 SVM 算法理论，之后提出了 HF-SVM 算法。主要介绍了 HF-SVM 的求解问题、优化目标和约束函数，然后讨论了常用的核函数，为了减少活动识别过程中的计算代价，本文提出了硬件友好型核函数。然后讨论了本文求解 SVM 模型用到的 SMO 算法，并使用 SMO 算法进行了活动识别。

第五章：基于组合分类器的活动识别技术。由于现实中要识别的活动种类是多样的，因此活动识别问题是一种多分类问题。本章首先介绍了多分类问题常用的策略一对一策略和一对多策略，本章讨论了两种策略的方法和缺点，然后基于一对一策略，提出了 SV 算法来组合多个分类器得到最终的预测结果。

第六章：总结展望。对本文的主要工作做了总结并对后续研究工作的着手点与展开进行了展望。

1.5 本章小结

本章主要介绍了活动识别的研究背景、研究意义和研究现状。分析了当前基于智能手机传感器的活动识别技术存在的问题，提出了基于层次聚类的特征选择算法和硬件友好型核函数的支持向量机算法的多类别活动识别技术。然后介绍了本文的研究内容和本文的组织结构。

万方数据

第2章 相关技术综述

2.1 智能手机传感器技术

智能手机内置了大量的传感器来测量动作状态、位置变化和各种周边环境条件。这些传感器的一些是通过硬件生成的，一些是通过软件合成的[12]。软件合成的传感器将多个硬件传感器获取到的数据进行分析处理和加工得到自己的数据，因此也被称为虚拟传感器或者合成传感器。智能手机内置的传感器根据功能和作用不同，分为三大类：

（1）动作传感器

包括加速度计、陀螺仪、重力和矢量传感器。

（2）环境传感器

测量各种周边环境参数（如周围的气温、气压、光照强度和湿度等），包含气压、光照和温湿度度传感器等。

（3）位置传感器

这些传感器测量的是传感器物理位置的变化，包括方向等传感器等。

在活动识别领域，经常使用动作类传感器来完成活动相关的信息数据的采集，用到的较多的传感器是动作传感器，如加速度计、重力传感器、陀螺仪等。

2.2 统计学习理论

统计学习(Statistical Learning)是一门对样例数据构建概率统计模型并运用学习模型对数据进行预测、分类和分析的学科，也称为统计机器学习[13]。

2.2.1 损失函数和风险函数

监督学习问题的任务是在给定的假设空间中选取算法模型f作为决策函数。对于给定的输入X，由学习到的模型给出相应的预测值$f(X)$，这个输出的预测值$f(X)$与真实值Y，并非是一致的，因此可以用损失函数 Loss Function）或者代价

万方数据

函数（Cost Function）来度量预测出现错误的程度，损失函数是模型预测值$f(X)$与真实值 Y 的非负实值函数，可以记作$L(Y,f(X))$ 。

统计机器学习中经常用到的代价函数有以下几种[20][14]：

（1）0-1 损失函数（0-1 Loss Function）

$$L(Y,f(X)) = \begin{cases} 0, & Y = f(x) \\ 1, & Y \neq f(X) \end{cases} \tag{1}$$

（2）平方损失函数 （Quadratic Loss Function）

$$\mathrm{L}(Y,f(X)) = (Y - f(X))^2 \tag{2}$$

（3）绝对损失函数 （Absolute Loss Function）

$$L(Y,f(X)) = |Y - f(X)| \tag{3}$$

（4） 对数损失函数（Logarithmic Loss Function）或对数似然函数（Log-likelihood Loss Function）

$$L(Y,P(Y|X)) = -\log P(Y|X) \tag{4}$$

损失函数值越小的算法模型越好。由于样本（X，Y）来自于样本空间，服从联合分布 P(X，Y)，所以损失函数的期望是

$$R_{exp}(f) = E_P[L(Y,f(X))] = \int_{\chi\times y} L(y,f(x))P(x)\,dx\,dy \tag{5}$$

损失函数的期望是风险函数（Risk Function）也称作期望损失（Expected Loss）。

对于给定的训练数据集$\mathrm{T} = \{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ ，学习到的算法模型$f(X)$关于训练样本数据的平均损失是经验风险（Empirical Risk）也叫作经验损失（Empirical Loss），记作：

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i)) \tag{6}$$

依据伯努利大数定理，当$N \to \infty$ 时，经验风险几乎接近于期望风险。

2.2.2 训练误差与测试误差

假设从训练样本数据中学习到的模型是$Y = \tilde{f}(X)$ ，训练误差是算法模型在训练样本数据集上的平均损失函数

万方数据

$$R_{emp}(\tilde{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \tilde{f}(x_i)) \quad (7)$$

测试误差是模型在测试样本数据集上的平均损失函数

$$e_{test} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i, \tilde{f}(x_i)) \quad (8)$$

训练误差体现了学习到的算法模型在已知的样本数据集上的分类能力，训练误差较小则经验风险较小。测试误差体现了学习到的算法模型在未知的样本数据集上的分类能力，也称为模型的泛化能力。对于给定的两个学习模型，测试误差小的模型具有更好的泛化能力。

2.2.3 正则化与过拟合

按照经验风险最小化的原理，具有较小经验风险的学习模型是最优的算法模型，因为模型的经验风险小说明模型对训练数据所服从的分布学习的越充分。

由于训练样本来自于样本空间，当样本容量足够大的时候，训练样本集趋近于样本空间，按照经验风险最小化理论，学习到的模型在样本空间上也必然具有较好的分类能力，因此能保证较好的学习效果。比如逻辑斯蒂回归（Logistic Regression，LR）分类模型用到的极大似然估计(Maximum Likelihood Estimation，MLE)就是经验风险最小化的例子。

但是当训练样本的容量相对很少时，使经验风险最小化很容易导致过拟合现象。过拟合现象是指学习到的模型包含过多的参数，以至于出现学习到的模型在训练样本数据集上分类的性能很好，但是在未知的测试数据集上分类的效果很差。

依据奥卡姆剃刀（Occam's Razor）原理，我们应优先选择较简单的模型[14]。为了防止经验风险最小化学习到的模型发生过拟合现象，可以在经验风险项的基础上增加一个正则项或者惩罚项。正则项（或惩罚项、惩罚因子）是模型复杂度的单调递增函数，模型复杂度越大，正则项的值越大。

因此我们求解的目标变为

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \quad (9)$$

也就是说我们希望通过使经验风险项和正则项整体最小，来寻找最优模型的解。

万方数据

2.2.4 交叉验证与保持方法

如果可用的原始数据集中样本的个数充足，较简单的方法是将样本数据随机的划分成训练集，验证集和测试集三部分。其中，训练集用于模型的训练，验证集则用于模型选择，从中选择具有最小验证误差的模型为最优模型。测试集用于对学习到的模型的泛化能力做最终的评估。

在实际中算法女应用中可以利用的原始数据集的样本个数是有限的。当样本数目较少时并不能充分反映样本空间中样本的分布情况。较常用的评估模型性能的方法有以下两种。

保持方法：将原始数据集随机划分成两个互不相交的数据集合，分别称为训练集和测试集。在训练集上训练模型，并选择具有较小训练误差的模型为最优模型，在测试集上评估模型的性能。这种方法可能因训练数据较少而使模型方差变大，或者模型学习不充分而使模型性能下降。

交叉验证是算法实践中较常用的方法。在该方法中，每个样本数据记录用于训练的次数相同，并且恰好检验一次。常用的交叉验证方法有以下几种：

1.简单交叉验证

简单交叉验证的方法是：随机将已给的数据分成两部分，一部分用做训练集，另一部分用作测试集（例如，70%的数据用于训练，30%的数据用于测试）。然后在不同参数下用训练集训练算法模型，在测试集上测试各个模型的泛化能力，从中选出测试误差小，泛化能力强的模型。

2.S 折交叉验证

S 折交叉验证（S-fold cross validation）的方法如下：首先随机的将可用的原始数据集划分成 S 个大小相等互不相交的数据子集，然后从中选择一个子集进行用于测试，其余 S-1 个子集用于模型训练。这个过程可以重复 S 轮，使每一轮的测试集互不相同。可以从这 S 个不同的模型中选择一个测试误差最小的模型为最优模型。

也可以将简单交叉验证和 S 折交叉验证融合起来确定学习到的模型的最优参数，我们可以对相同的参数进行一次 S 折交叉验证，用 S 个模型的平均测试误差

万方数据

的大小来衡量参数选取的优劣。

2.3 活动识别相关技术

2.3.1 识别过程概述

数据采集人员（或志愿者）按照研究人员的要求，将内置了各种传感器的设备（这里是智能手机）携带在身上，并按照研究人员的要求进行一些规定的活动动作。智能手机内置的加速度计和陀螺仪等传感器完成活动相关的信号数据的采集和传输。采集到的数据信号是连续的数字信号，为了从数字信号中得到可以用于模型训练的数据，需要对采集到的数字信号进行数字信号预处理并按照一定的提取算法进行特征提取。

数据的预处理包括去除噪声（去噪），滤波，抽样，离散化和二元化，比例缩放，量纲统一等操作，具体选择何种预处理技术因使用数据的特点和系统自身而异。对于用智能手机的加速度计和陀螺仪等传感器进行人体活动识别，需要的预处理过程是噪声处理。利用智能手机的传感器采集数据会受到一些（比如身体抖动等）外部环境因素的影响，从而导致采集的信号会带有很多噪声。由于采集到的实时的数据信号很长，直接处理并不是很方便，我们需要对噪声处理后的数据信号进行进一步的加工处理（比如抽样），把它分成一个个容易处理的小片段。常用的方法是含有50%重叠的固定窗口法。

对于每一个小片段我们进行特征的提取。我们可以按照设计好的特征提取算法计算每个片段对应的各个特征的值，从而实现特征矩阵。但是，有时候，我们并不清楚那些特征与样本的label值有较强的相关性，或者，我们按照特征提取算法得到的特征的个数很多，直接处理计算量会很大，因此我们也需要对提取好的特征进行特征的融合或特征的选择来确定具有最佳分类性能的特征属性集。使特征选择过程筛选出来的特征属性集可以有效的去除冗余特征，提高识别率。之后，把我们处理后的样本送入模型进行训练，从中选取泛化能力强的较强的模型来进行识别工作。

万方数据

2.3.2 数据预处理

利用智能手机内置的加速度计所采集的人体加速度信号是混杂的，包含了仪器噪声，重力加速度，和人体颤抖等的干扰。为了保证学习到的模型具有较好的识别率，具有较小的模型方差需要进行噪声过滤等预处理。常见的预处理技术有：去噪，分片，归一化，平滑，维规约，离散化和二元化等。

2.3.2.1 去除噪声和平滑

数字滤波器经常用于对采集到的数字信号进行去噪声和平滑等处理操作。数字滤波器是一个离散时间系统，他按照预定的算法，将输入信号序列转换为输出信号序列[32]。数字滤波器具有精度高，可靠性程度高，可程控改变特性或复用，便于集成等优点，按照所处理的信号的维数，可以分为一维、二维或多维数字滤波器[12]。

从滤波器单位冲激响应的长度来分，可将数字滤波器分为两种，其中的一种是有限冲激响应（Finite Impulse Response，FIR）滤波器，还有一种就是无限冲激响应（Infinite Impulse Response，IIR)滤波器[32]；从频域响应的角度来分类，可将数字滤波器分为低通滤波器（Low Pass Filter）、高通滤波器（High Pass Filter）、带通滤波器（Band Pass Filter）、带阻滤波器（Band Stop Filter）。FIR 滤波器系数的计算主要有三种窗函数法、频率采样法和最优化方法；IIR 滤波器系数的计算从原形滤波器出发，常用的圆形滤波器包括巴特沃斯滤波器、切比雪夫滤波器和椭圆滤波器[12]。

