中南大学

硕士学位论文

基于多特征的图像分类决策树生成方法研究

姓名：胡樱

申请学位级别：硕士

专业：计算机应用技术

指导教师：罗三定

20070401

摘 要

如何合理而又高效地组织图像数据、并结合图像底层特征，将人工智能及知识发现等技术运用于图像分类中，是计算机视觉研究领域的一个热点问题。

本文在利用计算机视觉观察和理解世界、并以某种程度的智能完成特定任务的研究中做了一些基础工作，对图像特征提取、分类特征选择及分类规则提取等理论和方法进行了研究，促进了图像分类技术与机器视觉、数据库及知识发现等技术的结合。

本文所做的研究工作包括以下几个方面：

图像分类中需要选取出具有平移、缩放和旋转不变性的特征，本文首先描述了一些常用的图像颜色、纹理和形状特征，并对这些特征进行了分析

然后，分析了决策树算法中的两种分类特征选择标准，提出了基于类间特征整体分布关系的特征选择标准。通过基于类间差异的分类误差定义，给出了若干分类特征相关的概念定义，为后续决策树构建研究建立理论基础。提出了基于样本整体分布的决策树生成算法，并对算法进行了性能分析和优化，同时也讨论了多类对象数目的增减对决策树的影响。

接着分析了特征空间中不同区域内的分类规则，通过在决策树的构建过程中引入先验知识，提出了带误差权值的决策树、以及建立在其基础之上的分类规则提取方法。

最后，通过实验对本文提出的分类方法进行验证，并结合实验结果评价了分类算法的错误率、精度和可理解性。

关键词：图像处理，分类误差，决策树，多特征分类

ABSTRACT

It's a hot problem in computer vision research area that how to organize the image data reasonably and effectively and combining the image low figures how to use artificial intelligent and knowledge discovery in the images classification.

In order to observe and understand word with computer vision well and then accomplish the certain task in some extent intelligence, some base work has done in this paper. It is researched that the theory and method of image figure extraction, classification figures choose and rules acquisition.

The researches of this paper are as follows:

Firstly, color, texture and shape figures of image are discussed. And the figures with shift, scale and rotation invariance are used as the initial figure set.

Secondly, two indexes of classification figures choose in decision tree algorithm are analyzed and the figures choose standard based relationship of figure distribution among various classes is proposed. Through the classification error definition based on distinct among classes, the concepts about classification figure are defined and as the theory basis for the decision tree generation.

Thirdly, it proposes the decision tree generation algorithm based on samples total distribution, then analyses the performance of algorithm and optimizes the algorithm. At the same time, it also discusses the influence of decision tree with the fluctuation of objects number.

Fourthly, it analyses the classification rules of different fields of figure space. In the process of decision tree generation, the prior knowledge is inducted. And the classification rules extracted method with the weighted decision tree is proposed.

At last, this paper will testify the accuracy of the classifying thought and evaluate the error rate, precision and understandability of the classifying algorithm by a group of images.

KEY WORDS: image process, classification inaccuracy, decision tree, multi-figure classification

原 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。尽我所知，除了论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，也不包含为获得中南大学或其他单位的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

作者签名：胡樱 日期：2007年5月10日

关于学位论文使用授权说明

本人了解中南大学有关保留、使用学位论文的规定，即：学校有权保留学位论文，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以采用复印、缩印或其它手段保存学位论文；学校可根据国家或湖南省有关部门规定送交学位论文。

作者签名：胡樱 导师签名 日期：2007年5月10日

第一章 绪论

1.1 选题背景

随着基于 Internet 的图像数据海量增长和广泛应用，图像分类技术与计算机网络、机器视觉、数据库及知识发现等技术的结合日益增强。如何合理、高效地组织图像数据，并结合图像底层特征，将人工智能及知识发现等技术运用于图像分类中，已成为研究的新热点。图像分类其本质还是从内容分析入手，寻找一个较好的方法来缩小底层特征和高层语义之间的联系[1]。可以发现，在缩小图像领域范围的基础上，综合考虑多种特征，学习自动建立每类图像的特征模型，并利用特征模型间的差异达到区分的目的。

图像的自动分类在许多应用领域都是一项关键任务，这些领域包括基于内容的图像检索（Content Based Image Retrieval, CBIR）、因特网数据过滤、医学应用、基于机器视觉的工业检测等。随着计算机技术以及遥感、军事、气象、工业检测等图像应用领域的飞速发展，图像技术的应用尤其是图像识别和分类技术的应用，在以下方面得到了快速发展。

遥感：遥感是指远距离不接触而获取物体信息的一种技术。遥感信息主要指通过遥感卫星所拍摄到的图像，通过对遥感图像的处理和识别，在军事上可以帮助军方快速的找到机场、公路、桥梁等关键设施，也可以让人们随时了解荒漠植被的生长情况以及森林火警等。

生物特征识别：随着安全意识的增强，人们根据人类自身生物特征的唯一性，提出了根据生物特征对人进行识别的图像处理技术。人类所具有的生物特征包括指纹、虹膜等。通过图像采集设备获取人的这些生物特征，并对这些特征进行图像处理，然后将处理结果和预先存贮在数据库内的信息进行比对就可以实现识别的功能。

图像检测：随着自动化技术的飞速发展，人们对生产的自动化要求越来越高。产品在生产的过程中，不可避免的会产生次品，如果单靠工人进行视觉判断，由于工人的疲劳和不同的主观判断标准，有可能会漏掉真正的次品，同时手工挑选的速度远远低于生产速度，从而单靠工人的手工挑选远远无法满足生产的自动化要求。因此产品的自动化检测技术随之产生，通过对生产线上的产品一一采集图像，并对其进行实时图像处理，可以自动判别产品的合格与否，大大的提高了生产率。

将图像处理技术和人工智能技术相结合的智能系统除了应用在图像数据库

检索上、卫星或红外图像的分析、医学诊断上的细胞分类，在基于机器视觉的分拣系统[2]中也能产生很大的作用，自动提取待分拣物体的模式，当然这种应用是比较简单的，但是必须根据实际应用的特点挖掘合适的分类模式。

对于利用视觉进行自动识别分类的系统来说，如何处理获得的图像信息，从中提取出能用于分类的有效信息，是一个机器视觉系统的核心。通常在实际应用中，针对不同的需要，对物体的分类有不同的要求。在一个基于机器视觉的分拣系统中，对物体的识别实际上是对所获得的图像进行分析，得到物体之间进行区分和识别的关系和准则。基于特征学习的图像分类是本文研究的重点。

1.2 研究现状

1.2.1 图像处理

随着计算机技术的发展，数字图像处理技术也得到了快速发展，它的理论和方法更加完善，其精确性、灵活性和通用性也大大提高，目前已经广泛应用于通信、医疗、遥感、宇宙探测、工农业生产等领域。

随着图像处理技术的深入发展，人们开始研究如何用计算机系统解释图像，实现类似人类视觉系统理解外部世界，被称为图像理解或计算机视觉。图像理解是由模式识别发展起来的方法，该处理输入的是图像，输出的是一种描述。这种描述不仅是单纯的用符号做出详细的描绘，而且要利用客观世界的知识使计算机进行联想、思考及推论，从而理解图像所表现的内容。计算机理解的过程包括图像预处理、图像描述、图像分析和理解。图像的正确解释离不开知识的引导，它属于人工智能的范畴，与思维科学体系的其他学科有着密切的联系。

图像识别的过程包括图像预处理、图像分割、特征提取和图像分类。而前三个过程是数字图像处理技术重点研究的内容。

图像预处理：目的是采用一系列技术改善图像质量，增强图像的视觉效果，或者将图像转换成一种更适合于人或机器进行分析处理的形式。通常改善方法有两类：一类是不考虑图像降质的原因，只将图像中感兴趣的特征有选择地突出，而衰减其次要信息；另一类针对图像降质的原因，设法补偿降质因素，从而使改善后的图像尽可能地逼近原始图像。第一类方法能提高图像的可读性，它是为了某种应用目的而去改善图像质量的，改善后的图像不一定逼近原始图像，其结果只能增强某种信息的辨别。这类方法通常称为图像增强技术，第二类方法能提高图像质量的逼真度，一般称为图像复原技术。

图像分割：图像分割是一种重要的图像处理技术，它是将一幅图像分解为若干互不交迭的有意义区域的集合，每个区域的像素有着相同的特性。图像分割的质量直接影响着计算机对图像的识别分类的效果，它在多数自动图像模式识别中

是一个基本的预备性步骤。图像分割算法一般是基于亮度值的两个基本特性之一：不连续性和相似性。第一类性质的应用途径是基于亮度的不连续变化分割图像，比如图像的边缘；第二类性质的主要应用途径是依据事先制定的准则将图像分割为相似的区域。基于亮度值的上述两个基本特性，图像分割方法大体上也可以分为两大类：一类是基于边缘的分割，即通过边缘检测得到图像属性变化较大的物体边缘像素，然后利用这些边缘像素将图像分割；另一类是基于区域的分割，即利用像素间的相似性来区分图像中有意义的区域。图像的分割对于图像分类而言，其目的是要将图像中感兴趣的部分或者能充分代表图像内容的区域分割出来，降低图像分析的数据量，同时也可以提高图像分析的质量。

图像特征提取：图像特征的分析与提取是图像识别分类系统的重要组成部分，它是从输入模式中抽取有利于识别分类的一组指标，这些指标表征了模式之间的差异与特殊性。模式的特征提取是模式分类的关键，模式分类器通常根据特征做出决策，所以特征提取直接影响到整个模式识别系统的复杂程度和识别精确性。特征对于不同类别的模式而言，应有明显的差异，或者说它们携带着较多的类别信息，采用这些特征分类能获得较低的误识率及较细的分辨层次。在实际模式识别问题中，特征过多会影响模式分类的速度和效果，所选择的特征应可能少而精。特征提取与识别对象的各种物理的、形态的性能都有联系，它依赖于人对被识别对象的认识，往往是经验性和技巧性的，没有统一的方法和理论可循。

1.2.2 数据分类

所谓模式是对某些感兴趣的客体的定量或结构的描述，模式类是具有某些共同特性的模式的集合。模式识别是根据研究对象的特性或属性利用以计算机为中心的机器系统运用一定的分析算法认定它的类别，系统应使分类识别的结果尽可能地符合真实。模式识别是研究一种技术，依靠这种技术，机器将自动地把待识别模式分配到各自的模式类中。模式识别是一个很宽的研究领域，可以把具有一定思维的或物质的模型统称为模式。模式识别的根本任务是对输入模式作分类。显然图像是一种模式，图像分类是一种特定的模式识别，可称为图像识别。

模式识别技术以及分类方法[3]的理论与应用都可以应用在图像分类上，通过几十年的研究分类理论已经获得大量的成果[4]，下面对各种分类方法及应用进行综述。

（1）决策树方法

利用信息论中的互信息（信息增益）寻找出数据集中具有最大信息的字段，建立决策树中的每一个结点，再根据字段的不同取值建立树的分支的过程，即建立决策树。国际上最有影响的决策树方法是 Quinlan 研究的 ID3 方法。

决策树（Decision Tree）是较早应用于数据挖掘分类问题的一种方法。在数

据量较大时，决策树方法能较快地构造出分类器；树的结构可以很方便地转化为SQL语言形式的陈述，以便用来更有效地访问数据库；而IF-THEN规则可以较容易地从树结构转化而来，因此研究相对较多。在文献[5-9]给出的已经是比较成熟的结果。

绝大多数决策树方法分两步构造分类器：树的生成和树的修剪[10]。在树的生成阶段，分拆准则决定了当前分拆属性和分拆点。给出合适的分拆准则是非常重要的，现有的决策树方法中使用两种分拆准则：ID3准则和CART准则。ID3算法采用信息论方法，减小分类对样本的平均测试值。它的属性选择是基于可能性假设，即决策树的复杂性与消息传递的信息量（熵）有关。

可伸缩性算法是目前数据挖掘分类问题的一个研究热点和难点，因为机器学习的算法只有具有了伸缩性，才可以真正成为适用于数据挖掘的算法。基于决策树的分类算法实现可伸缩性，比起其他的算法要相对容易一些，目前最快的可伸缩分类算法是被称为RainForest[7]的决策树算法。

（2）粗糙集分类方法

粗糙集理论是八十年代初Pawlak.Z[8]针对Firege.G的边界域思想提出的，基于给定训练数据内部的等价类，用上下近似集合来逼近数据库中的不精确概念。这一理论从新的角度对知识进行了定义，它把知识看作是对论域的划分，认为知识是有粒度的，从而引入代数学中的等价关系来讨论知识。用于分类，可以发现不准确数据或噪声数据内在的结构联系；用于特征归约，可以识别和删除无助于给定训练数据分类的属性；用于相关分析，可以根据分类任务评估每个属性的贡献或意义。其主要思想是在保持分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。

粗糙集理论研究与应用只限于对数据给出的知识问题进行处理，对于文本和连续图像问题的处理尚待研究。由于随机集理论在图像处理中获得了成功应用，所以粗糙集理论与随机集理论结合的进一步研究有望使它在图像中应用[11, 12]。

（3）神经网络分类方法

神经网络结构复杂，每个神经元通常采用非线性的作用函数，当大量神经元连成一个网络并动态运行时，构成了一个复杂的非线性动力学系统，具有不可预测性、不可逆性、多吸引子等特点。

它模拟人脑神经元结构，以MP模型和Hebb学习规则为基础，建立了三大类神经网络模型。前馈式网络，以反向传播模型、函数型网络为代表，用于预测、模式识别等方面。反馈式网络，以Hopfield离散模型和连续模型为代表，分别用于联想记忆和优化计算。自组织网络，以APT模型和Kohonen模型为代表，用于聚类。

标准 BP 算法收敛速度慢、易陷入局部极小的缺点[13]。BP 网络分类器的设计本质上是无约束的非线性最优化问题，分类器是否有效，取决于 BP 网络本身是否根据特征维数 N 和类别数 C 具备良好的自适应性。算法和网络结构的优化，使得分类器设计变得复杂。

（4）贝叶斯分类方法

贝叶斯分类方法是基于统计学的分类方法，它的主要思想是：用训练集的类分布来作为每个类别的概率分布，由训练集每个类别的先验概率，利用概率论和数理统计学中著名的贝叶斯定理来估算对于某个特定的样本，它属于每个类别的概率值（后验概率），后验概率最高的那个类别即为这个样本的类别。贝叶斯方法还可以分为朴素贝叶斯方法和可以处理属性间存在相互依赖关系的贝叶斯信念网络。

目前在大规模数据的分类问题中，朴素贝叶斯分类的类条件独立的假设很难满足，研究较多的是贝叶斯信念网络[14]，但是贝叶斯信念网络中评估函数比较难选，学习训练的复杂度比较大，这些问题都有待解决。

（5）支持向量机方法

支持向量机（Support Vector Machines, SVM）起源于统计学习和运筹学的最优化理论，它研究如何构造学习机，实现模式分类问题。其最大的特点是根据 Vapnik[15]结构风险最小化（Structural Risk Minimization, SRM 准则）原理构造决策超平面使每一类数据之间的分类间隔（Margin）最大。SVM 的思想就是在样本数目适宜的前提下，选取比较好的 VC 维，使经验风险和置信值达到一个折中，使每一类别数据之间的分类间隔（Margin）最大，最终使实际风险变小。对于线性不可分数据，按照 Cover 定理[16]，将低维空间不可分数据映射到高维空间中。支持向量机根据 SRM 准则尽量提高学习机的泛化能力[17]。即由有限的训练集样本得到小的误差，仍然能够保证对独立的测试集保持小的误差。另外，由于支持向量算法是一个凸优化问题，所以局部最优解一定是全局最优解，这是其他学习算法所不及的[18, 19]。

SVM 是小样本两类问题的最优方法，对于多类问题是构建多个两类 SVM 分类器，并与其它多类决策方法结合来解决。但是这样需要训练的 SVM 分类器个数太多，存在过拟合问题[20]。

（6）模糊集合论方法

利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。模糊性是排中律的一种特例，由于概念本身没有明确的外延，故而某一对象是否符合这一概念的划分就有不确定性，模糊数学正是从这一不确性中确立隶属规律，用适当的隶属函数来确定这种不确定性的隶属度。基本的隶属度

凭经验或领域专家给出，主观性强。模糊性是客观存在的，系统的复杂性越高，模糊性越强，这是 Zadeh 总结出的互克性原理。模糊系统的学习问题，其可理解性和精确性是需解决的问题[21]。

（7）遗传算法

一般的遗传算法由四个部分组成：编码机制、控制参数、适应度函数、遗传算子。初代种群产生之后，按照适者生存和优胜劣汰的原理，逐代（generation）演化产生出越来越好的近似解。在每一代，根据问题域中个体的适应度（fit2ness）大小选择和变异，产生出代表新的解集的种群。这个过程将导致种群像自然进化一样的后生代种群比前代更加适应环境，末代种群中的最优个体经过解码（decoding），可以作为问题近似最优解。遗传算法是模拟生物进化过程的算法，由三个基本算子组成：

选择，是指从一个旧种群（父代）中选出生命力强的个体，产生新种群（后代）的过程。

杂交，是选择两个不同的个体的部分进行交换，形成新的个体。

变异，对某些个体的某些基因进行变异。

遗传算法已在优化计算和分类机器学习等方面发挥了显著的作用。通常利用遗传算法的搜索特性为其他分类算法的参数做优化选择。

综上所述，这些分类算法存在训练复杂度大，或者可理解性差的缺点。对于图像这类多模式多特征，数据量大的分类问题，则要求分类算法能够易于实现并且算法的适应性和学习时间低的特点。

1.2.3 决策树的生成算法

对于具体的决策树生成算法，目前主要有以下几类：

（1）ID3 算法（Iterative Dicho to mizer 3)

Quinlan 的 ID3 算法是国际上公认的最早有影响的决策树算法。ID3 算法是基于信息熵的决策树算法，它是根据属性集的取值分类。

ID3采用自顶向下不回溯的策略搜索全部的属性空间，它建立决策树的算法简单，深度小，分类速度快。然而，ID3对于大的属性集则执行效率下降快，准确性降低，并且学习能力低下。

（2）C4.5 算法

C4.5 算法是 ID3 算法的改进，利用信息增益率选择测试属性。能处理连续型属性，还允许训练样本集中出现属性空缺的样本。它采用了一种归纳学习的机制。

（3）SLIQ 算法（Supervised Learning In Quest）

由于 ID3 和 C4.5 算法只能对于小的数据集有效，以及分类的准确性值得商

榷，对于此有人提出一些改进的算法 SLIQ。

SLIQ 算法的采用将属性分片，将其中的类常驻内存，随后由一关键字进行索引，每个数据集通过每个属性列表一个入口的连接来加以表示，类别列表的大小由训练样本数决定。

当记录集达到无法全部放入内存、以至于需要与外存进行交互时，由于产生较大的 I/O，此时算法的效率下降。

（4）SPRINT 算法（Scalable Parallelizable Induction of Decision Trees）

此算法利用了一个不同的属性列表数据结构。该数据结构保存类别与连接关键字 ID 信息，当一个结点进行分解时，相应属性列表也被分解到各个子结点中；而当一个结点进行分解时，列表的纪录顺序也被保留，因此分解后的列表无需再排序。此算法的设计思想使得它能够易于实现并行计算，从而具有较好的扩展性。

但是 SPRINT 算法也有认为存在的缺点：为每个结点都保存属性表，这个表的大小有可能是数据库中原始数据大小的好几倍。维护每个结点属性表的 Hash 表的开销很大（该表的大小与该结点所具有的纪录成正比）。

（5）RainForest 算法

RainForest 算法框架关注于提高决策树算法的伸缩性，该框架可运用于大多数决策树算法（例如 SPRINT 和 SLIQ），使算法获得的结果与将全部的数据放置于内存所得到的结果一致，也就是说该算法能根据内存的大小自适应的调整决策树算法的具体过程。他保持一个 AVC 集合（属性、值、类别），并描述每个类别的分布。

