电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

（电子科技大学图标）

论文题目 改进的 K-近邻模式分类

学科专业 模式识别与智能系统

学　　号 **201221070545**

作者姓名 梁 洲

指导教师 朱 宏　教 授

分类号 _____ 密级 _____

UDC [注1] _____

学 位 论 文

改进的 K-近邻模式分类

（题名和副题名）

梁 洲

（作者姓名）

指导教师 朱 宏 教 授

电子科技大学 成 都

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 模式识别与智能系统

提交论文日期 2015.05.12 论文答辩日期 2015. 05.18

学位授予单位和日期 电子科技大学 2015 年 6 月 28 日

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

IMPROVED K- NEAREST NEIGHBOR CLASSIFICATION

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Major: **Pattern Recognition and Intelligent System**

Author: **Liang Zhou**

Advisor: **Prof. Zhu Hong**

School : **School of Automation Engineering**

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，也不包含为获得电子科技大学或其它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。

作者签名：_____ 日期： 年 月 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、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。本人授权电子科技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应遵守此规定）

作者签名：_____ 导师签名：_____

日期： 年 月 日

摘　要

在模式识别、机器学习、数据挖掘等领域中，*k*-近邻分类由于其直观、简单等特点有着广泛的研究和应用背景。其应用范围已包括生物信息认证、图像的分类、人脸识别等领域。

本文是在分析*k*-近邻算法的基础上，对*k*-近邻分类算法的改进进行了研究。主要研究工作如下：

(1)提出基于类内近邻距离加权的改进伪近邻分类算法。考虑测试样本的多个近邻，距离近的对归属该类的影响较大，因而拥有的权值较大。对每个类别，测试样本得到在该类经过距离加权的伪近邻，再运用近邻分类算法进行分类，从而提高分类精度。

(2)提出了基于类均值的最近邻分类算法。在分类过程中，分类精度容易受到离群点的影响。采用基于类均值的最近邻分类，利用每类样本的类均值信息，降低了离群点对分类精度的影响。

(3)提出了基于局部均值的近邻分类算法。在采用近邻分类的过程中，训练样本的数量比较少从而导致分类的准确率降低。为了提高近邻分类的分类性能，利用测试样本在每类训练样本集的*k*个近邻的均值信息，从而提高在小样本下的分类精度，同时防止训练时过拟合。

关键词：模式分类，*K*-近邻准则， 局部均值，类均值

ABSTRACT

K- nearest neighbor classification is widely used in pattern recognition , data mining ,machine learning and other fields for its intuitive and simple. Its application has been involved in bioinformatics certification, classification images, face recognition and other fields.

This article is based on k-nearest neighbor analysis to improve k- nearest neighbor classification algorithms. There are three main works as follows:

This paper is mainly based on three aspects about k- neighbor classification:

1. The neighbor classification based on distance weighted intra-class neighborhood was proposed. This classification method make use of the neighboring samples of each class and the near sample have more effect on the class, thus the sample have greater weighted distance. Then this method use the neighbor classification rule to improve the classification accuracy.

2. The nearest neighbor classification based on class mean was proposed. In the process of classification the accuracy of nearest neighbor classifications is easily influenced by outliers. The nearest neighbor classification based on class means make full use of the class means information and reduce the impact of outliers on the classification accuracy.

3. The nearest neighbor classification based on local mean was proposed. The small number of training samples result to low classification accuracy in the process of nearest neighbor classification. In order to improve the performance of nearest neighbor classification, this classification method make full use of local mean information of test samples and improve the classification accuracy in a small sample, while preventing over-fitting.

Keywords： Pattern classification, K-nearest neighbor classification, Class means, Local means

目 录

第一章 绪论

1.1 研究背景与意义

近几十年来，随着信息技术的迅猛发展和互联网的高速普及使得人们获得和存储数据的能力得到逐步的提高。面对不断增长的数据，传统的数据分析方法已经不能满足人们对数据信息的要求。为了解决数据和信息之间的不对称，需要一种能够使用智能的分析方法从海量数据中提取有价值的信息技术。面对着大数据时代信息知识的挑战，因此诞生了数据挖掘技术。数据挖掘技术是一门多学科的交叉领域。它涉及机器学习，大数据处理、模式分类和数据存储等多个领域的理论和技术。

模式分类是数据挖掘的重要研究方向，也是数据挖掘的重要任务之一。在模式识别领域中机器学习方法是一个研究的热点。常用的机器学习方法有 k-近邻分类、k-均值聚类等，这些算法的性能优劣主要取决于样本数据之间的相似度。在 k-近邻模式分类中，由于样本数据集的类别分布、样本特征之间关系，采用传统的距离度量使得计算不精确，因此降低了分类精度。而且当在人脸识别等高维的样本特征空间分类的时候，采用 k-近邻分类的精度严重下降，导致计算的时间复杂度与空间复杂度变大。因此在实际的模式分类过程中，采用合理的距离度量可以提高分类精度。

在模式识别的实际应用领域中，待分类数据样本集的概率密度是不可知的。并且大多数数据样本集的概率密度也与实际应用的数据样本形式不一致，因此非参数估计在实际的分类过程中有着广泛的应用。近邻分类方法属于一种典型的非参数分类方法，其具有较高的分类性能，因此广泛应用在信息检索与分类、DNA 序列预测、人脸识别、生物信息认证、图像分类、车牌检测和人工智能等诸多领域[1]。

在实际应用中，常用的模式识别系统的组成如下图 1-1 所示。模式识别系统通常又分为两大部分：一部分完成分类系统的训练工作。即通过数据样本确定分类器的各项参数，从而完成分类器的设计任务。另一部分实现对测试样本点的分类。在模式识别过程中分类决策是对待分类数据样本进行决策的过程。下图 1-1 就是常用的模式识别系统框图。

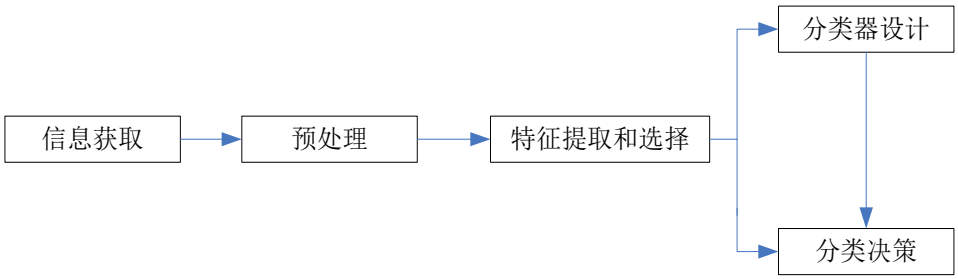


图1-1 模式识别系统

1.数据获取

在数据获取阶段，采集的数据信号需要传输给计算机。为了使采集到的数据转化为计算机能够识别的数据信息在传输给计算机之前需要数据转换。通常采集的输入信号有下面 3 种类型：

(1) 二维图像 比如指纹、地图这一系列的对象；

(2) 一维波形 比如正弦波形等实验数据等；

(3) 逻辑值 对某参量是与否的判断。比如事件的发生与否，可以用逻辑值 0 和 1 来表示。通过对数据的采集以及前期的数据处理，可以通过采用矩阵或者是向量表示二维图像或一维波形。

2.预处理

进行预处理不但能降低噪声信号对测试的影响而且又能对原始信号加强。能对仪器造成数据采集测量或者是因其他不可避免的因素所造成的退化现象进行复原。由于原数据可能混有大量的噪声，因此在进行数据处理之前去噪声是有必要的。由于采集的原始数据量很大，维数很高，因此使得计算过程中的复杂度很高，因此先对数据样本集的维数进行预处理。

3. 特征提取和选择

通过对原始样本数据进行变换，能够得到表现模式分类的重要特征。测量空间是由原始的数据组成的。在模式分类过程中进行的空间称为特征空间。在特征空间中的样本通常以向量的形式表示，一般在特征空间中表示为一个点。特征提取过程中通常提取属于同一个分类中的不同的样本，并且特征值非常相近。属于同一类别的样本的特征值具有相似性，而不同类的样本的特征值则有较大的差异性。

4.分类决策

分类决策指的是利用统计方法将特征空间中分类的对象进行分类。通过从样本训练集中得到一个分类的决策标准，使用这种标准进行分类决策时的分类准确率更高。

5.分类器设计

在模式分类系统中分类器的作用是：用提取到的特征信息对每一个等待分类的对象进行标记。

1.2 研究现状

k-近邻规则分类方法最早是由 Fix 和 Hodges 在 20 世纪 50 年代首次提出[2]，自此以后 *k* -近邻分类规则就在模式识别等领域得到了深入的研究。在 *k*-近邻规则分类方法提出之后，Cover 和 Hart 发表了《Nearest Neighbor Pattern Classification》，从而在理论上证明了 *k*-近邻分类的误差率近似接近最优的贝叶斯分类误差率[3] 。*k*-近邻分类以其简单、直观的优点，大量的研究者投身其中，从而使得 *k*-近邻分类的相关研究取得了重大的突破。自从《Nearest Neighbor Pattern Classification》发表后，吸引了大量的学者从事 *k* -近邻分类规则的理论研究。关于 *k* -近邻分类规则的研究方向主要分布在以下几个领域中：小样本下的分类精度、分类距离度量的选择、高维空间下数据的降维研究、*k* 值的选取以及稀疏表示在近邻分类规则的实际应用等[4]。

在样本集的数量比较小的时候，*k*-近邻规则分类的误差率比较高。Wagne 在对分类过程中误差率上做了大量的工作并取得了很大的成果。并且，Wagne 对剪辑近邻分类法的敛散性方面进行了研究[5]。Devroye 等人在对近邻分类的改进展开了深入的研究。Devroye 所著的《A Probabilistic Theory of Pattern Recognition》书中，对经典贝叶斯分类规则、*k*-近邻算法等算法进行了详细的论述。并对各类算法的敛散性、误差率、收敛率、距离度量的选择和维数的降低等方面进行了详细的阐述[6]。

在使用近邻规则分类的过程中，计算复杂度是影响分类结果的一个重要因素。复杂度分为空间复杂度和时间复杂度[7]。剪辑近邻法与压缩近邻法是较早提出的快速算法之一。Wai Lam 等提出了对数据样本集的进行优化和样本重要信息提取，并通过集成概念原型方法减少了计算后的存储容量大小[8]。McNames 提出了采用 PCA 分析方法来构建分类搜索树。在每一次搜索过程中，在树的节点上朝着最大方差方向搜索，从而加快算法的收敛速度[9]。

近邻分类算法当样本的数量趋近于无穷或者样本的数目非常大的时候，这时分类的精度非常高。当样本数据的数量比较小的时候，融合距离估计的 *k*-近邻分类算法的分类准确率要高于基于自适应增强(AdBoost)的分类器[10]。融合之后的分类算法的分类准确率也高于使用其他距离量的分类算法，而且同时还降低了维数灾对分类的影响[11]。与此同时，Hamamot[12]等人也把通过把自助法技术和近邻分相结合，与传统的自助采样技术不同。Hamamot 提出通过对训练样本的局部信息组合后进行自助的采样，避免了当训练样本的数目比较小时导致分类的精度不高的问题。

在采用近邻分类算法进行分类时，距离度量的选择对分类结果的影响也特别重要。在模式分类过程中，距离度量的选择会影响到分类的准确率。Jacobs 等人提出选择向量的相关性来进行度量而非传统的如欧式距离等来度量。并且提出了采用非典型点来代替边界点来捕捉类的结构，并且取得了较好的分类效果[13]。Ricci 和 Avesani 提出了在近邻分类过程中采用局部的距离度量。该方法的思想首先是从样本数据集中提取出有用的特征信息，然后通过融合学习算法从而提高分类的准确率 [14]。

自从最近邻分类规则的方法提出之后，大量的学者把研究方向集中在了近邻分类规则的改进上。*k*-近邻算法其实质是最近邻算法的扩展，对测试点距离最近的 *k* 个近邻的类别进行判断[15]。Dudan 提出了基于距离加权的近邻分类算法，在该算法中加入了不同距离度量的近邻对分类的结果产生不同的分类效果[16]。*k*-近邻分类算法在实际中已得到了广泛的应用，例如采用近邻搜索方法对海量数据的处理[17]，采用融合近邻分类技术进行人脸识别[18,19]。利用 *k*-近邻规则对语音信息进行分类[20]，利用 *k*-近邻分类算法进行地理信息系统分类[21]，以及对移动的飞机目标进行自动分类识别[22]等。

以上对最近邻分类规则及其派生分类规则以及对应分类规则构成要素的研究。总的说来，都属于 *k*-近邻分类规则的研究范畴，因而不管怎么改进，仍不能摆脱 *k*-近邻分类规则本身所特有的缺点。虽然距离加权的 *k*-近邻分类规则融合了距离权值信息，但对经过加权处理的 *k* 个近邻，最后也是用 *k* 近邻分类规则进行分类。取测试样本点 *x* 的 *k* 个近邻，通过对这 *k* 个近邻中大多数属于哪一类，就把未知样本 *x* 归为哪一类。从前面的分析可知，*k* 近邻分类规则有其局限性，在研究中发现一个新的分类规则，其借鉴 *k* 近邻分类规则，本质上但又有很多差异，而规则的分类性能却好于 *k*-近邻分类规则。进行的改进的 *k*-近邻分类规则的研究正是这方面的探索。

1.3 本文主要研究工作

对改进的 *k*-近邻分类规则进行研究时，其最大的技术难点是把 *k*-近邻分类规则的渐进分类误差概率表达为一个解析式，因为只有把规则的渐进误差概率表达为解析式，才能与贝叶斯分类规则[23]的分类误差率进行比较。最后与贝叶斯分类规则的分类误差率对比，对改进的 *k*-近邻分类算法进行评价，并分析影响分类误差率的因素。

本文的主要内容是在 *k*-近邻分类的基础上，围绕改进 *k*-近邻分类的方法展开研究。在论文中要达到的目标是：基于概率密度估计的非参数方法，进行一个新的模式分类规则的研究。寻找一个分类性能优于最近邻分类规则、*k*-近邻分类规则和距离加权的 *k*-近邻分类规则的分类规则，并以最优贝叶斯分类器为参照对象，运用概率理论，分析所提出分类规则的分类性能；以机器学习库的标准数据集为测试对象，通过计算机仿真来验证所提出分类规则的分类性能。

正如前面在国内外研究现状分析中所介绍的那样，人们对近邻分类规则及派生分类规则、*k*-近邻分类规则和距离加权的 *k*-近邻分类规则进行了多方面的研究，其许多理论和分析方法也可借鉴用于所进行的改进的 *k*-近邻分类规则的研究。

在 *k*-近邻分类存在的这些问题中，本文主要围绕 *k*-近邻分类的三个方面：如何提高 *k*-近邻算法的分类精度；怎样采用一个合适的距离来度量 *k*；如何降低 *k*-近邻分类的在计算时的复杂度；怎么提高分类过程中的搜索速度等方面展开本文的内容。

(1)对数据样本的相似性进行了阐述。改进的 *k*-近邻分类算法就利用了这样的特性。在对改进的最近邻分类过程中，离测试点近的点所具有的权值就更大。扩展的最近邻分类算法一般适用于在样本数目特别大的时候，这时近邻分类的性能能够得到显著的提升。

(2)样本数据集受到离群点的影响使得分类的准确率下降。采用类均值改进的 *k*-近邻分类可以提高分类准确率。因此在对基于类均值的扩展最近邻分类，融合了分类样本的平均值信息，从而降低了离群点对分类精度的影响。并且防止了训练数据的过拟合，不但提高了近邻分类的精度而且在样本的分布不均匀的情况下也有很好的分类性能。

(3) 在采用近邻分类的过程中，训练样本的数量比较少，那么分类的准确率将会降低。在样本数量不足的时候，为了提高的分类的准确率采用了基于局部均值的 *k*-近邻分类。采用这类改进分类方法不但能降低离群点对近邻分类精度的影响，而且还可以防止训练数据集的过拟合，显著地提高分类器的分类性能。

本文的在研究改进的 *k*-近邻分类规则，其主要特色与创新之处是：

(1) 它利用了测试样本的近邻信息，并且融合了不同距离的近邻点对分类准确率的影响；

(2) 它是对同类样本中的 *k* 个近邻进行加权，而不是对测试样本的个近邻大多数情况下不同类进行加权，因而与只有个近邻都为同类的可作拒绝的近邻分类规则相比，它避免了拒绝类的存在；

1.4 本文的组织结构

本论文在原有*k*-近邻模式分类的研究的基础上，通过对现有算法的缺点和不足提出了一种基于类均值和局部均值的改进的*k*-近邻分类方法。论文的主要内容分成六个章节，主要结构安排如下：

第一章：绪论。简要阐述了*k*-近邻模式分类算法的研究意义以及背景。介绍了*k*-近邻模式分类算法在国内外的发展以及现况。针对*k*-近邻模式分类的不足，提出了改进的*k*-近邻模式分类算法。

第二章：本章简要介绍了最近邻算法和*k*-近邻算法并且分别对它们的分类误差率分析。从而由最近邻算法引申出*k*-近邻算法，并对*k*-近邻算法的误差率进行了阐述。

第三章：本章首先对近邻分类过程中的距离度量进行了介绍。然后对 *k*-近邻分类算法的特征加权和加权投票进行了研究，采用了基于距离加权的伪近邻分类算法。通过 UCI 数据集仿真测试并与传统的 *k*-近邻分类算法的分类准确率进行了比对。

第四章：本章在原有*k*-近邻模式分类的研究的基础上，通过对现有算法的缺点分析提出了一种基于类均值的改进的*k*-近邻分类方法。同时利用类均值的改进方法来实现距离度量，用自适应的方法来确定类均值的权值。并利用UCI数据库做出数据仿真实验验证分析进行比对。

第五章：采用基于局部均值的 *k*-近邻分类主要融合了局部平均值信息对测试样本点进行分类，进而提高模式分类的准确率。通过结合 BP 神经网络采用自适应的方法自动寻找最优的权值。改进后的算法融合了样本数据集的整体信息，从而使得改进的最近邻分类的分类准确率要高于传统的近邻分类。

第六章：是论文的总结和展望章节，通过前面几章的介绍，对前面的研究加以总结，通过基于改进的*k*-近邻算法—基于局部均值的*k*-近邻算法和类均值的近邻算法改进了近邻分类的分类性能。这类方法不但利用每类样本数据点的局部信息

而且又利用了每一类样本数据点的全局信息。这种分类方法主要在于选取适当的近邻参数和权值参数，在实际的应用中很容易实现。

第二章 近邻算法概述

k-近邻分类在模式识别、机器学习和数据挖掘领域中属于一种非常简单、有效的分类方法[24]。近邻分类是一种基于实例的分类算法。在模式分类系统中常用的分类方法有：基于决策树分类法、SVM算法和经典的贝叶斯分类法等[25,26]。近邻分类与上述分类算法相比显著的特点就是一种基于实例的懒惰学习算法。下面将分别阐述最近邻分类和 k -近邻分类算法。

2.1 最邻算法

最近邻分类算法在数据挖掘分类算法中属于比较经典算法之一。它的主要思想就是通过找到测试样本点距离最近的那个类别，测试样本就属于那一类分类样本。如果有 n 个类别 $w_1, w_2, ..., w_n$ 的分类问题，每类别的样本 N_i 个，那么规定 w_n 类的判定函数为

$$g_i(x) = \min_k \left\| x - x_i^j \right\| \qquad k = 1,\ 2,\ N_i \tag{2-1}$$

其中 x_i^k 的角标 i 表示 w_i 类， N_i 个样本的第 k 个。

按照 2-1 式，

$$g_j(x) = \min_i g_i(x) \qquad x \in w_j \tag{2-2}$$

因此最近邻算法另外的一种表达就是：令 $S^n = \{x_1, ..., x_n\}$ 。距离测试点 x 在集合 S^n 中最近的点称为 x' 。那么根据最近邻分类算法就把测试样本点 x 分到了所属于的类别。