2.3.2.2 分片

在人体进行活动的过程中，内置各种传感器的智能手机放置在人体身上不停的进行着数据采集工作，所以采集到的信号长度往往都比较长，得到的信号数据并不能直接用于后续的处理。在进行活动识别之前，首先需要对采集到的数据信号进行分割成小片段处理，常用的几个分片方法有滑动窗口、自顶向下、自底向上和滑动窗口与自底向上。研究工作中常用的方法是50%重叠的滑动窗口。

万方数据

2.4 本章总结

本章介绍了首先介绍了智能手机及其内置的传感器技术。智能手机的传感器技术决定了基于智能手机的活动识别的数据的形式和特点，是展开算法研究的基础。传感器的到的数据经过预处理得到的一些特征属性。这个特征属性是每个分片得到的时域频域信号的一些统计量值。这些值是连续的。应用传统的算法需要进行离散化处理，虽然离散化技术已经很成熟，但是连续数据的离散化将导致信息的丢失。本章第二部分讨论的是统计学习的相关理论知识。这些理论技术在后面的章节中或者是理论基础或者是用到的方法。

万方数据

第3章 基于层次聚类的特征选择算法

当样本数据包含的特征属性特别多，很容易产生“维灾难”问题[22]。样本的维度很高时，这些特征属性之间经常是相互关联的，这些相互关联特征的出现给传统的学习算法带来许多挑战或困难，它们不仅给学习算法的性能带来严重影响，而且还可能引入噪声数据，从而干扰学习过程。

特征选择是统计学、机器学习和数据挖掘、文本分类、语音识别等领域中的经典问题，目的是为了解决算法模型在大规模数据的实际应用中的计算和性能问题。特征选择是从样例的特征属性集中选择出一个特征子集，把样本数据由高维变成低维数据，同时降低特征属性之间的冗余信息，使分类器在特征子集上分类的性能最好。随着上世纪90年代新技术的提出和发展，尤其是大数据时代数据的规模和维度越来越大的背景下，特征选择问题引起了越来越多人工智能领域研究人员的广泛关注。

3.1 特征选择技术

3.1.1 特征选择过程

特征选择的过程如下图3.1所：

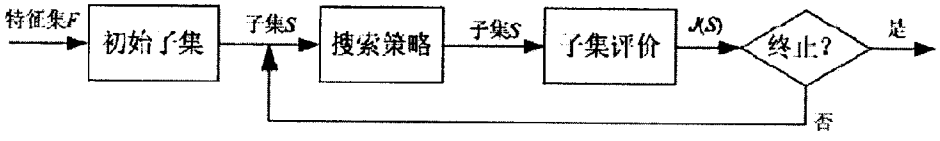

图 3.1 特征选取过程

特征选择的过程大体上有前向搜索和后向搜索两种，前向过程初始时刻特征子集为空，然后按照一定的搜索策略，利用评价函数计算的结果每次选择最合适的特征加入到特征子集中。而后项搜索策略初始时刻特征子集等于特征的全集，然后根据评价函数的值从特征子集中逐步移除某个特征，直到达到一定的停止条

万方数据

件。

特征选择算法评价函数的具体实现有多种度量方式，但是总体来说评价函数可以分为三大类型：

（1） 从特征集中找到一个特征子集，是这个特征子集在某个评价函数下得到的值最大（意味着所含有的信息最大）；

（2） 设定一个阈值，在评价函数取值大于阈值的前提下，找到最小特征子集。

（3） 从特征集中找到某一个特征子集，使所选取的特征子集的评级函数值最大，同时所包含的特征的数目最少。

这三种方式体现了特征选择时侧重的不同的目标。第一种方式，着重所选择的特征子集在评价函数下尽可能少的信息损失。第二种方式着重在在满足给定阈值情况下最小子集。第三种方式是两者的折中，兼顾了评价函数的值（信息量）和特征子集的个数。

特征选择的过程包括确定生成子集的搜索过程所使用的方法策略，然后根据样本数据的特点选择合适的评价函数，确定生成过程的终止条件，实验验证特征选择的效果。终止条件可以是设定的一个阈值，评价函数或者特征子集中特征的个数达到一定标准后就可停止特征子集的搜索过程。

3.1.2 特征选择的方法

特征选择算法既可以在分类学习算法的预处理阶段完成，也可以成为算法学习的组成部分。根据特征选择算法分类学习中进行的时机，可分为 Embedded、Filter 和 Wrapper 这三种类型[10]。Embedded 类型的算法模型的特征选择与分类学习模型的学习过程同时进行，在模型训练学习的过程中动态选择合适的特征。这样的方法两种，一种是算法进行的每一次迭代都会进行特征选择，比如使用 C4.5 或者 ID3 算法的决策树；另一种是通过控制正则化项，通过参数稀疏的方法实现特征选择。

Filter 类型的算法模型，先进行特征的选择，这是作为预处理步骤，然后利用

万方数据

选择的特征来训练模型。Filter类特征选择算法的相似之处是特征选择过程的每步都是选择与类别相关程度最高，且与已选择特征冗余性最低的特征。这种预处理方式在某些情况下特别有优势，如处理大规模数据或在线数据等，利用特征选择后的数据训练模型减少模型训练过程中的计算量。Wrapper 类型将特征选择算法作为算法模型训练学习过程的一个组成部分，将分类器的分类性能作为特征重要性的评价标准，选择能使分类器性能最优时分类器使用的的特征子集为最优特征子集。

3.1.3 评价函数

评价函数是对特征选择过程中所选择的特征或子集的优劣程度进行评估的尺度函数[10]。评价函数的选择直接决定了特征选择算法选择的特征，也直接决定了学习到的模型的性能，因此选择合适的评价函数进行特征选择非常重要[21]。

特征选择用到的评价函数有多种形式，主要包括距离度量、一致性度量、信息度量和依赖性度量等。

（1）距离度量：按照距离公式形式的不同，距离度量又可细分为类间距离和概率距离两种。类间距离通常指是指几何空间的距离定义，如欧氏距离和马氏距离等。Relief [15]及其变种 Relief F 就是使用欧氏距离来衡量特征子集的重要性程度。概率距离采用概率统计形式来计算类间距离和类内距离，其中类间距离越大，且类内距离越小，则类别间的可分离性就越高[16]。

（2）一致性度量：如果两个数据样本的特征值相同，但所属类别不同，则称它们是不一致的（否则就是一致的）[17]。一致性度量标准的优点是能获得一个较小的特征子集，但它对噪声数据比较敏感，且只适合离散特征。

（3）信息度量：基于信息熵等信息论技术来度量分类类别的不确定性程度。信息熵能很好地量化特征相对于类别的不确定性程度，因此它在特征选择算法中得到广泛关注，如 Yu 和 Liu 等利用对等称不确定性[18]来度量特征间的相关性和冗余性。除信息熵外，常用的基于信息的度量有信息增益、互信息和最小描述长度等。

万方数据

（4）依赖性度量：通过计算样本类别与样本特征属性之间的统计相关性来度量某个特征对于确定分类类别的重要性程度。当前有 t-test、F-measure、Pearson 相关系数、概率误差、Fisher 分数、线性可判定分析[14]和最小平方回归误差[19]等度量方法。

3.2 改进的基于层次聚类的特征选择算法

在基于层次聚类的特征选择算法中，互信息和信息熵经常被用来度量候选类与标签类之间的距离。互信息是信息论的有用度量，表示一个随机变量在另一个随机变量中所包含的信息，是已知一个随机变量的情况下，另一个随机变量不确定的减少。互信息常用于离散型、标称型或已知分布的连续性。在活动识别领域，传感器采集到的数据经过预处理后是连续型的。虽然现在可以通过不同的方式计算连续特征的信息熵，如核密度，Parzan 窗口等，这样的处理方式增加了模型学习和应用的复杂度，所以如果要计算互信息就必须把连续性型数据离散化。虽然目前可用的特征离散化方法有很多，如等频率算法、等宽区间算法、最小描述长度（MDL）[14]和布尔推理算法等，但是连续数据的离散化会导致信息的丢失，从而学习到的模型对样本的分类会产生一些错误。

为了更好地应用基于层次聚类的特征选择算法，本文改进了基于层次聚类的特征选择算法。本文提出的评价函数基于依赖性度量的思想。首先选择使用 SNN 来度量已选择的特征子集内部的距离，这个距离表明已选择的特征子集所包含的冗余信息的多少；使用皮尔逊相关系数[14]（Pearson's correlation）来度量候选特征和类别之间的相关性，这个度量来衡量候选特征对确定分类类别的重要性。为了综合考虑最大类别相关性和最小冗余信息这两种度量方式，本文提出了新的评价函数。本文提出的评价函数能够很好的应用于连续性样本数据。

3.2.1 改进的距离度量

3.2.1.1 基于 SNN 的类间距离度量

基于相似度和欧氏距离的度量标准虽然应用广泛，但是存在很大的缺点。比

万方数据

如说，用于学习的模型中有一些噪声数据，如果待预测的采集到的样本数据，与某一个噪声相似系数很高（或欧氏距离很小），那么模型就可能受这条噪声数据的干扰而做出错误的分类。共享最近邻相似度 SNN 则可以很好地避免此类问题。它基于以下原理：如果两个数据点都与一些相同的点相似，则即使直接的相似度量不能指出，他们也相似。

在本文中，将使用共享最近邻的评价函数。共享最近邻算法。所使用的邻近性度量是欧式距离。

SNN 算法的算法描述如下表 3.1。

表 3.1 SNN 算法描述

计算共享相似度的步骤：
(1)：找出所有点的 k-最近邻。
(2)：if 两个点 x 和 y 不是相互在对方的 k-最近邻中 then
(3)：　　similarity$(x, y) \leftarrow 0$
(4)：否则
(5)：　　similarity$(x, y) \leftarrow 0$ 共享近邻的个数
(6)：endif

本文中我们选用 SNN 来度量选择的特征子集内部的距离。SNN 相似度值越大，说明两个数据越相似，说明被度量的其中一个特征属性所包含的另外一个特征属性的信息越少。换言之，知道了其中一个特征属性，另一个特征属性的不确定性减少的越小。因此 SNN 值越大对应于互信息值越小。

3.2.1.2 基于皮尔逊相关性的依赖性度量

皮尔逊相关系数可以用来衡量两个连续型的样本数据之间的相似性。皮尔逊相关系数的定义如下表 3.2。

万方数据

表 3.2 皮尔逊相关系数定义

$$\mathrm{corr}(x,y)=\frac{\mathrm{covariance}(x,y)}{\mathrm{standard\_deviation}(x)\times\mathrm{standard\_deviation}(y)}=\frac{s_{xy}}{s_x s_y}$$

其中：

$$\mathrm{covariance}(x,y)=s_{xy}=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$\mathrm{standard\_deviation}(x)=s_x=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$$

$$\mathrm{standard\_deviation}(y)=s_y=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$$

$$\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k \quad \text{是}x\text{的均值}$$

$$\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k \quad \text{是}y\text{的均值}$$

本文中用皮尔逊相关系数来度量候选特征与样本分类类别的距离。根据定义知道皮尔逊相关系数的取值为$[-1,1]$。如果皮尔逊相关系数为 1，则说明候选特征与类别的正相关性，皮尔逊相关系数为-1，则说明候选特征与类别的负相关性，皮尔逊相关系数为 0，则说明候选特征与类别的不相关性。在本文中我们希望选择与分类类别最具相关性的特征，而正相关和负相关都是具有相关，因此我们真正使用的度量是皮尔逊相关系数的绝对值。用[0,1]之间的取值来衡量特征与类别之间的相关程度。

3.2.2 改进的评价函数

基于层次聚类的特征选择算法中，互信息被用来度量候选特征f 与类别C之间的距离。互信息是信息论领域一种有用的度量。他可以看成是一个随机变量中

万方数据

包含的关于另一个变量的信息。是一个变量已知的情况下，另一个变量减少的信息量。互信息这种距离度量方法适合离散型和标称型数据类型或者分布已知的连续型随机变量。在活动识别领域，样本数据是由智能手机或者智能穿戴设备内置的传感器采集而来，采集到的信号经过去噪声、切片和傅里叶变换之后得到样本数据，这样的样本数据是分布未知连续性的。因而，互信息这种度量距离的方法并不适合活动识别领域。