在各种分类模型中，基于决策树的分类模型以其如下特有的优点广为人们采用：(a) 决策树方法结构简单，便于人们理解；(b) 决策树模型效率高，对训练数据量较大的情况较为适合；(c) 决策树方法通常不需要训练集外的知识；(d) 决策树方法具有较高的分类精确度。但缺点是在构造决策树的过程中需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。因此，本文利用决策树的思想构造分类器，利用类间关系作为特征选择的标准，解决传统决策树算法多次读取数据导致低效的缺点。

1.3 关键问题及研究思路

1.3.1 研究问题的描述

图像的内容属性包括高层的语义特征和低层的视觉特征。图像的语义特征指图像表示的主题意义、故事、场景以及图像中物体对象的概念、姿态、时空关系等语义信息。图像的视觉特征指图像的颜色、纹理、图像中物体对象的形状、位置、方向、位置关系等直观的视觉特征。图像的特征不仅应具有可视性、不变性

和独立性，而且应与图像处理的方法有关，反应的是人类视觉有关图像的共性。视觉特征是计算机视觉技术能利用起来进行分析的图像内容，因此一般利用这些最能符合图像内容和理解的特征来进行描述和分类。

图像包含了现实世界的实体，它们所组成的图像集合则包含着这些实体的变化、相互关系以及隐藏在其中的各种模式、演化规律等信息。图像识别的特点是模式类别极多、识别系统庞大并且在实际应用中经常需要增加可识别的模式类别。图像分类是模式分类（Pattern Classification）在图像处理中的应用，它完成将图像数据从二维灰度空间转换到目标模式空间的工作。分类的结果是将图像根据不同属性划分为多个不同类别的子区域。一般地，分类后不同的图像区域之间性质差异应尽可能的大，而区域内部性质应保证平稳特性。

图像的处理和分类的一个重要研究目的就是利用迅速发展的计算机、人工智能、信号处理等技术，建立能模仿人类视觉系统的计算机图像处理与理解系统，实现对外界景象的自动分析与解释。现代智能理论和技术在近几十年中得到了持续的关注和长足的发展，动力来源正是其广阔的实际需求、无限的应用前景和与其他学科之间的渗透能力。基于数据的机器学习是现代智能技术中十分重要的一个方面，主要研究如何从一些观测数据（样本）出发，得出目前尚不能通过原理分析得到的规律，并利用这些规律去分析实际观测的数据，从而得到有价值的决策、估计或预测等结论。通过学习算法、认知模型研究人类学习的计算理论和实验模型。

图像识别系统包括三个部分：特征提取方法、特征选择方法和训练方法。在统计模式识别的方法中，这三个部分是独立计算的过程。其中，特征提取和特征选择的概念非常相似，而且可以用同样的模式可分性度量准则进行衡量。但它们之间的关系决不可混淆。因为特征提取是一种样本空间的映射变换过程[22]而特征选择是直接对特征数据进行处理的过程。从广义上讲，特征提取是一种通过寻求变换映射获取少量最有效特征的方法。特征选择则不是寻求某种映射变换，而是在特征空间中直接比较某个或者某组特征对分类的贡献大小，挑选贡献大的特征，剔除贡献小的特征。而训练的目的则是利用已提取和选择的特征数据构成分类器。构成分类器的训练过程是在特征空间进行的，其训练的结果是直接产生判别规则。统计模式识别中的特征选择是一种按各个特征对分类能力贡献的大小进行排序的过程，它一方面强调各个特征的贡献是有等级区分的，而另一方面在构成分类器的训练过程中却同时使用这些特征，把它们当作平等的关系。但是在面向实际的一个突出问题是在构成多模式类的分类器时，很多算法在解决较少模式类的分类问题时具有较好的效果，但处理极多模式问题的时候不尽人意。

因此，进行图像数据的分类主要是从图像识别系统的三个部分进行研究，提

取和选择能描述图像实质的特征，并利用这些特征训练出分类规则，以获得较好的分类效果，并且使处理的速度尽可能快。其目的是构建计算复杂度较小、判决精度较高和可理解性较好的分类器。

1.3.2 课题研究思路

图像分类的研究其目的是要从大量的图像数据中抽取出一定的模式和规则，用于其它图像的识别和类别预测。而且图像是一个由多种特征描述的事物，而且图像所表现的内容也比较丰富。对图像的分类其主要的三个过程是课题需要研究的重点及内容。本文针对这三个方面的问题展开：

（1）特征提取

图像的底层特征是能够从图像数据中分析得来，对于图像特征的描述，尤其是图像颜色、纹理和形状三类主要的特征描述已经有了很多成熟的表示方法。既然用于图像的分类，图像的特征就需要能满足不同形态的图像在该特征上的尺寸、旋转和平移不变性，而且该特征能很好地描述一类图像的特点。这些特征必须是容易描述和提取的，并且尽可能的能提取大量的特征以避免由于特征提取的不完备使得后面两个步骤不好开展。

（2）特征选择

在众多的图像特征中降低分析的维数，则需要从提取的特征集合种选择少而精的特征用于图像类特征模型的建立，并且用于图像的分类。而不同的分类算法有着各自的优势与不足，适用于具有不同特点的数据。不存在某种方法能适合各种特点的数据。分类器的构建首先要减少其处理的数据量，选择对分类效果最好的特征，并且要使得分类算法的计算复杂度低，描述简洁易于实现和理解。而且图像作为一种数据量大的分类问题，选择尽可能少的特征并能保证分类精度的特征选择算法是研究的目标。同时，对于实际的应用，应该使得算法的时间复杂度低，便于理解，且实现简便。

本文利用各个图像类在特征分布上的相互关系来进行特征选择，基于决策树的思想，设计错误率尽可能低并且实现简单的特征选择算法。

（3）图像分类

利用选择的特征进行图像的分类，其实质是如何进行特征的融合[23]。多维特征根据特征数值划分成了不同的特征空间，在各个空间内，图像分类的结果和可信度因图像类之间的特征关系不同而有所不同。在图像类间特征值分布相交叠的区域，其分类结果可利用其他特征的融合来获得较高的分类可信度。本文采用带误差（或置信度）的决策树来获得分类的规则，对每个结点上的特征进行匹配获取树分支的分类误差，从根结点沿着较小误差分支到达的叶结点就是分类结果。引入先验特征知识，对决策树的生成进行指导，特征匹配时对分类特征进行

加权。

1.4 本文工作安排

本文主要研究数据挖掘中的分类问题。然后通过图像多特征分类来验证提出的分类方法。第二章讨论了在图像分类中，图像的特征提取方法和分类中的图像特征应用；第三章首先分析决策树算法中特征选择标准，提出从分类特征的特点出发进行分析，讨论了分类误差的计算方法，为基于样本类整体的特征分布为对象的决策树构建算法提供理论依据；第四章讨论决策树构建中特征选择函数的改进、性能分析和对象增减影响决策树的重建问题，论述了分类规则的提取方法，并结合先验知识来指导决策树的生成，产生启发式分类决策；第五章通过对图像特征的分析，利用本文讨论的分类方法进行分类，验证分类效果的准确率，并且图像是具有模糊性的一种分类，需考虑机器精确性和实际识别模糊性的矛盾。最后，对本文工作做出总结，提出下一步研究的内容。

第二章 面向分类的图像特征

图像分类就是要找到图像中的同类景物在相同的条件下具有的相同或类似的光谱信息特征和空间信息特征，并且分析不同类景物在特征上的差异。从某种意义上，对于目标任务而言，图像特征提取的方法对分类器的构建起着重要的作用。

2.1 图像特征提取与分类

由于人们感知的主观性，对于一个给定特征并不存在唯一的、最好的表示方法。对于每一类特征都存在着很多不同的表示方法，这些方法从不同的角度来刻画该类特征。图像中物体的特征主要有两类：一类用于描述和识别单个物体或物体的某个部分，如颜色、纹理及形状、结构。其中，颜色、纹理特征强调物体表面的颜色、灰度变化，形状、结构特征强调物体的外形，如弯曲、大小等；另一类用于描述物体的景物组成或物体的三维信息，如左右、远近、透视等。图像特征的分析与处理，涉及到模式识别、计算机视觉、图像理解等诸多研究领域。

在图像分类中，根据不同图像的特点，提取的特征应该是具有旋转、平移和尺度不变性的。同时还要考虑提取的图像特征对分类算法复杂性和一个图像识别分类系统的实用性等的影响，因此，图像特征提取应该考虑以下几个方面：

（1）特征的代表性。在分类中，需要提取出同类图像中具有典型代表性的特征作为整个类的一个识别标准，这些特征能准确描述图像要表达的内容。不能很好描述图像内容的特征是没有用的，比如最大灰度，在一般的拍摄条件下，它的值都有可能是255，那么对于分类而言，所有类的图像就都会具有这样的特征，进行特征提取分析就毫无意义。

（2）特征的不变性。在计算机的存储表示中，图像是个二维数组。对同一个物体从不同的角度获得的图像就具有差异，但是不能因为分析的数据有了变化而对图像中表现的实质内容进行错误的识别。因此，用于分类的特征应该具有旋转、平移和尺度不变性的特点。

（3）特征的针对性。特征提取需符合具体分类问题的要求。对于不同的自动图像分类的应用，应该提取不同的特征。有些图像分类需要提取全局特征，有些需要对局部特征进行分析提取。比如在不同的拍摄条件下，物体的形状大小特征在图像中的表现可能是相似的，具有缩放的物体间分类，利用面积特征作为分类特征显然不可分；而在黑白摄像机的拍摄下，色彩信息的提取也无意义。

（4）特征提取的可操作性。一些满足以上特性的特征可能涉及到过于复杂

的提取算法，难以实现或计算量过大，预处理或分类计算耗时较大，实际应用中不满足实时性要求。因此，应优先选取尽可能简单的特征提取方法，表达方式也力求简洁明了。

综上所述，针对不同的图像分类问题，提取合适的特征能减少数据处理的计算量，也影响着分类器构建的效果，最终影响的是分类结果的准确率。

下面将对颜色、纹理、形状等二维信息特征的一些典型的描述方法进行阐述和分析。为后续章节中分类特征选择提供初始的特征集合。

2.2 图像颜色特征

颜色是彩色图像最显著、最直观的物理特征，一般用颜色直方图来表示，非常稳定，具有旋转、平移和尺度不变性，表现出相当强的鲁棒性。颜色特征的定义比较明确，抽取也相对容易，具有很好的判别能力，并且非常稳定，对于旋转、平移、尺度变化，甚至各种变形都不敏感。

2.2.1 颜色空间

图像的颜色可以有多种表示方法。为了准确提取能够表征原始视频颜色信息的一组颜色，提取算法必须在符合人类视觉系统的生理特征和符合人类观察经验的视觉感知特征的颜色空间内进行。常用的颜色空间有 RGB、HSV、HMMD、YCbCr：

（1）RGB 模型

RGB 空间模型是最常用的一种颜色模型。色觉的产生需要一个发光光源，光源的光通过反射或透射传递到眼睛，根据人眼的结构，所有颜色都可看作是 3 个基本颜色——红(R，red)、绿（G，green）、蓝（B，blue）——的不同组合。以 RGB 三个参数为坐标，我们可以得到一个单位立方体来描述 RGB 颜色模型。常用的 RGB 模型是规定 R、G、B 的值都在[0，255]之间，(0,0,0)表示黑色，(255,255,255)表示白色，一共可以表示 2^{24} 种颜色。由于 RGB 模型是面向硬设备的最常用的颜色模型，所以通常其它颜色空间的模型都是通过 RGB 空间转换得到的。

为了去除光照对颜色的影响，需要对颜色分量做归一化处理：

$$\begin{cases} R = \dfrac{R}{R+B+G} \\ G = \dfrac{G}{R+G+B} \\ B = \dfrac{B}{R+G+B} \end{cases} \tag{2-1}$$

（2）HSV 模型

HSV 模型与人类的视觉感知系统有较好的一致性，这个模型的参数分别是：色调（H，hue）、饱和度（S，saturation）、明度（V，value），RGB 模型向 HSV 模型的转换公式如下：

$$h=\begin{cases}\arccos\dfrac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}, & B\le G\\ 360-\arccos\dfrac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}, & B>G\end{cases} \tag{2-2}$$

$$s=\frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)} \tag{2-3}$$

$$v=\frac{\max(R,G,B)}{255} \tag{2-4}$$

（3）HMMD 模型

HMMD 颜色空间所得到的参数就是色调(H，hue)、最大值（M，max）、最小值（M，min）以及它们的差值（D，different），转换公式如下（H 求法同公式（2-2））：

$$\begin{cases}Max = \max(R,G,B)\\ Min = \min(R,G,B)\\ D = Max - Min\end{cases} \tag{2-5}$$

（4）YCbCr 模型

Y 代表亮度，Cb 和 Cr 代表色度和饱和度，通过式（2-6）可以完成 RGB 向 YCbCr 的转换。

YCbCr 颜色空间多用于视频图像的压缩领域，因为 Y 分量是比较重要的数据，所以可将其的取样比例设定的较高，而 Cb、Cr 分量只取部分数据。

$$\begin{cases}Y = 0.299R+0.587G+0.114B\\ Cb = -0.1687R-0.3313G+0.5B+128\\ Cr = 0.5R-0.4187G-0.0813B+128\end{cases} \tag{2-6}$$

2.2.2 颜色特征的量化

对颜色特征的描述方法有很多。对于每个像素，可以用各个颜色空间的分量来表示。如灰度值 Gray、红色分量 R、蓝色分量 G、绿色分量 B、色调 H、饱和度 S、明度 V 等等。另外还有些比较特别的参量在进行图像分类时能表现出类与类之间的差异，RGB 分量之间的差，分量间的大小关系都可以作为图像的特征。

为了提高计算处理的速度，通常要减少参数的等级数量，将多数量的等级映射到少数量的等级，这样可以获得其它的特征描述参量。

（1）8 主色

颜色空间的压缩采用主色聚类的方法[24]，即将256维的RGB颜色空间压缩到8维的以主色为分割域的颜色空间，规定主颜色为：红、橙、黄、绿、青、蓝、紫、黑等八种颜色。实验证明这几种颜色可以很好的代表图像颜色的分类。

首先根据颜色值表找到这八种颜色对应的RGB值，分别为：

红（R=255，G= 0，B= 0）、橙（R=255，G=153，B= 0）

黄（R=255，G=255，B= 0）、绿（R= 0，G=255，B= 0）

青（R= 0，G=255，B=255）、蓝（R= 0，G= 0，B=255）

紫（R=153，G= 0，B=255）、黑（R= 0，G= 0，B= 0）

然后采用最短距离法进行颜色聚类，公式如下：

$$cd(p,i)=\sqrt{(R_i-R)^2+(G_i-G)^2+(B_i-B)^2}\ (i=0,1,\cdots,7) \tag{2-7}$$

计算某一像素点$p(R,G,B)$与主色$i(R_i,G_i,B_i)$的色差，得到与p点距离最近的主色j，将p点聚类到主色j集合，即j的计数器加1。

（2）HSV空间的量化

根据人类的视觉分辨能力，将HSV空间的颜色分量简化，把H分成8份，S和V各3份，并且可以将3个颜色分量合成为一维特征矢量，即$l=9H+3S+V$[25]。于是，依据l的取值范围，HSV颜色空间的特征等级数降为72级。

$$H=\begin{cases}0,h\in[0,20]\cup[316,359]\\1,h\in[21,40]\\2,h\in[41,75]\\3,h\in[76,155]\\4,h\in[156,190]\\5,h\in[191,270]\\6,h\in[271,295]\\7,h\in[296,315]\end{cases},\ S=\begin{cases}0,s\in[0,0.2]\\1,s\in(0.2,0.7]\\2,s\in(0.7,1]\end{cases},\ V=\begin{cases}0,v\in[0,0.2]\\1,v\in(0.2,0.7]\\2,v\in(0.7,1]\end{cases} \tag{2-8}$$

对于图像描述而言，颜色直方图是最常用的颜色特征提取方法。它能简单描述一幅图像中颜色的全局分布，特别是用于描述那些难以自动分割和不需要考虑物体空间位置的图像。但是这种方法不能描述图像颜色的局部分布和每种色彩的空间位置。计算颜色直方图需要将颜色空间划分成若干个小的颜色空间，每个区间成为直方图的一个数量级，这个过程称为颜色量化。各种颜色特征用直方图来描述。

（3）颜色矩

另外一种简单而有效的颜色特征提取方法是颜色矩[26]。图像中任何的颜色分布均可以用它的矩来表示。由于颜色分布信息主要集中在低阶矩中，因此，仅采

用颜色的一阶矩、二阶矩和三阶矩就足以表达图像的颜色分布。

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \tag{2-9}$$

$$\sigma_i = [\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - \mu_i)^2]^{1/2} \tag{2-10}$$

$$s_i = [\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - \mu_i)^3]^{1/3} \tag{2-11}$$

式中，p_{ij}是图像中第j个像素的第i个颜色分量。

2.3 图像纹理特征

纹理指的是某种具有一致性的视觉图案，这种一致性是由于一个领域内像素灰度、颜色等的规律的空间分布所形成的，纹理是所有物体表面材质的内在属性。在纹理中包含物体表面结构组织、其与周围环境之间关系方面的重要信息纹理通常被看作图像的某种局部性质，或是对局部区域中像素之间关系的一种度量。

提取纹理特征的方法[27, 28]也有许多，主要可分为统计法、结构法和模型法。所谓统计法就是根据人们的直观视觉感受，从心理学因素出发，形成纹理特征的一种表示方法。结构法是指从图像的结构观点出发，认为纹理是由纹理元素按照某种特定的排列规则联合起来构成的。在模型法中，纹理是通过某种概率分布或基函数的线性组合来表示的，模型的系数用来描述图像的纹理信息。

纹理特征的表示主要分为基于空间性质、基于频域性质、基于结构感知性质的纹理表示，通常采用的是基于空间性质的纹理表示对图像的纹理进行描述。

2.3.1 Tamura 纹理特征

基于人类对纹理的视觉感知心理学研究，Tamura 等人提出了 6 种纹理描述特征[29]，即粗糙度（coarseness）、对比度（contrast）、方向度（directionality）、线像度（linelikeness）、规整度（regularity）和粗略度（roughness）。前三个分量应用比较广泛，下面给出其描述方法。

（1）粗糙度

粗糙度测量纹理的间隔尺度或力度，与图像的分辨率有关，分辨率大则纹理粗。首先计算图像中大小为$2k \times 2k$个像素的活动窗口中像素的平均强度值；对于每个像素，分别计算它在水平和垂直方向上互不重叠的窗口之间的平均强度差；取平均强度差最大的 k 值，最后计算粗糙度如下：

$$F_{crs} = \frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n} 2^k f(i,j) \tag{2-12}$$

其中 k 满足：$\max_{x,y}\sum_{i=1}^{m}\sum_{j=1}^{n} f(i,j)/2^{2k}$ 且$k = 0,1,\cdots,5$。

（2）对比度

对比度依赖于像素的灰度分布，可测量图像中局部的灰度变化对比度通常与灰度的动态范围及图像中边缘的尖锐程度等有关。

$$F_{con}=\frac{\sigma}{[\mu_4/\sigma^4]^{1/4}} \tag{2-13}$$

其中 μ_4 是图像灰度的 4 阶中心矩，σ 是图像灰度的标准方差。

（3）方向性

方向性主要描述纹理是如何沿某些方向散布或集中的。图像局部区域的朝向可借助梯度模板计算，对应与梯度向量的角度。通过卷积计算得到每个像素处的梯度向量；对所有像素的梯度向量进行离散化处理得到直方图信息，最终描述了纹理的方向性。

2.3.2 基于共生矩阵的纹理表示

应用统计的方法来描述图像纹理的典型代表是利用灰度共生矩阵。它是图像中两个像素灰度级联合分布的统计形式，能较好地反映灰度级相关性的规律。

灰度共现矩阵是描述在 θ 方向上，相隔 d 像元距离的一对像元，分别具有灰度层 i 和 j 的出现概率，其元素可以记为 $p(i,j\mid d,\theta)$，$p(i,j\mid d,\theta)$ 的计算公式为：

$$p(i,j\mid d,\theta)=\#\left\{\left[(k,l),(m,n)\right]\in(Z_r*Z_C)\mid k-m=\theta,\mid l-n\mid=d,f(k,l)=i,f(m,n)=j\right\} \tag{2-14}$$