假设 N 个样本下的平均误差率为 $P_N(e)$ 。并且样本 x 的最近邻为 x' ，那么此时的平均误差率

$$P_N(e) = \iint P_N(e \mid x, x') p(x' \mid x) dx' p(x) dx \tag{2-3}$$

此时定义最近邻的渐近平均误差率为 P。当 $N \to \infty$ 时， $P_N(e)$ 的极限，记作

$$P = \lim_{N \to \infty} P_N(e) \tag{2-4}$$

此时存在以下关系式

$$P^* \le P \le P^*(2 - \frac{c}{c-1} P^*) \tag{2-5}$$

P^* 表示贝叶斯误差率， c 为分类的类别数。在分析最近邻分类算法的误差率之

前，先讨论贝叶斯分类规则的误差率。定义 $w_m(x)$ 为

$$P(w_m \mid x) = \max_i P(w_i \mid x) \tag{2-6}$$

则有贝叶斯条件错误率

$$P^*(e \mid x) = 1 - P(w_m \mid x) \tag{2-7}$$

因此贝叶斯错误率为

$$P^* = \int P^*(e \mid x) p(x) dx \tag{2-8}$$

2.2 *k*-近邻算法

k-近邻分类本质上是最近邻分类的一种扩展。现在期望利用 *k*-近邻分类规则对测试数据点 *x* 完成正确的分类。在测试样本点 *x* 附近找 *k* 个近邻点，就将测试样本点分为 *k* 个近邻点所含最多的那一类别。*N* 个训练的样本中，N_1 个属于类别 ω_1，N_2 个属于类别 ω_2 类，…，有 N_c 个属于类别 ω_c。如果 $k_1, k_2, \cdots, k_c$ 分别属于 $\omega_1, \omega_2, \cdots, \omega_c$ 类，那么可以将判别函数定义为

$$g_i(x) = k_i, i = 1, 2, ..., c$$

其决策规则为：若

$$g_j(x) = \max_i k_i$$

则决策 $x \in w_j$。

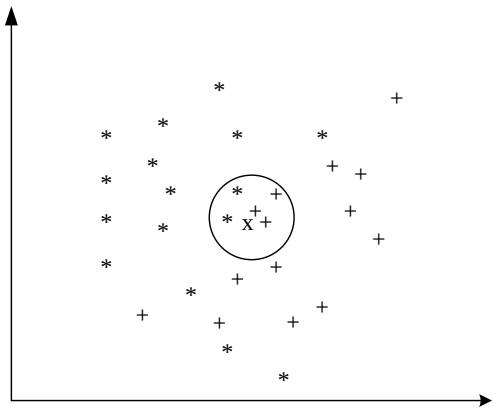


图 2-1 *k* 近邻示意图

如图 2-1 如示，图中 *x* 为测试样本点，分别有两类已知类别信息。从图中可以得到两类样本的分布情况。在使用 *k*-近邻分类算法进行分类时，当 $k = 5$ 的时候，根据 *k* -近邻分类的规则，测试样本点包含有 3 个正号类，2 个星号类，因此测试样

本点 x 被分为正号类。

在 k-近邻分类算法中，如果 $k=1$，那么 k-近邻分类算法就退化成了最近邻分类算法。下图 2-2 分给出了当 $k=1, k=2$ 和 $k=3$ 时，采用 k-近邻分类方法的分类结果。测试样本点根据距离最近的类别标号进行分类。在图 2-2(a)中，离测试样本点距离最近的是一个负数类，因此测试样本被分到负数类中。在图 (b)中，测试点的包含两个近邻点，分别为 1 个正号类和 1 个负号类。这时使用 k-近邻分类不能准确地将测试样本分类。在图 2-2(c)中，测试点的包含 3 个近邻点,分别为 2 个正号类和 1 个负号类。根据近邻分类原则，该点被分为正号类。

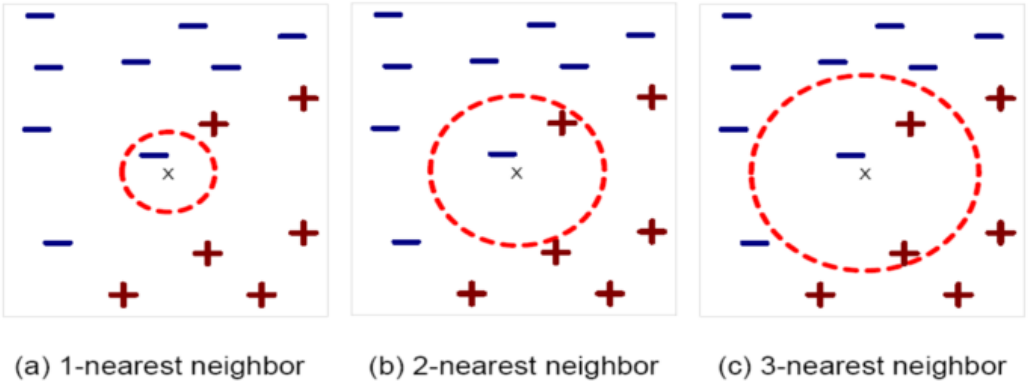


图 2-2 测试样本的最近邻、2-近邻和 3-近邻

由上一小节最近邻的条件误差率：

$$P_N(e|x,x') = P(\omega_1|x)P(\omega_2|x') + P(\omega_2|x)P(\omega_1|x') \tag{2-9}$$

当 $N \to \infty$ 时，x 与最近邻 x' 非常接近，可得到

$$P(\omega_i|x') \doteq P(\omega_i|x)$$

将其代入上式(2-9)可以得到：

$$P_{N\to\infty}(e|x,x') = P(\omega_1|x)P(\omega_2|x') + P(\omega_2|x)P(\omega_1|x) \tag{2-10}$$

将上面的式子推广到 k 近邻算法。假设 x 属于 ω_1，而 k_1 小于等于 $\frac{k-1}{2}$，则发生这样事件的概率为

$$\sum_{j=0}^{(k-1)/2} C_k^j P(\omega_1|x)^j P(\omega_2|x)^{k-j} \tag{2-11}$$

式中 $C_k^j = \frac{k!}{j!(k-j)!}$。

同理可以得到当 x 属于 ω_2 时的情况。所以在给定的 x 的条件误差率为

$$P_{N\to\infty}^{k}(e\mid x)=P(\omega_1\mid x)\sum_{j=0}^{(k-1)/2}C_k^jP(\omega_1\mid x)P(\omega_2\mid x)^{k-j}$$
$$+P(\omega_2\mid x)\sum_{j=0}^{(k-1)/2}C_k^jP(\omega_2\mid x)P(\omega_2\mid x)^{k-j} \tag{2-12}$$

其中第一项和第二项分别为对于 $x\in\omega_1$ 和 $x\in\omega_2$ 的条件误差率。因此上面的子可以改写为一般形式

$$P_{N\to\infty}^{k}(e\mid x)=P(\omega_1\mid x)\sum_{j=0}^{(k-1)/2}C_k^jP(\omega_1\mid x)P(\omega_2\mid x)^{k-j}$$
$$+[1-P(\omega_i\mid x)]\sum_{j=(k-1)/2}^{k}C_k^jP(\omega_i\mid x)^j[1-P(\omega_i\mid x)]^{k-j}\quad i=1,2,... \tag{2-13}$$

在这样的情形下，贝叶斯的条件误差概率为

$$P^*(e\mid x)=\min[P(\omega_1\mid x),P(\omega_2\mid x)]=\min[P(\omega_1\mid x),1-P(\omega_1\mid x)] \tag{2-14}$$

由此可以得给定的 x 的条件误差率为

$$P_{N\to\infty}^{k}(e\mid x)=P^*(e\mid x)\sum_{j=0}^{(k-1)/2}C_k^jP^*(e\mid x)^j[1-P^*(e\mid x)]^{k-j}$$
$$+[1-P^*(e\mid x)]\sum_{j=0}^{(k-1)/2}C_k^jP^*(e\mid x)^j[1-P*(e\mid x)]^{k-j} \tag{2-15}$$

为了得到渐近平均误差率 P_k，可以对所有 x 求期望。定义贝叶斯条件误差率 $P^*(e\mid x)$ 的函数 $C_k[P^*(e\mid x)]$ 为大于 $P_{N\to\infty}^{k}(e\mid x)$ 的最小凹函数，那么也就是说对所有的 x 有

$$P_{N\to\infty}^{k}(e\mid x)\le C_k[P^*(e\mid x)]$$

同时，根据 Jenson 不等式

$$P=E[P_{N\to\infty}^{k}(e\mid x)]\le E\left\{C_k[P^*(e\mid x)]\right\}\le C_k\left\{E[P^*(e\mid x)]\right\} \tag{2-16}$$

当 k 值增加时，$P_{N\to\infty}^{k}(e\mid x)$ 的值减小。函数 C_k 随 k 减小而减小。由此可以推出

$$P^*\le P\le C_k(P^*)\le C_{k-1}(P^*)\le\cdots\le C_1(P^*)\le 2P^*(1-P^*) \tag{2-17}$$

其中最后一项是两类分类近邻分类算法的错误率的上限，也就是 c=2 的情况。当 k 的值不断增加并且趋于无穷的情况下，k-近邻算法就成了最优算法。k-近邻算法的分类准确率提高的前提是需要样本数量 N 无限大的时候才可以实现。因此，在使用 k-近邻算法分类的时候，选择较大的 k 值可以提高分类的准确率。同时测试样本的 k 个近邻点都得非常靠近测试点 x，这样才可以保证 $P(\omega_i\mid x')$ 和 $P(\omega_i\mid x)$ 近似相等。因此在使用 k-近邻算法进行分类的过程中，对于近邻点的 k 值的选择应该综合考虑。

不管是近邻算法还是 *k*-近邻算法，都拥有简单，易实现的特点。根据上面的讨论，可以得到其误差率

$$P^* \le P \le P^*\left(2-\frac{c}{c-1}P^*\right) \tag{2-18}$$

因为 P^* 常规来说是非常小的，假如把上面的表达式中第二项忽略掉，那么上面的表达式可以简化成：

$$P^* \le P \le 2P^* \tag{2-19}$$

由此可以看出，近邻分类的误差是介于贝叶斯的误差率 P^* 和贝叶斯误差率的两倍 $2P^*$ 之间。由于 *k*-近邻分类算法简单，易实现等特点，因此广泛应用在信息分类，数据挖掘等领域。图 2-3 为近邻分类的误差率分析图。

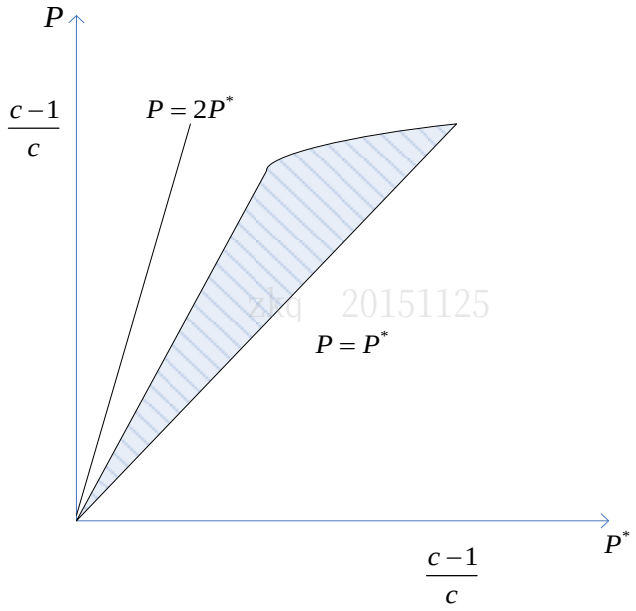


图 2-3 近邻分类的误差率分析图

2.3 *k*-近邻分类的优缺点

近邻算法在分类过程中很重要的一个环节就是选择近邻点的个数，也就是选择 *k* 值的大小。在分类过程中，如果选择不同的 *k* 值，那么近邻算法会产生不一样的分类结果。

如果采用的 *k* 值比较小，就意味着使用较小的数据集合进行预测，导致学习的误差率会降低。当输入的数据样本集比较接近或者相似的训练样本才会影响分类的结果，那么就会导致学习过程中的估计误差变大。*k* 值的减小表示着分类过程中的模型更加复杂，在分类过程中容易产生过拟合；如果采用的 *k* 值较大，在分类过程中意味着用大量的近邻点信息进行分类判断。这样会造成分类的近似误差

增大。

如果 $k=n$，那么没有太多的研究意义。因为不管样本的数据结构是什么，都只是简单的预测测试点周围的所有数据点，计算复杂度特别高。这样的模型没能提取出测试点周围的有用信息，而造成计算的浪费。因而在分类算法应用中，k 值的选择为一较小的奇数。

观察下图 2-4 的例子，可以得出，对于未知样本 x 通过 k 近邻分类算法，可以得到 x 应属于圆点类。然而对于测试样本 y，通过 k 近邻分类算法可以得到 y 应属于三角形的结论，而这样的结论却与概率分布相矛盾。

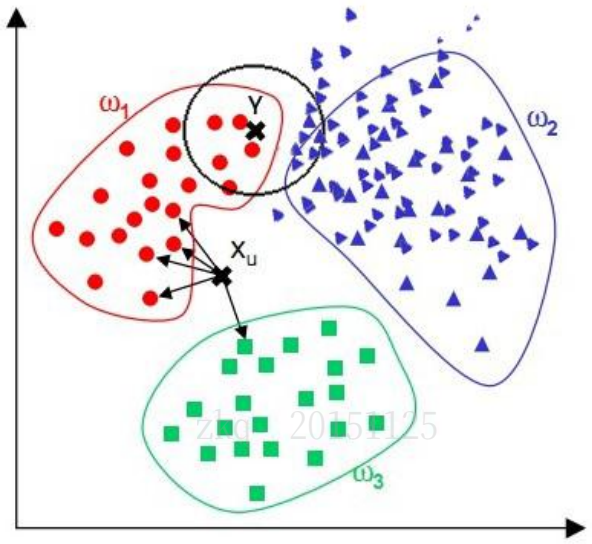


图2-4　测试样本的*k*-近邻分类

上图使用 *k*-近邻分类时有个重要的缺点。这类样本并不接近目标样本，而数量小的这类样本很靠近目标样本。这个时候，认为该测试样本属于数量小的样本所属的一类。但是 *k*-近邻算法却不关心这个问题，它只关心哪类样本的数量最多，而不把距离远近考虑在内。因此，可以采用融合距离加权的方法来改进 *k*-近邻算法。距离测试样本点近的点所具有的权值越大，相反距离测试样本点距离远的点所具有的权值则越小。融合距离加权的方法可以避免因为样本过大导致误判的情况。

从算法实现的过程可以发现，该算法存两个严重的问题。第一个是计算过程中需要存储所有的训练样本数据。第二个是需要进行繁重的距离计算。

k-近邻算法拥有以下的优点：

(1) 算法过程简单、有效而且易于理解。

(2)计算时间和空间线性于训练集的规模。

k -近邻分类算法同样有着以下的缺点：

(1) *k* -近邻分类算法是一种懒散学习法。在运算上比一些积极学习的算法要慢很多。

(2)类别评分不是规格化的。

(3)输出的可解释性不强，例如决策树的可解释性较强。

(4)当数据样本集的分布不均衡时，如果一个样本数据的数量很大，而其余的数据样本的数量很小的时候，就会增加分类误差率。这时进行分类时，对于一个测试数据样本，该测试样本的 *k* 个近邻中类别的样本数量占大多数。由于这种算法只考虑最近邻的样本数据，当有一类的数据样本的数目很庞大的时候，分类结果并不是接近真实的分类值。这时可以考虑采用添加不同的权值的方法来改进分类算法。

(5)分类过程中的计算的复杂度比较大。可以采用优化的方法是对数据样本集在分类之前进行数据结构优化。

2.4 *k*-近邻分类的改进

在分类过程中，近邻分类算法有着两个主要的不足。一个是在分类中需要对所有的样本点的信息进行储存。其次是在分类过程中距离度量的计算导致算法的复杂度很高[27]。因此对近邻分类提出了两种改进的方向：

(1)在分类判决之前，首先对样本数据集进行预处理。通过对数据集的分布进行优化，将与测试点分类关系不大的样本点从数据集中进行剔除。在分类的过程中将计算的范围锁定在测试样本点的领域附近，这样可以加快分类的速度并且减小计算复杂度。

(2)在样本数据集中选出对分类结果影响较大的近邻点，从而使样本集的数目减少。 *k* -近邻分类算法的改进方法主要可以分为加快分类速度、对训练样本库的维度、相似度的距离公式优化和 *k* 值确定四种类型。下面将分别讨论近邻分类的改进方向。

1. 快速搜索近邻法

近邻分类算法的一个缺点就是数据的计算量非常大，通常采用快速搜索近邻法来解决。快速搜索近邻法的基本思路是采用分而治之的思想，通过把样本数据分成一些不相交的子集，然后在对子集进行搜索。快速搜索近邻法对于最近邻分类

算法和 k-近邻算法都很适用。

如果用 $m=\{x_1,x_2,\cdots,x_N\}$ 表示为样本集，近邻分类就是在测试点 x 附近找出距离最近的 k 个近邻点。当 $k=1$ 的时候，也就是在最近邻分类的情况下，以后可以扩展到 k 个近邻分类的情况。快速搜索近邻法主要分成两步：

第一步就是把样本数据集进行 m 级分解，形成树状数据结构。

第二步用搜索算法找到测试样本点的最近邻。图 2-5 是距离选择为欧式距离时的图示。

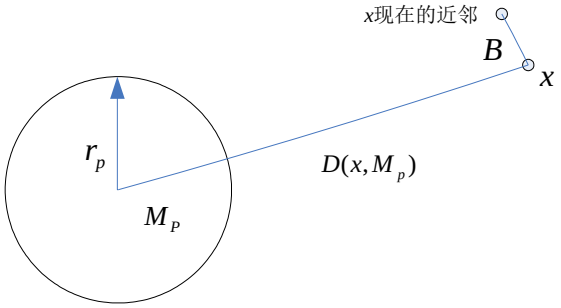


图 2-5 判断某子集是否可能为近邻

2.剪辑近邻法

通过经验可以对分类器的分类误差率进行估计即采用对测试样本集对分类的误差率进行估计[28]。在估计的过程中，选择不同的方法会导致不同的分类结果。如果用全部测试数据样本来设计分类器和估计误差率，由于缺少之间的独立性会影响分类误差率。

如果将样本数据集分成两个独立的集合，分别为训练样本集和测试样本集，并且采用训练样本集来设计分类器，使用测试数据集估计误差率。当两个数据集样本在独立的条件下，对误差率则较为准确。从错误率的基本分析思路则推出了剪辑近邻法。下面将详细进行介绍。

剪辑近邻法的基本思想是将分类过程分成两部分：假设要把 N 个样本分成 C 类。采用集合 $M^N=\{M_1^{N_1},M_2^{N_2},\cdots,M_c^{N_c}\}$ 表示，那么每一类别可以表示为 $M_i^{N_i}=\{x_i^k\}$ $(i=1,2,\cdots,c;k=1,2,\cdots,N_i)$。该算法的第一步是利用已知类别的数据样本集 M^N 中的样本先进行预分类，并删除掉被错分类的数据样本，把剩下来的数据样本重新组成样本集 M^{NE}。那么新生成的数据样本数量比原始数据样本数量少。这样的数据操作称为剪辑，第二步利用剪辑后的数据样本集 M^{NE} 和最近邻分类规则对测试样本 x 进行分类。

下面是最近剪辑算法的实现过程：

(1) *begin* initialize $j \leftarrow 0, D \leftarrow$ *data* set,n $\leftarrow$原型点个数