本文中将使用皮尔逊相关系数来计算候选特征f与分类类别C之间的距离$d(f,C)$。在每次迭代过程中都要计算候选特征f 和已选特征子集之间S的 SNN 距离$D(f,S)$，即，已选特征子集包含f的 k 近邻的个数。同时还要计算将候选特征f和已选特征子集S合并之后已选特征子集之间的类内距离$D_{new}(S)$

$$D_{new}(S) = D_{old}(S) + D(f,S)$$

由于$D(f,S)$表明的是已选择特征子集包含候选特征f的共享最近邻的个数，表明的是引入的冗余信息的大小，而$d(f,C)$表明的是候选特征与分类类别的依赖性度量，这个度量越大，说明分类类别与之越相关。我们希望选择的特征既具有很大的相关度，又具有较少的冗余信息.我们需要将两种距离折中，因此引入评价函数$J(f)$：

$$J(f) = \frac{d(f,C)}{D_{new}(S)+\lambda} \tag{10}$$

其中λ 是惩罚因子，目的是寻求两种距离的较好折中，得到最优的评价函数$J(f)$

3.2.3 搜索策略

生成特征子集的搜索算法有完全搜索、启发式搜索、随机搜索三大类[10]。

完全搜索主要有宽度搜索(Breadth First Search)、分支界限搜索(Branch and Bound)、定向搜索(Beam Search)、最优优先算法(Best First Search)这几类算法。

启发式搜索主要有序列前向选择（Sequential Forward Selection)、序列后向选择(Sequential Backward Selection，SBS)、双向搜索(Bidirectional Search)、增 L 去 R 选择算法（Plus-l Minus-R Selection，LRS)、序列浮动选择(Sequential Floating Selection，SFS)、决策树(Decision Tree Method，DTM)等算法。

万方数据

随机算法主要有随机产生序列选择算法(Random Generation plus Sequential Selection，RGSS)、模拟退火算法(Simulated Annealing，SA)、遗传算法(Genetic Algorithms，GA)等。

基于层次聚类特征选择算法就是建立在聚类算法思想的基础之上的，他的搜索策略与聚类相似，即不断将选择的特征子集与单个候选特征进行融合操作，直至融合的特征数目达到指定的阈值时结束。与传统聚类算法不同之处是，这种方法处理的个体不是样本数据点而是特征属性列。

为了尽可能的减少基于层次聚类的特征选择算法在特征选择过程中的复杂度。我们可以根据特征选择的实际情况，合理的选择前向搜索或后项搜索。当样本包含的特征个数较多，采用前向的搜索策略将使模型的搜索过程变得较慢。因为前向搜索过程需要在每次迭代过程中选择当前最优的候选特征加入特征子集，如果选择的特征个数较多，必然搜索需要的迭代过程就较长。由于后项是每轮迭代删除特征，因此如果选择的特征个数远远多于淘汰的个数时应使用后项搜索。

3.2.4 特征选择算法流程

基于层次聚类的特征选择算法是一种自底向上的算法，思想与基于启发式搜索的序列前向选择的搜索策略相一致。初始时刻已选择的特征子集为空，每次选择的过程中都选择当前性能最好的特征。

基于层次聚类的特征选择算法的具体实现流程如下表 3.3 所示。运用特征选择算法之前，首先对样本数据集实施必要的预处理（例如归一化），将初始特征子集置为空并初始化相关参数。

万方数据

表 3.3　改进的层次聚类算法

基于层次聚类的特征选择算法

输入：样本数据集，特征集合 F，分类类别 C

输出：选择的特征子集

（1） 初始化相关参数。将初始特征子集置为空，$D(S)$=0

（2） 对特征集合 F 中的每一个特征 f,计算他与分类类别之间的距离 $d(f,C)$，计算 f 与已选特征子集之间的距离$D(f,S)$ ，计算出我们定义的的评价函数的值$D(f,S)$

（3） 从（2）中选出 最优的$f = \arg\max_{f \in F} J(f)$ 将该特征加入到选择的特征子集中

（4） 更新已选特征子集内部的距离

$$D_{new}(S) = D_{old}(S) + D(f,S)$$

（5） 更新特征集合$F = F - f$

（6） 当 特征子集的特征个数$|S| <$ δ 重复（2）~（4）

终止

返回选择的特征子集

第一轮迭代，利用皮尔逊相关系数的计算公式去计算每个候选特征 f 与分类类别f 之间的距离$d(f,C)$，选择距离最大的那个候选特征加入到特征子集中。

之后，基于层次聚类的特征选择算法在每一次迭代的过程中选择最大的$J(f)$将其与已选择特征子集S进行合并，以组合成新的特征子集。随后更新已选特征子集的类内距离$D(S)$，以加快聚类过程的速度。这个合并聚类过程一直循环迭代，直到候选类 S 中的特征个数超过预先设定的阈值时结束。

3.3 实验结果

基于层次聚类的特征提取算法有三个参数分别是，选取的特征个数 C，最近邻的个数 K，评价函数中保持最大相关性和最小冗余平衡的λ 。这三个参数的选

万方数据

取都会影响特征选取的过程。

3.3.1 对比实验

为了比较基于层次聚类的特征选择算法与文献[10]提出基于层次聚类和互信息的特征选取算法 ISFS，以及典型的特征选取算法 CFS 算法所获的特征子集的算法性能，我们将用 UCI 机器学习存储库中的 2 个样本数据集（Internet Advertisements 和 Mushroom）来评测。Internet Advertisements 数据集用来对互联网上的广告图片进行分类和识别。他的特征是从图片的长、宽、高，是否含有某些域名和短语等方面提取出来。他的数据集的特征个数，和数据量（样本个数）如下表 3.5。Mushroom 数据集给出了蘑菇的一些特征，比如孢子的颜色，茎的形状，寄生环境（落叶还是草）等来分类蘑菇是否有毒。

这 2 个数据集的简要信息描述如下表 3.5。

表 3.4 本章使用的评测数据集描述

序号	数据集名称	样本个数	特征个数	类别数
1	Internet Advertisements	3279	1558	2
2	Mushroom	8124	22	2

文中使用的这两个数据集来源于现实应用领域，因此有些样本数据的某些特征属性可能由于种种原因有缺失。对于特征缺失数据的处理，目前已经提出了许多方法：（1）直接删除带有缺失数据的样本数据，这样的做法会使可以用来训练模型的样本数目减少；（2）将缺失数据作为一个特殊值处理（在某些情况下可以用 0 补缺）；（3）使用统计插值方法对缺失数据进行补齐操作（比如离散型数值可以统计该特征出现次数最多的值，连续性的可以统计该属性的均值）；（4）使用粗糙集理论对数据进行补齐。在本次实验中，Internet Advertisements 数据集，在关于图片的宽度，高度等连续性的值，我们用对应属性的均值来做为缺省的值，对于离散型和标称型数据，将统计出现次数最高的那个数值。Mushroom 数据集由于都是离散值，缺省是用出现次数最多的那个值来代替。

万方数据

我们在 windows 10 平台下,使用 python 2.7.13 及 numpy，scikit-learn 等工具包进行下面的实验。Scikit-learn 含有常用的机器学习算法模型的工具包。

我们首先使用 Mushroom 数据集来说明特征选择的必要性。由于该数据集主要用于关联规则的发现和提取。在这个数据集中，每一个特征属性都与蘑菇是否有毒有很强的相关性。关于这个数据集的一些官方基准的规则如下表 3.7。

表 3.5 Mushroom 规则基准准确率

规则	解释	准确率
odor=NOT（almond.OR.anise.OR.none）	气味非杏仁味或茴香味或无气味	98.52%
spore-print-color=green	孢子颜色是绿色	99.41%
habitat=leaves.AND.cap-color=white	栖息在落叶上并且帽子颜色是白色	100%

上表的准确率是 UCI 开源数据集官方网站提供的基准准确率。可以发现只用少个几个特征就可以达到很高的准确率。

我们尝试从 mushroom 数据集中提取 1-22 个特征，将提取到的特征对应的属性列从原始数据集中抽取出来，构建成新的数据集，然后我们使用十折交叉验证的方法，来计算不同数目的特征分类的平均准确率。于是我们得到下面的图 3.2

万方数据

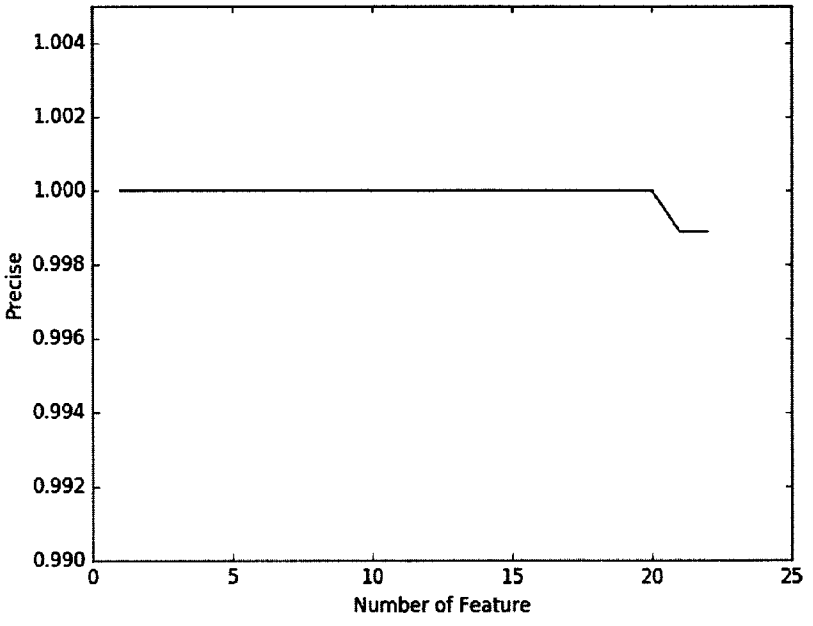

图 3.2 不同个数的特征的模型准确率

图 3.2 很好地说明了，当我们仅用少数几个特征（大于等于 1，小于等于 20）时，学到的模型的平均准确率为 100%，当我们使用 21 个特征或者 22 个特征的时候准确率反而降低。这是因为使用的特征越多模型越复杂，错误分类的可能性越大。因此对原始数据集进行特征选择对于获得较高准确率的模型是很重要的。

下图 3.3 是 Internet advertisement 数据集进行的实验。原始数据集有 1558 个特征，本次试验我们选取的特征的个数是 10-30 个，从图中可以看到，一方面并不是特征的个数越小越好。另一方面，模型并没有随着选择的特征数目的增多准确率提高。因此当在使用模型分类数据集的时候，应当进行特征选择，既可以减少模型训练的复杂度，也能保证模型具有较高的分类准确率。