其中，若图像水平方向和垂直方向各有 N_C 和 N_r 个像元，且每个像元被灰度量化为 N_m 层，则 $Z_C=\{1,2,...N_c\}$，$Z_r=\{1,2,...N_r\}$ 分别为图像的水平空间域和垂直空间域。若 $G=\{1,2,...,N_m\}$ 为量化灰度层集，函数 f 表示图像中每个像元具有 N_m 个灰度层中的一个值 G，即 f：$Z_r*Z_C\to G$。k、m 和 l、n 分别在所选计算窗口中变动，#表示使大括号成立的像元对数。通过以上定义，可以发现灰度共现矩阵是一个对称矩阵，其阶数由图像中的灰度层数决定。这个矩阵是距离和方向的函数，在规定的计算窗口或图像区域内统计符合条件的像元对数。

由灰度共现矩阵总结的一组参数可以用来描述图像的纹理特征，其包括：

（1）能量：　$E(d,\theta)=\sum\{p(i,j\mid d,\theta)\}^2$　　(2-15)

（2）熵：　$H(d,\theta)=-\sum_{i,j}\{[p(i,j\mid d,\theta)]-\log[p(i,j\mid d,\theta)]\}$　　(2-16)

（3）惯性矩：　$I(d,\theta)=\sum_{i,j}(i-j)^2\,p(i,j\mid d,\theta)$　　(2-17)

（4）相关：　$C(d,\theta)=\dfrac{\sum_{i,j}(i-u_x)(j-u_y)p(i,j\mid d,\theta)}{\sigma_x\sigma_y}$　　(2-18)

其中 $\begin{cases}u_x=\sum_i i\sum_j p(i,j\mid d,\theta)\\ u_y=\sum_j j\sum_i p(i,j\mid d,\theta)\end{cases}$，$\begin{cases}\sigma_x=\sum_i(i-u_x)^2\sum_j p(i,j\mid d,\theta)\\ \sigma_y=\sum_j(j-u_y)^2\sum_i p(i,j\mid d,\theta)\end{cases}$。

（5）局部平稳： $$L(d,\theta)=\sum_{i,j}\frac{1}{1+(i-j)^2}p(i,j\mid d,\theta) \tag{2-19}$$

能量是对灰度分布均匀性的度量。灰度分布越不均匀，$p(i,j)$相差越大，能量也越大；反之，当$p(i,j)$越均匀时，能量越小。相关度用来描述$p(i,j)$矩阵中行或列元素之间的相似程度，是灰度线性关系的度量。

2.4 图像形状特征

在计算机视觉中，相对于颜色或纹理等底层特征而言，形状特征属于图像的中间层特征，是图像几何特征的反映，形状可能比颜色和纹理包含更多地语义信息，所传递的语义往往也更具体和更准确。形状特征作为刻画图像中物体和区域特点的重要特征，是描述高层视觉特征（如目标、对象等）的重要手段，而目标、对象对获取图像语义尤为重要。

2.4.1 几何常量特征

通常从物体的轮廓出发得出的一些简单的特征参量具有平移、比例、旋转不变性等特点，如周长、面积、复杂度、轮廓的顶点数和凹点数[30]。

（1）周长： $$l=|B| \tag{2-20}$$

其中 B 为图像中边缘像素点的集合。

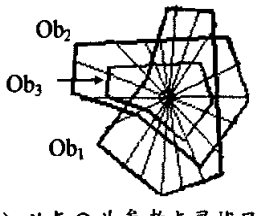
（2）面积： $$S=|O| \tag{2-21}$$

其中 O 表示图像中物体像素点的集合。

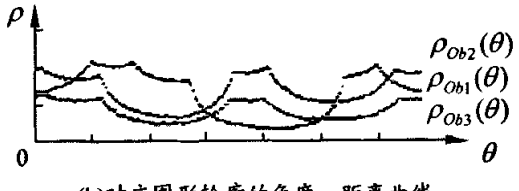
（3）复杂度： $$c=l^2/S \tag{2-22}$$

（4）占空比：区域面积与区域最小外接矩形面积之比。

（5）轮廓凸凹点数的计算可以通过对轮廓曲线进行处理。首先获得物体的质心，然后从质心为起点沿不同的角度的射线方向寻找目标外围轮廓，可以得到一序列关于θ的距离值$\rho(\theta_1)$，$\rho(\theta_2)\cdots\cdots\rho(\theta_n)$。$\rho(\theta)$将构成反映目标外围轮廓形状的一维曲线。而且轮廓曲线具有旋转不变性，轮廓凸凹点数可以分析轮廓曲线二阶导数较小的点的个数。图 2-1(b)是三个相似物体的轮廓曲线。同样，轮廓曲线也能作为一种形状特征。



(a) 以点 O 为参考点寻找目标轮廓



(b)对应图形轮廓的角度—距离曲线

图 2-1 外轮廓的角度—距离曲线

2.4.2 形状轮廓矩特征

矩在图像识别和分析中应用广泛，有 Hu 矩、Zernike 矩、复数矩等。MK Hu[31] 利用二阶和三阶中心矩构造了七个不变矩，利用这七个不变矩可以描述一幅图像的区域形状，且具有旋转、平移和尺度不变性。Hu 矩是经典的矩描述方法，相应公式如下：

（1）对于图像 $f(x,y)$，则其 $(p+q)$ 阶矩定义为：

$$m_{pq}=\sum_x\sum_y x^p y^q f(x,y) \tag{2-23}$$

（2）$(p+q)$ 阶中心矩为：

$$\mu_{pq}=\sum_x\sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{2-24}$$

其中 $\bar{x}=\dfrac{m_{10}}{m_{00}}$，$\bar{y}=\dfrac{m_{01}}{m_{00}}$。

（3）零阶矩 $m_{00}=\sum_x\sum_y f(x,y)$ 表示密度的总和，求得物体的质量。一阶矩 m_{10} 和 m_{01} 分别除以零阶矩后的得到的是物体的中心坐标。中心矩 μ_{pq} 反映物体区域中灰度中心分布的度量。μ_{20} 和 μ_{02} 分别表示物体区域围绕通过灰度重心的垂直和水平轴线的惯性矩.。若 $\mu_{20}>\mu_{01}$，则可能是水平方向拉长的物体。μ_{30} 和 μ_{03} 的幅值可以度量物体对于垂直和水平轴线的不对称性。如果是完全对称的形状，其值为零。对 $(p+q)$ 阶中心矩规范化，即 $\eta_{pq}=\mu_{pq}/\mu_{00}^r$，$r=(p+q+2)/2$。

（4）Hu 矩的七个不变矩定义为：

$$\begin{cases}
\phi_1=\eta_{20}+\eta_{02}\\
\phi_2=(\eta_{20}-\eta_{02})^2+4\eta_{11}^2\\
\phi_3=(\eta_{30}-3\eta_{12})^2+(3\eta_{21}-\eta_{03})^2\\
\phi_4=(\eta_{30}+\eta_{12})^2+(\eta_{21}+\eta_{03})^2\\
\phi_5=(\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})\times[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]\\
\qquad+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})\times[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]\\
\phi_6=(\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]+4\mu_{11}(3\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03})\\
\phi_7=(3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})\times[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]\\
\qquad+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})\times[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]
\end{cases} \tag{2-25}$$

而在实际应用中有时要对这七个矩进行修正：

$$t_1=\phi_1,t_2=\phi_2,t_3=\sqrt[5]{\phi_3^2},t_4=\sqrt[5]{\phi_4^2},t_5=\sqrt[5]{\phi_5^2},t_6=\sqrt[5]{\phi_6^2},t_7=\sqrt[5]{\phi_7^2} \tag{2-26}$$

2.5 图像特征分析

2.5.1 图像底层特征总结

前面几节列举了图像的三类基本底层特征。对于每一类图像特征而言有多种表达方式，是从不同角度模拟人类感知的主观性。表 2-1 简单地描述了各种图像底层特征各自的优点和局限性。

表 2-1 各种图像特征优缺点分析比较

特征类型	图像特征	优点	局限性
颜色	主色，主色调，直方图，颜色空间分量，颜色矩	提取方便，自动提取，计算量相对较少，图像全局特征。	不能提供基于目标的局部信息，在描述图像语义方面具有一定局限性。
纹理	共生矩阵，Tamura 纹理	提取方便，提取图像的全局和局部特征。	特征提取和匹配的计算来量大，图像的语义描述有限。
形状	几何常量，矩不变量	根据物体的形状生成高层特征，针对性强。	自动提取比较麻烦，依赖于好的图像分割算法。

通常在基于内容的图像检索、网络数据过滤等应用中由于图片的规格大小不一，一般使用颜色和纹理特征；而在图像检测和生物识别等应用中，形状特征就起到了很重要的作用，这些应用主要是从单独的个体来进行分析的，而且很多应用中使用的是黑白的摄像机，个体大小一般是统一的。不同的应用对全局特征和局部特征的要求也不同。

2.5.2 图像特征与分类

从统计决策理论来看，图像分类在数学上就是对呈现统计可变的数据做出决策的过程。同一类图像的多个样本在某个特征上的值是存在一定的规律的，获得某种分布曲线，而各个类的特征值分布曲线之间的关系正体现了图像内景物的相似性或差异性。这种规律性的实质正是进行图像分类需要提取的信息。

一般来说，对于多个图像类的分类，单一的特征是很难完成的。也就是说在一个特征中，多个类的特征分布曲线之间是有关联的，而不是分离开的。特征对于分类而言，从统计分布来看有以下两种情况：

（1）一个特征可能将多个类分成若干子集，子集中各个类之间利用其他特征能进一步区分。即各个类在特征中的分布曲线呈现相对独立的情形，能通过多个特征分布一一区分开来。

（2）还会存在这样的情况，所能给出的特征不能将所有类完全区分，这些类在特征的分布上或多或少具有相关性，而这种特征分布间的相关性使得在利用统计理论进行分类的时候要考虑它引起的分类误差。

第二种情况中多特征分类中基于误差分析进行分类方法研究是本文一个重

点研究内容。如果同类大样本的数据构成的统计分布曲线是满足正态分布的，那么，利用特征进行的分类引起的误差可以利用贝叶斯的误差估计方法进行计算，而不满足正态分布的类间关系的误差计算需进一步研究。分类误差的分析以及类与类之间在特征上的分布关系都将在第三章详细介绍。

当然，这些图像的特征之间有些具有相关性，比如直方图和颜色矩，几何常量中面积周长与复杂度之间存在某些相关，而纹理和颜色却是互补的两类特征。由于在对多类物体进行检测时，并不清楚他们在哪些特征上同类物体具有相似性或者类间差异很大，因此，这些特征都作为初始的特征集合。但是，在每次进行分类特征分析的时候，在同等分类效果的特征中应该选取与已选特征相关性小的特征作为分类特征。

因此，特征的提取效果和如何应用这些特征进行分类，以及基于多特征的分类方法都是需要做的一些工作，下面两章将对这些问题进行研究。

2.6 本章小结

图像特征的描述是基于特征的图像分类的重要基础，对于分类模型结果的好坏有着本质性的影响。本章分析了特征提取与分类的关系后，描述了颜色、纹理和形状等主要视觉特征的一些表示方法，由于个别特征不一定能直接表示各类图像的具体实质，因此，基于特征的图像分类首先是要获得一个特征集合，尽可能地贴近图像的实质，并能完整地区分图像类别。最后，通过总结图像底层特征的特点，分析了各类特征与分类的关系。本章工作为后续的图像分类的特征选择作铺垫。

第三章 决策树分类特征选择

特征选择是在原数据集中找到一个适于分类算法的子集。这一子集应该满足不显著降低分类精度，不影响类分布以及具有稳定、适应性强的特点[32]。本章通过分析多类物体的特征分布来对特征选择进行讨论，获得决策树生成中的特征选择指标。

3.1 一般决策树分类特征选择指标

3.1.1 信息增益

基本的决策树算法 ID3/C4.5 使用信息增益（Information Gain）作为特征对结点进行划分的指标。选择具有最高信息增益的特征作为当前结点的测试特征，该特征使得对结果划分中的样本分类所需的信息量最小，并且导致期望熵降低。

按照信息论的定义，设 S 是 s 个数据样本的集合，具有 m 个不同的类 C_i。设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息为：

$$I(s_1,s_2,...,s_m)=-\sum_{i=1}^{m}p_i\log_2(p_i) \tag{3-1}$$

其中 p_i 是任意样本属于 C_i 的概率，并用 s_i/S 估计。

设特征 A 具有 v 个不同的离散值 $\{a_1,a_2,...a_v\}$，可以将 S 划分成 v 个子集；S_j 包含 S 中这样一些样本，它们在 A 上具有值 a_j。设 s_{ij} 是子集 S_j 中类 C_i 的样本数。根据 A 划分子集的熵为：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+...+s_{mj}}{s}I(s_{1j},...,s_{mj}) \tag{3-2}$$

熵值越小，子集划分的纯度越高。那么特征 A 上获得的信息增益为：

$$Gain(A)=I(s_1,s_2,...,s_m)-E(A) \tag{3-3}$$

计算每个特征的信息增益。具有最高信息增益的特征选作给定集合 S 的测试特征。创建一个结点，并对特征的每个离散的值域范围创建分支。

3.1.2 基尼指数

如果决策树是二叉树，通常用基尼指数作为划分的标准。设数据集 S 的类别 C 有 m 个。那么其基尼指数就是：

$$Gini(S)=1-\sum_{i=1}^{m}p_i^2 \tag{3-4}$$

其中 p_i 是类别 c_i 出现的频率。如果特征 A 将数据集 S 分成两部分 S_1，S_2。那么这个划分的基尼指数就是：

$$Gini(S) = (S_1 / S) * Gini(S_1) + (S_2 / S) * Gini(S_2) \quad (3\text{-}5)$$

选择基尼指数最小的特征对结点数据进行划分。决策树是二叉树时，设离散型特征 A 有 v 个值，则特征 A 有 2^v 种划分数据集 S 的方法，其中一个划分方法的基尼指数最小，为特征 A 的最佳划分。然后从所有特征的最佳划分中选出基尼指数最小者，作为树结点的最佳划分。

3.1.3 缺点分析

上两小节描述的决策树特征选择指标是从离散的特征值个数的基础上计算信息增益或者基尼指数来选择树结点的划分特征。信息增益和基尼指数的方法都是从特征值的值域划分范围出发进行特征选择的较好方法，但是各个特征值的离散化程度，它直接影响了决策树的复杂程度。从特征值的划分范围进行决策树的分支生成的方法使得所构建的决策树具有以下的劣势：

（1）缺乏伸缩性，由于进行深度优先搜索，所以算法受内存大小限制，难于处理大训练集。

（2）在训练数据量大时容易导致树的尺寸过大，树的高度很大，导致产生的规则过于复杂难于理解。

（3）过度剪枝可以使得决策树变小，但是却丧失了决策树的预测准确率。

虽然为了解决这些问题有了很多的研究，但是由于从特征值域范围划分子集的局限，随着样本增多不可避免会引起以上问题。所以在保证算法良好的运算效率同时，要使得决策树尺寸尽量小，并保证有高准确性和良好的解释性。

为了克服以上决策树的缺点，本文提出从特征的类间统计分布关系出发来进行特征选择，并且研究决策树结点分支划分和特征选择的指标。本章后面部分将对基于类间特征分布关系的特征选择方法所涉及的问题进行阐述，并且通过分析特征的特点给出了若干分类特征相关的定义，为本文的核心内容决策树的生成方法研究提供理论依据。

3.2 分类的特征关系

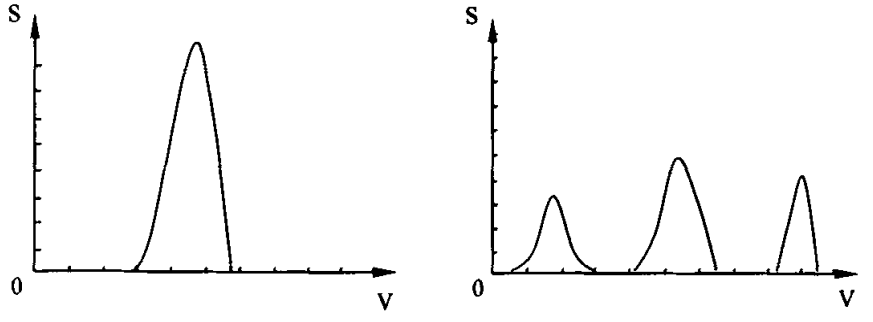
3.2.1 单个类的特征值分布

分类模型中提取的特征集合应该很好的描述图像的特征，并且各类的特征集合能相互区别。相同类的图像样本在某个特征上表现出一定的特征关系，且样本的特征分布也存在一定的规律。

下面给出共同特征和个性特征的定义：

定义 3.1 $\{x_i, i=1,2\cdots n\}$ 为一训练样本集合，$\{\xi_j, j=1,2\cdots m\}$ 为一组特征集合，$P(x_i,\xi_j)$ 表示样本 x_i 对于特征 ξ_j 的特征值，集合 $\{P(x_i,\xi_j), i=1,2\cdots n\}$ 表示样本对于特征 ξ_j 的特征值集合；如果对于任意样本 x_i，在特征 ξ_j 上的值 $P(x_i,\xi_j)\neq Null$ 且 $\{P(x_i,\xi_j), i=1,2\cdots n\}$ 的标准方差 σ 小于给定的阈值 T 时，则称特征 ξ_j 是样本集合 $\{x_i, i=1,2\cdots n\}$ 的共同特征。

定义 3.2 $\{x_i, i=1,2\cdots n\}$ 为一训练样本集合，$\{\xi_j, j=1,2\cdots m\}$ 为一组特征集合，$P(x_i,\xi_j)$ 表示样本 x_i 对于特征 ξ_j 的特征值，集合 $\{P(x_i,\xi_j), i=1,2\cdots n\}$ 表示样本对于特征 ξ_j 的特征值集合；如果对于特征 ξ_j，特征值集合 $\{P(x_i,\xi_j), i=1,2\cdots n\}$ 能划分成有限的 M 个子集，且子集内值域范围小于给定阈值 T_A，子集间的取值距离大于给定阈值 T_B，则称特征 ξ_j 是样本集合 $\{x_i, i=1,2\cdots n\}$ 的个性特征。



(a) 共同特征　　(b) 个性特征

图 3-1 单个类共同特征和个性特征在特征分布上的体现

如图 3-1 所示，图中横坐标 V 表示特征值的度量等级，纵坐标 S 表示同类的多个样本在某个特征值上的统计数目。本文中的 V 和 S 在坐标中都表示此意。3-1(a)图单一峰值表示同类物体的特征值在该特征上集中于某个值，具有共性。3-1(b)图多个峰值表示同类物体的值在该特征上分布在某几个区域，该类物体能够根据特征值分成多个子集，在其他区域上的取值都不符合该类物体的特点。对于个性特征而言，特征统计的分布在进行与其他类的特征比较中要根据多峰值分割成多个子集，每个子集分别与其他类的特征分布进行比较。

这两个定义很好地描述了某个类训练集样本在特征分布上的特殊情况，体现了样本特征值在该特征上的特点。并且样本特征值分布越集中表示样本在该特征上表现越相似，而且这样的特征适合于用来描述该类样本。特征的选择就是要提取能表现同类图像样本相似的特征，共同特征的分析是选取分类特征的重要步骤。