(2) 构造 D 的全部*Voroni* 图

(3) *do* $j \leftarrow j+1$;对每一个点 x'_j

(4) 找出 x'_j 的*Voroni* 近邻

(5) *if* 不是和 x'_j 属于同一类，则标记 x'_j

(6) *until* $j = n$

(7) 删除没标记的点

(8) 构造其余点的*Voroni* 图

(9) *end*

3.压缩近邻法

压缩近邻法压缩样本的思想是对现在有数据样本集的数据结构进行优化从而产生一个新的数据样本集[29]。

压缩近邻法的算法的实现是通过两个存储器 Store 和 Grabbag [30]。分别存放集原始的数据样本集和新生成的数据样本。

剪辑近邻法的算法的实现可以以下分成三步完成。

(1)初始化 Store 把原始数据样本集存储到Grabbag 。从 Grabbag 中随意选择一个数据样本存入 Store 中，然后作为新数据样本集的起始数据。

(2)样本集生成 在 Grabbag 中取出第 i 个样本。采用近邻法分类法对 Store 中的当前数据样本集进行分类。如果分类过程是无误的，则将该测试样本存储 Grabbag 中，并且对 Grabbag 中存储的所有数据样本重复第一阶段和第二阶段。

(3)结束过程 若 Grabbag 中的数据样本在第二阶段时没有存入 Store，那么此时算法将停止，否则从第二阶段继续开始。

2.5 分类系统评价指标

当完成一个分类器的设计后，需要知道该分类器的分类性能。如何评价该分类器性能，设计的分类器能否满足设计要求。分类误差率是衡量一个分类器的重要标准。如果分类误差率特别高，那么设计出的分类就无法在实际中应用。

模式分类系统的评价指标包括对分类误差概率的估计，分类误差率和置信区间。分类误差率的置信度是对分类器多次在试样本集上分类结果的可靠性评价。因为样本数目是有限的，分类器进行训练时，数据样本不但要作为训练数据而且同时也要作为测试数据。

分类系统设计训练过程的最后阶段：基于一组选定训练特征向量的优化分类器已经设计完成，现在的任务是评价设计的系统分类错误概率的情况。

2.5.1 分类误差概率估计

对于给定M类的分类任务，由 N 个测试样本构成的测试分类器。N_i表示每一类中的测试样本数量， $\sum_{i=1}^{M} N_i = N$， p_i为类别ω_i的分类误差概率。k_i个样本分类的误差率：

$$prob\{k_i\} = \binom{N_i}{k_i} p_i^{k_i} (1-p_i)^{N_i - k_i} \tag{2-20}$$

在 (2-20)中，由于 p_i的值是不确定的。当 $\hat{p}_i$为最大值时，得到 p_i的近似估计值 $\hat{p}_i$。

$$\hat{p}_i = \frac{k_i}{N_i} \tag{2-21}$$

因此，分类的总误差率估计为：

$$\hat{p}_i = \sum_{i=1}^{M} p(\omega_i) \frac{k_i}{N_i} \tag{2-22}$$

从数理统计知识可以由二项式分布的性质可以得出：

$$E[k_i] = N_i p_i$$

因此可得出：

$$E[\hat{p}_i] = \sum_{i=1}^{M} p(\omega_i) p_i = p \tag{2-23}$$

使用小的数据集对分类器的分类指标进行评定时，那么估计的误差率是不准确的。图 2-6 为 p 的 95％置信区间与 $\hat{p}$ 和 N 的关系。

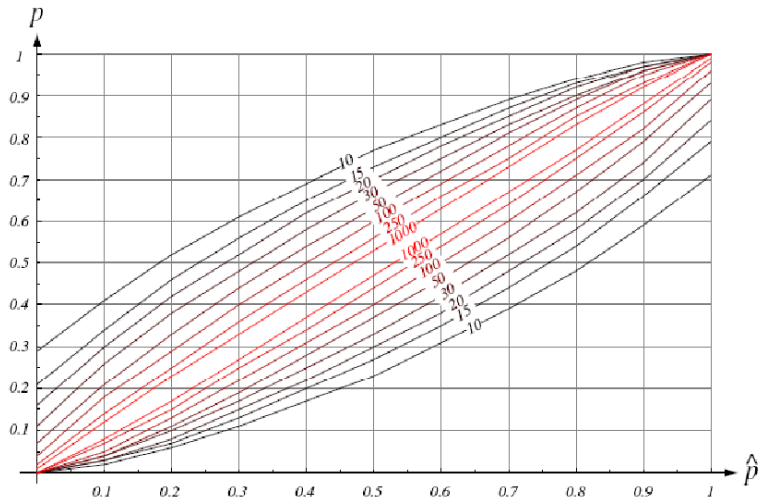


图 2-6 误差置信区间

2.5.2 分类误差率

分类误差率也是对分类系统评价的一个指标。系统的分类误差率表示的是平均分类误差率。训练样本的分类误差率为：

$$Err_{Train}=\frac{No.of\ \ misclassfied\ classfication\ on\ training\ samples}{No.of\ \ training\ \ samples}$$

测试样本的分类误差率为

$$Err_{Test}=\frac{No.of\ \ misclassfied\ classfication\ on\ testing\ samples}{No.of\ \ testing\ \ samples}$$

如果有 m 个样本集，可以计算出平均误差率 e 与方差 S：

$$S^2=\frac{1}{M}\sum_{i=1}^{M}(e_i-e_{av})^2$$

$$平均误差率=\frac{各样本的误差率之和}{样本数}$$

识别率作为评判分类系统的一个标准，识别率和分类误差率的关系如下：

$$Err_{Train}=1-Acc_{Train}$$

$$Err_{Test}=1-Acc_{Test}$$

2.5.3 置信区间

$\bar{x}$ 表示分类的平均误差率，n 表示测量的次数，通过采用方差 σ^2 的无偏估计 S^2 来代替 σ^2，置信水平为 $(1-\alpha)$，枢轴量为 T ，则有：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2-24}$$

$$T = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t(n-1) \tag{2-25}$$

$$P\{-t_{\frac{\alpha}{2}}(n-1) < \frac{\bar{x} - \mu}{S / \sqrt{n}} < t_{\frac{\alpha}{2}}(n-1)\} = 1 - \alpha \tag{2-26}$$

置信区间如下表示：

$$(\bar{x} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1), \bar{x} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)) \tag{2-27}$$

2.6 本章小节

本章节中简要地介绍了最近邻分类和k-近邻分类算法，并分别对两类近邻算法的误差率进行分析。在k-近邻分类算法进行分类的时候，通过k-近邻分类算法的介绍，分析现有k-近邻分类算法的优点与不足，并简要介绍了分类系统的评价指标。

第三章 基于距离加权的伪近邻分类

在模式分类算法中，*k*-近邻分类算法由于具有简单、高效等特点广泛应用在模式识别和分类等领域。*k*-近邻分类算法从训练数据样本集中选出距离测试样本点最近的 *k* 个最近邻点，通常选取欧几里德距离作为距离度量。然后再依据 *k* 个最近邻样本点所属于的具体类别进行分类决策，从而确定测试样本点应该被分到哪一类。

本章主要分析 *k*-近邻分类算法容易受到离群点和数据分布的影响，因此采用了距离加权的伪近邻改进算法。采用伪近邻改进算法对训练样本消除噪声，以提高分类器的分类准确率。

3.1 近邻分类的距离度量

k -近邻算法的关键在于找到测试样本点的近邻。如何搜寻测试样本的近邻、近邻的判别标准是什么、以及采用什么样的距离来度量。这一系列问题便是下面要介绍的距离度量表示法。在特征空间中两个样本数据的距离反应出两个样本数据之间的相似性程度。*k*-近邻分类模型所采用的特征空间一般都是 *n* 维实数的向量空间。常用的距离可以是欧式距离，也可以选择其它距离，因此就有不同距离度量的考虑，下面就来具体介绍常用距离度量的表示法。

1. 欧氏距离

欧式距离的又叫欧几里德距离。欧几里德距离的度量表示的是在 m 维空间中两个点之间的真实距离[32]。在欧氏空间中两点 $x=(x_1,...,x_n)$ 和 $y=(y_1,...,y_n)$ 之间的距离为：

$$\begin{aligned} d(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \\ &= \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \end{aligned} \tag{3-1}$$

在二维空间平面上两点 $a(x_1,y_1)$ 与 $b(x_2,y_2)$ 的欧氏距离：

$$d_{ab}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{3-2}$$

在三维空间平面上 $a(x_1, y_1, z_1)$ 与 $b(x_2, y_2, z_2)$ 间的欧氏距离：

$$d_{ab} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \tag{3-3}$$

在 n 维向量上两向量 $a(x_{11}, x_{12}, \cdots, x_{1n})$ 与 $b(x_{21}, x_{22}, \cdots, x_{2n})$ 间的欧氏距离：

$$d_{ab} = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2} \tag{3-4}$$

如果采用向量运算表达式为：

$$d_{ab} = \sqrt{(a-b)(a-b)^T} \tag{3-5}$$

2. 曼哈顿距离

曼哈顿距离是定义在欧氏空间中的直角坐标系上两点分别对坐标轴上的投影的距离总和[33]。在二维坐标点 $a(x_1, y_1)$ 与点 $b(x_2, y_2)$ 之间的曼哈顿距离为：

$$d_{ab} = |x_1 - x_2| + |y_1 - y_2| \tag{3-6}$$

通俗来讲，从曼哈顿街区如果要从一个十字路口开车到另外一个十字路口，所行驶的距离不是他们两点之间的直线距离，而是实际行驶距离。这个距离就是曼哈顿距离。两个 n 维向量 $a(x_{11}, x_{12}, \cdots, x_{1n})$ 与 $b(x_{21}, x_{22}, \cdots, x_{2n})$ 间的曼哈顿距离如下式示：

$$d_{ab} = \sum_{k=1}^{n} |x_{1k} - x_{2k}| \tag{3-7}$$

3．切比雪夫距离

切比雪夫距离是在向量空间中的一种距离的度量。假设有两个向量或者是两个点 p 和 q，它们的坐标分别为 p_i 和 q_i，两点之间的切比雪夫距离表示为：

$$D_{chebyshev}(p, q) = \max_i(|p_i - q_i|) \tag{3-8}$$

同时也等于 L_p 度量的极值：

$$\lim_{k \to \infty} (\sum_{i=1}^{n} |p_i - q_i|^k)^{1/k}$$

所以切比雪夫距离又被称为 L_∞ 度量。在平面几何中，若二点 p 及 q 的直角坐标系坐标为 $p(x_1, y_1)$ 和 $q(x_2, y_2)$，那么切比雪夫距离为：

$$D_{chebyshev} = \max(|x_2 - x_1|, |y_2 - y_1|)$$

在 n 维空间中，向量 $a(x_{11}, x_{12}, \cdots, x_{1n})$ 与 $b(x_{21}, x_{22}, \cdots, x_{2n})$ 之间的切比雪夫距离表示为

$$d_{ab}=\max_i(|x_{1i}-x_{2i}|)$$

另外的一种等价的表达形式：

$$d_{ab}=\lim_{k\to\infty}\left(\sum_{i=1}^{n}|x_{1i}-x_{2i}|^k\right)^{1/k}$$

4．闵可夫斯基距离

闵可夫斯基距离不是一种距离，而是一组距离的定义。闵氏距离的定义如下，在空间中向量$a(x_{11},x_{12},\cdots,x_{1n})$与向量$b(x_{21},x_{22},\cdots,x_{2n})$间的闵可夫斯基距离定义为：

$$d_{ab}=\sqrt[m]{\sum_{k=1}^{n}|x_{1k}-x_{2k}|^m} \tag{3-9}$$

在这个表达式中，m是一个可以变化的参量。

当$m=1$时，其表示的为曼哈顿距离。

当$m=2$时，其表示的就是欧氏距离。

当$m\to\infty$时，其表示的就是切比雪夫距离。

5．马氏距离

马氏距离是两个数据样本集的协方差距离，同时也是表示两个数据样本集相似度的方法[34]。假设有x_1-x_m个样本协方差矩阵表示为S，样本的均值向量记为u，那么样本向量X和u的马氏距离表达式如下：

$$D(X)=\sqrt{(X-u)^T S^{-1}(X-u)} \tag{3-10}$$

向量x_i与x_j之间的马氏距离表达式如下：

$$D(X_i,X_j)=\sqrt{(X_i-X_j)^T S^{-1}(X_i-X_j)}$$

如果样本数据向量是独立并且属于同样的概率分布，此时的协方差矩阵为单位矩阵，那么上面的马氏距离表达为：

$$D(X_i,X_j)=\sqrt{(X_i-X_j)^T(X_i-X_j)}$$

此时马氏距离就变成了欧式距离。当样本的协方差矩阵为对角矩阵的时候，马氏距离的表达式就退化成了欧式距离。马氏距离的优点在于与量纲无关，作为距离试题量可以排除变量之间的相关性的干扰。

6. 巴氏距离

在统计学中，巴氏距离是表示两个测试样本的离散度，连续概率分布的相似性。巴氏距离与衡量两个统计样本或种群之间的重叠量 Bhattacharyya 系数有关联。与此同时， Bhattacharyya 系数用来测量中的类分类的可分离性。

对于离散概率分布 p 和 q 在同一域 X，定义：

$$D_b(p,q) = -In(BC(p,q))$$

其中：

$$B_c(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

是 Bhattacharyya 系数。

当分布属于连续概率分布， Bhattacharyya 系数为：

$$B_C(p,q) = \int \sqrt{p(x)q(x)}dx \tag{3-11}$$

在 $0 \le B_C \le 1$ 并且 $0 \le B_D \le \infty$ 的时候， 巴氏距离 B_D 并没有服从三角不等式。

对于多变量的高斯分布， $p_i = N(m_i, P_i)$

$$D_B = \frac{1}{8}(m_1 - m_2)^T P^{-1}(m_1 - m_2) + \frac{1}{2} In(\frac{\det P}{\sqrt{\det P_1 \det P_2}}) \tag{3-12}$$

协方差的分布 $P = \frac{P_1 + P_2}{2}$ 。

计算巴氏系数涉及集成的基本形式的两个样本的重叠的时间间隔的值的两个样本被分裂成一个选定的分区数，并且在每个分区中的每个样品的成员的数量，在下面的公式中使用

$$\text{Bhattacharyya} = \sum_{i=1}^{n} \sqrt{(\sum a_i \cdot \sum b_i)} \text{。} \tag{3-13}$$

7．夹角余弦

夹角余弦一般又被称为余弦相似度，采用向量空间中两个向量之间夹角的余弦值表示两向量差异大小 [36]。两向量之间的夹角越小，那么余弦值就越大，相似度就越高。

在二维空间中，向量 $a(x_1, y_1)$ 与向量 $b(x_2, y_2)$ 的夹角余弦公式：

$$\cos\theta = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2}\sqrt{y_1^2 + y_2^2}} \tag{3-14}$$

两个 n 维样本点 $a(x_{11},x_{12},...,x_{1n})$ 和 $b(x_{21},x_{22},...,x_{2n})$ 之间的夹角余弦：

$$\cos\theta=\frac{a\cdot b}{|a||b|}$$

对于两个 n 维样本点 $a(x_{11},x_{12},...,x_{1n})$ 和 $b(x_{21},x_{22},...,x_{2n})$，采用夹角余弦来衡量这两个样本之间的相似程度，即：

$$\cos\theta=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^2}\sqrt{\sum_{k=1}^{n}x_{2k}^2}} \tag{3-15}$$

对于下图3-1的测试样本采用*k*-近邻分类算法，如果采用欧几里德距离，那么其距离度量在二维空间中的轨迹就是一个圆形，此时将测试样本将归到”1”类中。如果采用投影距离，测试样本距离“2”类更近，因此将归于”2”类中。但从图中概率密度分布来看，测试样本的应该属于类别“2”的概率更大。在图3-1中，采用欧氏距离度量，目标样本(图中黑点)，通过1-近邻分类，目标样本应分为第一类。采用水平距离(*x*轴上)度量，目标样本通过1-近邻分类，目标样本应分为第二类。从数据结构观察，目标样本分为第二类更为合理。在图3-1中，决策面(这里为一条直线)垂直于水平面(水平轴)，在确定目标样本的邻域时，其邻域范围应在决策面正交方向收缩(即水平方向收缩)，与决策面平行方向伸展(即垂直方向伸展)，这就要求采用相应的距离度量，使之达到这种效果，就需要距离度量自动地适应这样的需求。因此在进行*k*-近邻分类时选择不同的距离度量对于分类的结果有着重要的影响。

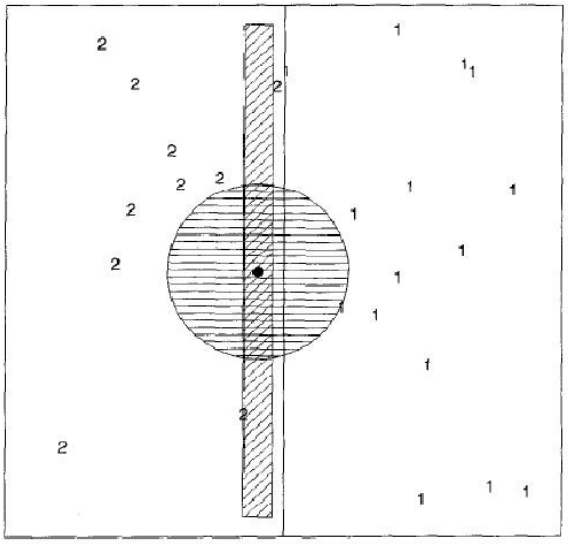


图3-1　不同的距离度量对分类的影响

3.2 距离加权的伪近邻算法

在近邻分类规则中，测试数据样本的k个近邻在分类过程中被默认地被认为对分类影响具有相同的权值。而忽略了这k个近邻与测试样本 x 距离的长度有关系。改进的方法就是对于这k个近邻，根据近邻点与测试样本 x 距离的不同而分别赋于不同的权值。距离越近，那么近邻点对分类性能的影响就越大，其所代表的权值就越大。这个思想最开始是由 Dudani首先提出的[37]。在改进的近邻算法中通过和距离较远的近邻点进行比较，距离近的近邻点在分类过程中的权值就更大。距离测试数据样本点最近的近邻点的权值为权值1，距离它最远的近邻点的权值为0，剩下的近邻点的权值则根据其到测试样本的距离呈现线性变化。

学者已经在理论上证明了k–最近邻的渐进性能是优于最近邻分类的。k–最近邻算法最大的优点是简洁直观。但是当样本数量比较小的时候，采用k–近邻分类的误差率比较高。

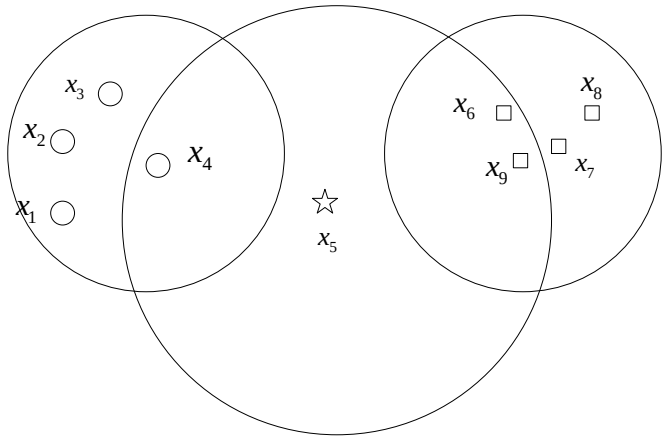


图3-2 互近邻图

传统的k–近邻分类算法是利用数据样本之间的距离来对测试样本进行分类。也就是找出对测试数据样本影响最大的前k个近邻点。采用这样的分类方法只是通过简单地考虑数据样本个体之间的相似性，而没有考虑数据样本之间互为近邻的情况。如图3-2则清楚地描述了数据样本之间由于互近邻而产生的影响作用，从图3-2可以看出，样本点 x_4 的三个近邻点分别是 x_1 、 x_2 和 x_3 。样本点 x_9 的三个近邻点分别是 x_6 、 x_7 和 x_8 。而样本点 x_5 的三个近邻点分别是 x_4 、 x_6 和 x_9 ，但是 x_5 的5个近邻却不在 x_4 、 x_6 和 x_9 的3个近邻中。究其原因是因为 x_4 、 x_6 和 x_9 三个样本点距离 x_5 很远。从而导致 x_5 是一个孤立的奇异点。也就是说 x_5 是一个异常的噪音数据。