万方数据

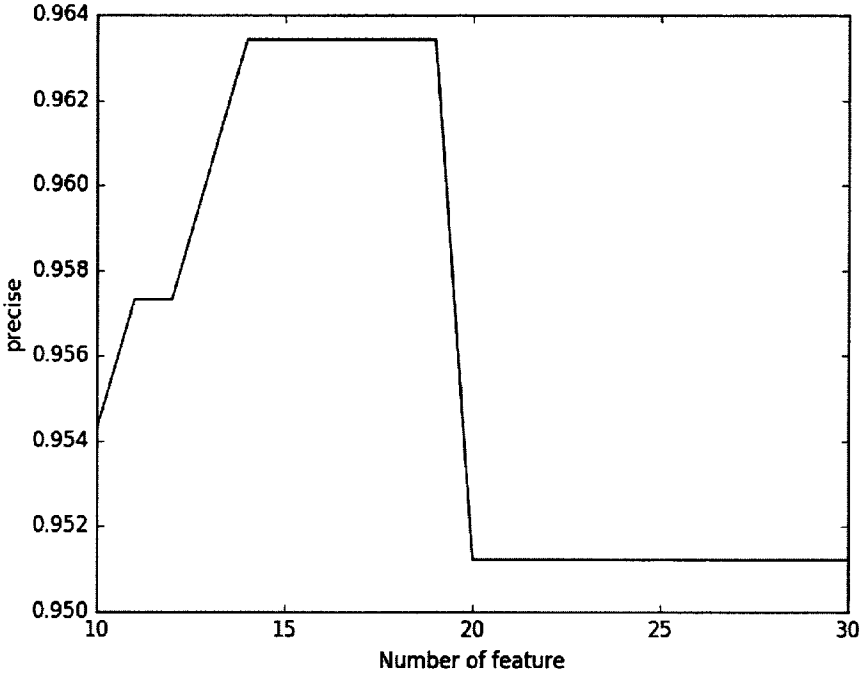

图 3.3 Advertisement 特征个数与准确率

Scikit-learn 的朴素贝叶斯分类器（Naïve Bayes Classification，NBC）有三种，分别是高斯贝叶斯分类器，伯努利贝叶斯分类器，多项式贝叶斯分类器。我们首先测试 Internet Advertisement 数据集在三种贝叶斯分类器下的准确率如下表 3.6 所示。

表 3.6 三种分类器性能评测

分类器	GaussianNB	BernoulliNB	MultinomialNB
准确率	0.772866	0.974085	0.902439

为了与其他特征提取算法具有较好的对比度，我们使用伯努利分类器来完成算法，这是因为伯努利分类器与文献[11]中提出的 ISFS 算法实验结果表中提到的原始分类结果较一致。

我们实验过程中选择的特征的数目是从 10-22 中选取，而由于 advertisements

万方数据

数据集的特征数目较多，考虑到实验的复杂性，我们从500-1200，每次增加50个来求最优分类准确率，结果如下表3.7所示。

表 3.7 二个数据集对比评测结果

数据集	原始	基于层次聚类特征选择	ISFS	CFS
Internet Advertisement	97.41%	96.15%	97.37%	97.18%
Mushroom	99.89%	100%	98.33%	98.33%

通过上面的实验，可以发现，在Internet Advertisement 数据集上，基于层次聚类的特征选取算法的分类准确率性能接近于ISFS，略优于CFS，在Sonar数据集上略优于ISFS和CFS，在Mushroom数据集上略低于ISFS和CFS，基于层次聚类的特征选择算法比较适合连续性特征值的数据，因此具有一定的使用价值。

3.3.2 活动识别数据集

我们使用 UCI 上关于人体活动识别的数据集来进行本章所提算法的实验验证。该数据集是由30位年龄在19-48周岁的志愿者采集的[9]。这些志愿者随身携带（Samsung Galaxy S2）智能手机来进行6种活动（分别是行走，上楼，下楼，坐下，站起，平躺）。使用智能手机内置的加速度计和角速度仪，以50Hz的频率采集得到了原始的三轴线性加速度和三轴角加速度。并有这些数据处理进行特征提取得到该数据集。该数据集有561个特征，10299条样本。

该数据集的特征属性如下表3.8所示，每一个特征属性都是提取到的统计值。

万方数据

表 3.8 活动识别数据集提取的特征示例

特征属性名	描述
tBodyAcc-mean()-X	身体加速度X轴时域均值
tBodyAcc-std()-X	身体加速度X轴时域方差
tBodyAccJerk-mean()-X	身体加速度X轴反射信号均值
fBodyAcc-mean()-X	身体加速度X轴频域信号均值
fBodyGyro-mean()-X	身体旋转向量X轴频域均值
angle(X,gravityMean)	重力均值和X轴的夹角

3.3.3 活动识别实验结果

我们首先讨论λ 和K的合理取值。实验通过固定其他两个参数，改变另一个参数的方法。

（1） λ 的合理取值

由评价函数$J(f)$ 的定义可知λ的取值关于$J(f)$是单调递减函数。其变化曲线如下图3.4所示。

万方数据

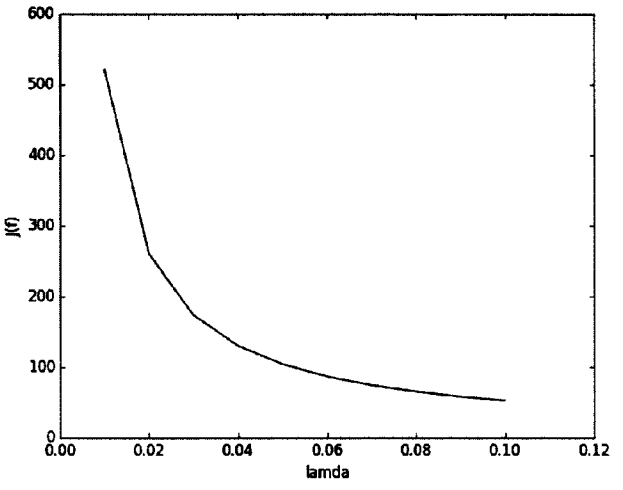

图 3.4 评价函数变化曲线图

本组实验假定选取 10 个特征（C=10），计算共享最近邻 SNN 距离时用到的 K 近邻是 100（$K = 210$），则我们得到以下结果

表 3.9 λ取值与特征子集选择

λ 取值	选取的特征（编号）	特征子集内部距离	$J(f)$
0.02	512, 65, 450, 69, 70, 521, 298, 503, 57, 62	0	1195.141103
0.04	512, 65, 450, 69, 70, 71, 72, 298, 57, 62	4	598.192248
0.06	65, 67, 69, 70, 71, 72, 74, 209, 222, 62	9	399.208196
0.08	65, 67, 69, 70, 71, 72, 74, 209, 222, 62	13	299.715349
0.1	65, 67, 69, 70, 71, 72, 74, 209, 222, 62	17	240.018987

可以看到，随着选取的λ得增大，评价函数呈双曲线下降。结合各候选特征与分类类别之间的皮尔逊相关系数的值，λ 应在$\lambda < 0.1$之间选取。

万方数据

(二) K 的合理取值

本组实验，假定选取 10 个特征(C=10)，假定评价函数$J(f)$ 中平衡最大相关性与最小类间距离的$\lambda = 2$ ，改变计算共享最近邻 SNN 时 K 的取值，则得到以下结果表 3.10 所示：

表 3.10 K 取值与特征子集选择

K 取值	选取的特征（编号）	特征子集内部距离	$J(f)$
10	512, 65, 66, 69, 70, 295, 450, 210, 62, 223	0	2.892891
60	65, 450, 69, 70, 298, 53, 57, 538, 316, 62	0	2.675628
110	512, 65, 450, 69, 70, 521, 298, 57, 62, 37	1	2.563577
160	65, 67, 69, 70, 72, 298, 209, 57, 74, 222	7	1.954307
210	65, 67, 69, 70, 71, 72, 74, 209, 62, 222	13	1.554372
260	65, 67, 69, 70, 71, 72, 74, 209, 62, 222	17	1.235045
310	65, 67, 69, 70, 71, 72, 73, 209, 222, 62	25	1.025697
360	65, 69, 70, 71, 72, 73, 75, 209, 222, 62	36	0.959168
410	65, 67, 69, 70, 71, 72, 73, 75, 209, 62	37	0.782631
460	65, 67, 69, 70, 71, 72, 73, 75, 209, 62	37	0.782631

K 是定义的 SNN 中使用到的最近邻的个数。因为该数据共有 561 个特征，因此 K 的取值不能太小。否则，大多数特征都将不是另一个特征的 K 近邻，从而导致共享最近邻 SNN 大多数都为 0。

从上表中数据可以看出，当 K 较小（K<60）的时候，因为候选特征的 SNN 包含已选特征子集中特征个数为 0，使得特征子集内部距离始终无零，此时特征选取的每轮迭代过程都是由候选特征与分类类别的相关性决定。随着 K 增大，

万方数据

SNN 不为 0，特征选取的过程综合了候选特征与分类类别的相关性和已选特征子集中的冗余度。210<K<260 时，选择出的特征子集达到稳定。随着 K 进一步增大，特征子集之间的冗余性在决策中起到更大的作用。

我们固定选取 K=250，$\lambda = 0.01$ ，每次只改变选取的特征的个数。为了简化实验的复杂度，我们固定的选取步行和爬楼梯这两种活动。我们将步行的样本作为正样本，爬楼梯的样本当做负样本。然后，用我们的算法模型从构建的样本中选取特征，将选取到的特征属性对应的样本列提取出来，构建成新训练样本。然后用 scikit-learn 工具包中的 svm 模块去训练模型。测试样本同样提取步行和爬楼梯两种活动得到，我们分别计算其分类准确率。然后绘制了如下图 3.5。

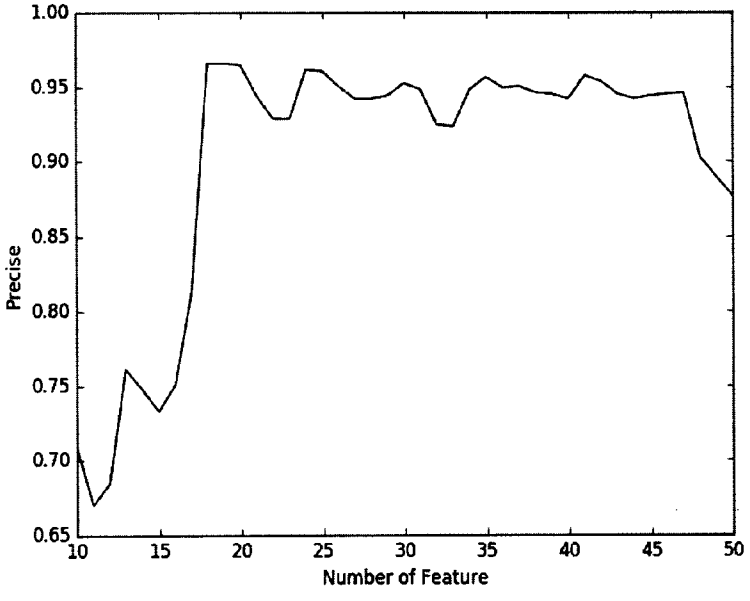

图 3.5 特征个数与分类准确率

从上图 3.5 可知，在特征个数由 10-50 变化的过程中，准确率并不是单调的增加或减少。因此，如果要对每一个二值 SVM 分类器寻找最优解。也就是，对于每一个二值分类器，必须使用网格搜索的办法，遍历所有可能的特征的个数，然后取最优解。在本文中，为了简化模型的复杂度，我们选取了 19 个特征，此时二值 SVM 分类模型的准确率为 96.59%。

万方数据

3.4 本章总结

本章首先介绍了特征选择的过程、方法分类、生成子集的策略和评价函数。在活动识别领域，传感器采集到的数据经过预处理之后是连续性的。虽然现在可以通过不同的方式计算连续特征的信息熵，如核密度，Parzan 窗口等，这样的处理方式增加了模型学习和应用的复杂度，所以如果要计算互信息就必须把连续性型数据离散化。虽然目前可用的特征离散化方法有很多，如等频率算法、等宽区间算法、最小描述长度（MDL）和布尔推理算法等，但是连续数据的离散化会导致信息的丢失，从而学习到的模型对样本的分类会产生一些错误。

本文基于前人的研究结果，结合活动识别领域采集到的数据的特点，提出了基于层次聚类的特征选取算法。该算法使用一种启发式搜索策略，逐步选取最合适的特征加入到特征子集中。在选择的过程中使用皮尔逊相关系数来计算候选特征与分类类别之间的相似性，使用候选特征与特征子集的共享最近邻个数，来度量候选特征加入特征子集之后特征子集的冗余信息。

经过与其他算法的对比试验，证明本章提出的算法具有一定的应用价值。

万方数据

第4章 改进的硬件友好型支持向量机

支持向量机（Support Vector Machine，SVM）是由 Vapnik 等于 1995 年首次提出的。作为机器学习与数据挖掘领域较新的，具有坚实数学背景及系统理论基础的机器学习分类算法，支持向量机备受关注。SVM 可以很好的应用于高维数据，避免了维灾难问题，因此在实践中得到了很好地应用，如手写数字识别，生物特征识别，疾病诊断，入侵检测，视频图像处理等领域。