3.2.2 类间的特征关系

对于某个特征而言，各个类的特征值分布之间存在着某些联系，而且这些联系也是它们之间相似或差异的一种表现。如果类间的特征值分布重叠较多那么在

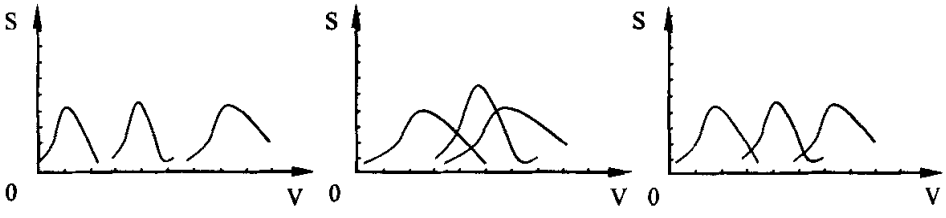
该特征上类与类具有相似性；如果特征值分布处于不同的值域，那么在该特征上类与类是有差异的。在某些特征上，各个类可能具有相似性，特征值的分布较为集中；也可能某些类相似，还有些类的特征值分布具有独立性。因此，多个类之间在特征上的分布关系有：

（1）对于特征ξ_{jm}，特征值集合两两互不相交；

（2）对于特征ξ_{jm}，至少两个类特征值集合大部分相交；

（3）对于特征ξ_{jm}，存在两个类取值区域小部分相交。

对于多特征而言，不同类之间在不同的特征上的分布对于分类是个多维的关系。而且各个类在特征上的关系也会是有交叉的，例如在特征 a 上，类 x 与类 y 相似而与类 z 有很大差异；在特征 b 上，类 x 与类 y 差异而与类 z 相似。因此，各个类在特征上的分布关系是分类特征提取的重要依据。



(a) 无交集　(b) 大部分相交　(c)少量相交

图 3-2 类间特征值分布关系

3.3 分类误差

3.3.1 分类误差的定义

上一节分析了各类样本特征值分布的类间关系，在类间特征值分布交叠的区域的分类判断就存在一些误差。因此，分类误差的大小也决定了分类结果的可信度，同时也是进行分类特征选择的一个标准。下面给出分类误差的定义。

定义 3.3 $\{x_i, i=1,2\cdots n_i\}$和$\{x_j, j=1,2\cdots n_j\}$是两类训练样本集合，对于特征ξ_j，根据两个类在特征ξ_j上的特征值集合的交集情况，分类到第 i 类的错误率e_{ij}称为第 i 类在特征ξ_j上相对于第 j 类的分类误差。

由于各个类训练样本的数目不同，误差计算时要求每个类的特征曲线的面积积分是相同的，因此将特征的统计分布进行归一化处理，那么各类特征分布曲线的面积积分为 1。而类 i 的样本分布统计值 $f_i(x)$归一化公式为：

$$f_i^*(x) = f_i(x)/N(i) \tag{3-6}$$

其中 $N(i)$表示第 i 个类的样本个数。

3.3.2 误差分割点的设定

从类间特征关系来看，误差的分析主要从特征值分布上来分析两个类特征值分布之间交叉的程度。

对于符合正态分布的特征值曲线，如下图 3-3 所示，分类误差从交点处计算，即：

$$e_{ab}=\frac{s1}{S1},e_{ba}=\frac{s2}{S2} \tag{3-7}$$

其中 *S1* 是曲线 *a* 的积分面积，*S2* 是曲线 *b* 的积分面积。

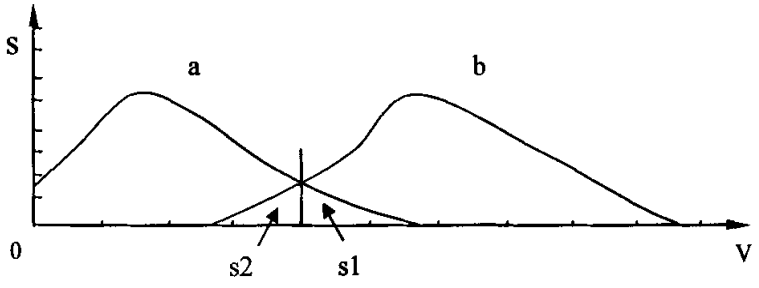
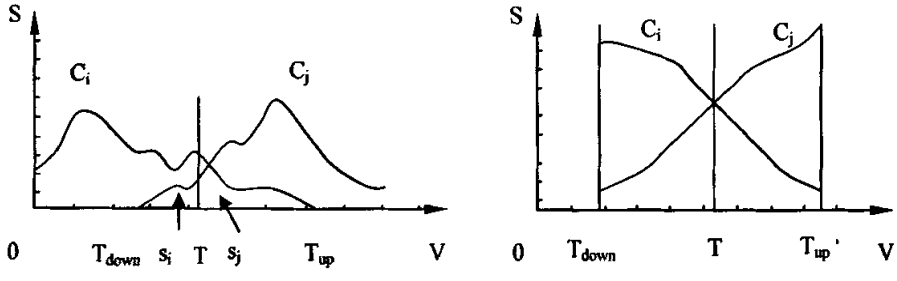


图 3-3 分类误差计算

对于不规则分布特征值曲线，如图 3-4(a)所示，两类的特征值在$[T_{down},T_{up}]$处有交集，因此要判断在此区域两类分类的误差。C_i 的误差 e_i 等于 s_i 区域面积除以 C_i 曲线的面积，C_j 的误差 e_j 等于 s_j 区域面积除以 C_j 曲线的面积，分割点 T 在$[T_{down},T_{up}]$区域内滑动，两个类的误差也相应变化，为了使 e_i+e_j 最小，要选择合适的 T。

分割点 T 的选取方法：在$[T_{down},T_{up}]$区域内取不同的 T 计算得出两类的误差面积曲线如图 3-4(b)所示，在两个面积曲线的交点处分割可使得两个误差面积的和为最小，该交点即为误差计算的分割点 T。分类误差的计算公式为：

$$\text{分类误差}=\text{类误差面积}/\text{类分布曲线的积分面积} \tag{3-8}$$



（a）特征值分布 （b）面积曲线

图 3-4 分类误差

3.3.3 类间误差分析表

基于特征的分类是多个特征的融合来分类，单个特征通常不能将所有的类完全分开，或者只能将这些类分成若干个子集。两个类在特征分布上的关系表现了该特征对于分开这两个类具有多大的效用，根据 3.2.1 节给出的特征类间关系，通过误差分析可以将类间误差大的类合并成一个子集。

定义 3.4 对于特征ξ_j，如果类 i 和类 j 的分类误差$e_{ij} > T_{ij}$，则类 i 和类 j 称为在特征ξ_j上的相似类。

相似类是类间关系的一种表现，但是其不具有传递性。如图 3-2(b)的分布来看，第二类分布分别与其他两个类的分布有较大的交集，分别与这两个类构成相似类，但是与之相似的两个类之间并没有相似关系。但是对于分类而言，这三个类无法获得很低的分类误差，需要看作一个子集。但是对于每个特征，可以先从类两两之间的关系着手来分析，根据分类误差的计算，每个特征可以得到一张类间误差分析表。

假设有 n 个类，对于特征ξ_j，类 i 与类 j 进行分析，在该特征上类 i 的分类误差为e_{ij}，类 j 的分类误差为e_{ji}，这样各个类两两之间的特征关系用分类误差来表示，对于每个特征，都可以获得一个误差分析表。

表 3-1 特征ξ_j的分类误差分析表

ξ_j上的分类误差	C_1	C_2	C_3	……	C_{n-1}	C_n
C_1	0	e_{12}	e_{13}	……	$e_{1(n-1)}$	e_{1n}
C_2	e_{21}	0	e_{23}	……	$e_{2(n-1)}$	e_{2n}
C_3	e_{31}	e_{32}	0	……	$e_{3(n-1)}$	e_{3n}
……	……	……	……	0	……	……
C_{n-1}	$e_{(n-1)1}$	$e_{(n-1)2}$	$e_{(n-1)3}$	……	0	$e_{(n-1)n}$
C_n	e_{n1}	e_{n2}	e_{n3}	……	$e_{n(n-1)}$	0

从分类误差分析表可以获得在特征ξ_j上类与类之间的差别关系，并且可以得到两个类之间区分的特征。如果分析表中第 C_i 行的值很小，那么可以说该特征能将类 C_i 分类出来。因此，可以根据每个特征的误差分析表进行分类特征选择和分类规则的制定。

3.4 图像特征的匹配

当图像的特征确定以后便确定了一个特征空间，图像的识别便是通过在此特征空间中定义某种形式的距离测度完成。其特征的匹配程度就是图像的分类误差。图像的特征具有特殊性，在描述上能用数字描述也能用其中抽象的符号来代表。

3.4.1 抽象的符号特征

对于一个标示性或者抽象符号的特征作距离的比较，在特征 k 上的特征值为 ξ，图像特征 x 的匹配公式为：

$$d^{(k)}=\begin{cases}0, x_k=\xi\\ 1, x_k\neq\xi\end{cases} \tag{3-9}$$

由于抽象的符号特征的匹配应该是完全匹配的，因此，其图像的匹配误差只有 0，1 两个值。

3.4.2 数字化特征

数字化的特征就可以用待测图像的特征值与各个类别的特征值进行比较，根据距离远近来分类。

对于落在两个子集之间的待识别样本，其分类的误差是根据它的特征值到各子集在该特征上的数学期望 ξ 的距离来计算的。将距离值归一化到[0，1]的区间[33]，这样需要一个特征值的距离测度阈值 t_k。由于样本之间的差异在超出一定范围时可以直接视为他们的差异是很大的能被区分，那么在阈值以外的距离值可以都用 1 来表示，无需具体表示。因此，数字化特征的距离值公式改为：

$$d^{(k)}=\min(|x_i-\xi|/t_k,1) \tag{3-10}$$

3.5 分类特征的选取

3.5.1 共同特征和个性特征

共同特征描述的是类中样本在某特征上的共性，表示样本特征值在某个值域上的集聚性。

设 C_i 类有 N_i 个样本，样本 t 的第 k 个特征值为 x_t^k。根据方差分析方法，使得类内误差最小，类间误差最大[34]。由样本均值和标准差的公式(3-11)和(3-12)可得到 C_i 类的 k 个特征的均值 $(u_i^1,u_i^2,...,u_i^k)$ 和标准方差 $(\sigma_i^1,\sigma_i^2,...,\sigma_i^k)$。

$$u_i^k=\frac{1}{N_i}\sum_{j=1}^{N_i}x_j^k \tag{3-11}$$

$$\sigma_i^k=\sqrt{\frac{1}{N_i-1}\sum_{j=1}^{N_i}(x_j^k-u_i^k)^2} \tag{3-12}$$

如果 $\sigma_i^k<T_\sigma$，那么特征 k 是类 C_i 的共同特征，且样本特征值集中在 $[u_i^k-\sigma_p^k,u_i^k+\sigma_l^k]$ 区间。

个性特征表示所有同类样本在某特征的有限个区域内具有集聚性。对于每个特征 k，把所有同类样本的取值进行桶聚类[35]。设类 C_i 的特征 k 的初始类数目

$N=0$，第 N 类的特征值范围为 $[a,b]$，则新加入的样本点值 Pot 到类 N 的距离计算公式为：

$$D_N|Pot,[a,b]| = Max(|Pot-a|,|Pot-b|) \tag{3-13}$$

如果 $Pot \in [a,b]$，则 $D_N|Pot,[a,b]|=0$。给每个距离排序，把新样本聚类到最小的 D_N 的类。给定样本聚类的取值的最大宽度为 T_1，即 $b-a<T_1$。对新样本加入的类修改特征值范围。如果 $D|Pot,[a,b]|>T_1$，建立新类，其取值范围为 $[Pot,Pot]$。否则给每个距离排序，把新样本聚类到 D_N 最小的类。如果 $Pot<a$ 则 $[a,Pot]$；如果 $Pot>b$ 则 $[Pot,b]$。这样根据特征 k 类 C_i 被分成 N_k 个类，每个类的取值范围用 $Pot_k[N_k]$ 记录。对 $Pot_k[N_k]$ 进行二次聚类，距离小于 T_B 的类合并。

个性特征将一个类的特征值分布分成了多个子区域，可以将其分成多个子类来进行与其他类的误差分析。因为分类误差的设定主要是从符合正态分布的分布曲线引申出来的，而且可以考虑到某些子类具有的独特性，无需与其他类进行误差的分析。

3.5.2 分类相关特征定义

与单独一个类的特征分布相对应，类与类之间在特征上的表现也存在相关性，而这些具有独立性而且能将各个类分开来的特征对于构造分类器具有重要的作用。因此，从类与类之间在特征分布上的关系对分类的影响可以得出下面的定义：

定义 3.5 $\{x_{1i},i=1,2\cdots n_1\},\{x_{2i},i=1,2\cdots n_2\},\cdots,\{x_{Ni},i=1,2\cdots n_N\}$ 分别为 N 个类的训练样本集合，$P(x_{ki},\xi_j)$ 表示样本 x_{ki} 对于特征 ξ_j 的特征值，集合 $\{P_k\}=\{P(x_{ki},\xi_j),i=1,2\cdots n_k\}$ 表示第 k 类样本对于特征 ξ_j 的特征值集合；对于第 k 类训练样本在特征 ξ_j 上的特征值集合 P_k，如果存在 $\forall e_{ki},i\neq k$ 有 $e_{ki}<T_e$（T_e 为给定的误差阈值），则称特征 ξ_j 是第 k 类训练样本集合 $\{x_{ki},i=1,2\cdots n_k\}$ 的显著特征。

定义 3.6 $\{x_{1i},i=1,2\cdots n_1\},\{x_{2i},i=1,2\cdots n_2\},\cdots,\{x_{Ni},i=1,2\cdots n_N\}$ 分别为 N 个类的训练样本集合，$P(x_{ki},\xi_j)$ 表示样本 x_{ki} 对于特征 ξ_j 的特征值，集合 $\{P_k\}=\{P(x_{ki},\xi_j),i=1,2\cdots n_k\}$ 表示第 k 类样本对于特征 ξ_j 的特征值集合；对于第 k 类训练样本在特征 ξ_j 上的特征值集合 P_k，如果存在 $\forall\{P_i\},i\neq k$ 有 $\{P_i\}\cap\{P_k\}=\phi$，则称特征 ξ_j 是第 k 类训练样本集合 $\{x_{ki},i=1,2\cdots n_k\}$ 的独特特征。

这两个定义给出了如何寻找最佳的特征来进行分类，只有使分类误差最小的特征才能作为分类的特征。当然对于多个类而言，并不是所有的特征都能够使它们之间的差异比较大，在临界处的分类误差比较小。这样就要选出能够把所有类区分开，并且使得分类的决策规则具有最小的误差率。分类误差小即表示类之间可分。

各个特征在分类中的置信度不同，特征的作用结果也不同。特征集合中除去

分类无关特征后的特征子集中，单个特征不可能把所有的类都给区分开，某些特征只能将所有类粗分成一些子集，再利用其他的特征进一步将子集分成一个个的类。下面给出基于分类作用的分类定义：

定义 3.7 对于特征ξ_j，如果它是各个类的显著特征，则特征ξ_j为完全分类特征。

定义 3.8 对于特征ξ_j，将该特征上的相似类合并成多个子集，且子集间的分类误差$e_{sq} < T_{sq}$，则特征ξ_j为可分类特征。

如果存在完全分类特征，仅一个特征就可以将所有的类区别开来，这样，每个类在该特征上的值域分布就是分类的取值范围。当然，由于训练样本的不完备，可能处于值域范围附近的某些值也可能是某个类的特征值，因此，还需要参看其他特征。

如果不存在完全分类特征，需要从分类误差较小的特征中选取一些特征用于分类。由于相似类的存在使得某些特征将各类分成了多个子集，这些子集之间也存在分类的关系。因此，我们要寻找子集间分类误差小且经过多次分类后能将各个类分别开来的特征集合。

3.5.3 分类无关特征的剔除

对于初始的特征集而言，因为要包含能描述物体的所有特征，所以特征集是很大的。某些特征本来不是这个分类问题的特征，只不过因为对分类问题本身的领域知识不了解而被加入进来，它们对分类没有任何帮助。或者这些特征在各个类中的值域在很相近的区域，是所有类共有的特征。因此，为了加快特征分析的进度首先要剔除掉多余的无关特征，没必要在无关的特征上做分析。

定义 3.9 对于特征ξ_j，如果对于每个样本x_{ki}的特征值$P(x_{ki},\xi_j)$为空或者在该特征上各个类两两互为相似类，则特征ξ_j为分类无关特征。

对每个特征都获得一张分类误差分析表，表内误差的均值大于某个给定阈值T_e，那么可以知道在这个特征上，各个类之间的类间误差都比较大，且各个类在该特征上都具有相似性，故可以剔除该特征。

3.5.4 分类冗余特征

所谓冗余特征是多个特征具有相同的分类结果，选取分类置信度最高的特征作为分类特征，其他的特征是该分类结果下的冗余特征。

特征与特征之间具有相关性，具有相同性质的特征在描述物体时的差异不大，而且通常初始特征集为了尽可能不丢失描述物体的特征，会把所有能描述的特征放入初始特征集合。或者当物体之间差异较大时，某些类物体在多个特征上都具有独立的特征值，即类与类之间的差异在多个特征上都得到了体现。一般来说，没必要选取多个冗余特征增加算法的复杂度。因此，选取分类置信度最高的

作为分类特征，其余的特征作为备选特征，按分类置信度的高低放入备选特征集合。

3.5.5 特征的划分能力

所谓特征的划分能力是指该特征能将各个类分开的程度，能使各个类获得的分类置信度越高，并且能够将类集分成的子集越多，特征的划分能力就越强高。

特征的划分能力具有如下的两种情况：

（1）某个类的显著特征集合不为空，那么该集合中的特征具有较强的划分能力，能通过该特征独立分出一个类，并且该类的分类置信度最高；

（2）任何类都不存在显著特征，那么需要多个特征的融合来分类，并且这些特征能够使分类结果的置信度最高。这样说明对于每个可选择的特征来说都存在相似类，需要合并相似类将其作为一个整体来分类。

如图 3-2(b)中讨论的特征在类间的关系，说明类的相似性不具备传递性。类 x 与类 y 相似，类 y 与类 z 相似，而类 x 与类 z 不相似。但是，特征作为分类来说，这三个类都是无法利用它来区别开来的，它们的特征值区间内具有不确定性，因此需要将它们都合并成一个子集，并且检查类两两之间的分类误差，将类间分类误差小的关系做标记，当已经将子集中的其他类分类好了的时候，如果这个分类误差最小可以用该特征来区分这两个类。分别根据每个特征，合并所有的子集并计算子集间在某特征上的分类误差，将子集与之相关的所有分类误差的和作为该子集在特征上的分类误差，并且子集的分类误差就体现了特征的划分能力。

3.5.6 基于样本整体分布的特征选取指标

通过上面几节对整体样本分布的分析，根据特征的划分能力，本文提出以下几条在决策树构建中分类特征选择的指标：

（1）类间相关度最小；

（2）分支数目最多；

（3）类间距离最大。

3.6 本章小结

本章首先分析了一般决策树特征选择的指标，给出了一系列有关的定义，提出了基于类间特征分布的进行特征选择方法与标准。

利用统计的方法获得各类样本的特征值分布，讨论了类间分类误差的计算方法，利用距离公式的归一化来处理在特征值分布之外的样本特征值的分类误差。针对分类特征对决策树的影响给出了相关概念，分析了在不同的类间特征分布关系下特征的分类能力，提出了决策树构建中新的分类特征选取的指标。为下章基于样本整体分布的决策树生成算法设计提供了理论依据。

第四章 基于样本整体分布的决策树算法

前两章对图像特征和样本的特征分布关系进行了详细阐述，并且给出了分类特征选择的依据和指标，本章则重点研究讨论基于这些依据和指标的决策树生成算法。

4.1 数据分类

4.1.1 分类器的构造

分类是数据挖掘领域的一种非常重要的方法。它是指依据所分析对象的特征或属性，建立不同的组类来描述事物。分类通过分析训练数据样本，产生关于类别的精确描述。这种关于类别的描述通常由分类规则组成，可以用来对未来的数据进行分类和预测。

解决分类问题的关键是构造分类器。从抽象的角度来看，分类器的构造是对样本集的统计分析及学习，并从中获取未知的、隐藏于数据中的重要信息——分类判别函数。然后选定一种带有参数的模型函数，通过学习确定模型函数的参数，用它来直接或间接地拟合或逼近分类判别函数。