在基于投票的近邻分类规则里，待分类数据样本的k个近邻点被默认地认为对于分类具有相等的权值[38]。而忽略了k个近邻点与测试样本数据x的距离的远近。

基于此提出的一个改进的想法就是对于这k个近邻点，根据近邻点和测试数据样本x 之间的距离的远近从而赋不同的权值。近邻点与测试数据样本之间的距离越近，则该近邻点对分类的影响就越大，那代表的权值就越大。反之距离远的点所代表的权值就越小。

这个算法假定所有的实例对应于 n 维欧氏空间 n 中的点。更准确地来说就是把任意的实例 x 表示为下面的特征向量：

$$< a_1(x), a_2(x), \cdots, a_n(x) >$$

$a_r(x)$ 为 x 的第 r 个属性。两实例 x_i 和 x_j 之间的距离为 $d(x_i, x_j)$，其中

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

说明：

1. 在最近邻分类过程中，目标函数的值是离散值或者是实值。

2. 近邻分类算法的返回值 $f'(x_q)$ 为对 $f(x_q)$ 的估计，它就是距离 x_q 最近的 k 个训练样例中最普遍的 f 值。

3.2.1 特征加权

对于某个特定的类别，数据空间中往往存在与其不相关的属性，表现为在空间上该类别的样本是分散的，只有在某些低维的子空间上才是“密集的”[39]。因此，为了更加准确地度量两个样本之间的相似度，因此考虑各个属性对不同类别的重要程度。假设训练样本集合中存在一个属于类别为 Y 的样本 X，特征加权方法就是在计算待分类样本与 X 的相似度时。将与类别 Y 相关的属性赋予较大的权重，否则赋予较小的权值。使用一种基于属性权重大小与同类样本投影到该属性上的分布离散程度成反比思想的特征加权方法[39]。直观地说，如果训练数据集上同一类别的样本在某一属性上分布得越集中，该属性的重要程度就越高，就赋予该属性较大的权重。设 $\text{Ceuter}_k =< c_{k1}, c_{k2}, c_{kd} >$ 为训练数据集上第 k 类样本的中心，采用公式(3-16)计算，其中 $Num(k)$ 表示训练数据集上属于第 k 类样本的数目

$$c_{kj} = \frac{1}{Num(k)} \sum_{y_{i=k}} x_{ij} \tag{3-16}$$

那么，第 k 类样本属性 j 的权重 ω_{kj} 使用公式(3-17)计算。这里 x_{ij} 表示训练样本 X_i 中属性 j 的取值，Y_i 表示 X_i 的类别，△是为避免分母为 0 而引入的一个很小的数值，实验中一般取△=10^{-4}。

$$w_{kj}=\left(\sum_{m=1}^{d}\left(\frac{\sum_{y_i=k}[(x_{ij}-c_{kj})^2+\Delta]}{\sum_{y_i=k}[[(x_{im}-c_{km})^2+\Delta]]}\right)^2\right)^{-1} \tag{3-17}$$

以欧氏距离为例，对于给定待分类样本 $X'=\{x'_1,x'_2,...,x'_d\}$ 和训练数据集上属于第 k 类的样本 X_i，两者基于上述特征加权方法的相似度使用公式(3-18)算:

$$dis(X',X_i)=\sqrt{\sum_{j=1}^{d}\omega_{kj}(x'_{ij}-x_{ij})^2} \tag{3-18}$$

基于神经网络的前馈学习网络来学习样本各属性权重的灵敏度法，具体算法过程如下

第一步：使用前馈学习网络对训练数据集中每个样本进行学习，以样本的各个属性作为输人，样本的类别属性作为输出。当达到指定的训练精度或者迭代次数以后，学习结束。

第二步:使用建立好的神经网络对训练数据集中每个样本的类别进行预测，将所预测的样本 X 的类别以 p_i^0 表示。

第三步:依次去除每一个输人属性 $x_{ij}(j=1,2,...,d)$。此时该神经网络所预测的样本 X_i 的类别以 p_i^j 表示。

第四步:使用公式(3-19)计算每个属性 j 的权重 ω_i，其中 s_j 采用公式(3-20)计算。s_j 中的 n 表示训练样本集合中的样本个数。

$$\omega_j=\frac{S_j}{\sum_{l=1}^{d}S_l} \tag{3-19}$$

$$S_j=(\sum_{i=1}^{n}\frac{\left|p_i^0-p_i^j\right|}{p_i^0})/n \tag{3-20}$$

从公式(3-20)可以直观地看出，当属性对分类的重要程度很高时，其对应的权重也相应地增大。

3.2.2 加权投票

在基于加权投票的改进 k-近邻算法中，待分类样本的类别，而是对集合内的样本赋予不同的权重来体现各自对未知样本的影响。

加权投票方式中最直观的做法是将相似度转化为相应的权值。使用欧几里德距离时，近邻集合内与待分类样本较近的样本在投票时被赋予较大的权重[40]。随着距离的逐渐增大，权重也相应地减小。给定待分类样本 X，设 $NN=\{X_1,X_2,...,X_k\}$ 是按照距离降序排列的 X' 近邻集合，即有 $dis(X',X_1)\le dis(X',X_k)$。那么近邻集合内，样本 X 进行投票的权重为

$$\omega=\begin{cases}\dfrac{dis(X',X_k)-dis(X',X_i)}{dis(X',X_k)-dis(X',X_1)} & \text{if } \mathrm{dis}(X',X_k)\neq \mathrm{dis}(X',X_1)\\ 1 & \text{if } \mathrm{dis}(X',X_k)=\mathrm{dis}(X',X_1)\end{cases} \tag{3-21}$$

根据上述权重，待分类样本 X'的类别为

$$y'=\arg\max_{y}\sum_{x_i\in NN}\omega_i\times\delta(y=y_i)$$

其中 y_i 表示近邻集合 NN 中样本 X_i 的类别，$\delta(y=y_i)$ 是一个狄拉克 δ 函数。当 $y=y_i$ 时，函数的取值为 1，否则为 0。

由于在样本分布稀疏的区域中，k 个近邻样本组成的局部邻域比较大，因此待分类样本的近邻选择容易趋向于样本分布比较密集的区域，也就是说分布密集区域内的样本更容易被选中作为待分类样本的近邻。这样无疑使得分类结果偏向于样本个数较多的类别。针对以上问题，通过加大待分类样本与数据稀疏区域内样本的相似度来降低训练数据集上样本分布不均对相似度造成的影响，学者提出了一种基于密度的 k–近邻改进算法[41]。算法中待分类样本 X' 与其近邻 X 的距离 $dis(X',X)$ 为:

$$dis(X',X)=\frac{\|X'-X\|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|X-X_i\|}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|X'-X_i\|}} \tag{3-22}$$

其中 $\|X'-X\|$ 表示 X' 与 X 的欧氏距离，n 表示训练样本的个数。类似的算法还有同样基于密度的加权投票法。值得注意的是，近邻集合内的样本权重不会为 0。然而，虽然特征加权和投票加权方法在一定程度上能够有效提高 k–近邻算法的分类精度是以牺牲算法分类效率为代价得到的，这两种方法提高了 k-近邻算法分类精度但计算复杂度很高。

基于距离权重的伪近邻算法：集合 $S=\{s_1,s_2,...,s_n\}$ 有 n 个数据样本。存在一个值 $k\in\{k\mid 0<k<n,k\in Z\}$，对于其中每一个数据样本 s_i、都拥有与之相对应的 k 最近邻集合 $N_i=\{t_1,t_{2,}...,t_k\}$，$N_i\subseteq S$。如果 s_i 与 s_j 两个数据样本之间互相都是属于 k 最近邻，那么 $s_i\in N_j$ 并且也属于 $s_j\in N_i$。反之则称 s_i 和 s_j 不是互为 k 最近邻关系。如果两个均拥有互相 k 最近邻关系的数据样本 s_i 和 s_j，则称 s_i 为 s_j 的互 k 最近邻。

距离加权的 k 最近邻分类是将训练数据集中选出还未分类的数据样本之后进对测试样本进行分类处理。以降低数据样本中的伪近邻点对于分类的影响。在得出测试样本的 k 最近邻集合之后，再采用 k 最近邻分类距离加权的投票原则进行分类。在基于权重的 k 最近邻算法所采用的距离度量相似度为欧几里德距离。在最后的分类中使用下式作为距离权重计算公式。

$$dist(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{id} - x_{jd})^2} \tag{3-23}$$

在表达式中， x_{id} 代表数据样本 x_i 在第 d 维的属性。

权重的表示采用 $\omega_i = 1/d_i$ 来计算。d_i 表示 x_i 与 k 最近邻 t_i 之间的欧几里德距离。S 代表等待分类的数据样本的数目。k 则代表初始时刻选择的最近邻点的个数。$T = \{t_1, t_2, ..., t_n\}$ 表示 n 个具有类别的训练数据的集合。

第一步：利用式 3-23 计算出待分类数据样本 s 和训练数据样本集 T 的 k 个最近邻集合 $N = \{t_1, t_2, ..., t_k\}$。

第二步：判别测试数据样本 s 是否属于 $t_i (i = 1, 2, ..., k)$ 的 k 个最近邻。

第三步：对其余的近邻点同样使用距离加权的分类方法。即通过计算出每个类别所拥有的权值，然后得到每个分类数据类别的权值之比。

第四步：选出近邻点中权重最大的那一类别作为测试数据样本 s 的类别，将 s 所属的类别进行输出。

3.3 仿真测试

实验中所采用的环境：CPU为Intel(R)Core i5，2.4GHz，内存2G，开发平台Matlab2010(b)。仿真实验测试的过程中使用UCI标准机器学习库的数据。表3-1则是UCI数据集的样本信息表。表中列出了各类数据集的样本数量，样本维数和样本的类别。表中有11类数据集，实验中利用交叉验证方法的方法，然而采用交叉验证的目的是为了得到可靠稳定的模型。把数据集平均的分成10份，在分类过程中，用90%的数据为训练数据，剩下10%为测试数据。循环重复这样的过程十次，并且保证每份数据即要做测试数据也要作为训练数据。最后把十次分类的结果取平均值得到平均分类准确率，并且以平均分类准确率作为最后分类的结果。表3-1则列出了仿真测试过程中用到的UCI数据集的样本信息。

表3-1 UCI数据集信息

数据集	样本数	样本维数	类别数
Iris	150	4	3
Wine	178	13	3
Monks	124	2	2
Ecoli	336	7	8
Glass	214	9	7
Heart-statlog	270	13	2
Diabetes	768	8	2
Vehicle	848	18	4
Ionnsphere	351	34	2
Liver	345	6	2
Segment	2310	19	7

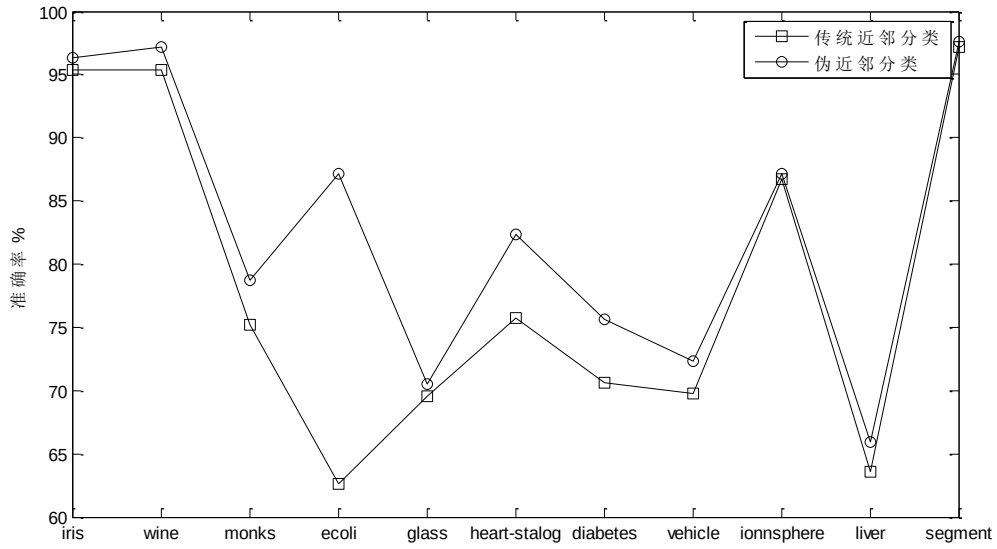


图3-3 分类准确率对比

图3-3为11个测试数据集的改进算法和传统算法分类准确率对比图。从下表3-2的测试结果中可以看出，在这11个测试数据集的实验中，基于权重的伪k–近邻算法具有最高的分类准确率。与传统的k–近邻算法相比，基于权重的伪k–近邻算法具有最高的分类准确率在这测试数据集上具有较高的分类准确率。所以改进的k–近邻算法有效地提高了分类的准确率，能够优化后的数据样本之间的联系更为紧密，并且还降低了离群点对分类的影响。从基于权重的k–近邻算法和传统的k–近邻

算法的分类准确率的对比中可得出，基于权重的*k*–近邻算法的分类准确率总体优于传统的*k*–近邻算法。

表3-2 分类的准确率对比

数据集	*k* 近邻分类(%)	带权值的伪近邻分类(%)
Iris	95.4	96.37
Wine	95.33	97.2
Monks	75.23	78.70
Ecoli	62.62	87.15
Glass	69.56	70.55
Heart-statlog	75.74	82.39
Diabetes	70.57	75.63
Vehicle	69.78	72.35
Ionnsphere	86.70	87.12
Liver	63.56	65.88
Segment	97.12	97.15

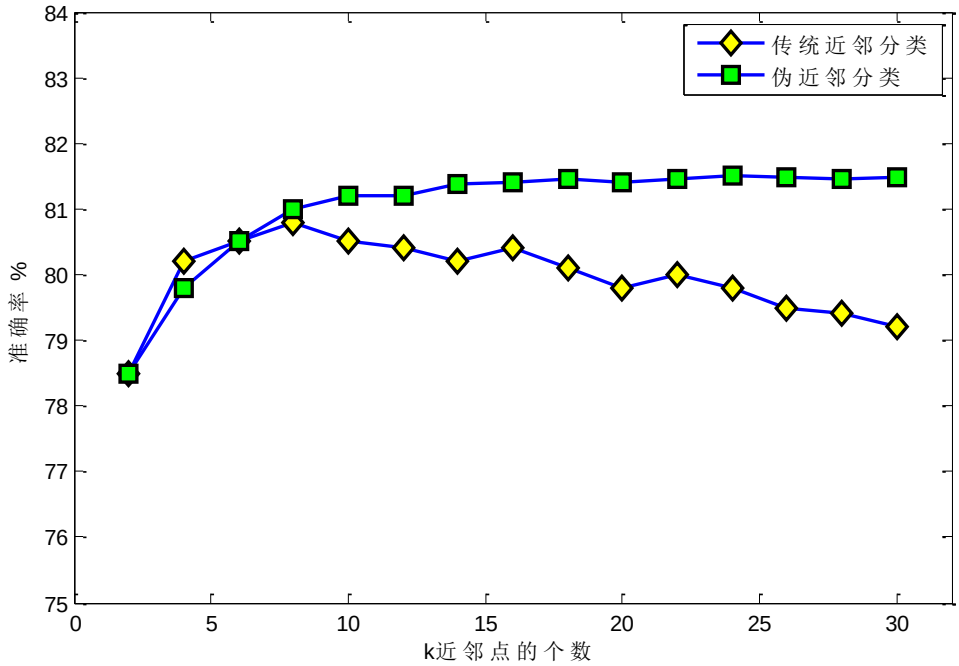


图3-4 分类平均准确率

图3-4是改进的分类算法和传统分类算法在测试数据样本上的平均分类准确率。从上图可以发现基于权重的k最近邻算法在总体上具有较好的分类准确率。当k取值超过7时，基于权重的k最近邻算法的分类准确率基本稳定。相反，原始的基于权重的k最近邻算法受k值的影响较大。随着k值的增加，分类器的分类性能变差，由此得出k值的变化对分类准确类的影响因素很大。综合上面的数据分析，基于权重的k最近邻算法具有以下的优点：

(1)分类的误差率在总体上有了一定的降低；

(2)对于高维的数据样本也有较好的分类表现。同时还降低了不相关的属性值对于分类的结果的影响；

(3)可以将其他分类算法与k–近邻算法进融合；

(4)改进的k–近邻算法方法简单，实现起来较为容易。

从图3-5可以看出，测试数据样本点位于类别1样本较为密集的区域。在采用k-近邻分类的时候，对于测试样本点当近邻点的数目3的时候。采用传统的近邻分类规则进行分类时，由于测试数据样本点的3个近邻点中有2个属于类别2，根据投票分类原则测试样本点被分为类别2。但如果采用基于权重的伪k–近邻算法进行分类，在每类训练数据样本周围取3个近邻，那么测试样本点的属于类1，则测试样本点应被判为类1。而从测试点所处的样本点分布来看，测试点分为类别1则更为准确。

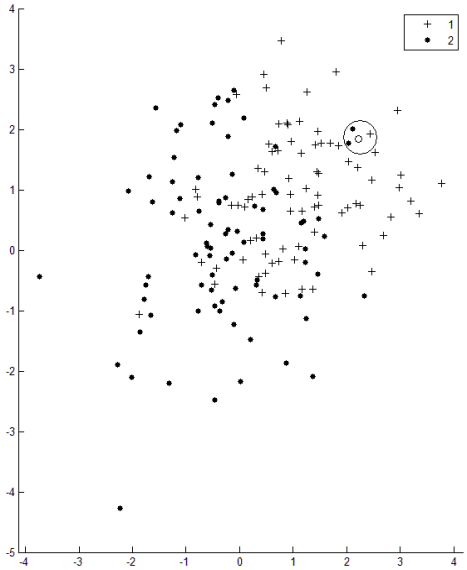


图3-5两种分类方法在进行分类时的差异

3.4 本章小节

本章对近邻分类过程中的距离度量进行了介绍，选择不同的距离对于分类结果有着重要的影响。然后对 k-近邻分类算法的特征加权和加权投票进行了研究，采用了基于距离加权的伪近邻算法并通过实验仿真测试与传统的 k-近邻分类算法的分类性能进行了比对。传统的近邻分类算法只利用了测试数据样本和某一类数据样本集里的一个近邻点的相似性，而基于距离加权的伪近邻算法却利用了测试数据样本和某一类数据样本集中的多个近邻点之间的相似性。

第四章 基于类均值的近邻分类

基于类均值的 *k*-近邻分类不但利用了测试样本的最近邻点，而且还融合了数据样本集中的类均值。采用这种分类方法的分类误差率比传统的 *k*-近邻分类的误差率要低，并且拥有较好的分类结果。离群点的概念在目前的模式识别领域中到目前为止没有统一定义。各种不同的分类方法从各自的角度所给出离群点的定义都不相同。在数理统计学中，离群点是指在数值中远离平均数值的极大值和极小值。同时离群点也被称为歧异值。形成离群点的主要原因有：首先可能是数据在采样过程中形成的误差，比如数据记录上的误差，技术人员在记录过程中产生的错误等，都可能引入误差。在人口数据统计中，由于洪水，地震等自然灾害原因导致人口数量下降，从而对人口统计数量上造成影响。在基因 DNA 序列中由于基因突变导致基因序列的变化等情况形成离群点。在对于离群点的检测方面遇到以下的困难：

1.当数据样本的维度为非数值型时，在检测过程中需要对数据维度进行预处理等;