本章将改进硬件友好性支持向量机(Hardware-Friendly Support Vector Machine，HF-SVM)，并用改进的 HF-SVM 算法，利用第三章特征选择得到的样本进行行为识别。

4.1 SVM 技术

支持向量机根据样本数据是否线性可分，可以分成线性可分支持向量机，线性支持向量机，非线性支持向量机。对于线性可分的训练数据可以通过使分割超平面的间距最大化，得到线性可分支持向量机模型。对于近似可分的训练数据，可以通过引入正值的松弛变量来得到（软间隔）线性支持向量机。对于线性不可分时的训练数据，可以通过施加核函数变换，将线性不可分情况转换为线性可分情况然后寻求间隔最大的分割超平面。

如下图 4.1 所示，子图 A,C,D 是是线性不可分情况。此时无法找到一个分割平面（直线）将两类数据分割开来。图 B 是线性近似可分的情况，他虽然不是严格意义上的线性可分，但是通过引入松弛变量，放松对分割边界的要求，是可以找到分割超平面。

万方数据

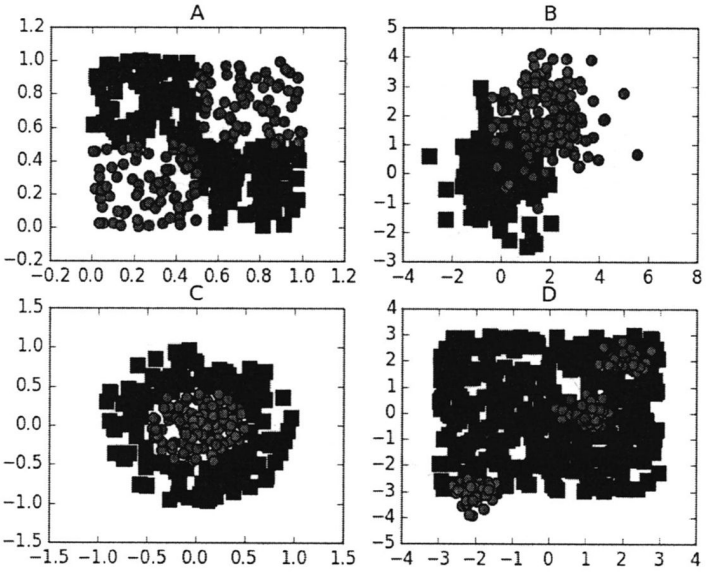

图 4.1　线性近似可分与线性不可分

假设给定特征空间的数据集$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i \in \chi = R^2$ ，$y_i \in \{-1,1\}$ ，$i=1,2,\cdots,N$

学习的目标是在特征空间找到一个分离超平面，能将实例分到不同的类。分离超平面对应的方程式 $w \cdot x + b = 0$，其中 w 是法向量，b 是截距。

对于所有正样本（类标号为 1）它位于分割超平面上方，对于所有负样本（类别号为-1），它位于分割超平面下方。因而有下面的两个式子：

$$w \cdot x_i + b \geq 1 \qquad \text{如果} \quad y_i = 1 \tag{11}$$

$$w \cdot x_i + b \leq -1 \qquad \text{如果} \quad y_i = -1 \tag{12}$$

概括起来则是

$$y_i(w \cdot x_i + b) \geq 1, \qquad i = 1,2,\cdots,N \tag{13}$$

因为分割超平面到正负两类数据间的间隔是$d = \frac{2}{|w|}$，从而可以推导出 SVM 的求解任务是：

$$\min_w \frac{|w|^2}{2}$$

万方数据

Subject to

$$y_i(w \cdot x_i + b) \geq 1, i = 1,2,\cdots,N \tag{14}$$

上式中求解的目标函数是二次的，而约束条件则关于w 和b 是线性的。因此上式是一个凸优化问题。可以利用拉格朗日乘子法去求解极值。

$$L_P = \frac{|w|^2}{2} - \sum_{i=1}^{N} \lambda_i \left(y_i(w \cdot x_i + b) - 1\right) \tag{15}$$

为了最小化拉格朗日函数，必须对L_P关于w 和b求导:

$$\frac{\partial L_P}{\partial w} = 0 \implies w = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{16}$$

$$\frac{\partial L_P}{\partial b} = 0 \implies \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{17}$$

因为上式中的拉格朗日乘子是未知的，因此仍然不能求解得到w 和b 的解。但是可以将式子（17）和（18）与式子（15）中的约束式结合起来，并利用 Karuch-Kuhn-Tucher（KKT）变换将（15）式中的不等式约束变换为以下等式约束

$$\lambda_i \geq 0$$
$$\lambda_i[y_i(w \cdot x_i + b) - 1] = 0 \tag{18}$$

约束（19）表明除非训练样本实例满足$y_i(w \cdot x_i + b) - 1 = 0$ 否则对应的拉格朗日乘子λ_i必须为0。而这些$\lambda_i > 0$ 的训练样本实例就是支持向量（如下图 4.2 显示了数据集、分割超平面、图中圈出了数据集的支持向量）。

万方数据

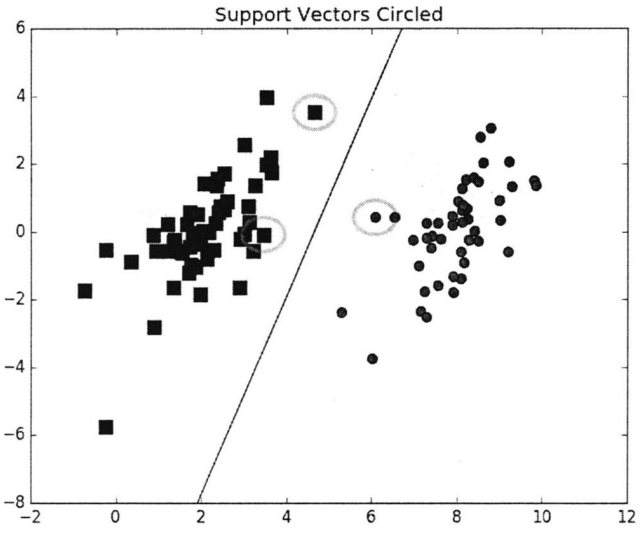

图 4.2 分割超平面与支持向量示意图

直接（16）和（17）带入（18）求解并不容易。可以求其对偶问题得到

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i x_j \tag{19}$$

由于实际中的数据点并不总是线性可分，可能大部分是可以分的但是含有一些噪声，也可能包含一些离群点。而这些离群点的出现总是能够影响分割超平面的选择。甚至在一些情况下，因为某个离群点的出现而导致找不到分割好平面，而我们又希望模型能够容忍一些离群点的存在，因此需要引入松弛变量ξ 来放宽约束条件，就出现了软间隔的分类器。

综合线性可分和线性近似可分（施加松弛变量）的情况，可以将上面的优化问题转化成下面这样的目标式子（下面式子中的α是上面式子中的λ，为了推导方面我们改变了记号）：

$$\min_{\alpha} \frac{1}{2}\alpha^T Q \alpha + \gamma^T \alpha$$

Subject to

$$0 \le \alpha_i \le C \ , i = 0,1,\cdots,N$$
$$y^T \alpha = 0 \tag{20}$$

万方数据

其中，$\gamma_i = 1, i = 1, \cdots, N, q_{ij} = y_i y_j K(x_i, x_j)$

上式的解是原问题的解。若求得上式的α，则分割超平面是

$$y(x) = \sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b \tag{21}$$

对于待分类的样本数据，其特征是x，将其带入分割超平面中，我们可以根据求得的$y(x)$的正负来分类，即预测的分类是$\text{sign}(y(x))$。若$\text{sign}(y(x)) > 0$把他归为$y = +1$这一类。反之亦然。

上述式子中其中一个是最小化的目标函数，一个是在优化过程中必须满足的约束条件。起初人们使用二次规划求解工具（Quadratic Solver）来求解上述问题，这是一种用于在线性约束下优化具有多个变量的二次目标函数的技术。但是这个求解过程需要强大的计算能力和存储能力的支撑。其实现也十分复杂。

4.2 改进的 HF-SVM 算法

传统 SVM 的实现方法都是基于通用计算机而设计，他的求解过程需要快速的处理器，较强大的高精度浮点运算能力和强大的模型数据存储能力的支持。然而在嵌入式设备的系统设计和智能传感器网络等应用场景中，系统的计算能力，存储能力和能量利用受到硬件水平的严重限制。因此传统的 SVM 算法模型的实现方法并不能直接用于嵌入式设备和传感器网络这样的硬件受限应用中。此外，在嵌入式设备和传感器网络应用中，我们常常需要把我们设计的算法实现由浮点型运算转换成整形运算。甚至为了减少运算造成的能量消耗，在设计算法的时候，研究人员通常会选择尽可能地避免硬件乘法，因为乘法相对于加法运算更消耗资源和能量。

一种在传感器网络中应用 SVM 的方法是，离线进行算法模型的训练，然后将学习到的模型的参数同步到各个节点。为了减少每个节点的运算量和能量消耗，通常会把求得的浮点型的参数通过截断或者舍入的方法转化成整形。当然，这样的近似和舍入必然会引起新的误差，但是这些误差时可以通过统计学的方法估计得到。即便这样，在线性不可分的情况下，每个节点仍要具备计算核函数（如径向基函数）的能力。有一种方法是查表法。预先将一些值计算成表格。每个节点

万方数据

再实际运算时，通过查表近似求解和函数的值。

这几年智能手机得到了快速发展，以高通骁龙 821 处理器为例[4]，他是一款 4 核 64 位处理器，单核速度可达 2.4GHz，下载速度可达到 600Mbps。现在手机的计算能力，计算精度，存储容量已经远远超越了非智能机时代。因此，基于当前的智能手机的硬件配置，我们可以改进传统的 SVM 算法模型，我们提出了一种硬件友好型核函数，然后用 SMO 算法求解最优值问题。

4.2.1 改进的目标函数和决策函数

由 4.1 节，传统的 SVM 模型带求解的优化问题是下面这样的目

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha + \gamma^T\alpha$$

Subject to

$$0 \le \alpha_i \le C \ , i = 0,1,\cdots,N$$
$$y^T\alpha = 0 \tag{22}$$

其中，$\gamma_i = 1, i = 1,\cdots,N, q_{ij} = y_i y_j K(x_i, x_j)$ 。其决策平面（即分割超平面）是 $y(x) = \sum_{i=1}^{N} y_i \alpha_i \ K(x_i, x) \ + b$。

现在我们限定α_i 是定点小数。所有的α_i 都具有形如$0.xxxxxx$的k位定点小数形式。定点小数的位数k我们可以指定。这样的限定可以保证我们求解出来的系数不需要截断或者舍入，并且由于是定点小数，在利用决策平面分类的时候，计算速度要更快，代价更小，能耗更低。

为了实现上述的限定，施加如下变换：

$$\beta_i = \frac{1-2^{-k}}{C} \cdot \alpha_i \tag{23}$$

上式中的k是定点小数的位数。那么目标式子将变成：

$$\min_{\beta} \beta^T Q\beta + s^T\beta$$

Subject to

$$0 \le \beta_i \le 1 - 2^{-k} \tag{24}$$

万方数据

其中$s_i = -\frac{1-2^{-k}}{C} \quad \forall\ i \in [1,2,\cdots,N]$ 。

决策分割平面是:

$$y'(x) = \sum_{i=1}^{N} y_i \beta_i K(x_i\ , x) \tag{25}$$

相比原始的分割超平面，偏置值b被移除，从原始方程中移除偏置值b，可以使我们变化之后的决策函数也没有偏置值。而这样的移除，理论证明是允许的，这是因为我们将要使用的硬件友好型核函数（属于拉普拉斯函数族）满足一些较弱的密度条件。