图 4-1 描述了分类器的典型构建过程：

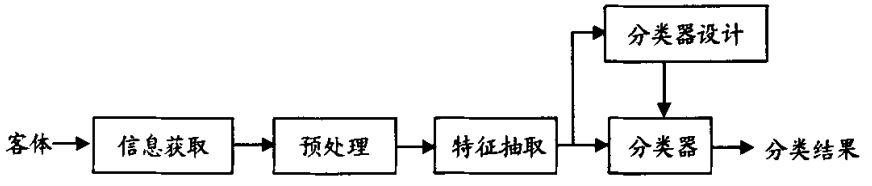


图 4-1 分类构建过程

4.1.2 分类方法的评价标准

对于某一具体的分类方法，可以采用以下标准来进行评价[36]：

（1）错误率：错误分类的事例数目与总事例的比率，用于标示模型正确预测新数据类标号的能力，是评价分类算法最重要的指标。

（2）速度：产生和使用模型的时间开销。某些情况下，这是一个重要的因素。

（3）健壮性：有噪声数据或空缺值数据时模型正确分类或预测的能力。

（4）伸缩性：对于给定的大量数据，有效地构造模型的能力。

（5）可解释性：学习模型提供的理解和观察的层次。

（6）模型的大小：即算法的紧凑性。

4.1.3 分类器的错误率

影响分类器错误率的因素有以下几个方面[37]：

（1）训练集的记录数量；

（2）属性的数目；

（3）属性中的信息；

（4）待预测记录的分布。

有两种方法可以用于对分类器的错误率进行评估，它们都假定待预测记录和训练集取自同样的样本分布：

（1）保留方法（holdout）：记录集中的一部分（通常是2/3）作为训练集，保留剩余的部分用作测试集。

（2）交叉纠错方法（cross validation）：数据集被分成k个没有交叉数据的子集，所有子集的大小大致相同。生成器训练和测试共k次；每一次，生成器使用去除一个子集的剩余数据作为训练集，然后在被去除的子集上进行测试。

4.2 决策树算法改进

4.2.1 基本过程

对一个分类问题或规则问题，决策树的构造是一个从上至下、分而治之的过程。这种方法将数据按树状结构分成若干分支，每个分支包含数据的类别归属共性，相当于分类发现中的类。这样可从每个分支中提取有用信息，发现数据中蕴涵的分类规则。如何构造精度高、规模小的决策树是较为核心的内容。

通常，理想的决策树可以分为以下3种：

（1）叶结点数量小；

（2）叶子结点深度最小；

（3）叶结点数最少且叶子结点深度最小。

决策树的初步生成是指由训练集建立决策树的过程[38]。这一过程可以简要地描述为以下几个步骤：

S1. 依据实际需求以及所处理数据的特性选择属性，包括作为分类依据的类主属性以及候选属性集；

S2. 在候选属性集中选择最有分类能力的属性作为当前决策结点（第一个决策结点称为根结点）的分类属性；

S3. 根据当前决策结点属性取值的不同，将训练数据集划分为若干子集；

S4. 针对S3中得到的每一个子集，重复S2和S3，直到最后的子集符合下面的三个条件之一：

（1）子集中的所有数据记录都属于同一类；

（2）该子集是遍历了所有候选属性得到的；

（3）子集中的所有剩余候选属性取值完全相同，不能根据这些候选属性进一步进行子集划分。

S5. 对满足子集中的所有数据（非单一）记录都属于同一类条件的所有叶子结点，依据其中的数据记录所属类别进行类别标识。

在决策树的初步生成过程中，将训练集作为输入，并由此得到一个基本的决策树。该决策树的每一个决策结点对应着一个分类属性，分支对应着依据该属性进一步划分的取值特征，叶子结点代表着各个类或类的分布。

4.2.2 最优算法

如 4.1.3 节所述，错误率是评价分类算法最重要的指标。因此，最优决策树应该能在当前状况下获得最低的错误率。

设初始状态时集合 $C:\{c_i \mid i=1\}$ 中唯一的元素 c_i 包含所有的待分类，而特征集合 Ξ 包含所有可用于分类的特征；在以建立最优决策树为目的的前提下，一元函数 OptRuler(·) 返回当前应当选取的分类特征，二元函数 Classify(·,·) 完成依据当前特征对 C 的划分操作。于是，建立最优决策树的过程可以描述为：

```
while (|c_i| ≠ 1 || |Ξ| ≠ 0) {
    ξ = OptRuler(C);
    Ξ = Ξ − ξ;
    {D_i} = Classify(C, ξ);
    C = (C − d_{i,j}) ∪ D_i
}
```

该循环结束后，可能出现以下三种情况：

（1）$|c_i|=1$ 且 $|\Xi|=0$，刚好可以建立完整分类的决策树；

（2）$|c_i|=1$ 且 $|\Xi|\neq 0$，有多余的特征未用于决策树的建立；

（3）$|c_i|\neq 1$ 且 $|\Xi|\neq 0$，利用当前的特征集无法建立完整的决策树。

4.2.3 改进的特征选取函数

4.2.2 节对建立决策树的最优算法进行了介绍。然而，该算法仅仅用于描述算法的基本框架，而对于选取当前分类特征的具体准则尚未涉及。因此，本节将对 OptRuler(·) 函数的设计方法进行讨论。

结合 3.3.2 节所述分类误差，本文依据同一特征值下任意两类之间的相关程度进行当前特征的选取。

设最优算法中的 while 循环运行到第 i 次时，对应的待分类集合为 $C^{(i)}:\{c_{i,j} \mid j=1,2,\cdots,n;n\leq|C|\}$；显然有 $|c_{i,j}|\neq 1$。对所有 $c_{i,j}$ 中所包含元素分别建立无向图，从而得到图集 $\{G_j=(V_j,\varnothing) \mid j=1,2,\cdots,n;n\leq|C|\}$。这些图均用邻接表表

示，记作$\{A^{(j)} \mid j=1,2,\cdots,n; n \le |C|\}$。

```
for (j=1; j<=|Ξ|; j++) {
    // 将{A^(j)}清零。
    for (x=1; x<=|c_{i,j}|; x++) {
        for (y=1; y<=|c_{i,j}|; y++) {
            // 若 v_x^(j) 与 v_y^(j) 的相关程度超过阈值 R_Threshold，则令 a_{x,y}^(j) = a_{y,x}^(j) = 1。
        }
    }
    // 计算{A^(j)}各个元素中连通分支的数目，
    //      用其和描述特征ξ_l对C^(i)中各个元素进一步划分的能力。
}
// 返回对C^(i)具有最强进一步划分能力的特征ξ_l。
```

初始的状态集 C 包含一个元素，该元素标记了所有待分的类。实际上集合 C 存储的是还需要细分的树结点。选取 C 中一个元素，根据给定的阈值 $R_{Threshold}$ 获得的最多连同分支的特征为 i，则特征 i 是该结点的分类特征，其连同分支就是它的下一层树结点。已经分过的结点从集合 C 中删除，将其子结点中仍需要分类的状态插入集合 C。当集合 C 为空时则表示所有的类都已经生成为叶结点，决策树构建完成。

该算法从类间最小的相关程度和最大区分度[39]两个方面来考虑的，这样构建的决策树叶结点深度小，且树的各个分支间的误差率小。

4.2.4 性能分析

在 1.2.3 节介绍的各种算法中，ID3 依据对样本分类熵的信息增益来选取当前特征。C4.5 利用信息增益率来选取当前特征。这些算法都是根据特征值的划分区间将所有样本分成若干子集，在各级子结点上对样本递归计算信息增益，这样需要不断统计样本的特征值，其计算复杂度高且需多次遍历各个样本特征值。并且连续特征值的离散化程度也是影响决策树构建速度和深度的重要因素。

本文采用的特征选择算法的时间复杂度与特征数目和类别数目有关。而在计算两类之间的相关程度时，由于使用的是类在该特征下的概率分布，因此不受样本数目的影响。特征选择算法中所涉及到的数据主要是每个特征中，各个类两两之间的相关程度，即两个类之间的分类误差面积之和。因此，对于一个 k 类的分类问题，对每个特征分析需要读取的数据为 C_k^2，其数据量远远小于对样本分析的数据量，且这部分数据完全可以存在内存中，相对于传统的决策树特征选取算法，其时间复杂度要低，速度快。类与类之间的相关程度影响着利用构建的决策树得出的分类结果的可靠性，这种相关度是利用 Bayes 方法进行分析的。在不确

定区域的分类结果分析是基于 Bayes 的误差分析，而且本文提出特征选择函数每次选取的都是满足最小误差率的特征，因此构建的决策树对于给定的样本数据而言是局部最优的。

4.3 决策树的分析

4.3.1 特征选取函数的优化

然而，对于OptRuler(·)函数的设计，4.2.3 节所述算法并不是十分的完善，其中还有一些需要进一步考虑的问题：

（1）当两个甚至多个特征同时具有最大的划分能力时，如何从中进行选取；

（2）当指定类的划分数目时，如何选取特征。

这些问题的解决依赖于在OptRuler(·)函数中引入其它的评价准则。

类的相关程度通常为正值，亦即两个类在同一特征下概率分布的重叠面积大小。显然，当两类不相关时，该重叠面积为零。为了能在这种情况下进一步对特征进行评价，本文引入类间距离的概念。

由于默认在某一特征下各个类的概率分布是存在的，因此，可以取各个类取值的数学期望作为类的位置，从而将两不相关类位置之间的距离定义为类间距离。显然，若该距离越大，决策时的不确定程度越小，决策的错误率亦会相应地降低。因此，对于分类而言，选取的特征应该能获得最大的类间距离[40]。

在上述特征选择算法中，选取的最强划分能力的特征是考虑的第一个找到的特征，而没有考虑同等划分能力下能获得最佳分类效果的特征。因此，在上述的特征选择中，对于具有最强划分能力的多个特征，选取类间距离最大的特征作为决策树结点的分类特征。

OptRuler(·)函数是对于给定的阈值 $R_{Threshold}$ 进行类间特征分析的，如果对于所有的特征而言，无法将所有的类在这个阈值下区分开来，那么对于仍需分析的状态集合 C 中的元素，应该适当增大阈值，使得建立完整的决策树。由于对象类的数目相对较小，而且已经作了初步的分类，那么阈值的调整可以根据仍需分类的对象间特征的最小相关度依次作为新的阈值，因为需要使得分类的错误率最小，那么树的分支间应该具有最小的相关度。

4.3.2 对象的增减

利用特征选择函数构建好一棵决策树后，随着分类问题的需要，待识别对象的数目出现了增减，那么已有的决策树则需要进行调整。本文的特征选择是根据各个类样本在特征上的分布关系获得的，分布曲线的变化使得它们在特征上的关系也发生了变化。

如果删除一个对象类，对于决策树而言，可以只从决策树的各个结点中删除

掉该对象的信息，并且删除该对象类的叶结点，如果删除该叶结点后只剩下其他一个叶结点，该结点需要合并到上层结点。但是，由于类与类之间的传递相关性，使得可以利用某个特征完全区分开来的两个类由于与某个类的特征分布都有很大的交集而合并成了一个子集分支。如果将这个使得它们相关联的对象类给删除了，那么原有的决策树就不一定是棵最优的决策树，只能说是棵能获得较低分类错误率的次优决策树。因为一棵最优的决策树不仅要错误率低，而且是叶结点的深度最小，即利用一个特征能获得最大的分支数目。

如果增加一个对象类，问题就变得有点复杂。如果新增加的类的特征分布与树结点中某个分支的特征分布相似，并且该类样本分布的加入使得分支之间的相关度在允许的范围内，则该类并入该分支，最后可以得到新对象类与已有的某个对象类的一个分支，那么只要寻找能最好区分这两个类的特征生成两个叶结点；或者该类在某个树分支结点的分类特征下是个独立的连通分支，那么只需要增加一个叶结点，不用寻找新的分类特征。

但是，如果该类样本特征分布的引入使得原来能分开的两个类与它的特征分布交叠较大而无法确定新增加的对象类应该属于哪个树分支。这样对于已经建立好的决策树，分支需要合并，树的结构也完全给破坏了，这时需要重新构建决策树。但是本文的建树过程简单时间复杂度低，而且只需要再计算该类与其他类在各个特征上的相关程度（类间两两误差之和），重建决策树也很快速。

4.4 分类规则的提取

通过特征选择构建了一棵最佳的决策树，在决策树的分支中，如果各分支包括的样本概率分布之间有重叠的面积，那么在这个重叠的特征区域内，需要判断分类结果并且分析其结果的误差。

决策树的各个结点选取了一些最有代表性且最能区分这些类的特征，这些特征作为描述这些类的知识。当然在各个类之间，这些特征的值域范围具有相关性，特征之间的融合构成了一个物体的描述，根据描述能告诉我们它的性质和归类。

因此，多个的特征之间取值的相互交织将特征空间划分成了不同的区域，每个区域表示了一类物体，这些区域有些与学习分析的物体有关，有些没有关系，这就需要对多维的特征空间进行分析以获得一个分类的准则。特征的融合也需要从分析特征空间关系出发来研究。

4.4.1 两个特征两个类的分类规则

首先分析利用两个特征进行两个类分类的情况。如图 4-2 所示，类 a 和类 b 的两种特征的样本分布曲线将二维特征空间划分成了 16 个不同的区域，样本的特征值分布在往右上的对角线上能比较容易的判断分类。但是在其它区域，根据

不同的特征其分类结果不同，而且分类结果的置信度也不相同，它取决于特征的分类能力的大小。满足分类能力大的特征其分类结果可信度要高些，反之要低些。

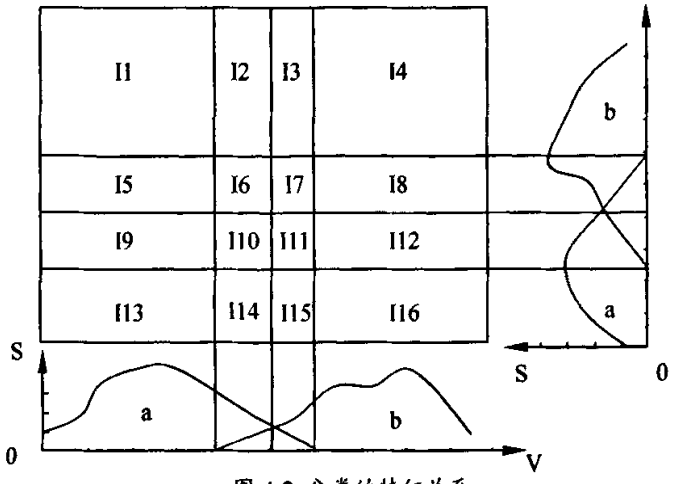


图 4-2 分类的特征关系

对于待测样本，它的两个特征值落在不同的区域其分类结果和置信度都不相同。假设类 a 和类 b 两个特征在分布曲线相交处的分类误差分别为 F1a、F1b、F2a 和 F2b，且对于每个特征而言，每个区域中两个类的分类误差之和为 1。很显然，两个类两个特征的分析是最简单的，根据类间的分类误差大小很容易判断分类的结果。这里，置信度的取值不限定在[0，1]范围，即比较两个类哪个置信度高就是哪个类。

表 4-1 各分类区域的结果分析

分类区域	分类结果	置信度
I1	不确定	N/A
I2	b	1+F1a
I3	b	1+ （1-F1b）
I4	b	1+1
I5	a	1+F2b
I6	F1a 和 F2b 大小确定	$\begin{cases} a: F2b+(1-F1a) \\ b: F1a+(1-F2b) \end{cases}$
I7	b	（1- F1b）+（1-F2b）
I8	b	1+（1-F2b）
I9	a	1+（1-F2a）
I10	a	（1-F1a）+（1- F2a）

I11	F2a 和 F1b 大小确定	$\begin{cases} a: F1b+(1-F2a) \\ b: F2a+(1-F1b) \end{cases}$
I12	b	1+F2a
I13	a	1+1
I14	a	1+（1-F1a）
I15	a	1+F1b
I16	不确定	N/A

如上述分析，不同的区域中哪个类的分类误差低则分到哪个类，且分类结果的置信度为$(1-e_1)+(1-e_2)$，即各个特征在此区域的可信度之和。

当然，待测样本的特征值也有可能根本没有落在这 16 个区域内，这样就要计算待测样本的特征值与每个类样本特征的距离。利用公式（3-10），为每个特征设定一个类间距离的阈值t_k，待测样本与每个类的最近距离与阈值比较来得到特征对每个类分类的可信度，即

$$d^{(k)} = \min(\min(|x-x_i|)/t_k, 1), x_i \in C_j^k \tag{4-1}$$

4.4.2 多特征两个类的分类规则

两个特征分类时处于两个类特征分布相交叠的区域，分类结果存在一定的误差，并且不属于特征分布区域内的待分类样本特征值根据邻近法来分类其可信度也不高。为了提高分类的可信度，需要利用其他特征来进行判别分类结果。如果在其他特征上分类的结果可信度更高一些，那么可以做出相应的判别。

这样，将分类特征集合中每个特征根据两个类间误差大小进行排序，作为待测样本进行特征匹配的顺序。一般来说，如果不存在不是两个类的样本，那么只要按照特征的顺序比较分类结果，若分类结果可信度为 1，则停止下一个特征的匹配，直接输出分类结果。对每个特征进行匹配，比较的分类结果可信度，具有最高可信度的分类结果即为输出结果。

如果待测样本的性质不确定，无法判别属于哪个类，那么扩展到利用 n 个特征进行两类物体分类，特征间的关系是个 n 维空间，则有4^n个区域，不同的区域分类结果也不相同。因此，要对每个特征进行匹配，获取可信度最高的分类结果，如果存在矛盾那么结果为不确定。简单来说，在所有的特征中，待测物体的特征值对于两个类来说分别可以找到一个置信度最高的分类特征。如果哪个类的分类置信度高，其分类结果就是哪个类；但是如果两个类分别存在置信度为 1 的特征，那么该物体在分类中它的特征表现出现了矛盾，即分类结果不确定。

4.4.3 多特征多个类的分类规则

分类特征选择的过程实际上就是分类决策树的构建过程。根据决策树每个结

点上的分类特征进行匹配，对于一个结点上的多个分支，待测样本的特征值总会落在某个分支上，或者是两个分支之间的误差区域，那么多类问题就转化为了两类问题。根据两个分支的样本在特征分布上的特点，每个分支获得一个分类误差，且两个分支的误差值和为 1。对于多叉树中分类结点的其他分支，这些分支以及其子结点所有的分支的分类误差都为 1，因为完全与那些类的特征值无关，这样类比较的数目也减少了。

利用决策树的分类规则为：

（1）获取待测样本的所有特征值；

（2）从根结点开始逐层与结点的分类特征作比较，对于完全不匹配的分支，其整棵子树上的分支的分类误差都为1。通过特征比较获得所有分支的分类误差，这样决策树就是一棵带分类误差的决策树；

（3）然后从根结点开始沿着误差最小的分支往下层到达的叶结点就是分类结果。其结果的误差率是从根结点开始所经历的所有分支中的分类误差的最大值。

4.5 先验知识的引入

4.5.1 先验特征

一个智能系统只有具备了足够的、正确的知识才能根据事实推导出正确的结论。知识的获取分成人工获取和机器自动获取。通过所提供的信息和知识，机器通过利用知识模型获得一些信息和表示。从大量样本数据中提取规律性的东西，也就是知识发现的过程。而图像数据信息的不完备性，为计算机学习提取规则带来困难。而人类的学习和常识推理很大程度上依赖于运用默认的一些信息的能力。在实际应用中，针对不同的需求，对物体的分类有不同的要求。而且机器自动分析对数据的处理精确度高，而人类在看待事物的时候通常具有模糊性，尤其是人类视觉在识别时是模糊的，而且物体相似性的界定也比较模糊。因此，要解决机器分析精确性和人类识别的模糊性的矛盾，需要引入启发式思想[41]。