2.当数据样本的维度是多维时，离群点的异常特征可能是多维度而并不是单一维度。

因此在本论文中，主要采用把与样本数据的平均值的距离超过 3 倍标准方差距离的点定义为离群点。

4.1 基于类均值的近邻分类

图 4-1 的来阐述数据样本的类均值在分类过程中对分类准确率的影响。图中的 4-1 左半部分表示了训练数据样本点的分布图，两个已知类别的类均值向量和测试样本点。根据图 4-1(a)中可知，测试数据样本点位于类别 1 训练数据样本点分布较为密集的区域中。但距离测试数据样本点最近的点为类别 2。如果采用最近邻的分类算法，那么该测试点将被归到类别 2 中。但是距离测试样本点最近的近邻点其实是类别 2 中的一个离群点。如果把该离群点作为近邻分类的一个近邻，那么就会造成分类错误。现在把每个类别的均值向量点也作为近邻分类的一个参考点。根据图 4-1 可以看到，这时采用的距离为测试样本点到类均值距离点的距离和最近邻距离之和。即采用组合距离，从图中 4-1 可以得到测试样本点到类 1 的组合距离小于类别 2 的组合距离，因此将测试样本点分为类别 1。由图 4-1 样本的概率密度

分布来看，将测试样本分为类别 1 是比较合理的。

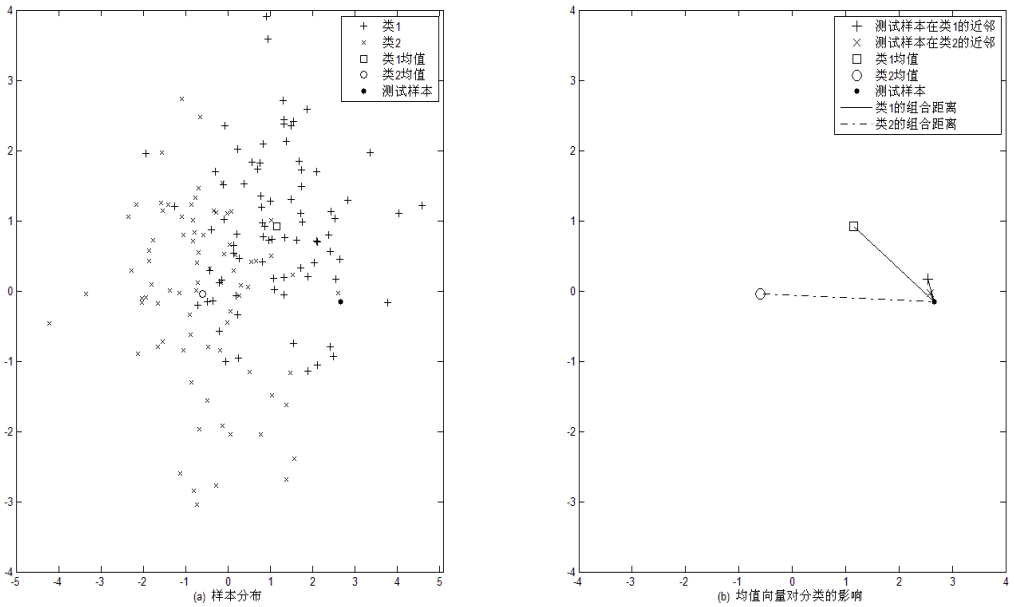


图 4-1 类均值对分类的影响

假设有N个训练数据样本，$N_1,N_2...N_m$分别对应的是类别$\omega_1,\omega_2...\omega_m$的训练样本数。$x_j^{(1)},x_j^{(2)},...,x_j^{(r)}$分别表示测试样本在不同的类$\omega_j$里的r个近邻，并且$X_j=\{x_j^i \mid i=1,2,...,N_j\}$代表类别$\omega_j$的训练数据样本集，$\mu_j$表示训练样本数据集在类别$\omega_j$中的均值向量。 那么可以得到：

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_j^i \tag{4-1}$$

基于类均值的最近邻分类方法算法过程如下：

(1) 计算测试样本在类别ω_j中的最近邻距离

$$d_j^1=\sqrt{(x-x_j^{(1)})^T(x-x_j^{(1)})} \tag{4-2}$$

(2) 计算测试样本和类均值向量μ_j在类别ω_j的距离

$$d_j^2=\sqrt{(x-\mu_j)^T(x-\mu_j)} \tag{4-3}$$

(3) 计算最近邻距离和类均值距离的组合

$$d_j=d_j^1+\omega*d_j^c \qquad 0\le\omega\le1 \tag{4-4}$$

(4) 采用最近邻分类方法进行分类，并满足以下条件：

$$d_c = \arg\min\{d_j\} \quad \text{j=1,2,...,m} \tag{4-5}$$

下面就是如何确定加权系数w*和近邻数k。本文中利用交叉验证的方法获得加权系数 w*，步骤如下所示：

1.把训练数据样本分为两部分即训练数据集和测试数据集。把数据样本集平均地分成10份，用9份数据样本为训练数据集，1份数据样本作为测试样本集，并且每一次的验证集不同。

2.使用基于类均值的k-近邻分类算法对测试数据集进行分类决策，从而得到误差向量$e_i(\omega)$。

3.重复第一步和第二步10次，得到分类的误差向量：

$$E_{cv}(\omega) = \frac{1}{10}\sum_{i=1}^{10} e_i(\omega) \tag{4-6}$$

4.距离加权系数 w*由下式得到：

$$E_{\min} = \min\{E_{cv}(\omega)\} \quad , \quad w^* = \arg\min(E_{\min}(w)) \tag{4-7}$$

在大多数情形下，基于类均值的k-近邻分类算法分类准确率高于传统的近邻算法。这种改进方法是对局部均值近邻分类的改进，因而具有局部均值近邻分类类似的分类性能。

4.2 实验数据

在本论文中使用了两类数据集。一种是采用加州大学欧文分校的 UCI 标准机器学习库里的数据集，另一种是人工模拟的数据集。下面分别对它们进行介绍。仿真所用的数据集假定所用数据集里样本类概率的先验概率相同，测试采用相同的数据样本集。在实验中测试数据样本为1000个。对每一个数据样本进行重复实验 100 次，每次实验后则重新产生新的数据。重复实验100次后将得到的平均值对应在置信区间上。

在对每个数据集进行测试的过程中，每一类数据集的分类误差率取其平均误差。实验测试过程中所使用的人工数据样本集为$I-\Lambda$， $I-4I$， $I-I$， *Ness*，Mixture-1和Mixture-2。在数据集中，μ_i表示数据样本类ω_i的均值向量。$\sum_i$ 为数据类的协方差矩阵。下面是对这4个人工数据样本集进行简介：

(1) $I-\Lambda$ 数据样本集由8维高斯分布的数据样本所组成，

$$\mu_1 = 0,$$

$$\mu_2 = [3.86, 3.10, 0.84, 1.64, 1.08, 0.23, 0.01]^T,$$

$$\sum\nolimits_1 = I_8 ,$$

$$\sum\nolimits_2 = diag[8.41,12.06,0.12,0.22,1.49,1.77,0.35,2.76]$$

其中 I_i 与 $diag[]$ 分别表示单位矩阵与对角矩阵。

(2) $I-4I$ 也是由8维高斯分布的数据样本所组成， 即

$$\mu_1 = \mu_2 = 0 ,$$

$$\sum\nolimits_1 = I_8 ,$$

$$\sum\nolimits_2 = 4I_8 。$$

(3) $I-I$ 数据由p维高斯分布的数据样本所组成，维数p可以变化

$$\mu_1 = 0 ,$$

$$\mu_2 = [1.56,0,...,0]^T$$

(4)Ness数据由p维的高斯数据样本构成：

$$\mu_1 = 0 ,$$

$$\mu_2 = [\Delta/2,0,...,\Delta/2]^T ,$$

$$\sum\nolimits_1 = I_p ,$$

$$\sum\nolimits_2 = \begin{bmatrix} I_{p/2} & 0 \\ 0 & \frac{1}{2} I_{p/2} \end{bmatrix}$$

(5) Mixture-1：由两个 p 维高斯数据样本组成：

$$f(x) = 0.5N(1.8,4\sum) + 0.5N(3,6\sum)$$

$$f(x) = 0.5N(2,4\sum) + 0.5N(2.8,6\sum)$$

(6) Mixture-2：由三个 p 维高斯数据样本组成：

$$f(x) = 0.3N(1,\sum) + 0.3N(3,9\sum) + 0.4N(3,6\sum)$$

$$f(x) = 0.3N(1.5,4\sum) + 0.3N(2.5,4\sum) + 0.4N(3.5,5\sum)$$

在论文中用到的UCI 标准机器学习数据集的数据特征信息在表4-1中。在使用UCI标准机器学习数据集分类的过程中，当其样本的数量比较小的时候可以通过采用交叉验证的方法来估计分类的误差率。在分类过程中，把数据集分成 m 份。在训练过程中用 $m-1$ 份数据来做训练，然后用另一份数据集来做测试。因此，每一份数据集都既是训练数据而且又是测试数据。在分类决策时，取 m 次实验之后的平均值计算。因此在仿真实验时就可以直接在特定的测试数据集进行上测试并对

分类器的分类误差率和性能进行对比。

表 4-1 用到的数据集特点

数据集名称	特征维数	样本数量	类别	测试样本
Pen	16	10992	10	3498
Letter	16	20000	26	4000
Thyroid	21	7200	3	3428
Optdigits	64	5620	10	1797
Wine	13	178	3	--
Iris	4	150	3	--
Banlance	4	625	3	--
Glass	9	214	6	--
Pima	8	768	2	--

4.3 仿真结果

对分类器的分类性能主要由两项分类性能进行评估：

(1) 数据样本的数量对分类准确率的影响：

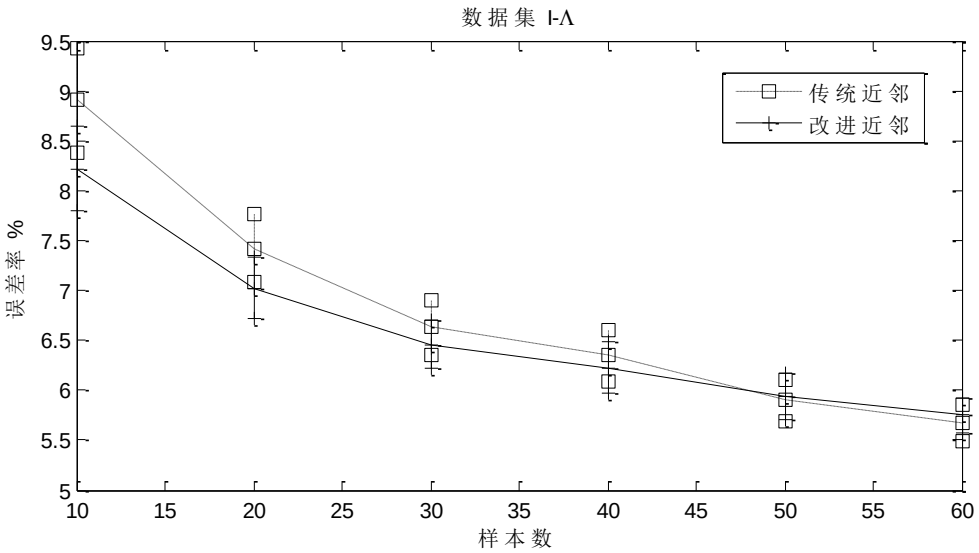


图4-2 $I-\Lambda$ 数据集分类误差率

$I-\Lambda$ 数据样本集由8维高斯分布的数据样本所组成。在图4-2中可以看到在分类时，$I-\Lambda$ 数据集样本的数量对分类准确率的影响。在分类过程中，随着样本数目的增加，不管是传统的近邻分类算法还是改进的近邻分类算法的分类误差率都得到降低。由图中可以明显得看到基于类均值的近邻分类算法的误差率要低于传统的近邻分类。

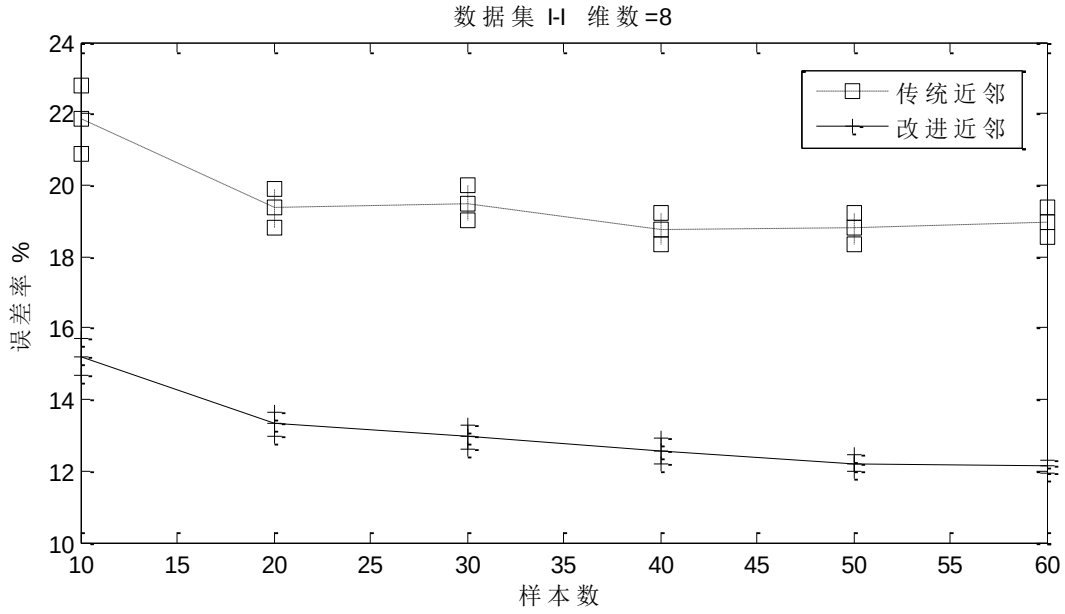


图4-3 数据集 $I-I$ 分类误差率

$I-I$ 数据由8维的高斯分布的数据样本所组成。样本数量从10到60增加，随着样本数目的增加分类误差率都有所降低。并且在 $I-I$ 数据集中，基于类均值的近邻分类算法的误差率要低于传统的近邻分类。

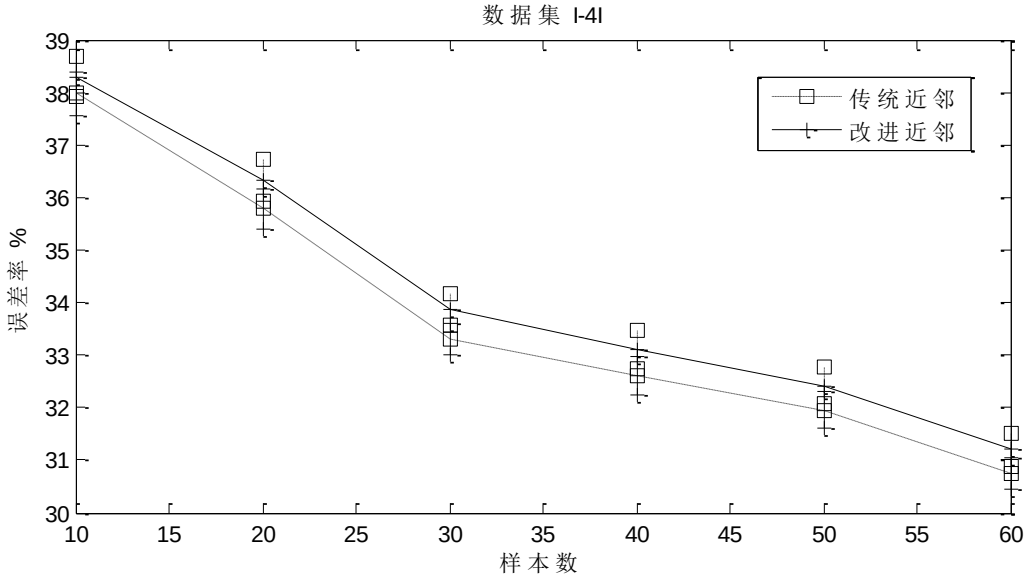


图4-4 数据集 $I-4I$ 分类误差率

在 $I-4I$ 数据集中，当样本数从10增加到60时，两种分类算法的分类误差率都随着样本数目的增加分类性能都有所改善。在 $I-4I$ 数据集中，基于类均值的近邻分类算法的分类性能要优于传统的近邻分类。

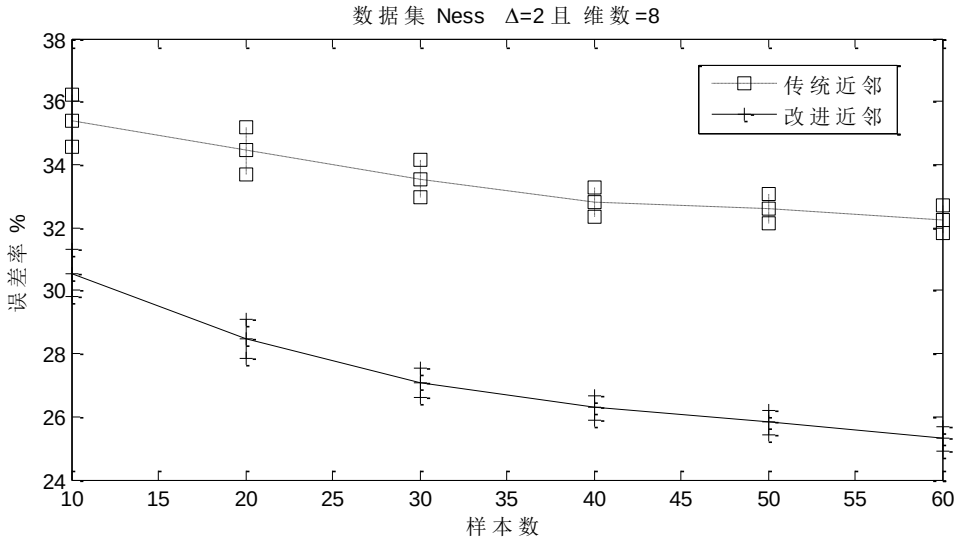


图 4-5 数据集 *Ness* 分类误差率

从上面的数据集测试结果中可以看到，分类器的分类性能随着样本数目的增加，分类误差率也随之下降。并且在大多数数据集的情况下，改进的近邻分类算法的分类误差率要低于传统的近邻分类。

(2) 数据样本的特征维数对于分类性能的影响

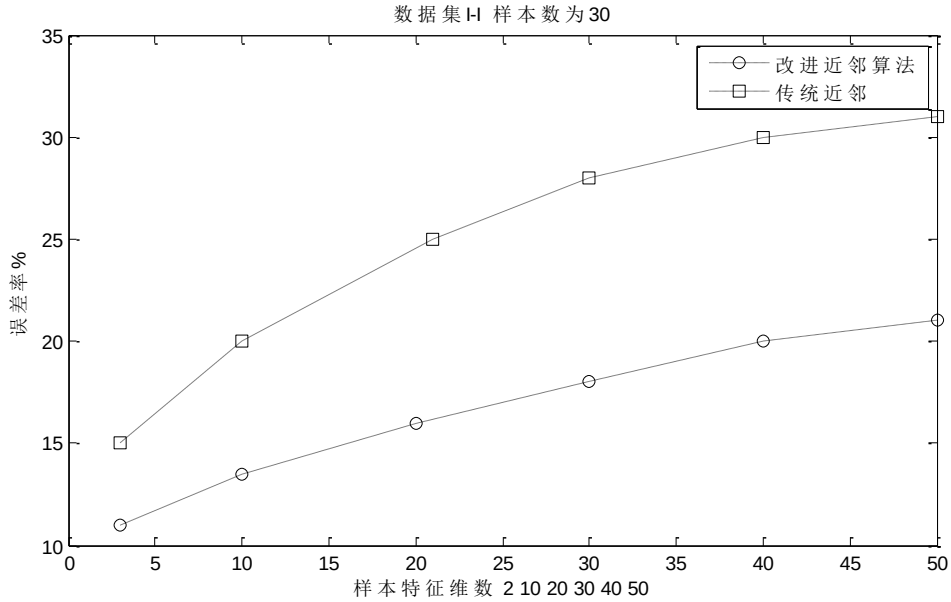


图 4-6 数据集 $I-I$ 维数变化对分类误差率的影响

在 $I-I$ 数据集中，数据集的样本数为 30，样本的特征维数从 2 开始一直增加到 50。在分类过程中，传统的近邻分类算法和改进的分类算法的分类误差率都随着样本维数的增加。但改进的近邻分类算法性能依然优于传统的近邻分类。