我们将使用改进的硬件友好型核函数（Hardware-Friendly Kernel, HFK）来实现低维线性不可分数据到高维线性可分数据的映射。

4.2.2 改进的硬件友好型核函数

在4.1节，我们讲到SVM的目标是寻找一个最大间隔的超平面将正负类样本实例数据分隔开。如前所述，对于线性不可分的训练样本，可以通过使用软间隔分类器来解决。但是在很多情况下训练样本数据往往是线性不可分的，也就是说正负类样本没有线性的边界。另一方面，我们知道在低维线性不可分的数据在维空间中可能是可分的。我们可以将低维变换成高维后，利用核函数来通过计算低维空间中的向量积来求解高维空间中的向量积。这种在低维空间求解高维空间向量积的技术有助于解决非线性的SVM。理论上满足Mercer定理的函数都可以用于解决线性不可分问题。

(1)常用的核函数[13]有以下几种:

线性核函数（Linear Kernel）形式如下:

$$k(x,y) = x^T y \tag{26}$$

(2)高斯核函数（Gaussian Kernel），也称径向基函数。对于数据中的噪音有着较好的抗干扰能力，同样，高斯核函数也有了很多的变种，如指数核，拉普拉斯核等。高斯核函数形式如下:

$$k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{27}$$

(3)指数核函数（Exponential Kernel）指数核函数是高斯核函数的变种，它

万方数据

仅仅是向量之间的 L2 距离调整为 L1 距离，这样的改动会对参数的依赖性降低，形式如下：

$$k(x,y) = \exp(-\frac{\|x-y\|}{2\sigma^2}) \tag{28}$$

(4)拉普拉斯核函数（Laplacian Kernel）

$$k(x,y) = exp(-\frac{\|x-y\|}{\sigma}) \tag{29}$$

(5) Sigmoid 核函数来源于神经网络，现在已经大量应用于深度学习，它是阶跃函数

$$k(x,y) = \tanh(\alpha x^T y + c) \tag{30}$$

根据前述公式(27)(29)，我们知道传统的高斯核函数和拉普拉斯核函数均需要做指数运算。目前，基于智能手机或者智能穿戴装备的用户活动识别算法模型的各个参数是离线训练得到的，这有助于减轻智能手机或者智能穿戴设备硬件资源不足带来的限制，同时可以减轻这些智能装备的能量消耗。但是在将模型参数传输到智能手机或者智能穿戴设备以后，所有的识别过程所需要的计算量、存储容量和消耗的电能都依赖于智能手举或者智能穿戴设备。活动的识别过程，主要是将传感器采集到的数据经过各种预处理得到样本，然后利用决策平面(25)，将样本数据带入到式子中，并判断$y'(x)$的符号的正负，若$y'(x)$是正值则说明样本在超平面之上，因而是正类，反之是负类。

然而由于我们的决策超平面中引入了核函数（常用的 SVM 技术中引入的是高斯核函数），尽可能减少核函数的计算所造成的资源和能源的开销很有意义。

虽然现在智能手机或者智能穿戴设备具备了计算诸如$\exp(x)$ 的能力。但是这样的计算无疑要消耗很大的资源。

这里引入一种硬件友好型核函数，其式子如下

$$k(x_i, x_j) = 2^{-\frac{\|x_i-x_j\|^2}{\sigma^2}} \tag{31}$$

这个核函数的收敛性文献[9]已经证明。这个核函数虽然也有指数，但是它是以 2 为底，他的值可以通过移位和加减法等运算得到，因而计算核函数的代价要

万方数据

比高斯核函数代价小得多。文献中使用的是曼哈顿距离（或者说是L1范数），由于在活动识别领域用于模型训练的样本实例数据是离散型的（或者说是连续性的），同时为了保持高斯核函数较好的抗噪性的优点，同时注意到推导最大间隔超平面的过程中使用的是欧几里得距离，在本文中将改用L2范式（或者称为欧几里得距离来计算）。

4.2.3 求解改进的HF-SVM

常用的求解传统SVM的方法很多，使用的最多的是序列最小优化（Sequence Minimal Optimization，SMO）算法求解[22]。

Platt 的SMO算法的思想是将大的优化问题分解成多个小的优化问题来求解。这些小问题往往很容易求解，并且对他们进行顺序求解的结果与将他们作为整体进行求解的结果是完全一致的。在结果完全相同的同时，SMO算法的求解时间会短很多。

SMO 算法的工作原理是：每次循环中选择两个α进行优化处理。一旦找到一对合适的α,那么就增大一个的同时减少另一个。这里所说的“合适”是指两个α必须要符合一定的条件，条件之一就是这两个α必须要在间隔边界之外，而其第二个条件则是这两个α 还没有进行过区间化处理或者不在区间边界上。

4.3 实验结果

4.3.1 对比实验

在上一部分中，我们为SVM引入了硬件友好型核函数。这个核函数有两个参数(k,γ)我们通过网格搜索的方法进行参数的寻优。

我们使用Internet Advertisement，Mushroom这2个数据里来做对比测试。这2个数据集中的缺省值，我们采用两种方法，一种是用均值代替（如Internet Advertisement数据集就是这样预处理），另一种是用出现次数最高的值代替（在Mushroom数据集就是这样预处理）我们首先用scikit-learn 中的SVM包求解。它使用径向基函数（即高斯核函数），和函数中的参数自动选取最优值。得到准确

万方数据

率如下表 4.1。

表 4.1 HF-SVM 与 SVM 对比实验

	SVM(高斯核函数)	HF-SVM
Internet Advertisement	97.41%	96.23%
Mushroom	99.89%	98.94%

表格 4.1 中，我们做了 HF-SVM 和高斯核函数的 SVM 的对比试验。在 Internet Advertisement 数据集上，我们使用了十折交叉验证的方法。每次选取 90%的数据用来训练模型，使用剩余的 10%的数据进行验证模型的分类准确率。为了较好的对比两个算法模型的准确率，我们并没有对这两个数据集做任何特征选取。我们均使用的是原始数据集。从表 4.1 中可以看到硬件友好型支持向量机的准确率比较接近高斯核函数的准确率，同时，考虑到硬件友好核函数在计算式的代价小于高斯核函数，因此硬件友好核函数仍具有一定的使用价值。

4.3.2 活动识别实验与分析

从活动识别数据集中任意选取出两类活动可以构建出一个样本。选定其中一类是正样本，则另一类是负样本。生成样本中选择 80%用于训练，20%用于测试。

下表是由步行和上楼梯组成的样本，该样本从数据集中抽取出来。令步行是正样本（类别是 1），上楼是负样本（类别是-1）。我们首先进行了特征的选择参数如下：从特征集中选取 150 个特征，使用的近邻个数 K=200，$\lambda = 0.001$。提取到的训练集和测试集的信息如下表 4.2 所示。

表 4.2 训练集和测试集信息

	样本个数	特征个数
训练集	2299	150
测试集	967	150

我们用提取到的特征对应的属性列构建新的样本，首先用 sk-learn 工具包的

万方数据

svm 模块训练模型，测得训练准确率，然后用测试数据测得测试准确率。然后我们用 HF-SVM 算法应用于提取到的训练数据来训练模型。此时，HF-SVM 的参数是模型迭代了 40 次，限制支持向量用 8 位定点小数，容忍误差 0.001，硬件友好型核函数的参数是 1.3。HF-SVM 算法学习到的模型的支持向量的个数是 32。然后分别测试训练误差和测试误差。得到下表 4.2（下表中使用的 SVM 的参数如下：）。

表 4.3 步行和上楼两种活动的准确率表

	训练准确率	测试准确率
高斯核函数的 SVM	98.40%	98.53%
改进的 HF-SVM	97.55%	90.29%

下表是模型的训练误差、测试误差、支持向量的个数随限定的定点小数的位数的变化曲线图，由下图可知，位数在 2-8 之间变化时，训练准确率和测试准确率都迅速的增加。大于 8 位置后训练准确率和测试准确率在很小的范围内变动。

万方数据

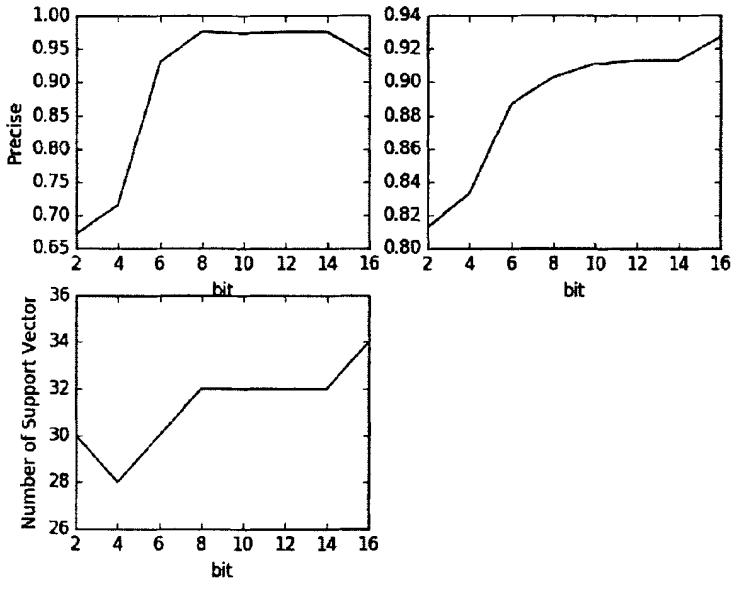

图 4.3 准确率和支持向量个数与位数 k 的关系

由上表 4.3 可知，如果可以单独选取某两类活动的一部分数据训练模型，利用剩余的数据测试模型，可以训练出具有很高测试准确率，泛化能力比较强的学习模型。但是实际上，活动识别问题是个多类问题，我们从中任意选取两类构建成样本训练模型，可以得到 15 个模型（在使用的 6 类活动的数据集的情况下），对于测试样本，并不清楚属于哪一类，尤其是实际活动的分类识别，并不知道样本究竟属于那一个分类器。我们只能组合所有分类器，通过提升所有分类器的性能来得到具有较高分类能力的活动识别算法模型。

4.4 本章总结

本章首先讨论了 SVM 算法模型的求解目标和约束式，然后讨论了 SVM 常用的求解方法。由于数据有线性可分和线性不可分两大类。为了解决线性不可分问题，通常需要引入核函数方法。在人体活动识别领域，由于智能手机等智能穿戴设备受计算能力、存储能力、电池续航能力等的限制，不能进行大规模的计算。为了

万方数据

进一步解决硬件资源的限制，我们提出了硬件友好型核函数。这个核函数与高斯核函数类似，它是以2为底的指数，计算机可以通过移位和加减法来完成指数运算，同时这个核函数也保持了高斯核函数的收敛性。本章的对比试验表明，硬件友好型核函数在计算代价小的同时，具有较好的算法分类精度。在应用活动识别数据进行实验中，我们提出的基于硬件友好型核函数的支持向量机算法保持了较好的训练误差和测试误差。但是由于活动识别问题不是简单地二分类问题。在下一章，我们将讨论如何通过组合各个分类器来得到具有强大分类能力的活动识别分类器。

万方数据

第5章 基于组合分类器的活动识别技术

用户活动识别研究的不是用户是否活动的问题，而是用户在进行什么活动。因此，用户活动识别问题并不是一个简单的二分类问题。文献[23]基于前进和垂直方向加速度信号的小波低频系数的能量特征对上楼、下楼和平地行走这三种行为进行分类。文献[24]对传感器得到的加速度信号进行数理统计分析，发现采集到的上楼、下楼和走路这三种活动相关的加速度信号的波形是可区分的，并提出根据波形的复杂程度来识别活动的技术。Preece 等[27]基于时频域特征对步行、上楼、下楼、贸跑、左脚单脚跳，右脚单脚跳和慢跑这 8 种行为进行分类识别。