所谓先验特征就是人们在识别物体时的经验知识，即某类物体必须具备的某些特征或者各类物体之间具有差异的某些特征。

在提取分类特征之前给定人类视觉理解的或者一些经验知识作为分类的先验知识，这些知识为分类特征选取和分类规则的制定起指导作用，为了获得更高的分类准确率，这样就首先给出一个类的共同特征知识和类间差异特征的集合。先验特征和机器分析的特征都是进行构建分类器的特征依据，因此需要将机器分析的结果和先验知识相融合，这样对于每个类的某些特征而言，机器分析和先验知识是否具有一致性，这两类知识在进行分类指导中具有怎样的关系，两者的可

信度如何，是特征融合需要考虑的问题[42]。

所以，本文将分析出来的特征和先验特征制定一个等级：

（1）是先验特征也是自动提取的特征；

（2）是自动提取特征但不是先验特征；

（3）是先验特征而不是自动提取的特征；

（4）分类无关的特征。每类特征具有一定的权值系数，其值为 d_i，且有 $d_1+d_2+d_3+d_4=1$，其中 $d_4=0$。

4.5.2 启发式分类特征选择

基于多特征的分类就是一个分类特征的融合问题。利用特征分类实现了一定的信息压缩，加快了处理速度。特征的组合是个很重要的问题。引入先验特征后，分类特征就必须考虑先验知识和学习训练出来的知识在分类中起到的重要程度。同时，由于人类对图像是别的模糊性，因此对于某些与机器精确学习得出的结果相矛盾的内容要进行识别和处理[43]。这些先验特征知识是通过人类视觉对图像类别区分的一种认识，也是在图像分类中分类效果的一种特殊的要求。同时，先验知识也可以包括某个类必须具备的特征。

通过对待训练的多个图像类的认识和初步的数据分析，能够获得一些认识和区别这些图像类的知识，这些知识对于机器精确的数据分析和分类规则的获取具有指导作用。在认识事物的时候，我们会注意到这类物体共同具有的特征，这些特征结合得到一个形象化的认识。当多个物体一起进行比较分析时，物体之间的差异也是首先获得的知识。那么，先验知识就包括物体类本身必须具备的特征和物体之间差异大的特征，而这些特征都应该作为分类特征。这些特征可以是根据特殊的视觉要求观察得来，也可以是通过经验数据分析得到。

对于各个类必须具备的特征集合，集合中的特征都是应该优先考虑用来分类的，分类特征选择是利用决策树来获得的，对每个特征依次比较类间的分类误差，因此，可以将先验知识中每个类必备的特征作为优先分析的特征，并且对其分析的阈值适当的放宽，使其能够对分类起到指导作用。

先验知识对于特征选择的指导具体方法是：

（1）将各个类必须具备的特征作为优先分析的特征，且在同等划分能力的条件下，选取先验特征作为树结点的分类特征；

（2）首先利用较小的误差阈值进行决策树的建立，如果先验特征没有被选为分类特征，则对这些特征分析时其误差阈值增大，重新构建另一棵决策树。

通过不同的阈值获得的分类特征与先验知识指导获得的分类特征，它们在分类时特征的权值系数是根据特征等级来获得的。如果用较小的误差阈值获得的分类特征中，其特征等级如等级 a 和 b；如果通过调整误差阈值获得的分类特征是

先验特征，那么其特征等级就是 c。

4.5.3 启发式分类规则

不同的特征分析顺序和误差阈值可以获得多棵决策树，每棵树选择的特征都是分类特征。在多棵决策树中，各个分支结点上不同的分类特征其特征等级不同，分类时对匹配结果作系数加权。待测样本从根结点开始对结点上的分类特征进行匹配，对于结点每个分支都会有个分类误差，且该误差是根据特征等级乘以了加权系数。如果某个分支的误差为 1，那么其下层的所有分支都为 1。这样，每棵决策树通过特征匹配变成了带加权误差的决策树。

从每棵决策树的根结点开始到达每个叶结点所经历的分支中，选择最大的分支误差作为该叶结点分类结果的分类误差，因此每棵决策树都可以找到一个分类误差最小的叶结点，这个叶结点获得的类就是该棵决策树的分类结果。将所有决策树获得的分类结果进行比较，选取误差最小的分类结果作为待测样本的识别结果。

4.6 本章小结

本章主要讨论了分类器构建中的特征选择函数的设计、完善与优化，并且与决策树算法进行了比较。决策树构建完成后，每个结点上的分类特征就构成了分类识别的特征空间。首先从简单的两类问题出发，给出了特征融合的方法，并分析了分类的距离函数设定。为了获得更高的分类准确率，考虑模糊性和精确性之间的关系引入了先验特征知识，将特征分成四个等级，不同等级的特征赋予不同的特征加权系数。先验知识对决策树的生成和分类规则提取进行指导。

第五章 图像分类的应用

基于对图像特征提取、分类特征选择和决策树生成算法的分析和讨论，本章着重对决策树的生成过程进行分析，并辅以具体的实例予以说明，通过该实例验证本文提出的算法。

5.1 实验平台

使用 Microsoft Visual C++ 6.0 开发工具建立了实验平台，图像识别分类系统包括特征提取，决策树生成和识别分类三个功能模块。系统建立分为两个阶段：训练阶段和识别阶段。首先输入多个类的训练样本，提取特征获得各类图像在各个特征上的特征值分布；然后根据类间的特征值分布构建决策树，获得分类特征；提取分类规则。对于待识别样本，将其特征值与决策树中相应的特征进行比较，获得误差最小的识别结果。图像识别分类系统处理流程如图 5-1 所示。

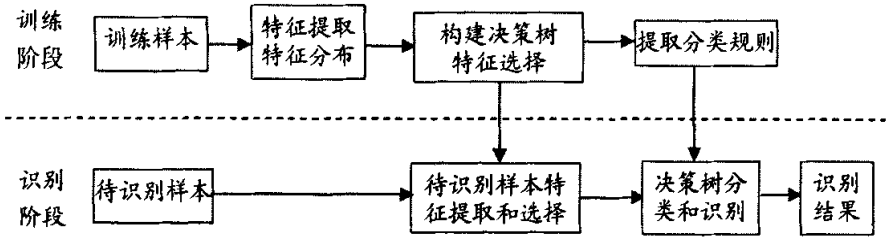


图 5-1 分类识别系统基本模型原理框图

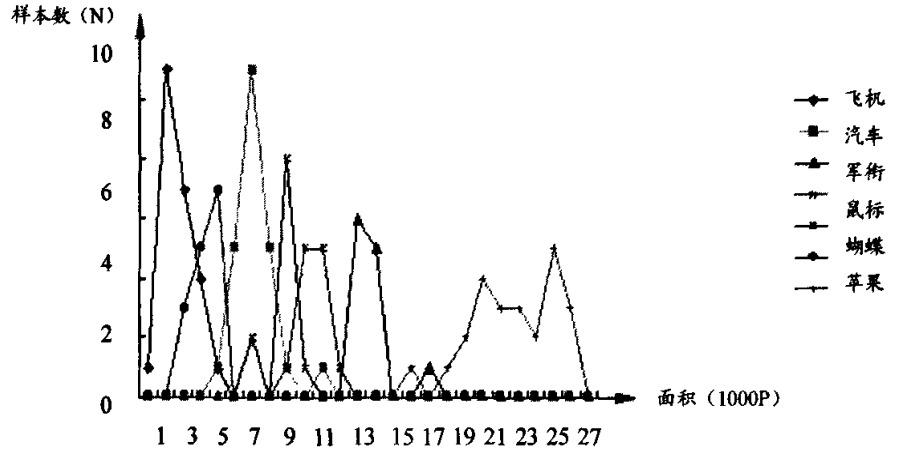
5.2 实例分析

本文选取了 7 类样本进行分析训练，训练集的样本分别为 C{飞机，汽车，军衔，鼠标，蝴蝶，苹果，手机}。这七类物体通过人眼视觉在特征上具有很大的差异，但是机器分析二维图像数据，得到的信息需要通过训练来提取分类规则。下图为实例的示例图片。

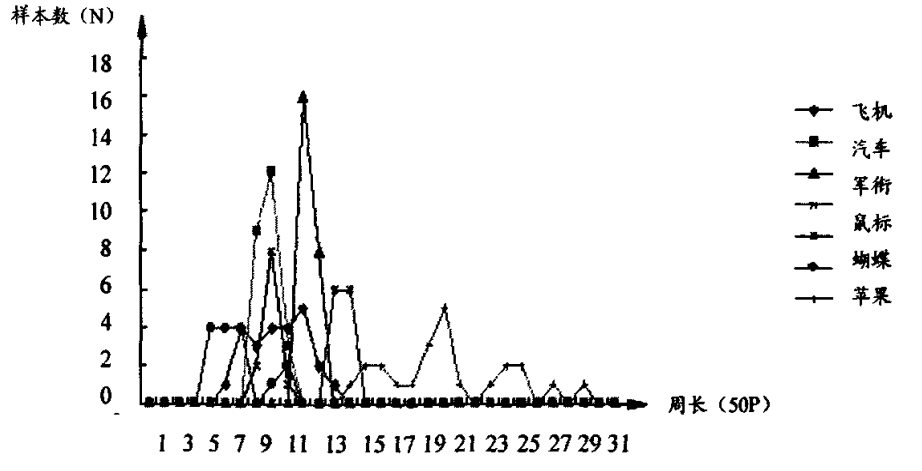
图 5-2 实验图片示例

5.2.1 图像特征提取

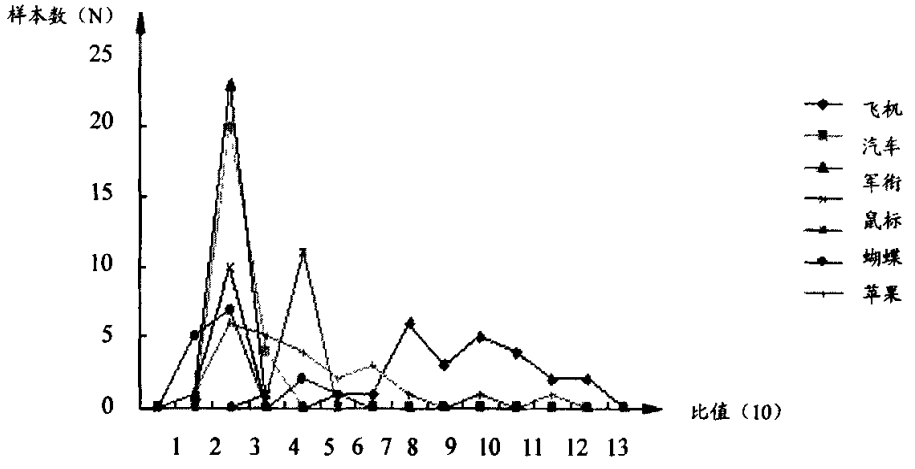
本文选取了图像的 19 个底层视觉特征作为初始特征，初始特征集合为Ξ。利用第二章给出的特征描述公式提取每个样本的特征值，对每个类的所有样本在各个特征上进行统计分布。为了降低统计的维数，使统计数组不至于过大，特征值的统计根据数值大小进行了量化，而且小数的统计也不是很方便，扩大若干倍成整数后再进行统计。对每个特征，各个类在特征上的样本统计分布如图 5-3 所示。图中的横坐标表示各个特征的区间等级，纵坐标内容表示对落在每个特征区间内的样本数目的统计值。图中各坐标单位字母分别表示：N=个数，P=象素点，G=色彩等级。其中色彩等级统计的特征的等级数分别表示为：红色、蓝色、绿色分量和平均灰度是 16 等级，主色和主色调是 8 等级。在计算类间误差时要求各类样本的积分面积一致，因此在进行构建决策树之前，对下述特征的统计值做了归一化处理，各个类统计值曲线的积分面积为 1。