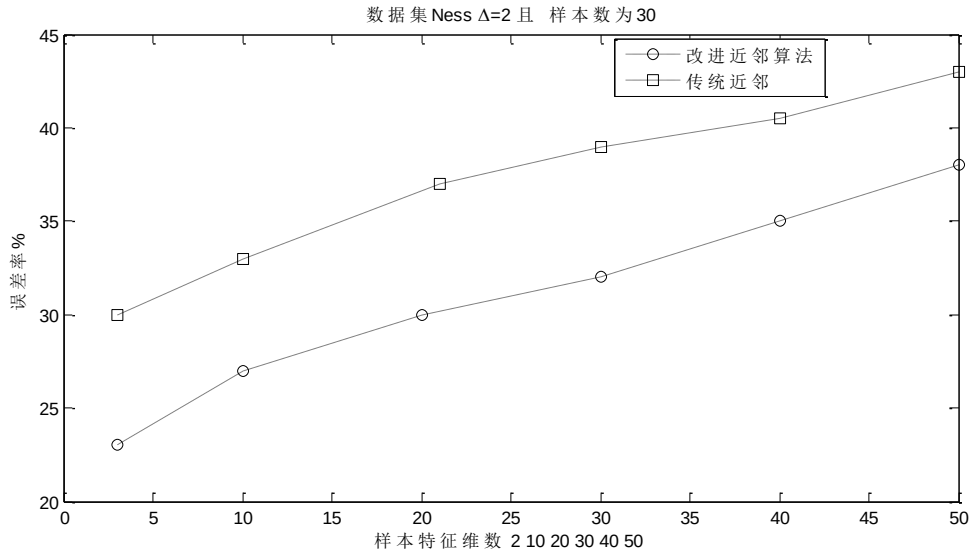


图 4-7 数据集 *Ness* 维数变化对分类误差率的影响

在 *Ness* 数据集中，数据集的样本数 30，$\Delta = 2$。样本的特征维数变化范围从 2-50。在分类过程中，传统的近邻分类算法和改进的分类算法的分类误差率都随着样本维数的增加分类性能降低，系统的分类误差率也随之增大。

从图 4-6 和 4-7 的仿真测试结果来看，当样本数据的维数增加时，基于类均值的 *k*-近邻分类算法与传统的 *k*-近邻分类算法的分类误差率都变大，导致分类准确率会降低。基于类均值的 *k*-近邻分类算法与传统的 *k*-近邻分类算法相比有较好的分类性能。下表 4-2 为两种分类算法在不同的数据集中的分类误差率对比。

表 4-2 分类误差率对比

数据集名称	分类方法	误差率%	置信区间%	距离加权系数
Pen	传统近邻分类	2.26	-	-
	类均值分类	2.24	-	w =0.013
Optdigits	传统近邻分类	2.0	-	-
	类均值分类	1.69	-	w =0.256
Thyroid	传统近邻分类	8.02	-	-
	类均值分类	7.96	-	w =0.013
Pima	传统近邻分类	32.71	32.01-33.8	-
	类均值分类	28.89	28.12-29.74	-
Wine	传统近邻分类	3.72	3.32-4.56	-
	类均值分类	3.30	28.35-29.60	-
Iris	传统近邻分类	5.24	4.32-5.57	-
	类均值分类	4.32	28.25-29.90	-
Glass	传统近邻分类	32.19	30.45-33.45	-
	类均值分类	31.34	30.74-32.82	-

4.4 本章小结

本章阐述了数据的类均值分布对分类准确率的影响，并介绍了基于类均值的 *k*-近邻分类算法。改进的 *k*-近邻分类利用了样本数据集的一些整体信息。通过仿真测试数据比较可以得到，基于类均值的改进最近邻分类的分类正确率要优于传统的最近邻分类。

第五章　基于局部均值的近邻分类

基于局部均值的*k*-近邻分类主要融合了局部平均值信息对测试样本点进行分类，进而提高模式分类的准确率。本章所采用的基于局部均值的分类方法，在对测试样本进行分类时，不但融合了*k*个近邻点与其数据集结构的相似性和分布性，还考虑了每一类别中*k*个近邻点的平均值。在进行分类的过程中，从每一类数据样本中选取到测试样本最近的*k*个近邻点，并且计算出*k*个近邻点的均值。对于测试样本点来说，在其周围的近邻点的分布可以说明数据结构的分布特征。在本章节通过数据仿真测试可以看到，融合了局部平均值的近邻分类较传统近邻分类相比在分类准确率上有所提升。实验结果数据表明，特别是在样本数据数量比较小的情况下基于局部均值分类方法拥有较好的分类性能。

5.1 局部均值算法

基于局部均值的*k*-近邻分类是一种改进的*k*-近邻分类。特别是在样本数据比较小的情况下，基于局部均值的*k*-近邻分类算法可以降低数据样本中离群点的存在的分类误差率。基于局部均值的*k*-近邻分类算法的主要流程是：在分类过程中，利用了训练数据样本的个数和距离。在每一类训练数据样本中选取与测试样本最近邻的*k*个样本，然后对每一类的*k*个样本数据取平均值得到平均样本。最后通过计算测试样本到每一类平均样本之间的距离，将测试样本分为距离最近的那一类。

设将要分类的数据样本集共有 C 类，第 i 类的数据样本有 N_i 个, 并且每个最近邻的数据样本设为 x_{ij} ，那么 $i=1,2,\cdots\cdots,C,\ j=1,2,\cdots\cdots,N_i$ 。从每一个分类的类别中分别取测试数据样本 x 的 k 个近邻。然后再计算出每一类 k 个样本的平均值 Y，可得到：

$$Y_i=\sum_{j=1}^{K}\frac{x_{ij}}{K},i=1,2,\cdots,C \tag{5-1}$$

最后计算 C 个平均样本与测试样本点 x 之间的欧氏距离 $D_i(x)=\|x-Y_i\|_2$，就把 x 分类到距离最小的那一类别中。

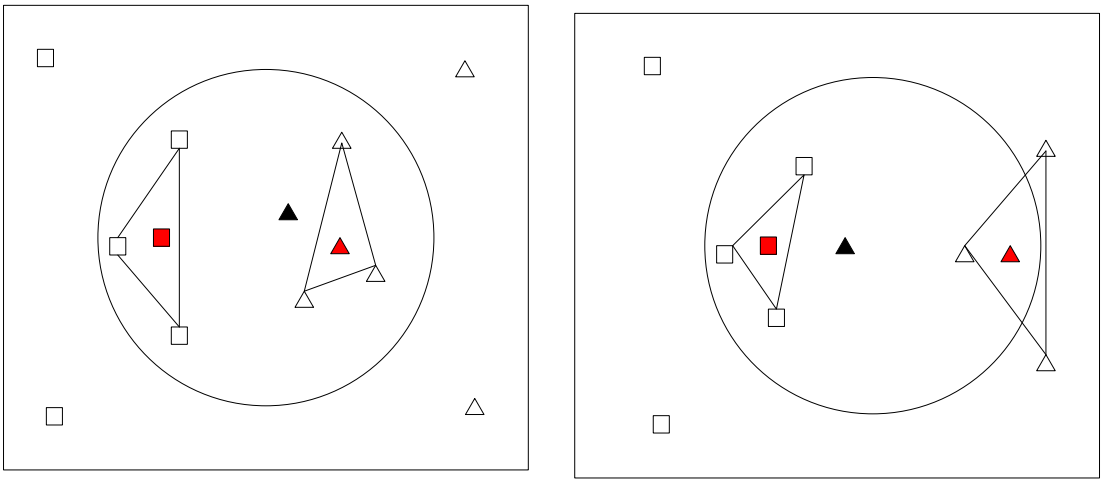
决策规则：

$$lable(x)=\arg\min_i\|x-Y_i\|_2 \tag{5-2}$$

如图 5-1 所示：样本数据集空间一共有两类数据正方形和三角形。测试数据样本为黑色三角形。在分类过程中，分别在每一类别中取距离测试样本点最近的 3 个近邻点。即 k=3，然后分别计算这两类的平均样本值，记为Y_1和Y_2。如下图 5-1 中所示。

再分别计算出Y_1和Y_2到测试样本点 x 之间的欧氏距离，设为若D_1和D_2：如图可以看到，当正方形和三角形都有 5 个样本时，黑三角与红三角距离更近，故将黑三角归类为三角形;

如图 5-1 可知，当正方形和三角形的样本数量不一致的时候，正方形有 5 个而三角形只有 3 个，此时黑三角与红正方形距离更近，故将黑三角归类为正方形。使用近邻分类就出现分类错误。由此可以看到这种分类方法虽然比传统的 k-近邻算法相比利用了距离的信息，提高分类的正确率。但是却无法解决数据样本由于分布不平衡而造成的分类误差。



(a)　　(b)

图 5-1 局部均值算法的分类过程

5.2 基于局部均值的近邻分类

基于局部均值的k-近邻分类算法在分类过程利用了测试样本周围的局部均值信息实现分类判别并融合了样本数据的类统计量。图5-2(a)中为数据的分布图。图5-2(b) 为融合了局部均值信息后的分类结果图。在图5-2中，取出测试数据样本在每类训练数据样本集里的3个近邻点。从图5-2(a)可以得到，根据图5-2(a)中可知，测试数据样本点位于类别2训练数据样本点分布较为密集的区域中，将测试数据样本归类为类别2比较合理。从图5-2(b)中可得，测试样本点和类别1中的局部均值向量的位置重合。因此当近邻点的数量选为3的时候，测试样本的将被分类为类别1。

由于局部均值向量的影响，测试样本点到类别1的组合距离大于测试样本点到类别2的组合距离。采用局部均值的*k*-近邻分类算法将测试样本点判断为类别2。从图5-2(a)中的样本数据的分布情况来看分类正确，并且避免了由于离群点的存在而导致分类精度降低。

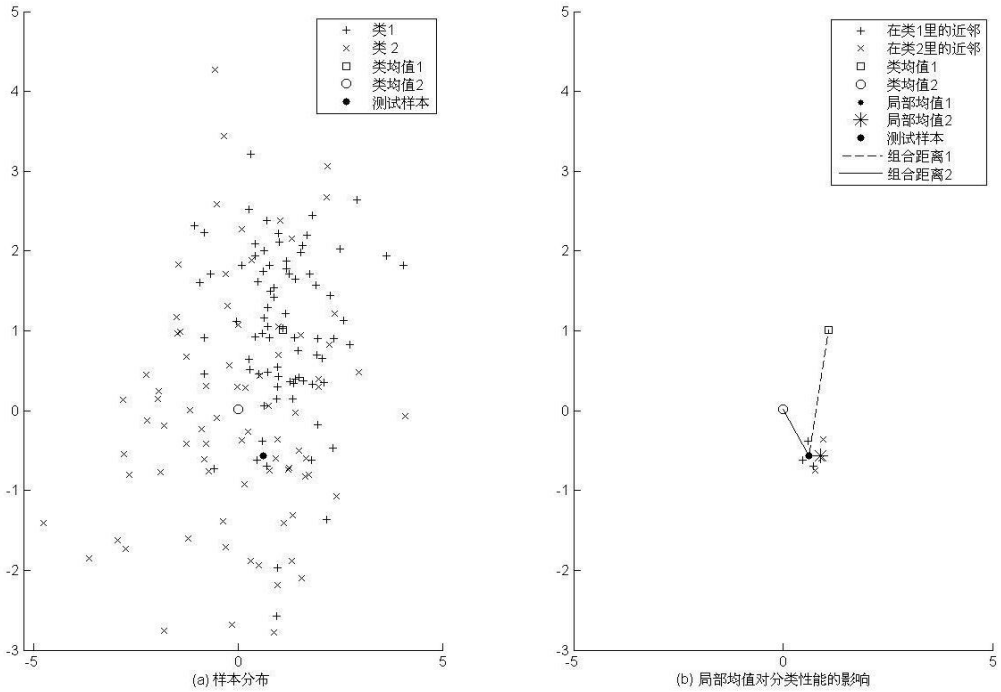


图5-2 局部均值对分类的影响

有N个训练样本，$N_1,N_2...N_m$分别代表对应类别$\omega_1,\omega_2...\omega_m$的训练样本的数量。$x_j^{(1)},x_j^{(2)},...,x_j^{(r)}$代表测试样本在类$\omega_j$中的$r$个近邻。$X_j=\{x_j^i \mid i=1,2...,N_j\}$代表属于类别$\omega_j$的训练样本数据集，$\mu_j$为训练样本集在类别$\omega_j$的均值向量，用$\sum_j$代表训练样本集在类$\omega_j$的协方差矩阵，那么可以得到：

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_j^i \tag{5-3}$$

和

$$\sum\nolimits_j=E[(x-\mu_j)(x-\mu_j)^T] \tag{5-4}$$

基于局部均值的改进最近邻分类算法过程如下所示：

(1)找到测试样本点在类别ω_j中的r个近邻，然后计算出测试点在类别ω_j的局部均值y_j：

$$y_j = \frac{1}{r}\sum_{i=1}^{r} x_j^{(i)} \tag{5-5}$$

(2)计算出测试点在类ω_j到其局部均值点y_j的距离，即局部距离d_j^b：

$$d_j^b = \sqrt{(x-y_i)^T(x-y_i)} \tag{5-6}$$

(3)计算出测试点x到类均值向量μ_j的距离d_j^a：

$$d_j^a = \sqrt{(x-\mu_j)^T\sum_j^{-1}(x-\mu_j)} \tag{5-7}$$

(4)通过两种距离的组合计算组合距离d_j^c：

$$d_j^c = d_j^b + \omega * \ d \quad 0 \le \omega \le 1 \tag{5-8}$$

(5)采用最近邻分类方法进行分类，如果测试样本分类为ω_c并满足以下条件：

$$d_c = \arg\min\{d_j\} \quad \text{j=1,2,...,m} \tag{5-9}$$

在实验中通过采用神经网络通过自适应的方法找出合适的权值w*。下面将使用BP神经网络寻找最优的权值进行介绍。

在实际的应用过程中，BP神经网络的应用范围已扩展到数据挖掘，经济预测、模式识别和模式分类等领域。

BP网络的神经元模型如图5-3所示

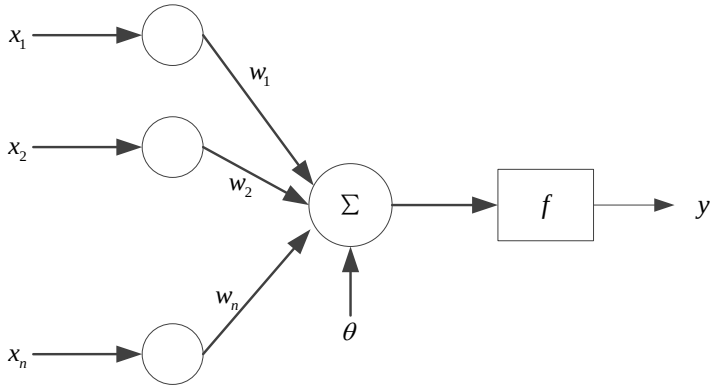


图5-3 BP神经元模型

在应用过程中经常使用的函数有logsig，tansig函数和线性函数。传输函数的输出为

$$a = \log\sin(W_P + b) \tag{5-10}$$

由BP神经元构成的二层网络如图5-4所示。在BP网络属于多层前馈神经网络的一种[42]。当网络的输出层的激励函数为S型函数，那么这个网络的输出范围将在(0,1)之间波动;当传输函数选择为线性函数时，网络的输出可以映射到任意值。

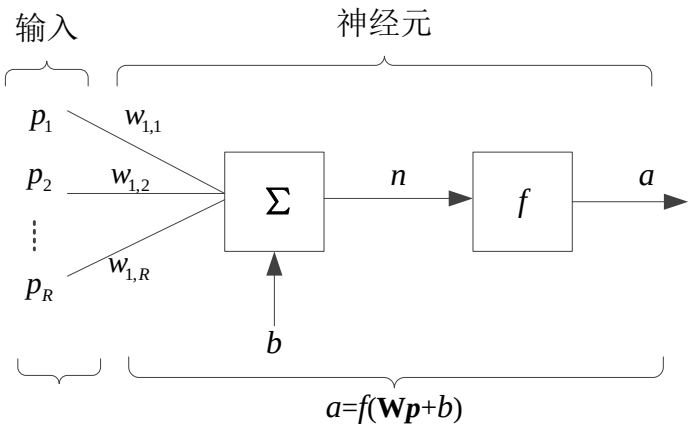


图5-4 BP神经网络模型

BP神经网络的结构确定好了之后，需要利用输入与输出的数据样本集对神经网络训练。在训练过程中不断地对神经网络的中的权值与阈值更新和修正，从而得到输入与输出之间的映射关系。下图5-5就是一个典型的单隐含层神经网络模型图。

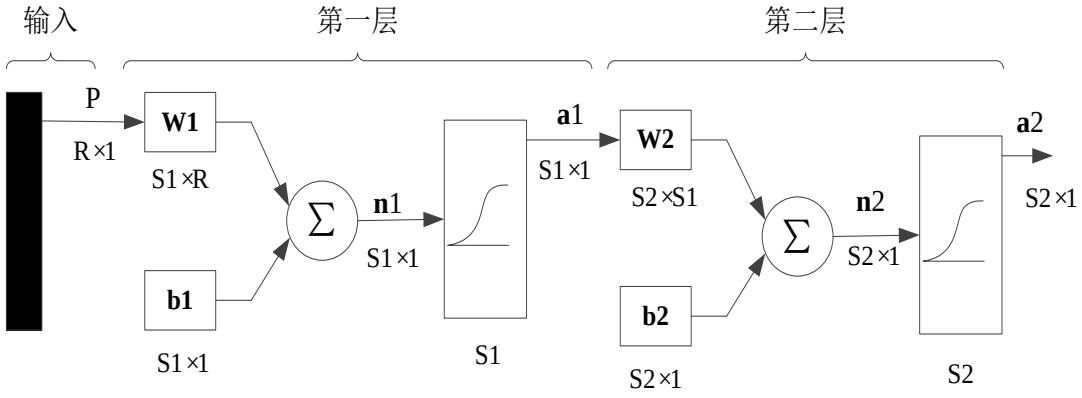


图5-5 单隐含层神经网络模型

BP 神经网络一般具有以下特点：

（1）从网络的结构图我们可以知道，BP 网络是多层网络构成的，同一层的神经元之间无连接。BP 网络多层的结构使其能够处理非线性问题，能逼近任意非线性函数，完成复杂的任务。

（2）网络中数据从输入层经过隐层传递到输出层，但是对权值和阈值的调整，沿着误差减小的方向，从输出层逐层向前，即误差的反向传播。

（3）网络中的传递函数必须是可微的。因为在误差反向传播时，使用了链式法则求导。BP 网络中一般使用 log-sigmoid 函数或线性函数作为传递函数。

BP神经网络的训练学习过程可以分成两步：

第一步是将学习数据样本输入到网络后，通过迭代前的权值与阈值，并且从神经网络的起始层开始分别计算出每个神经元的所对应的输出。

第二步是对神经网络中的阈值和权值重新更新和修正。通过从最后一层开始向前计算出每一层的阈值和权值对网络的总分类结果的影响。以此作为依据对每一个阈值和权值进行更新和修正。