5.1 多类别分类

处理多类别问题常用的方法大致可以分为两类：一种方法是针对多类别识别问题直接处理，训练处一种多类别识别模型，优化求解这么问题。另一种是将多分类问题按照一定的策略，转化成若干个二分类问题来处理。这种方法主要一对多（One-against-all，OVA）和一对一（One-against-One，OVO）两种策略[26]。基于智能手机的活动识别问题，受硬件资源的约束，算法模型不能够太过复杂，在实时识别分类的过程中不能因为模型的复杂而引入大量的计算。因此，下面将讨论常用的 OVA 策略和 OVO 策略。

5.1.1 OVA 策略

OVA 策略是 1994 年 Bottou 等[28]人所提出的，主要是将所需要求解的K类别分类问题拆分转化为K个二值分类问题来处理，其中第K个分类器，是以第K个类别的样本为正样本其余的所有样本为负样本构建成训练样本并学习得到的二值分类器。利用 OVA 策略在算法模型训练过程中必须训练K个分类器，每个类别的样本只在对应的分类器中当作正样本，在其他任何分类器中都是负样本。再利用多个模型进行分类测试或者识别的过程中，将测试样本分别送入相应的分类器，然后

万方数据

比较各个分类器的输出值，并选择输出值最高的取值为最终的分类结果。

OVA 策略存在的问题是：

（1）第K个分类器的训练样本是将第K类类别的样本数据做正样本，其余的样本作负样本，这样的处理方式可能会将样本变成非线性不可分的情况，增加了算法复杂度。

（2）在训练第K个分类器时，正样本的数目可能会比负样本的数目少很多，分类器将可能产生较大的误差。

（3）可能存在某些样本被多个分类器分为正类（多个分类器将它分为正类），而有些样本没有被任何分类器分为正类（每一个分类器都将他归为负类）。在这种情况下，分类的准确率将会下降。如下图，阴影区域被三个分割超平面分为正类，但是中间“？”区域没有被任何分类器分为正类。对于中间区域的样本分类的误差就会很大。

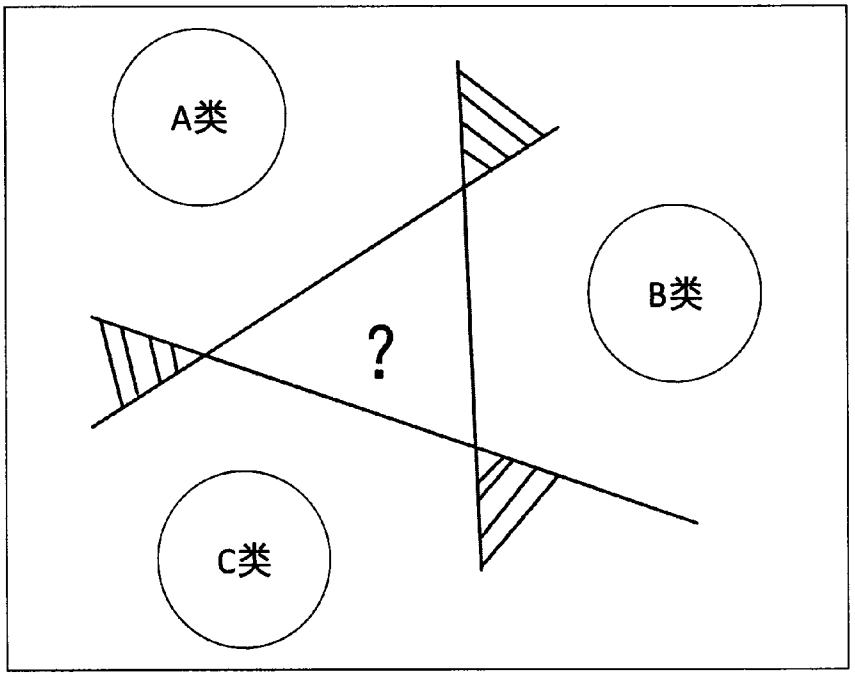

图 5.1 OVA 策略问题示意图

万方数据

5.1.2 OVO策略

OVO策略是Fnerr[28]等人提出的将K类多类别分类问题的策略。OVA策略将K类别多分类问题转化分解为$K(K-1)/2$个二值分类问题来处理,因此必须训练$K(K-1)/2$个二值分类器，然后组合所有分类器完成最终的分类[25]。

在训练模型的过程中,需要从K个类别中任意选取二个类别，将这两个类别对应的样本数据合并在一起构建成一个训练集，其中的一个类别视为正样本，另一个类别的样本视为负样本，然后训练学习得到一个二值分类的算法模型。由于选择具有任意性，因而这样的选择就有$C_K^2 = K(K-1)/2$ 种组合方式，那么就需要将这$K(K-1)/2$个二值分类器组合起来，根据每个分类器的输出值，采用某些策略进行K 类别分类。测试数据和识别样本数据的分类时，将每一个测试数据放入到$K(K-1)/2$二值分类器中，汇总各个二值分类器的分类结果，最后统计得到此样本属于各个类别的概率，将概率最大的那个类别作为预测的分类结果。

OVO策略存在的问题是：

（1）需要训练产生$K(K-1)/2$个分类器模型，这将会增加大量的训练时间和消耗大量的存储空间。

（2）此外，因为分类器数量巨大（$K(K-1)/2$ 个），并且每一个分类器的权重相同，如果用第K个样本进行测试时,会有$\frac{K(K-1)}{2}-(K-1)=\frac{(K-2)(K-1)}{2}$ 个分类器分类错误，这将会使模型产生巨大的误差。

由于OVA策略在正负样本不平衡等方面的问题,我们希望避免这样的问题，因此我们采用OVO策略，同时为了避免OVO策略因为各个分类器权重相同而产生的重大误差,我们这里给出一种基于投票方法和 Sigmoid 函数的分类器组合算法（Sigmoid-Vote ，SV）。SV算法用到的Sigmoid函数去求解虽然已经有一些研究者所使用。这些研究者的做法是，直接使用该函数作用于每个分类器的输出结果，从所有分类器中选择具有最大Sigmoid函数值的类作为最终的输出类。这样的做法虽然很好地避免的各个分类器直接投票时权重相同造成的问题，但是这样的做法容易受到噪声的干扰。可能出现某个噪声数据，他在某个分类器中拥有很大的输出值。但是在其他分类器中输出值却很低的问题。相比，上面的做法，SV

万方数据

算法对于使用 Sigmoid 函数之后的输出进行投票表决。能够很好的避免噪声的干扰。

5.2 SV 分类器组合算法

先前我们说，对于K类多类别分类，使用 OVO 策略将会构建$C_K^2 = \mathrm{K}(\mathrm{K}-1)/2$个分类器。做出最终的预测，我们需要综合各个分类器的输出结果。如果每个分类器的权重相同，如前所述，将会有$\frac{K(K-1)}{2} - (K-1) = \frac{(K-2)(K-1)}{2}$个分类器错误分类而使我们最终的输出预测类产生严重误差。

本节将讨论 SV 分类器组合算法。

我们知道每个 SVM 分类器的分类超平面是$y(x) = \sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b$，当我们把对应的测试样本数据输入到决策平面函数中，决策函数会输出一个连续型的数值。这个数值是分类器的输出结果。

首先引入 Sigmoid 函数：

$$f(x) = \frac{1}{1+e^{-x}} \tag{32}$$

Sigmoid 函数的定义域是$(-\infty, +\infty)$，值域是(0,1)，是如下图形的阶跃函数

万方数据

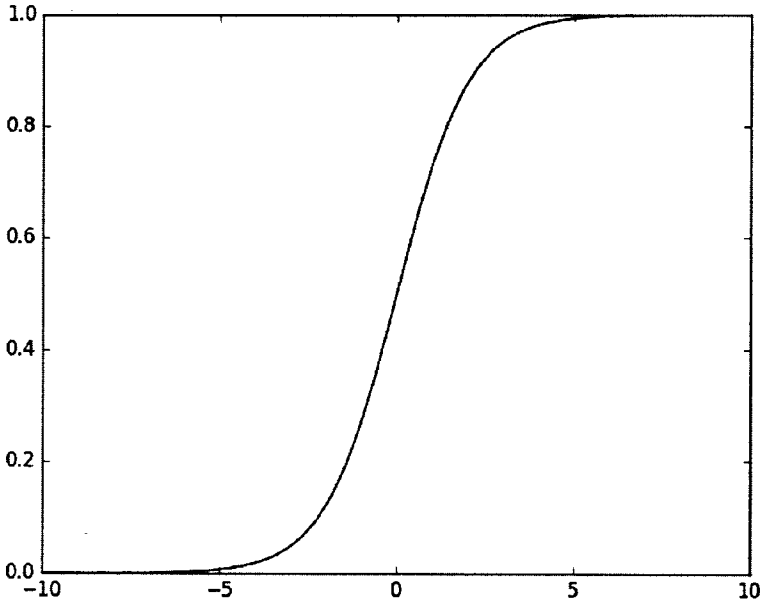

图 5.2 Sigmoid 函数

于是我们定义每个分类器的最终函数$C(x)$

$$C(x)=\begin{cases}1, & f(y(x))\geq t\\ 0, & \text{否则}\end{cases} \tag{33}$$

如果第k 个分类器最终函数$C(x)$为 1，则对第k个分类器产生的所对应的正例样本类投 1 票。最后我们统计最多的票的类，将这个类作为对应样本的最终预测类。如果有至少 2 个类得票相同（大于 0）或者没有一个类别得票大于 0，则比较$f(y(x))$，取值最大的。

5.3 实验结果

基于智能手机传感器的活动识别数据集共有 6 中活动，分别是行走，上楼梯，下楼梯，站起，坐下，侧卧。依据我们使用的 OVO 策略，我们构建了 15 个分类器。在每个分类器中有一个类是正类，另一个是反类，基于这种方法以及我们进行的特征选择，我们构建了 10 个样本，分别对应每个分类器。

万方数据

每一个 SVM 分类器均使用硬件友好型核函数。然后我们用每一个核函数去预测所有测试样本，并将决策函数的输出结果作用于我们的输出函数$C(x)$，组合各个分类器的输出，并投票，就可以得到最终的预测结果。

我们选定 K=8（限定每个支持向量用 8 位表示），组合分类器中的 t 是 0.5，用活动识别生成的测试样本进行试验测试，我们得到以下试验结果表 5.2。

表 5.1　组合分类器（8 位）分类结果

预测 / 真实	行走	上楼梯	下楼梯	站起	坐下	侧卧	召回率
行走	103	2	9	0	0	0	90.35%
上楼梯	1	98	37	0	0	0	72.06%
下楼梯	8	16	119	0	0	0	83.22%
站起	0	2	0	140	0	0	98.59%
坐下	0	0	0	4	108	0	96.43%
侧卧	0	0	0	0	3	139	97.89%
准确率	91.96%	83.05%	72.12%	97.22%	97.30%	100%	90.24%

文献[10]中的准确率召回率如下表 5.3，分析表中数据可以发现应用基于层次聚类特征选取和基于硬件友好型核函数的支持向量机，以及应用组合分类器算法，既减少了模型训练的复杂度，也具有较高的准确率。

万方数据

表 5.2 文献中分类器分类结果

Method	MC-SVM							MC-HF-SVM $k = 8$ bits						
Activity	Walking	Upstairs	Downstairs	Standing	Sitting	Laying	*Recall %*	Walking	Upstairs	Downstairs	Standing	Sitting	Laying	*Recall %*
Walking	**109**	0	5	0	0	0	*95.6*	**109**	2	3	0	0	0	*95.6*
Upstairs	1	**95**	40	0	0	0	*69.8*	1	**98**	37	0	0	0	*72.1*
Downstairs	15	9	**119**	0	0	0	*83.2*	15	14	**114**	0	0	0	*79.7*
Standing	0	5	0	**132**	5	0	*93.0*	0	5	0	**131**	6	0	*92.2*
Sitting	0	0	0	4	**108**	0	*96.4*	0	1	0	3	**108**	0	*96.4*
Laying	0	0	0	0	0	**142**	*100*	0	0	0	0	0	**142**	*100*
Precision %	*87.2*	*87.2*	*72.6*	*97.1*	*95.6*	*100*	***89.3***	*87.2*	*81.7*	*74.0*	*97.8*	*94.7*	*100*	***89.0***