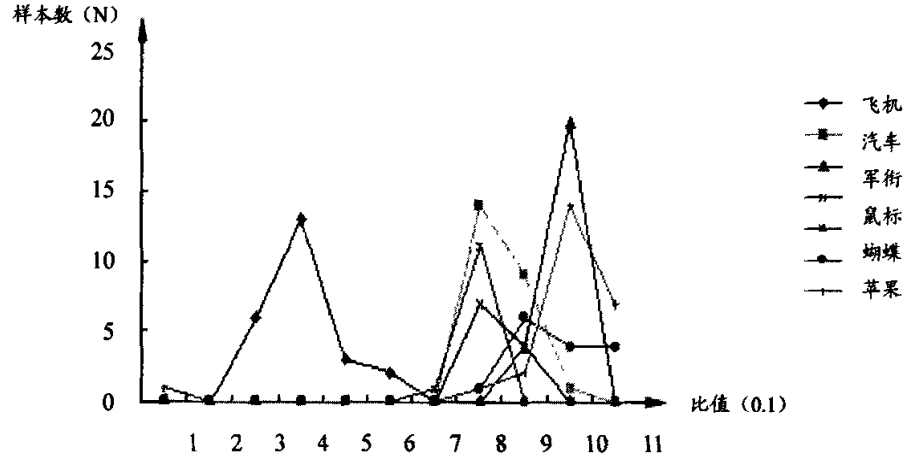
(a) 面积



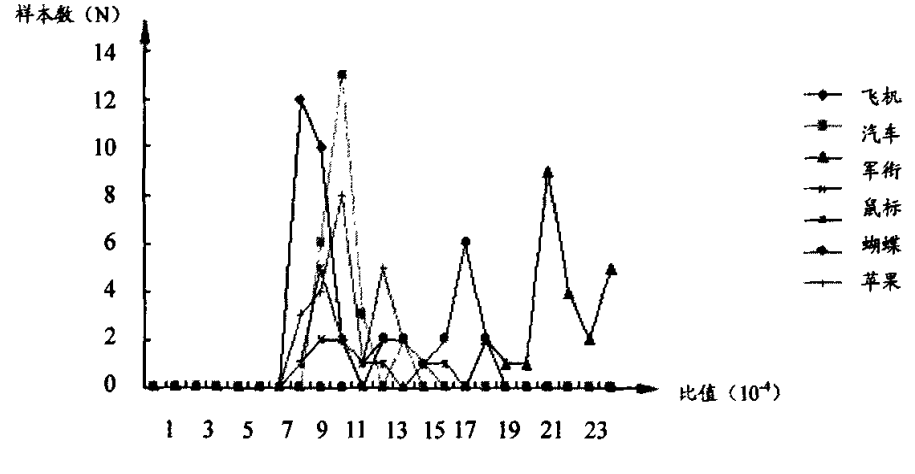
(b) 周长



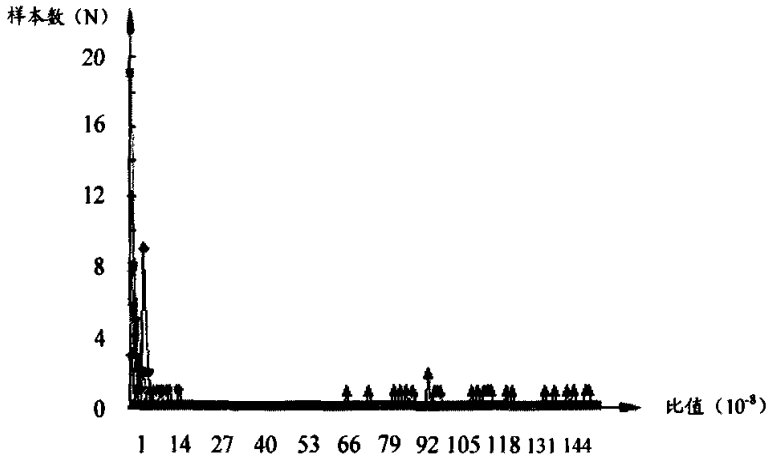
(c) 复杂度



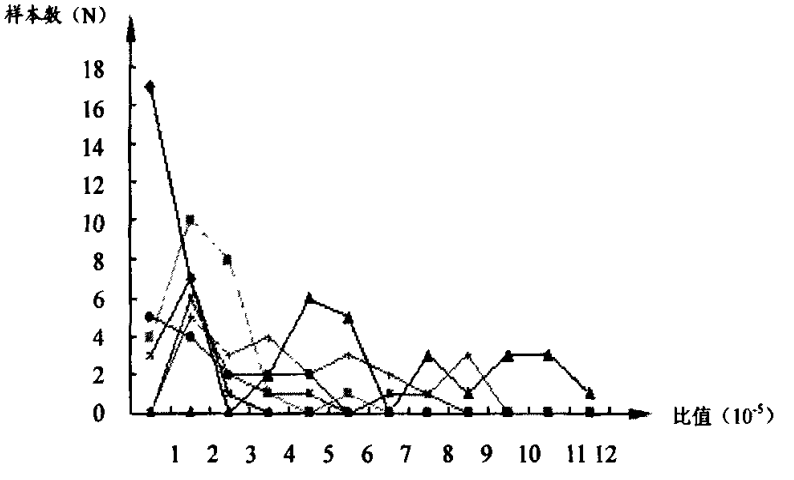
(d) 占空比



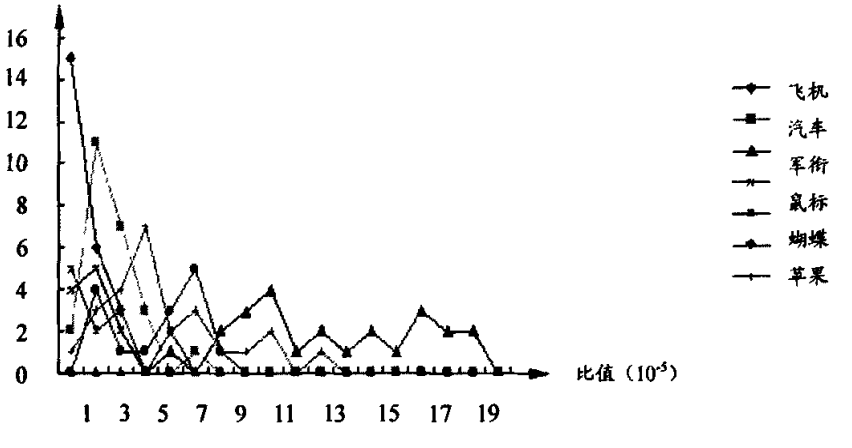
(e) 不变矩 1



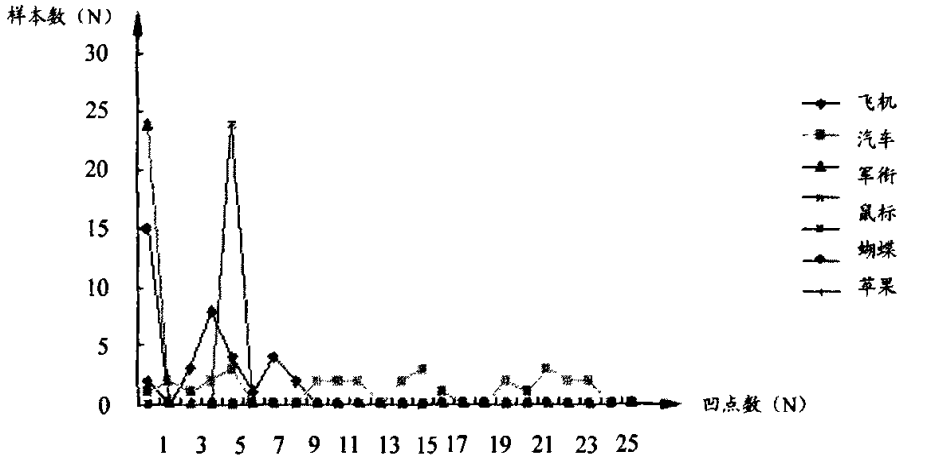
(f) 不变矩 2



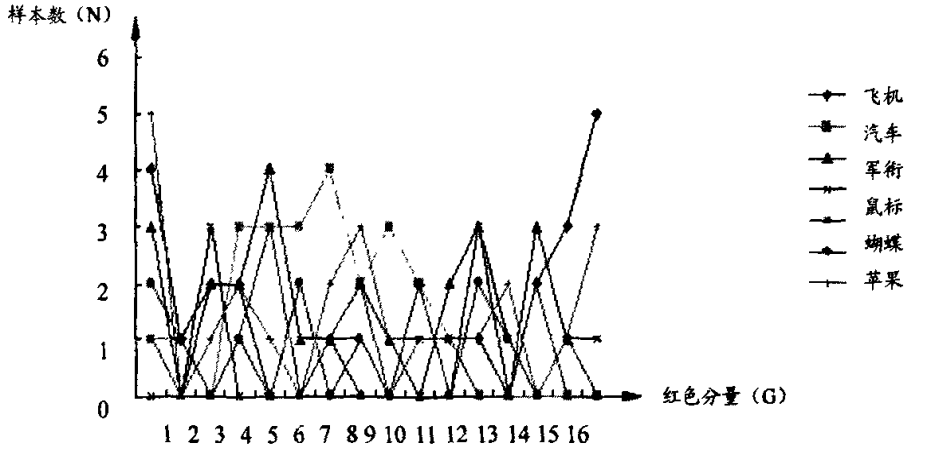
(g) 不变矩 3



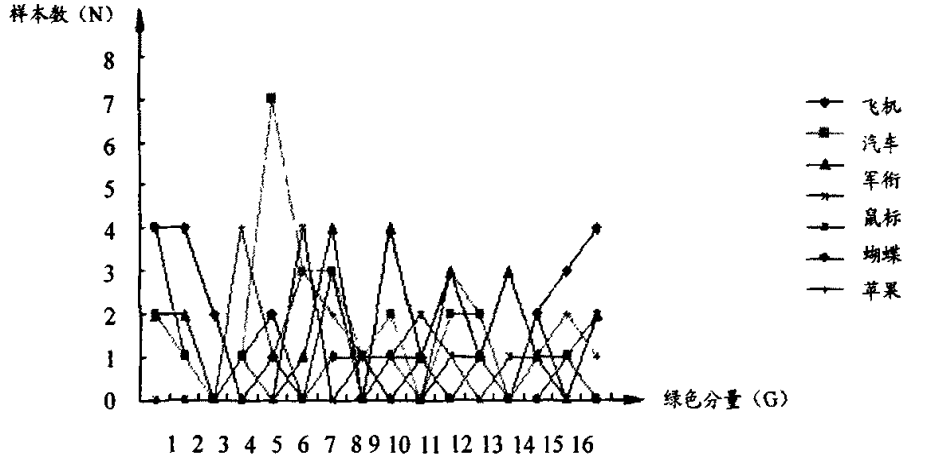
(h) 不变矩 4



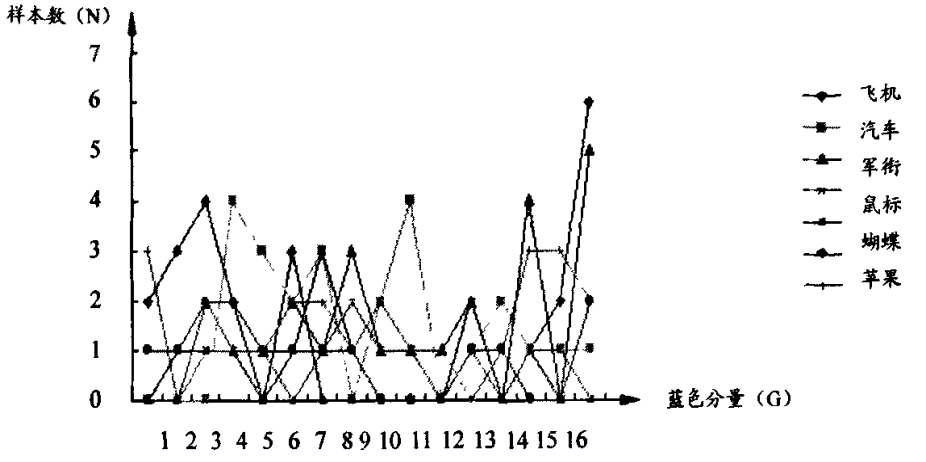
(i) 凹点数



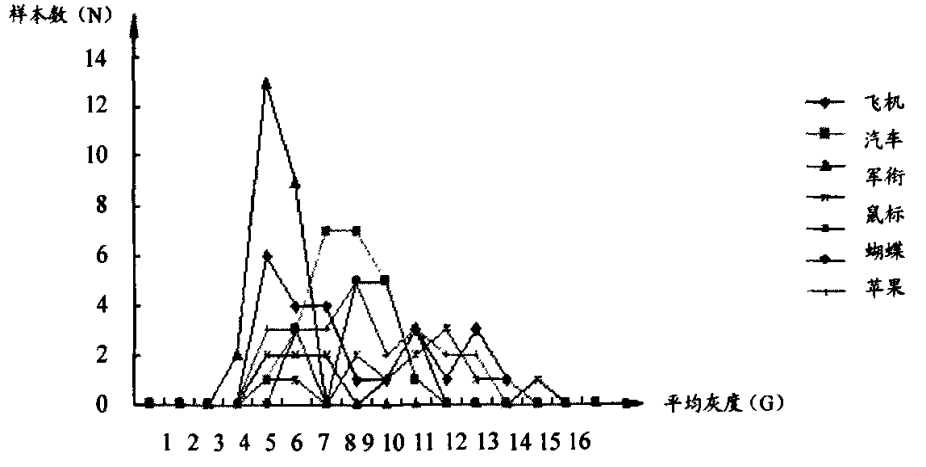
(j) 红色分量



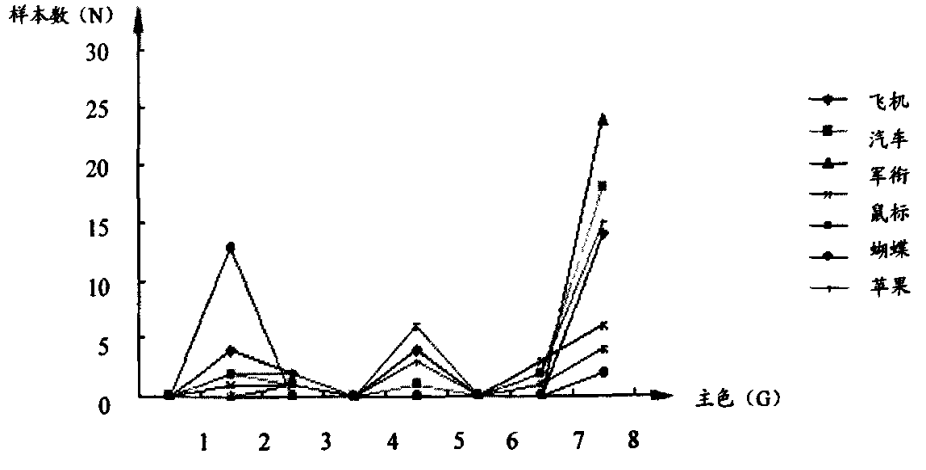
(k)绿色分量



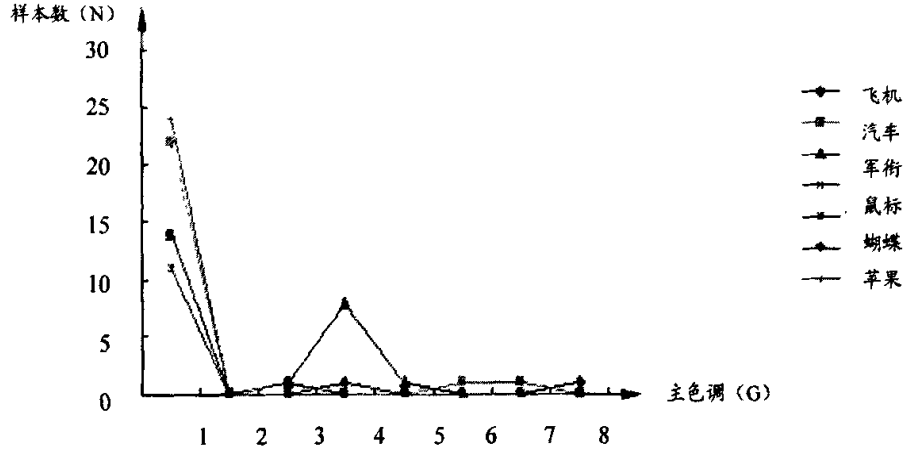
(l) 蓝色分量



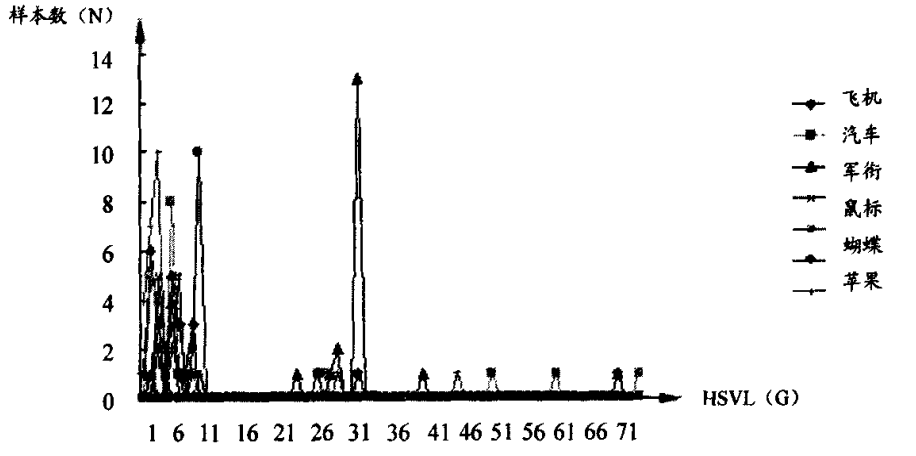
(m) 平均灰度



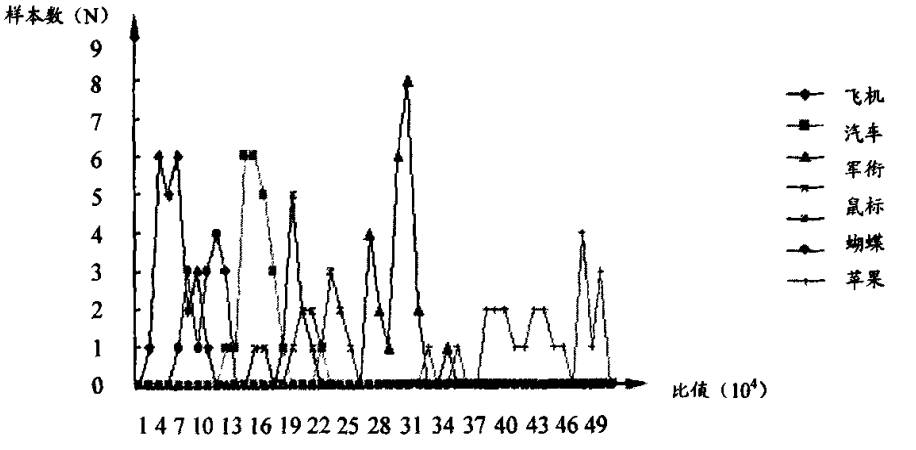
(n) 主色



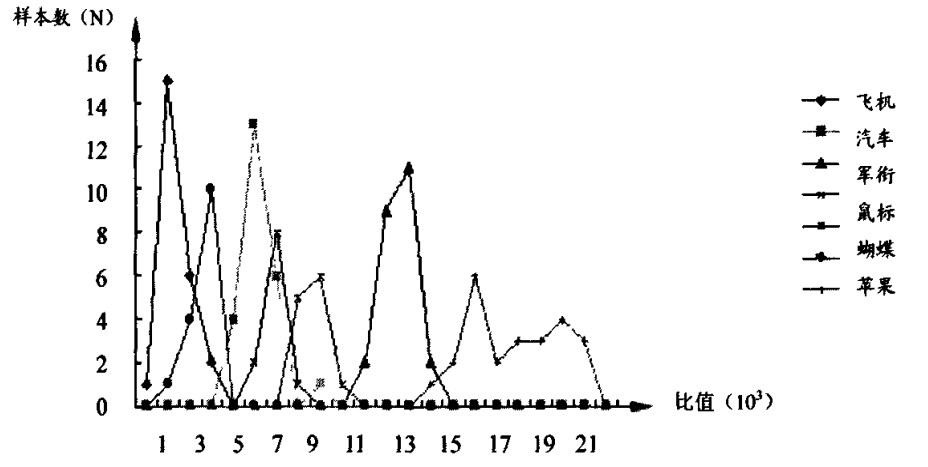
(o) 主色调



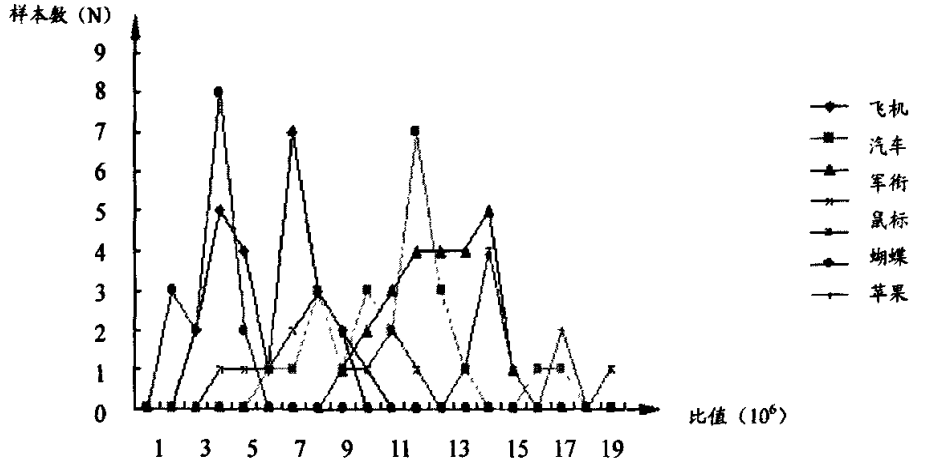
(p) HSB 一维参量



(q) 能量



(r) 熵



(s) 惯性矩

图 5-3 各类样本特征值统计分布

图 5-3 中，(a)~(i)是形状特征，通过人眼视觉这七类物体在形状上是最容易区别的，而且关于形状的特征描述具有尺度、旋转和平移不变性的特点。(j)~(p)是颜色特征，其他几个颜色特征由于是小样本，样本分布较广，因此将(j)~(m)中 256 级用等分法划分成 16 等级，由于每个类的样本比较少而且没有明显的颜色差异，故颜色特征对分类作用不明显。(q)~(s)是纹理特征，由于这三个特征的取值很宽，统计时也做了降维处理。

各个图像类在 19 个特征上的样本值分布关系体现了各个类的差异和相关性，必须从这些特征关系中提取出差异最大的信息，使得在不同的特征匹配顺序下，各个类都能从中剥离出来。

5.2.2 决策树生成

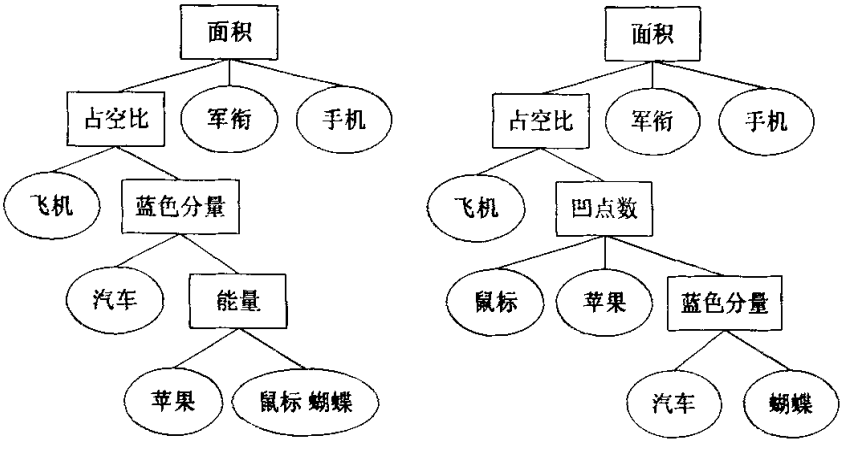
首先将各个特征上，每个类的统计分布值归一化，利用树型结构构造各个类

间的特征关系。树结构用邻接表来表示，其树结点和边的结构为：

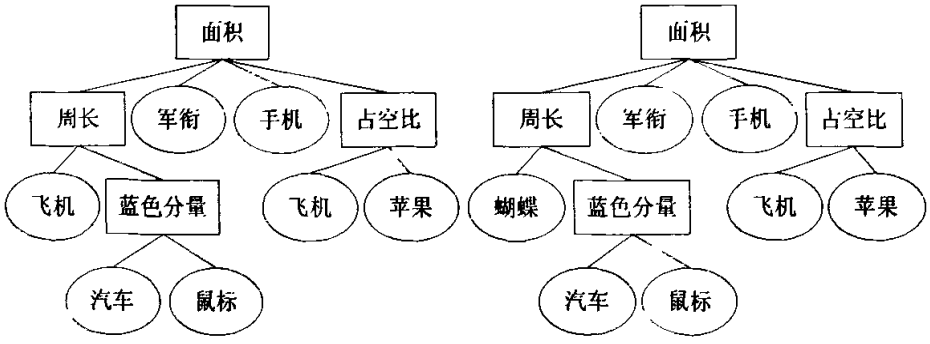
```
typedef struct STCNODE {
    int   iClassNum;   // 结点包含的类别数目
    int   *iClassflag; // 包含的具体类别
    int   *iFigure;    // 记录用于判断树分支的特征
    bool bYeFlag;      // 叶结点标记
} *PSTCTREENODE;
typedef struct STCEDGE {
    double   dError;  // 分支的误差
    int      iFigure; // 得到这条边的特征
    bool     bFlag;   // 边是否存在
} *PSTCEDGE;
STCNODE   stcTreeNode[13];
STCEDGE   stcTreeEdge[13][13];
```

初始的图像类为 7 个，二叉树的结点数为类别数的 2 倍减 1，即 13 个结点。且邻接表表示的边指定了树的方向性，二维数组的第一维表示父结点号，第二维表示子结点号。

利用图的连通性，将类间的分类误差作为连通的条件。设定一误差阈值 $R_{Threshold}$，从第一个特征开始进行连通性获取，将连通的类作为一个分支。然后从已经分出的分支中的类作下一个特征的连通性分析，直到所有的类都生成为叶结点。因此，阈值 $R_{Threshold}$ 的设定和特征分析的排序是影响决策树生成的两个重要因素。下面通过两个实验来比较不同的阈值和特征分析顺序下决策树生成的情况。



(a) T=0.03　　(b) T=0.035



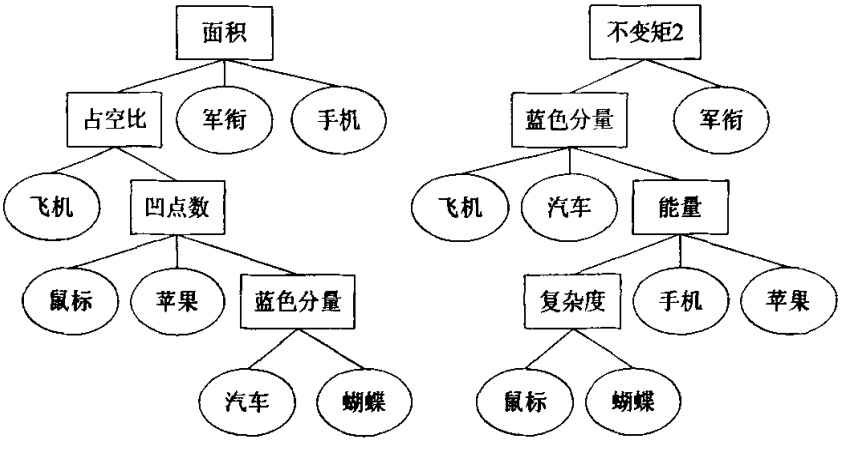
(c) T=0.045 (d) T=0.1

图 5-4 不同的误差阈值决策树的生成

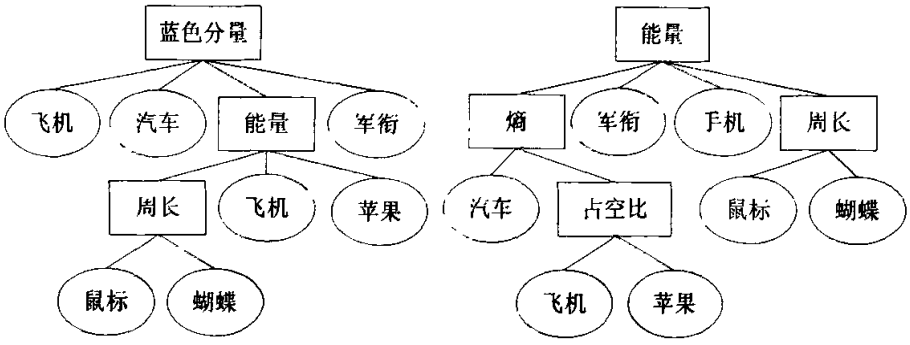
图 5-4 为第一组实验中，不同的误差阈值下决策树的生成情况。由该图可以看出，不同的误差阈值会影响决策树的生成：当设定误差阈值大于 0.045 时，其获得决策树没什么变化；而当误差阈值小于 0.03 时，只用一个阈值是无法将所有的类生成叶结点，需要增大阈值。根据各个类样本的特征值分布进行的分类决策树构建，要使得树的每个分支结点上的特征能获得最小的分类误差，因此需要不断调整阈值，使得所有的类都生成为叶结点。