重复第一步和第二步两过程，网络收敛时则停止网络的训练。由于误差之间的传递是通过每一级向前传递，并且不断地更新和修正每一层之间的权值与阈值。所以把这种算法叫做误差逆向传播算法。这种学习算法可以应用到有多个中间层的多层神经网络中，所以这样的神经网络被称为 BP 神经网络标准。所以标准的 BO 算法在实际的应用过程中也表现出一些不足，因此也出现了改进的算法比如变梯度法，牛顿算法和动量 BP 法等。

在数据进行输入到输入层之前需要将数据做归一化处理。常用的两种归一化处理方法：

min-max 标准化，它对原始数据起到一种线性变换的作用，使其结果被映射到[0 - 1]区间。转换函数如下：

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Z-score 标准化处理：

$$x_{avg} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{avg})^2$$

$$x_{new} = \frac{x - x_{avg}}{\delta}$$

图 5-6 展示了数据转化图

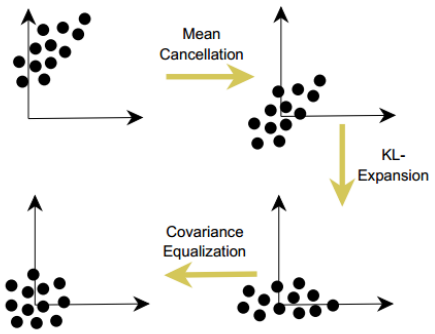


图 5-6 数据转变过程

在BP神经网络的建立过程主要输入层的设计、隐含层和输出层的设计以及激励函数的设计。

1.网络层数

在大部分神经网络的设计过程中首先确定的是网络层的数量。BP网络中拥有不同的隐含层。但是当在模式样本数量相对较大的情况下，就得减小网络的规模，增加隐含层的数量。在实际的应用过程中，隐含层的数目控制在两层以内。

2. 输入层的节点数

输入层中的作用是存储外部的所发出的数据。因此输入层的节点数目是由输入的维数决定。

3. 输出层的节点数

输出层的节点数目由输出的数据类型和数据大小这两个方面所决定。输出通过二进制的形式来表示，

(1)设样本数据集的数量为m，第 j 个节点的输出为1，可表示如下

$$O=\frac{[00...010...00]}{j}$$

如果输出的结果全为0，测试样本点不属于已知的任一个类别。

(2) 输出层的节点数目为 $\log_2^m$ 个。

4. 隐含层的节点数

一个两层的BP网络在拥有无限隐含层的节点的情况下，输入与输出之间可以实现非线性的映射。因此对于隐含层的节点数目的选择上，由于复杂性的原因到目前还没有找到较为合理的解析式。对于隐含层中的节点数目，一般采用下面的公式进行设计:

$$n=\sqrt{n_i+n_o}+a \tag{5-11}$$

在这个表达式中，n 代表隐含层的节点数；n_i 代表输入层节点数目；n_o 代表输出层节点数目；a 一般取 $1-10$ 之间的常数。

5. 传输函数

(1)阶跃传输函数。当输入的范围小于0的时候，输出将为0。当输入的范围大于0的时候，函数的输出为1.

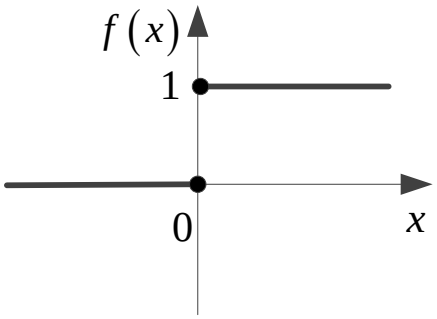


图 5-7 阶跃传输函数

(2) 线性传输函数可以实现输入和输出之间的线性映射。

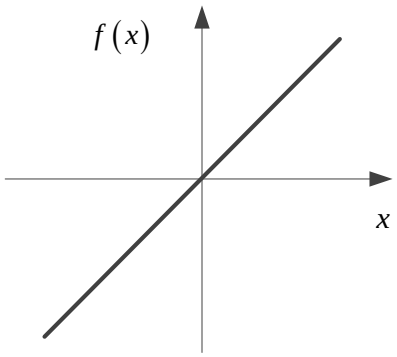


图 5-8 线性传输函数

(3) 对数S型传输函数。它是一个从实数域R到[0,1]的函数，表示状态连续型神经元模型，最常用的是单极性log-sigmoid函数，简称对数型函数，它本身及其导数都是连续函数，能很好的体现数学计算中的优越性

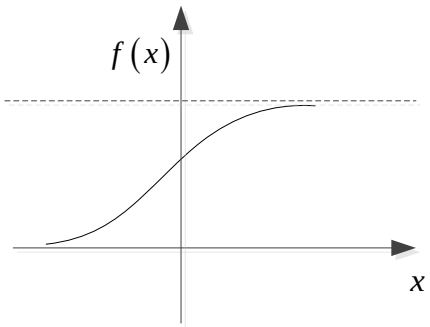


图5-9 S型函数曲线

S型函数在BP网络中应用范围比较广，图5-9为S型函数曲线：

$$f(x)=\frac{1}{1+e^{-x}} \tag{5-12}$$

在神经网络的训练过程中选择合适的训练算法也对网络的性能有前着重要的影响。下面将分别介绍常用的学习算法。

1，动量 BP 方法

动量 BP 法是在标准 BP 算法的权值更新阶段加入动量因子 γ（0<γ<1）：

$$\Delta \boldsymbol{W}_m(k+1)=\gamma\Delta \boldsymbol{W}_m(k)-(1-\gamma)\alpha \boldsymbol{s}_m(\boldsymbol{a}_{m-1})^T \tag{5-13}$$

$$\Delta \boldsymbol{b}_m(k+1)=\gamma\Delta \boldsymbol{b}_m(k)-(1-\gamma)\alpha \boldsymbol{s}_m \tag{5-14}$$

表示本次权值的调整方向不仅与本次计算的梯度有关，还与上一次的调整方向和幅值有关。动量因子的加入，使权值的调整有一定的惯性和抗震荡的能力，加快了收敛。

2.梯度下降法

一维梯度下降的公式可以写成：

$$W(t+1)=W(t)-\eta\frac{\partial E}{\partial W} \tag{5-15}$$

假设目标函数 E 可以被二次函数逼近，经泰勒展开式展开：

$$E(W)=E(W_C)+(W-W_C)\frac{dE(W_C)}{dW}+\frac{1}{2}(W-W_C)^2\frac{d^2E(W_C)}{dW^2}+... \tag{5-16}$$

其中，$\frac{dE(W_C)}{dW}=\frac{dE}{dW}|_{W=W_C}$

两边对W求导：

$$\frac{dE(W_C)}{dW}=\frac{dE(W_C)}{dW}+(W-W_C)\frac{d^2E(W_C)}{dW^2} \tag{5-17}$$

设$W=W_{\min}$，因此得出$\frac{dE(W_C)}{dW}\Big|_{W=W_{\min}}=0$

$$W_{\min}=W_c-(\frac{d^2E(W_C)}{dW^2})^{-1}\frac{dE(W_C)}{dW} \tag{5-18}$$

因此学习率$\eta=(\frac{d^2E(W_C)}{dW^2})^{-1}$。

但是在多维里，要获得一个最优学习率是非常难的一件事。令$\frac{d^2E(W_C)}{dW^2}=H$，那么 H 就不再是一个数而是矩阵，H 被称作为 Hessian 矩阵。

$$H_{i,j}=\frac{\partial^2 E(W_C)}{\partial W_i \partial W_j}$$

牛顿法和上式(5-18)类似：

$$\Delta W=H^{-1}\frac{dE(W_C)}{dW} \tag{5-19}$$

对于非线性共轭梯度法，这里介绍 Fletcher-Reeves 方法，算法流程：

已知 x_0

估计 $f_0=f(x_0),\nabla f_0=\nabla f(x_0)$

设置 $p_0\leftarrow-\nabla f_0,k\leftarrow 0;$

While $\nabla f_k\neq 0$

Compute α_k and set $x_{k+1}=x_k+\alpha_k p_k$

Evaluate ∇f_{k+1}

$$\beta_{k+1}^{FR}\leftarrow\frac{\nabla f_{k+1}^T\nabla f_{k+1}}{\nabla f_k^T\nabla f_k}$$

$$p_{k+1}\leftarrow\nabla f_{k+1}+\beta_{k+1}^{FR}p_k$$

$$k\leftarrow k+1$$

End(while)

计算 α_k 是根据 Strong Wolfe Conditions 来选择的。

$$f(x_k+\alpha_k p_k)\le f(x_k)+c_1\alpha_k\nabla f_k^T p_k$$

$$\left|\nabla f(x_k+\alpha_k p_k)\right|\le -c_2\nabla f_k^T p_k$$

其中 $0<c_1<c_2<0.5$。

BP 神经网络的学习过程如图 5-10 所示：

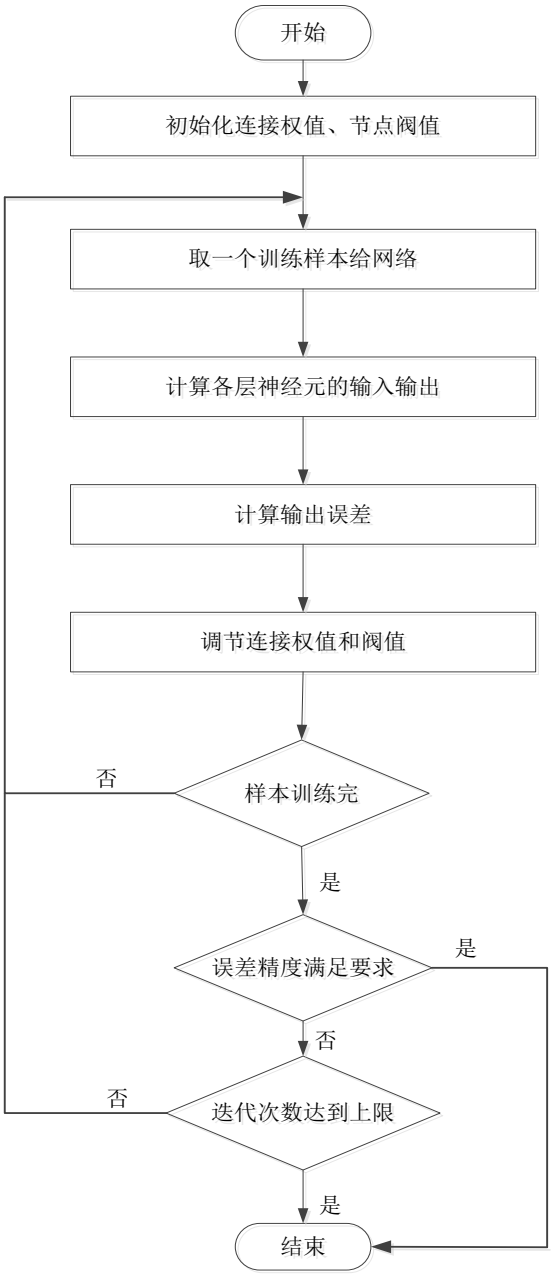


图5-10 BP网络训练流程

网络通过样本集的训练，最终得到了网络各层的权值和阀值。

5.3 实验数据方案

在实验过程中采用两种类型不同的数据样本集来测试传统 *k*-近邻分类和基于局部均值的 *k*-近邻分类算法的分类性能。在实验仿真测试过程中使用的第一种数据是 UCI 标准机器学习库的数据。UCI 标准机器学习数据库是由加州大学欧文分

校(University of California Irvine)提出的用于机器学习的数据库。第二种数据样本集是人工生成的数据集 $I-\Lambda$，$I-4I$，$I-I$ ，*Ness*，Mixture-1 和 Mixture-2。在实验仿真测试过程中，每个样本数据样本集中有 2000 个数据样本，而每一类中有 1000 个数据。对每一个数据样本进行重复实验 100 次，每次实验后则重新产生新的数据。

在仿真测试实验中，对传统 *k*-近邻分类和基于局部均值的 *k*-近邻分类两种不同的分类算法，主要从两项分类性能进行评估：

(1) 训练数据样本的数量对于分类性能的影响;

(2) 数据样本的特征维数对分类准确率的影响。

在实验中，每一类样本数据集中的数量范围从 10 到 60，数据样本集的特征维数范围 2 到 50。

5.4 仿真结果

1.训练数据样本的数量对于分类性能的影响：

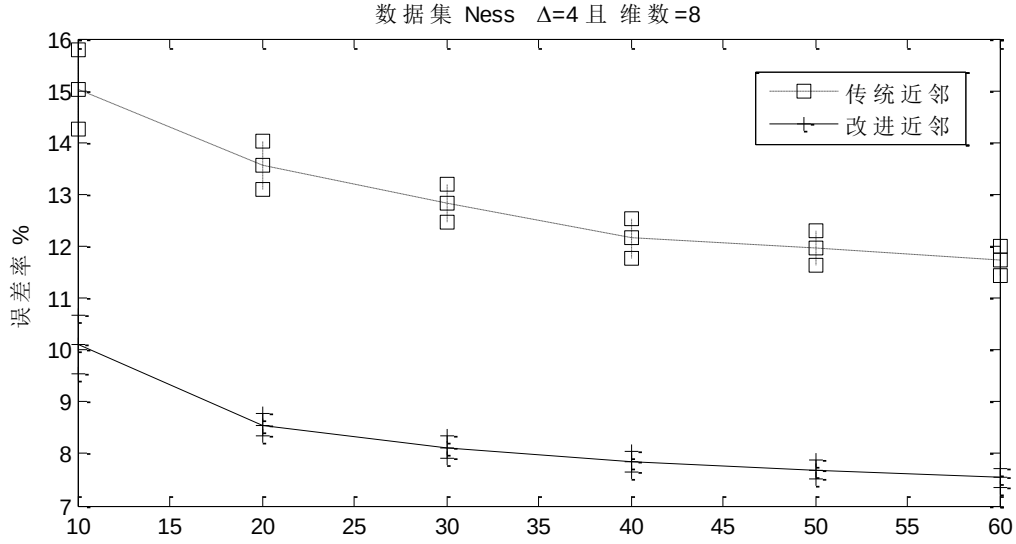


图5-11 样本数变化对数据集*Ness*的影响

对于样本集*Ness*,$\Delta=4$，并且维数为8。Ness样本集数目从10开始增加到60。从图5-11可以看到，当样本数目开始增加时，传统近邻分类算法和局部均值近邻算法的分类误差率都有所下降。并且从图5-11可以看到，改进的近邻分类的分类误差率低于传统的近邻分类算法。

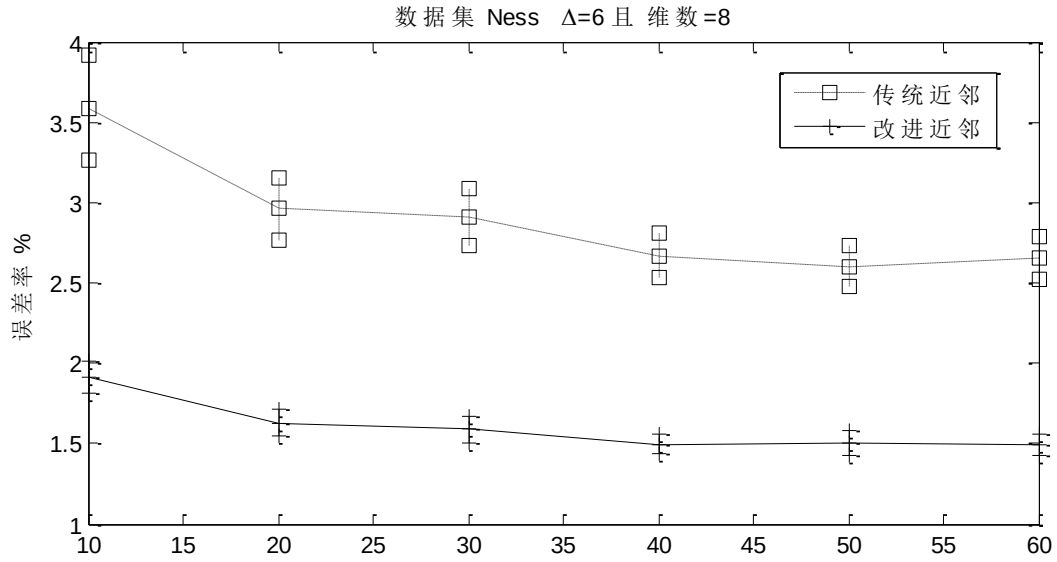


图5-12 数据集 *Ness* 分类误差率

此时测试同样采用*Ness*数据集，$\Delta=6$，此时的维数为8。从图5-12可以看到当样本数在10到20范围变化时，基于局部均值近邻分类的误差率下降较快。当样本数大于20，基于局部均值近邻分类的误差率基本上趋于稳定，并不随着样本数的增加而改变。在一定范围内，样本数的增加可以改善分类的性能。并且可以从上图得到改进后的近邻分类的分类误差率也同样低于传统的近邻算法。

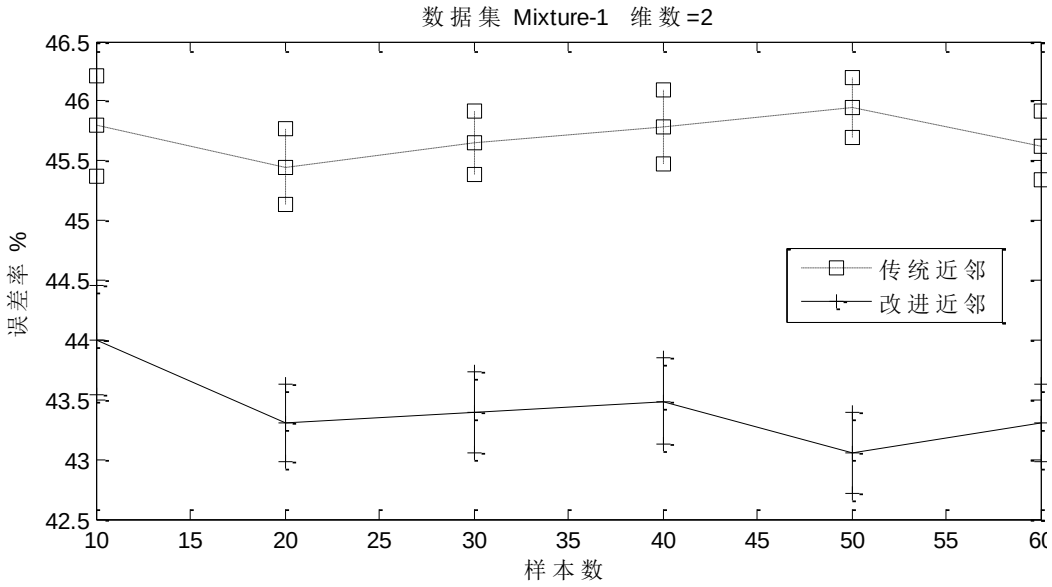


图5-13 样本数变化对分类误差率影响

在上图5-13中，测试采用Mixture-1数据集，并且数据集的维度为2。从图中可以看到，随着样本数目的增加，传统的近邻分类算法和基于局部均值的近邻分类算法的分类误差率都有所降低。但在样本数目增加的过程中，传统的近邻分类算法和改进的算法的分类误差率都出现波动。但从整体趋势来看，样本数据的增加有助于分类性能的改善。并且改进后的分类算法的分类误差率也低于传统的近邻分类算法。