5.4 本章总结

本章我们分析了常用的多类别分类的策略，并分析了 OVA 和 OVO 策略的优劣。基于 OVO 策略，我们提出了组合分类器的算法。通过实验验证，利用基于层次聚类的特征选取算法和基于硬件友好型支持向量机算法和组合分类器算法 SV，既可以有效减少模型训练的复杂度，同时保持较高的准确率，而且我们可以通过限定特征向量表示的位数，这对于其他智能穿戴设备进行人体活动识别提供了较好的支持。

万方数据

第6章 总结与展望

6.1 本文总结

随着各类穿戴智能产品的大量普及和移动互联网技术的深度发展，传统健康养老产业迈入了智能时代，衍生出大量的互联网创新产品与运营模式[1]。可穿戴智能产品带来的是一种新的健康生活方式，通过不断地量化分析从用户的生活中获取到的数据，帮助人们更好地进行保持健康生活状态[29]。

基于图像和视频分析处理的活动识别技术存在依赖外部设备，使用场景首先，用户隐私泄露，易受天气影响等问题。基于移动智能装备的活动识别技术通过智能装备（如智能手机）内置的传感器采集与用户活动相关的数据，用于对活动进行识别，因此避免了对外部设备的依赖，使得技术适用场景范围扩大；另外，由于这些传感器（如加速度计、心率传感器等）提供的数据包含的敏感信息较少且不具备直接的可读性，因此减少了用户隐私泄漏的风险，而种类丰富的传感器类型也为识别用户多样的活动提供了较为充分的信息。

传统的机器学习领域的一些算法在模型的学习和训练过程中需要强大的计算能力和大规模的存储能力。虽然模型的学习和训练过程可以离线去得到，学习到的模型参数在用于活动识别时仍然消耗大量的资源和能量，这在应用较多的智能手机上并不现实。另一方面，研究人员对于在嵌入式设备和传感器网络中应用支持向量机算法的模型进行了研究并提出了一些方法。但是这类算法虽然具有较低的计算代价和能耗，但是其计算精度并不高。

本文在前人研究的基础上，提出了基于层次聚类特征选择算法和硬件友好型支持向量机算法的活动识别算法，针对活动识别因为活动的复杂性导致活动识别是多分类的问题，本文提出了基于 OVO 策略的 SV 算法，来解决 OVO 策略因为各个分类器权重相同，而产生的较大误差的问题。本文改进的算法主要有：

A. 使用基于层次聚类算法的特征选择算法。基于层次聚类的特征选择算法是近几年提出的方法，在该算法活动识别领域尚未使用。与传统的聚类

万方数据

算法相比，该算法聚类的个体是不是样本实例而是特征属性列。然而基于层次聚类算法利用了互信息和关联系数两种评价函数来做为类内距离和类间距离的度量。这样的距离度量决定了算法只能应用于离散型和标称型数据类型的数据，在活动识别领域，能够使用的数据是由智能手机内置的传感器所采集，然后经过去除噪声和切片等预处理得到的。这样的数据是连续型。基于此，我们改进了算法的评价函数。我们使用皮尔逊相关系数来计算候选特征与分类类别之间的距离，使用共享最近邻（Shared Nearest Neighborhood，SNN）来计算候选特征与特征子集的距离。

B. 使用基于硬件友好型核函数的SVM算法。传统的SVM算法模型的训练需要强大的硬件运算能力和存储能力，在模型用于识别活动的过程中，利用决策函数（分割超平面）做分类需要进行大量的指数运算。现在智能手机虽然可以进行e^x这样的指数运算，但是相比2^n只需要一些移位和加减法运算，进行e^x这样的运算需要的计算能力要求更高。因此，本文中改进了硬件友好型核函数，使之既保持高斯核函数抗噪声的优点，也具有较小的计算代价。

C. 针对多类识别问题，提出了基于OVO策略的SV算法。多类别分类问题通常有一对一和一对多两种策略来转化成多个二分类问题。一对多策略往往由于正负样本不平衡数量差距很大而导致分类误差很大。一对一策略训练除的分类器个数比较多，需要组合多个分类器得到最终结果。本文基于一对一策略，提出SV算法，将多个分类器的输出结果作用于Sigmoid函数，然后通过每个分类器的投票得到最终输出。这个方法可以避免单独使用Sigmoid函数取最大值时易受到噪声的干扰而预测出错，也避免了多个分类器直接投票由于分类器权重相同导致的错误。

6.2 展望

基于智能手机传感器的活动识别技术吸引了越来越多的研究人员的注意，随

万方数据

着机器学习，深度学习，深度神经网络等领域的进一步研究，更多鲁棒性的算法被提出。本文受时间、条件等的限制，虽然取得了一些成果，但是要想推动基于智能手机传感器的活动识别技术的进一步发展，还需要更多的科研投入并在以下几个方面进一步改进：

（1）进一步提升算法模型的实时性。现有的基于智能手机传感器的活动识别技术的框架是，在具有强大计算能力的服务器上完成模型的训练，然后将模型的参数传输给用户的智能手机。智能手机完成数据的采集和预处理并用得到的模型参数进行活动识别。未来要进一步提升模型的实时性，这需要进一步研究具有更小计算存储资源要求的算法模型的研究。

（2）进一步研究更加能够用于复杂类别活动识别的多分类模型。过去的十几年，随着智能手机及其内置传感器技术的发展，研究人员已经对越来越多得活动进行了识别。但目前为止活动的种类还比较简单，另一方面随着活动种类的增多，模型的复杂度也会变大。未来要进一步研究能够适应智能手机硬件条件的多分类模型。

万方数据

参考文献

[1] 席恒,任行,翟绍果. 智慧养老:以信息化技术创新养老服务[J]. 老龄科学研究,2014,(07):12-20

[2] 孙泽浩. 基于手机和可穿戴设备的用户活动识别问题研究[D].中国科学技术大学,2016.

[3] 姚毓凯. 支持向量机关键技术及其在人体活动识别中的应用研究[D].兰州大学,2015.

[4] 黄亦凡,张冬阳. 论国产智能手机市场现状及发展[J]. 现代商业,2012,(06):55.

[5] 美国高通公司推出下一代高端骁龙移动处理器[J]. 微电脑世界,2013,(08):16.

[6] 李文洋. 基于智能手机传感器的行为识别算法研究[D]. 硕士学位论文. 2014

[7] Lafferty, J.D, Mc Callum, A. Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. Proceedings of the Eighteenth International Conference on Machine Learning, 2001,282-289.

[8] Davide Anguita, Alessandro Ghio, Stefano Pischiutta 等. A support vector machine with integer parameters[J]. Neurocomputing .2008,480–489

[9] Davide Anguita, Alessandro Ghio, Stefano Pischiutta 等. A Hardware–friendly Support Vector Machine for Embedded Automotive Applications [J]. International Joint Conference on Neural Networks .2007,1360-1364

[10] Davide Anguita, Alessandro Ghio, Luca Oneto 等. Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine[J]. Springer Berlin Heidelberg.2012

[11] 刘华文.基于信息熵的特征选择算法研究[D].博士学位论文.2010.

[12] Jorge Luis Reyes Ortiz.Smartphone-based Human Activity Recognition[C]. Springer International Publishing.79-91,2015

[13] 陈恺.基于移动智能终端的高速公路异常驾驶行为检测技术[D].硕士学位论文.2015.

[14] 李航.统计学习方法[M].北京:清华大学出版社.2012

万方数据

[15] Tom M.Mitchell. 机器学习[M].北京:机械工业出版.2015.

[16] Kira K, Rendell L. A practical approach to feature selection[C].Proc of the 9th International Conference on Machine Learning, 1992, 249-256.

[17] Kononenko I. Estimation attributes: analysis and extensions of RELIEF [C].Proc of the European Conference on Machine Learning, Catania, Italy, 1994, 171-182.

[18] Dash M, Liu H. Consistency-based search in feature selection[J]. Artificial Intelligence, 2003, 151(1-2): 155-176.

[19] Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy [J]. Journal of Machine Learning Research, 2004, 5: 1205-1224.

[20] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.

[21] 唐发明. 基于统计学习理论的支持向量机算法研究[D].博士学位论文.2005.

[22] 杨杰明.文本分类中文本表示模型和特征选择算法研究[D].博士学位论文.2013.

[23] Peter Harrington.机器学习实战[M].北京:人民邮电出版社.2015.

[24] Sekine M, Tamura T, Togawa T 等.Classification of waist-acceleration signals in a continuous walking record[J]. Medical Engineering&Physics,2000 22(4):285-291.

[25] Sekine M, Tamura T, Akay M 等. Discrimination of Walking Patterns Using Wavelet-Based Fractal Analysis[J]. IEEE transactions in neural systems and rehabilitation engineering.2002,10(3):188-196.

[26] 张文博.多类别智能分类器方法研究[D].博士学位论文.2014。

[27] 强琦.基于统计学习的多类别分类器研究[D].硕士学位论文.2006

[28] Preece S.J., Goulermas J.Y., Kenney L.P.J.等. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data[C]. IEEE transaction on Biomedical Engineering.2009,56:871-879.

[29] C.W.Hsu, C.J.Lin. A Comparison of Methods for Multiclass Support Vector Machines[C].IEEE transaction on Neural Networks.2002,415-425

万方数据

[30] Anthony D.Joseph. Activity-Based Computing[C].IEEE CS.1536-1528,2008

[31] 白琳.基于深度学习机制的人与物体交互活动识别技术[D].博士学位论文.2015

[32] 范昕炜.支持向量机算法的研究与应用[D].博士学位论文.2003.

[33] 程佩青.数字信号处理教程[M].北京:清华大学出版社.2013

万方数据

攻读硕士学位期间主要的研究成果

[1] 参与国家科技支撑计划项目《文物数字化保护标准体系及关键标准研究与示范》中的课题二“不可移动文物数字化保护关键标准研究与示范（以石窟寺为例）”，主要参与了不可移动文物数字化保护（以石窟寺为例）传输标准和交换标准的调研和研究工作，参加起草了相关标准的草案。

万方数据

致谢

时光飞逝，转眼间在浙江大学的两年半研究生生涯也即将到达尾声。在这将近三年的学习生活中，我开拓了视野，增长了知识和技能，同时也收获了友谊。在此论文即将完稿之际，我想对所有曾在我生活和学习中给予我帮助的导师、实验室老师、父母和同学表达我衷心的感谢。

首先我想要感谢我的指导老师邢卫副教授。在浙江大学这两年半的研究生生涯中，邢卫老师在学习和生活上给予了我极大的帮助和指导。研究生期间，邢卫老师帮助我制定了学习计划并指导我进行文献阅读，使我得到了成长。同时也是在邢卫老师的指导下，我顺利完成了硕士毕业论文的工作，为两年半的硕士生涯画上一个完美的句点，感谢这三年有邢卫导师的陪伴！

其次，我要感谢网络与媒体实验室给我提供了学习和科研的机会，感谢实验室在我研究生三年期间，组织的学习交流，学术分享，素质拓展等活动。这些活动让我受益匪浅，使我融入实验室团体，借鉴别人的经验快速适应科研生活。感谢实验室的鲁东明，许端清等老师，没有你们的指导和帮助，就没有这么融洽的网络与媒体大家庭。

再次，我要感谢我的父母，感谢你们的辛勤付出、支持和包容。是你们的无私的爱和付出支撑我走到今天。你们给予我的关心和帮助是我一直前进的动力！

最后，我要感谢实验室的同学和师弟师妹们。感谢李鹏飞、韩佳楠、黄澎江、林晓斌、林炀平、李艳蓉、潘海宽、祝凯林等同学，和你们一起度过的时光很开心。感谢李家豪、袁义军、吴奇轩等师弟师妹们，感谢你们为实验室带来的欢乐。

付浩

2017 年 1 月 1 日

万方数据