第二组实验，给定不同的特征顺序而不考虑特征获得最多分支数目，在误差阈值为 0.035 时生成的决策树情况。

特征分析的顺序不同，构建的决策树型式也有很大差异。只从每次分类的误差率考虑，图 5-5 中的决策树在分类中的影响不大。但是一棵最优的决策树，图 5-5(c)和(d)的决策树特征的划分能力更强，并且它们的根结点中的特征获得的类间距离来看，使用能量特征更优，(d)图的决策树更优。由于树的分支间是根据一个非 0 的阈值来确定的，不同的决策树做分类时其分类结果的可信度还是有差异的。



(a) 特征顺序 1 (b) 特征顺序 2



(c) 特征顺序 3　　(d) 特征顺序 4

图 5-5 T=0.035 下不同特征顺序的决策树

5.2.3 先验知识引入

现根据各个图像的特点以及对数据的初步分析给出先验知识，对于这个实例给出如下几个特征作为分类的指导：占空比、面积、复杂度。决策树的建立是对当前结点上的类做误差分析，不同的特征顺序也对分类产生影响。利用先验知识中对特征分析的排序做出一定的指导，目的是为了生成一棵最优的决策树，以便能获得最小的分类误差率。

根据给出的先验知识，将先验特征做优先分析，并且在允许的范围内尽可能使得这些特征都能作为分类特征而增大误差的阈值。因此，将先验知识的 4 个特征在特征分析中排在最前面，并且首先利用小误差阈值建树，如果不能使得所有类生成为叶结点，则增大误差阈值。启发式的决策树如图 5-6 所示。

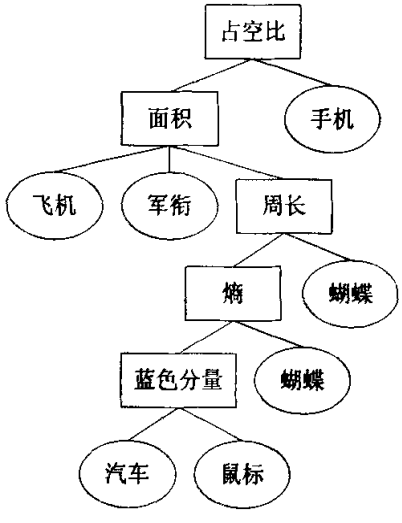


图 5-6 引入先验知识的决策树 T=0.001

图 5-6 的实验与前两组实验一样，是对同一组训练样本的特征数据进行分析。可以看出，通过先验特征的指导，对特征分析的顺序进行调整，使得在很小的误

差阈值 T=0.001 下能生成一棵最优的决策树，并且不需要调整误差阈值就使得所有的类都生成为了叶结点。与图 5-4 的(a)图比较，在没有调整特征分析顺序的决策树构建中 T=0.03 都没能生成决策树。启发式决策树的分类特征选择使得树的每个分支获得的分类误差更小，因此比没有先验知识指导生成的决策树更优。

5.2.4 分类结果分析

分类的误差率实验在不同大小的训练集上的整体测试，先验特征知识的引入使得决策树的构建不同，能够获得最优的决策树。针对不同知识下构建的决策树，分类的误差实验还比较了引入先验特征知识和不引入先验特征知识获得的分类规则对于分类结果的影响。实验结果如表 5-1 所示，相应的错误率比较曲线示意如图 5-7 所示。

表 5-1 分类判别错误率结果

训练集比例	40%	50%	60%	70%	80%
无先验知识	8.7%	8.3%	7%	6.3%	3.3%
有先验知识	8%	7.5%	5.4%	4.6%	2.7%

从分类结果上来看，本文给出的分类算法能够在小样本训练中能获得很好的分类效果。训练样本集比例的增大使得分类的准确率也有所提高，先验知识的引入，对每组测试都提高了分类的准确率。而且本文算法当样本数目很大时，其分类的特征关系分析更准确，决策树的构建更佳，分类结果应该也更好。

3%左右的误差率对于该方法来说是比较低了。产生误差的主要原因是由于样本不足，当待识别样本的特征值处于非特征分布的值域时，是利用类间的距离阈值来确定特征值与各类的相似性的，其阈值的设定对分类影响较大。

实验表明本文方法算法简单易于理解，决策树的构建方法简便，并且识别的错误率比较低，有很好的适应性。

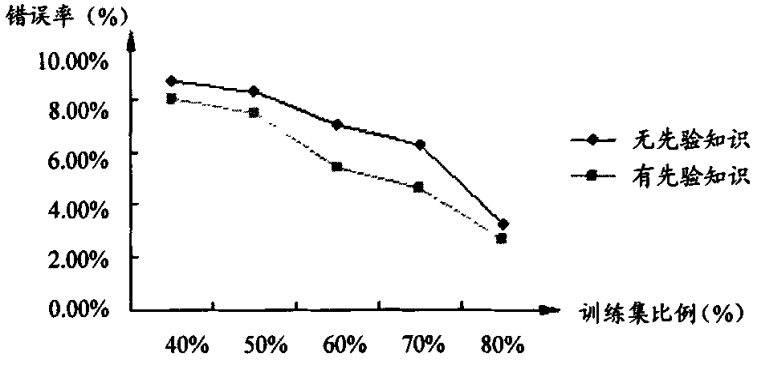


图 5-7 分类错误率曲线

5.3 算法评价

利用 4.1.2 节给出的分类算法的评价标准对本文提出的算法进行评价：

表 5-2 本文算法评价

错误率	实验证明利用本文算法能获得较小的分类错误率，精度较高。
速度	只需对样本数据做一次统计，分类器构建所需分析的数据较少，且这部分数据可以放入内存。
健壮性	利用特征距离的归一化来计算匹配误差，且距离计算是对样本统计分布的期望的距离，可以解决数值空缺和噪声点的影响。
可解释性	树型结构的分类器，易于提取分类规则，容易理解。
可伸缩性	样本数据量越大，误差计算更精确；分类器构建所需的数据也与样本数目无关，只与分析对象和特征的数目有关。
模型大小	以分析对象作为叶结点，且类别数目一般不大，所以构建的决策树的高度较小。

5.4 本章小结

本章分析了一个实例应用本文提出算法的实验过程。利用给定的样本图像提取初始特征值，通过分析特征的分布来构建决策树并且获取分类特征，生成启发式分类规则，并对分类结果进行分析，并给出了详细的分析步骤。比较了不同阈值和不同特征顺序下生成的决策树，先验知识的引入对构建一棵具有最小误差的决策树具有指导作用，并且实验证明先验知识的指导能提高分类结果的准确率。最后，利用分类算法的评价标准对本文提出的算法进行了评价总结。

第六章 总结及展望

6.1 工作总结

人工智能与机器学习实质上是根据实例得到一般概念，并将此概念运用到未知样本的归纳及演绎过程。图像的识别不仅需要根据图像的特点来设计训练过程，还需要考虑生成的分类规则的判别精度问题。图像识别的学习过程是从图像中提取底层视觉特征或高层语义特征，然后通过训练样本集归纳出分类准则。因此，图像识别学习系统的三个部分：特征的提取方法，特征选择方法和训练方法是研究的重点。

本文针对这三个部分展开研究工作，具体的研究工作总结如下：

（1）特征的提取方法很多，而对于图像的识别分类而言，选择的特征最好满足旋转、平移、尺度不变性。颜色、纹理和形状是最常用的三类特征，并且分析了图像特征与分类的关系

（2）针对决策树构建中特征选择的方法，通过分析原有决策树特征选择的劣势，提出了基于样本整体分布的特征选择指标。本文详细讨论了各个特征在单个图像类之中和类与类之间的表现关系，提出了一组特征分类相关的概念。选取类间误差最大和划分能力最强的特征作为分类特征，利用构建决策树的方法来实现分类特征的选择。并且讨论了对象数目增加时决策树的调整。

（3）针对各个类多特征分布所构成的特征空间，划分出多个分类可信度区域，分析了在不同特征空间区域，如何获得分类结果以及分析分类结果的可信度。通过分析两个类的特征分类关系，推导出多个类基于多特征的分类规则。对于特征分布构成的特征区域之外的区域，利用与各个类的最近距离与类间差异阈值的比较作为每个类利用某特征判别的可信度。这样可以解决特征匹配的不确定问题。

（4）引入了先验特征的概念，将分类特征分成了四个等级，不同等级赋予不同的特征融合系数，这样可以满足分类的特殊要求，也可以缓解在图像识别中机器学习精确性和人类视觉性之间的矛盾。提出了带加权误差的决策树算法。

（5）最后用一组图像对本文提出的算法进行验证和评价。

6.2 展望

智能技术是计算机技术中的一个研究重点，大量的不同的数据需要去分析和

处理，尤其对于图像数据的处理与分析需要进一步研究，很多特征表示只能表现图像内容的一个局部，也无法表示图像之间的细微差别。因此，首先要解决特征的不完备性问题。而且基于多特征的图像识别分类不仅仅依靠提取合适的图像特征表示，同时图像的处理中涉及的阈值问题也是一个需要研究的问题，不同的情况下阈值的大小是不同的，如何自适应阈值是在分类中很重要的问题。

在决策树的构建过程中，特征选择得顺序对于最后得决策树生成具有重要的影响，如何将在某特征上各类的连通个数的计算简化，使得分类决策树的构建最优也是需要进一步研究的问题。

在分类规则的表示方面，需要挖掘特征在各个类中的所有表现模式，不同的情况下分类规则具有特殊性，同时各个特征等级的系数设定以及误差阈值的确定是根据统计经验下得出的，还需要进一步通过数学的理论进行分析和提取。

最后，在实际的应用中，在基于计算机视觉的在线检测中，分类的特殊要求和分类的实时性要求都是需要研究的问题。在检测技术领域中的研究不仅具有深厚的理论价值，而且具有广阔的应用前景。同时，分类思想要尽量采用简单的算法实现，克服准确性和实时性的矛盾。

参考文献

[1] 肖政宏. 一种新的基于目标和特征的图像分类框架[J]. 计算机应用与软件, 2006, 23 (5): 105~107.

[2] 杨庚，王爱军. 分拣系统的软件结构与应用研究[J]. 计算机应用研究, 2003, (5): 80~82.

[3] 周晓宇, 李慎之, 戚晓芳, 等. 数据挖掘技术初探[J]. 小型微型计算机系统, 2002, 23 (3): 342~346.

[4] 方金城. 分类挖掘算法综述[J]. 沈阳工程学院学报（自然科学版）, 2006, 2 (1): 73~76.

[5] Shafer J, Agrawal Rand Mehta M. SPINT: a scalable parallel classifier for data mining[A]. in Proceedings of the 22nd VLDB Conference[C]. Bombay, India: 1996.

[6] Zaki M J, Ho C T, Agrawal R. Parallel Classification for Data Mining on Shared-Memory Multiprocessors[A]. in Proceedings of the 15th Int7 Conf, on Data Engineering[C]. March 1999: 198~205.

[7] Gehrke, Ramakrishnan R, Ganti V. Rainforest: A framework for fast decision tree construction of large dataset[A]. Proc. Int. Conf VLDB'98[C]. New York: Aug, 1998: 416~427.

[8] Pawlak Z. Rough Sets[J]. Journal of Computer Science, 1982, 11 (5): 341~356.

[9] GF Cooper, E Herskovtis. A Bayesian method for the induction of probabilistic network from data[J]. Machine Learning, 1992.

[10] Zhang J, Honavar V. Learning Decision Tree Classifiers from Attribute-Value Taxonomiesand Partially Specified Data[A]. in Proceedings of the International Conference on MachineLearning[C]. Washington DC: 2003: 880~887.

[11] 张文修，吴伟志. 基于随机集的粗糙集模型（I）[J]. 西安交通大学学报, 2000, 34 (12): 75~79.

[12] 张文修，吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学, 2000, 14 (4): 1~12.

[13] 闫河，唐德东，黄扬帆，等. 一种基于遗传算法的多类分类器设计方法[J].

仪器仪表学报, 2004, 25 (4): 414~417.

[14] Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle[J]. 1994, 10.

[15] Vapnik VN. The Nature of Statistical Learning Theory[A]. Spring-Verlag[C]. New York: 1995.

[16] Cover T M. Geometrical and statistical properties of systems and linear inequalities with applications in pattern recognition[J]. IEEE Trans. on Electronic Computers, 1965, 7 (19): 326~334.

[17] 王陈飞，肖诗斌. 基于 SVM 的图像分类研究[J]. 计算机与数字工程, 2006, 34 (8): 74-77.

[18] Tsai, Chih-Fong. Image mining by spectral feature: A case study of scenery image classification[J]. Expert Systems with Application, 2007, 7 (32): 135~142.

[19] Wing W Y Ng, Andres Dorado. Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error[J]. Pattern Recognition, 2007, 8 (40): 19~32.

[20] 朱晓霞，孙同景，陈桂友. 基于二叉树和 SVM 的指纹分类[J]. 山东大学学报（工学版）, 2006, 36 (1): 121~124.

[21] 刘建成，蒋新华，吴今培. 可理解模糊分类系统的分层演化学习[J]. 计算机工程与应用, 2006, 42 (5): 40~42.

[22] 赵云，喻炜. 应用结构模式识别技术设计图像分类器初探[J]. 青海大学学报(自然科学版), 2003, 21 (5): 51~53.

[23] 李长河，冯亚宁，石争浩. 图像匹配特征的一种融合表示[J]. 复旦学报（自然科学版）, 2004, 43 (5): 899~901.

[24] Han Jiawei. 1999 年数据挖掘论文集:数据挖掘中的知识分类[M]. 上海: 复旦大学出版社, 1999.

[25] 曹莉华，柳伟，李国辉. 基于多种主色调的图像检索算法研究与实现[J]. 计算机研究与发展，1999，36(1)：96~100.

[26] Stricker MA, Orengo M. Similarity of color images[A]. in: Proc. of SPIE:Storage and Retrieval for image and Video Databases III[C]. Feb 1995: 81~392.

[27] 向政权，马杰，夏定元. 基于内容检索的 BMP 图象特征提取[J]. 桂林电子工业学院学报, 2001, 21 (3): 56~60.

[28] 木妮娜，王素甫. 关于图像特征提取方法[J]. 新疆教育学院学报, 1997, 13

(35): 67~69.

[29] Tamura H, Mori S, Yamawaki T. Texture features corresponding to visual perception[J]. IEEE-SMC, 1978, 8 (6): 460~473.

[30] 邬浩, 潘云鹤, 庄越挺, 等. 基于对象形状的图象查询技术[J]. 软件学报, 1998, 9 (5): 343~349.

[31] HuM K. Visual pattern recognition by moment invariants[J]. IEEE Transom Information Theory, 1962, 170~179.

[32] 王娟, 慈林林, 姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27 (12): 68~71.

[33] Hong, Se June. Use of Contextual Information for Feature Ranking and Discretization[A]. IEEE Transactions on knowledge and data engineering[C]. 1997 (5): 718-730.

[34] 李昭阳，王元全，夏德深. 关于最佳鉴别特征维数问题的讨论[J]. 计算机学报, 2003, 26 (7): 825~830.

[35] 罗三定，肖飞. 不规则类圆形团块物体图像识别的新方法[J]. 中南大学学报（自然科学版）, 2004, 35 (4): 632~637.

[36] 罗可, 林睦纲, 郗东妹. 数据挖掘中分类算法综述[J]. 计算机工程, 2005, 31 (1): 3~6.

[37] 田金兰，李奔. 用决策树方法挖掘保险业务数据中的投资风险规则[J]. 小型微型计算机系统, 2000, 21 (10): 1034~1038.

[38] Provost F, Domingos P. Tree Induction for Probability-Based Ranking[J]. Machine Learning, 2003, 52 (3): 199~215.

[39] 程咏梅, 潘泉, 张洪才, 等. 计算机智能图像识别算法研究[J]. 计算机应用, 2004, 24 (2): 65~68.

[40] 苏小英，陈家琪. 基于间隔最大化的自动文本分类模型[J]. 计算机工程与设计, 2006, 27 (12): 2169~2173.

[41] 万钧, 钟亦平, 傅维明, 等. 启发式相关文本提取技术研究[J]. 小型微型计算机系统, 2004, 25 (4): 582~589.

[42] 林志贵, 徐立中, 严锡君, 等. 基于距离侧度的 D-S 证据融合决策方法[J]. 计算机研究与发展, 2006, 43 (1): 169~175.

[43] 杨杨，赵政. 模糊决策树在公共危机应急系统中的应用[J]. 计算机应用, 2006, 26 (10): 2457~2459.

致 谢

值此学位论文完成之际，我感慨万千，虽然研究生阶段的生活是短暂的，感觉是学生生涯中最艰难的一段日子，但是从一直指导我的导师和陪伴在我身边的伙伴那里学到了很多。

首先要感谢我的导师罗三定教授，他对待学术的一丝不苟的严谨作风和锲而不舍的求知精神，鞭策着我；他对待难题的态度和决心以及充满热情的人生态度，都让我受益匪浅，使我懂得了要开创性的工作和生活。不仅在学业上，他总是恪守着为人师表的本职，更是在生活上给与我关怀，使我同时保有快乐的心情和健康的身体。感谢沙莎教授在生活和学习上给我无微不至的关怀，象慈母也象朋友，她平和、亲切的待人态度也影响我以后的工作和学习。

非常感谢实验室的兄弟姐妹，尤其是 Apple 和芳子给与我的无私帮助。正是在这样一个温馨而融洽的大家庭中，我度过了学生时代的最后阶段，不仅顺利的完成了学习任务，也留下了许多美好的回忆。

感谢我的室友、旧同学和老朋友们，他们总是陪我欢笑陪我忧，在我苦恼的时候开解我，当我开心的时候祝福我。感谢男朋友谢峰对我论文工作的大力支持和帮助，使我能很好地完成本文的撰写。

感谢我的父母，感谢他们这么多年来对我的养育和教诲，我前进的每个脚步中都渗透着他们的汗水和心血，我永远爱他们。

最后，感谢论文的评审老师，我期待您的批评与指正。感谢论文的所有读者，若您能从中获得些许启发，我将感到不胜欣慰。

胡樱于中南大学计算机楼

二零零七年四月

攻读硕士学位期间的主要研究成果

项目开发

2006.9 至 2006.11，深圳华晶玻璃瓶有限公司玻璃原料杂质分拣系统。

论文发表

罗三定，胡樱. 基于样本分析的图像识别分类模型. 计算机应用研究[J]，已录待发.

基于多特征的图像分类决策树生成方法研究

作者： 胡樱
学位授予单位： 中南大学
被引用次数： 1次

本文读者也读过(1条)

1. 郝永宽 聚类分析在图像分类中的应用研究[学位论文]2008

引证文献(1条)

1.张勇.覃燕.李凯.陈建球 CTCS-3级列控系统车载设备人机界面信息的识别方法[期刊论文]-中国铁道科学 2010(4)