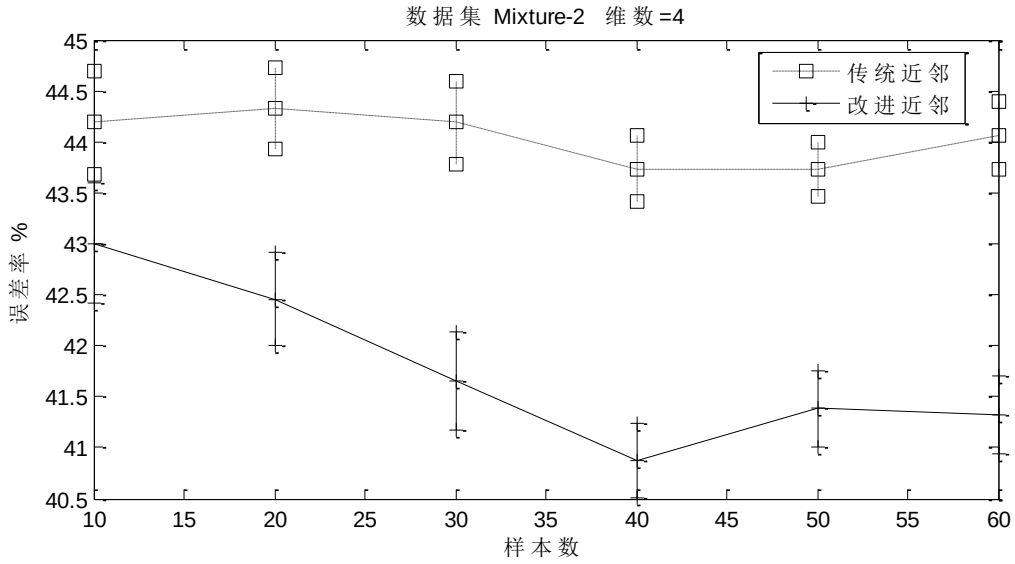


图5-14 样本数变化对分类误差率影响

此时测试采用Mixture-2数据集，并且数据集的维度为4。当样本数据集的数量增加的时候，改进的近邻分类算法的分类误差率有着明显的下降。当样本数量从10增加到40的时候，基于局部基值的近邻分类算法的分类误差率迅速下降。当样本数量在40到60之间增加的时候，改进后的近邻分类算法的分类误差率出现波动，分类误差率有所增加。但从整体变化上来看，数据集样本数量的增加使得改进后的分类性能有所提升。但对于传统的近邻分类算法，样本数据集的数目的增加对于Mixture-2数据集的性能影响不大。并且改进后的近邻分类性能也优于传统的近邻分类。

从上面的数据集测试结果中可以看到，在分类过程中分类器的分类性能随着样本数目的增加，有助于提高分类的正确率。并且在大多数数据集的情况下，基于局部均值的近邻分类算法的分类性能要优于传统的近邻分类。

2. 数据样本的特征维数对分类准确率的影响

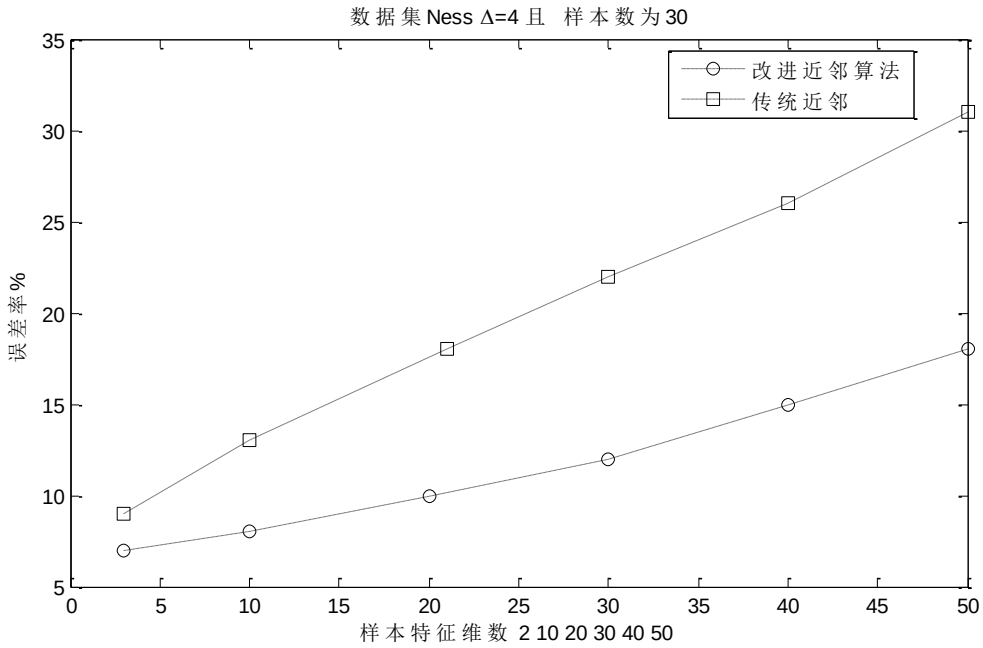


图 5-15 数据集维数变化对分类误差率的影响 1

图 5-15 测试中采用 *Ness* 数据集，样本集的数量为 30，$\Delta=4$。数据集的维度变化从 2 开始增加到 50。从图中可以看到当数据的维度增加时，传统的近邻分类和基于局部均值的近邻分类的分类误差率都增大。

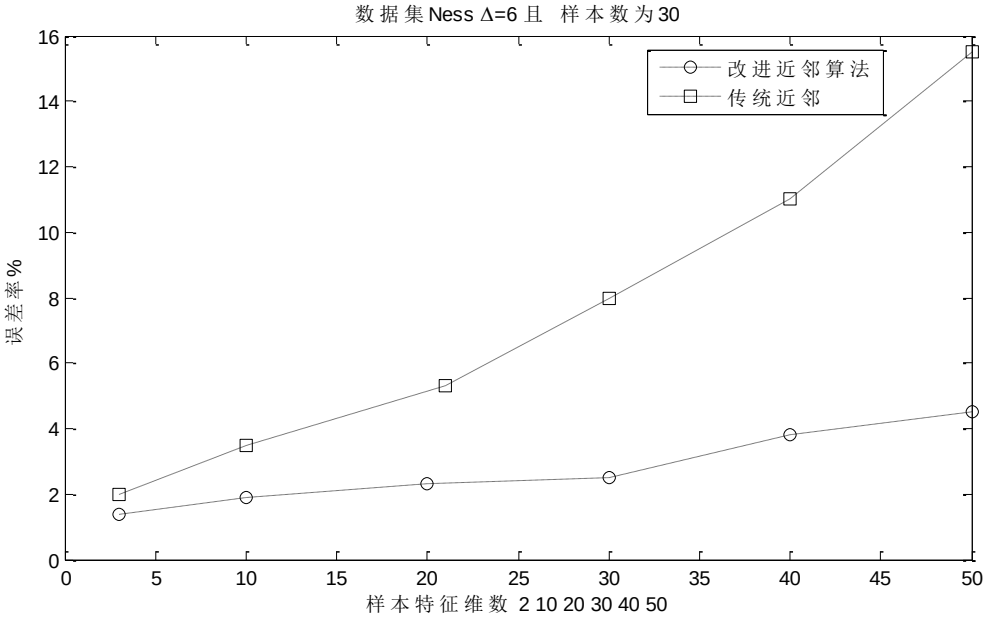


图 5-16 数据集维数变化对分类误差率的影响 2

在图 5-16 中，仿真测试采用的 *Ness* 数据集的样本数为 30，此时 $\Delta = 6$。样本的特征维数从 2 增加到 50。当数据样本的维数增加时，传统的近邻分类和基于局部均值的近邻分类的分类误差率都变大。当数据集的维数从 2 增加到 30 时，基于局部均值的的分类误差率变化比较缓慢。当维数大于 30 时，基于局部均值的的分类误差率增大。但对于传统的近邻分类算法，随着数据集的维数的增加，其分类误差率有着明显的增大。并且从图中可以看到，总体上基于局部均值的分类算法的分类误差率低于传统的近邻分类算法。

下表 5-1 是基于局部均值近邻分类与传统近邻分类误差对比表。在 7 个 UCI 测试数据集中，7 个数据集中基于局部均值的 *k*-近邻分类算法的分类误差率低于传统的 *k*-近邻分类算法；

表 5-1 分类误差对比表

数据集	误差率(%)		距离
	传统分类	局部均值分类	
pen	2.12 *k*=4	1.76	马氏距离
Letter	4.12 *k*=3	3.68	马氏距离
Thyroid	6.33 *k*=5	5.65	欧氏距离
pima	25.12	24.85	欧氏距离
Wine	30.72	4.78	马氏距离
Iris	2.67	2.12	马氏距离
Balance	17.31	16.45	马氏距离

在仿真测试实验中，对传统 *k*-近邻分类和基于局部均值的 *k*-近邻分类两种不同的分类算法，主要从两项分类性能进行评估：

(1) 训练数据样本的数量对于分类性能的影响

在分类过程中，从图 5-11 到图 5-14 研究了样本数量增加对分类误差率的影响。通过仿真测试结果可以看到，对于大部分测试数据集，当样本的数量增加时，使用传统的近邻分类算法和基于局部均值的改进近邻算法可以提高系统的分类精度。而且从图中的测试结果可得出基于局部均值的改进近邻算法的分类误差率高于传统的近邻分类。

(2) 数据样本的特征维数对分类准确率的影响。

从图 5-15，图 5-16 可以看出，当样本数据的维数增加时，基于局部均值的 *k*-近邻分类算法与传统的 *k*-近邻分类算法的分类误差率都变大，导致分类准确率会降低。基于局部均值的 *k*-近邻分类算法与传统的 *k*-近邻分类算法相比有较好的分类性能。

5.5 本章小节

本章采用了一种改进的近邻分类算法—基于局部均值近邻分类方法。算法的主要思想是对测试样本最近的局部均值点所对应的类别来进行分类。这种算法不但降低了离群点对分类精度的影响，而且还提高了分类的精度。为了验证算法的性能，在仿真测试中分别采用 UCI 数据集和人工生成的数据集与传统的 *k*-近邻分类做了大量的对比测试实验。实验数据表明，基于局部均值近邻分类方法在训练数据的数量和近邻点的个数选择变化时，其分类结果都明显优于传统的近邻分类。

第六章　总结和展望

6.1 总结

近邻分类是一种经典的非参数分类方法。在模式识分类领域中有着广泛的应用。本文分析了传统 *k*-近邻分类的优缺点后，根据 *k*-近邻分类的特点提出了基于局部均值的近邻算法和基于类均值的近邻算法。本文工作如下：

(1) 提出基于类内近邻距离加权的改进伪近邻分类算法。考虑测试样本的多个近邻，距离近的对归属该类的影响较大，因而拥有的权值较大。

(2) 提出了基于类均值的最近邻分类算法。在分类过程中，分类精度容易受到离群点的影响。采用基于类均值的最近邻分类，利用每类样本的类均值信息，降低了离群点对分类精度的影响。

(3)提出了基于局部均值的近邻分类算法。在采用近邻分类的过程中，训练样本的数量比较少从而导致分类的准确率降低。为了提高近邻分类的分类性能，利用测试样本在每类训练样本集的 *k* 个近邻的均值信息，从而提高在小样本下的分类精度，同时防止训练时过拟合。

本文采用的分类方法不同于传统的分类方法而是一种组合分类的方法。通过利用数据样本之间的互补信息来提高分类准确率。

6.2 展望

模式分类在数据挖掘领域中有着非常重要的应用价值。在模式分类中有诸多分类算法，其中*k*-近邻分类算法在实际分类过程中应用范围比较广。*k*-近邻分类算法由于效率高等特点从而在实际中有着广泛应用。虽然本文通过融合基于局部均值和类均值的方法提高了传统近邻分类的分类性能和准确率，但是在实际的分类应用中，仍有大量工作需要在后续的研究中继续完成。未来的改进研究方向大致可以从以下几个方面展开：

1. 其他算法与 *k* -近邻算法的融合

接下来的工作就是通过加强对其它分类算法的学习，尝试将其它的分类算法融合到 *k*-近邻算法计算中，以此来提高分类精度。

2. 距离度量的研究

在模式识别领域中对距离度量的研究非常关键，怎样依据不同的样本数据结

构从而选择适当的距离度量[43]。这也是是对 *k*-近邻算法进行改进的一个研究热门的方向。在距离度量中，对于各数据样本之间的相似性属性，还需要在未来的研究中进一步提升。

3. 对训练样本的优化

传统的 *k*-近邻算法一般在数据规模比较小时，能够取得较好的分类性能。在实际应用中如果数据样本的规模很大并且维度高，那么传统的 *k*-近邻分类算法将会耗费大量的时间计算量也特别大。把范围控制在测试数据样本最相关的那个小邻域上，然后再在小邻域上采用 *k*-近邻分类算法从而使算法的扩展性和适用性更强。

致　谢

研究生三年时光匆匆过去，即将告别校园生活，走上工作岗位。首先感谢我的导师朱宏教授，给我提供了优秀的学习平台，让我在科研道路上得到了锻炼和提高。朱老师平易近人，在科研上求真务实、态度严谨。三年的时间，我学到了很多知识使我终生受益。同时感谢教研室的张榆平老师、曾勇老师、杨忠孝老师和胡江平老师，对我的学习和生活给予的帮助和指导。

感谢教研室的所有的师兄师姐们对我的帮助，同时感谢和我一起学习的同学，让我在电子科技大学度过了三年丰富多彩的快乐时光。

衷心地感谢我的父母在我的求学路上的支持和鼓励。

最后感谢参加评阅的各位老师！

参考文献

[1] 边肇祺,张学工.模式识别（第二版）[M]. 清华大学出版社,2000,9-15.

[2] Tan S.-B. Neighbor-weighted K-nearest neighbor for unbalanced text corpus[J]. Expert Systems with Applications. 2005,667-671.

[3] Mejdoub M,Amar CB. Classification improvement of local feature vectors over the KNN algorithm[J]. Multimedia Tools & Applications,2013,64(1):197-218.

[4] Pan R, Dolog P, Xu G. KNN-Based Clustering for Improving Social Recommender Systems[J]. Lecture Notes in Computer Science, 2013.

[5] Aslam M W,Zhu Z,Nandi A K. Automatic modulation classification using combination of genetic programming and KNN[J].Wireless Communications IEEE Transactions on,2012, 11(8):2742 - 2750.

[6] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[J]. Lecture Notes in Computer Science, 2004.

[7] Jiang L, Cai Z, Wang D, et al. Bayesian Citation-KNN with distance weighting[J]. International Journal of Machine Learning & Cybernetics, 2013, 5(2):193-199.

[8] 苟建平. 模式分类的 K-近邻方法[D]. 电子科技大学,2013.

[9] Cover T. M. and Hart P. E.. Nearest neighbor pattern classification[J]. IEEE Trans.Inform. Theory,13(1):21-27,1967.

[10] Tan P.N,Steinbach M. and Kumar Introduction to Data Mining[M]. Person Education，2005,25-32.

[11] Domeniconi C,Peng J.and Gunopulos Locally adaptive metric nearest-neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24(9): 2002,112-130.

[12] Yang L. Distance-preserving projection of high-dimensional data for nonlinear dimensionality reduction[J].IEEE Transactions on Pattern Analysis and Machine Intelligence.2004：1243-1246.

[13] 余鹰,苗夺谦等. 基于变精度粗糙集的 KNN 分类改进算法[J]. 模式识别与人工智能,2012,25(4):617-623.

[14] Peng J.Heisterkamp D. R. and Dai H. K. Adaptive quasiconformal kernel nearest neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004: 656-661.

[15] Tan, Pang-Ning, Steinbach, et al. Introduction to Data Mining[J]. University of Minnesota, 2006.

[16] Zhang M, Zhou Z. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038–2048.

[17] Tan P. N, Steinbach M. and Kumar V. Introduction to Data Mining[J]. Person Education，2005,12-18.

[18] 王茜，杨正宽.一种基于加权 KNN 的大数据集下离群检测算法[J]. 计算机科学，2011, 38(10):177-180.

[19] 李荣陆,胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展,2004,(04):539-544.

[20] 从爽.面向 MATLAB 工具.箱的神经网络理论与应用,合肥,中国科学技术人学出版社.1998，56-60.

[21] Nello Cristianini,John Shawe Taylor. 支持向量机导论 [M]. 李国正,王猛曾,华军译. 北京：电子工业出版社,2004.3,110-125.

[22] K. Wong W, Cheung D W, Kao B, et al. Secure KNN computation on encrypted databases.[J]. Proceedings of the Acm Sigmod International Conference on Management of Data, 2009: 139-152.

[23] 张著英,黄玉龙,王翰虎. 一个高效的 KNN 分类算法 [J]. 计算机科学,2008,170-172.

[24] Song Y,Huang J,Zha H,et al. KNN:Informative K-Nearest Neighbor Pattern Classification[J]. Knowledge Discovery in Databases Pkdd, 2007:248--264.

[25] 桑应宾. 基于 K 近邻的分类算法研究[D]. 重庆大学,2009.

[26] 刘晓东,刘国荣,王颖,席延军. 散乱数据点的 k 近邻搜索算法[J]. 微电子学与计算机. 2006，23-26 .

[27] 黄剑华,丁建睿,刘家锋等. 基于局部加权的 Citation-KNN 算法[J]. 电子与信息学报，2013, (3):627-632.

[28] Liu W, Chawla S. Class Confidence Weighted KNN Algorithms for Imbalanced Data Sets[J]. Lecture Notes in Computer Science, 2011, 6635(2):345-356.

[29] Paredes R. and Vidal E. Learning weighted metrics to minimize nearest-neighbor classification error[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,11-14.

[30] Peng J., Heisterkamp D. R. and Dai H. K. Adaptive quasiconformal kernel nearest neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004，656-661.

[31] 陈振洲,李磊,姚正安. 基于 SVM 的特征加权 KNN 算法[J]. 中山大学学报:自然科学版，2005,44(1):17-20.

[32] Lei Z, Jian Wei L, Xiong Lin L. KNN and RVM Based Classification Method:KNN-RVM Classifier[J]. Pattern Recognition & Artificial Intelligence, 2010.

[33] Chen T. S. and Chang C. C..Diagonal Axes Method (DAM):A Fast Search Algorithm for Vector Quantization[J]. IEEE Trans. Circuits and Systems for Video Technology,1998,555-559

[34] He-ping G, Yong-xia J, Bai-ming F, et al. An Improved KNN Text Categorization Algorithm Based on DBSCAN[J]. Science Technology & Engineering, 2013.

[35] Min R, Stanley D A, Yuan Z, et al. A Deep Non-linear Feature Mapping for Large-Margin kNN Classification[J]. Proc.of the Int.conference on Data Mining, 2009:357-366.

[36] 宫秀军,刘少辉,史忠植.主动贝叶斯网络分类器[J].计算机研究与发展,2002,574-579.

[37] Hu Q,Yu D,Xie Z. Neighborhood classifiers[J]. Expert Systems with Applications,2008，34(2):866–876.

[38] AYM,AWD,BDT. Local Discriminative Distance Metrics Ensemble Learning[J]. Pattern Recognition,2013,46(8):2337–2349.

[39] 贾宇平,付耀文,庄钊文. 基于 *k* 近邻决策分布图的决策层融合目标识别[J]. 系统工程与电子技术. 2005,121-123.

[40] 聂方彦. 基于 PCA 与改进的最近邻法则的异常检测[J]. 计算机工程与设计. 2008，252-254.

[41] 邱天爽,杨春晖. 一种基于改进近邻分类器的人脸识别方法. 信号处理[J]. 2008,54-57.

[42] 周开利,康耀红. 神经网络模型及其 MATLAB 仿真程序设计[M] 清华大学出版社,2000.

[43] 曾勇. 广义近邻模式分类研究[D]. 上海交通大学, 2009.

攻读硕士期间取得的研究成果

[1] 杨忠孝,梁洲,张陈方等.智能型直流无位置传感光伏水泵驱动控制器[P]专利编号:ZL2014205572,中国,2014