图像高斯混合模型的判别学习方法

陈雪峰

2009 年 10 月

中图分类号：TP 391
UDC 分类号：

图像高斯混合模型的判别学习方法

作 者 姓 名　　陈雪峰
学 院 名 称　　计算机学院
指 导 教 师　　贾云得教授
答辩委员会主席　　
申请学位级别　　工学博士
学 科 专 业　　计算机应用技术
学位授予单位　　北京理工大学
论文答辩日期　　2009 年 12 月

Discriminative Learning Approach for Gaussian Mixture Modeling of Images

Candidate：Xuefeng Chen
Supervisor：Prof. Yunde Jia
Department：Beijing Institute of Technology
Date：December，2009

图像高斯混合模型的判别学习方法

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的，复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签　　名：　　日期：

导师签名：　　日期：

摘要

高斯混合模型（Gaussian Mixture Model）是统计模式识别中一类重要建模工具。基于高斯混合模型的图像识别方法，具有形式灵活、识别速度快、抗干扰能力强、识别准确率高等优点，成为图像识别领域中一种重要的建模方法，在文档分析与识别、图像与视频检索、生物特征识别与认证、目标检测与跟踪、医学图像分析与识别、智能交通、智能监控等领域得到广泛应用。

高斯混合模型学习方法按照学习目标不同可分为生成学习和判别学习。大量的研究表明，判别学习使模式识别系统的识别性能显著提高，学习效果明显好于传统的生成式学习。很多研究机构都开展了图像统计建模的判别学习研究工作，并提出了许多学习方法。本文面向图像识别问题，研究图像的高斯混合模型判别学习方法，用于从训练样本中获得判别能力强的高斯混合模型，实现有效分类。本文研究内容包括基于贝叶斯分类器的判别学习准则、基于判别学习的模型选择方法和判别学习智能优化算法。

本文提出了最大最小后验伪概率软目标学习方法（Soft target based MMP Learning with Data Selection，SoftDS-MMP），用于学习基于后验伪概率（Posterior Pesudo-probability）的贝叶斯分类器。SoftDS-MMP 对每个模式类的正样本和反样本的后验伪概率分别定义相应的软目标，用该软目标度量分类器在训练集上的分类损失。SoftDS-MMP 通过最小化分类损失，同时最大化两个软目标之间的距离，获得分类器的最优参数集合。本文还进一步利用软目标进行数据选择，从训练集中移出和插入训练样本，压缩训练数据集，提高训练速度。在数据选择过程中，对于那些后验伪概率远超过其相应软目标值的训练样本，在一定的训练周期内暂时将其移出训练集。与基于硬目标的学习方法相比，SoftDS-MMP 降低了过学习风险，提高了训练速度。

本文在 SoftDS-MMP 判别学习框架下，提出了一种高斯混合模型成份个数选择方法。该方法将 Soft-MMP 目标函数结合到贝叶斯模型选择框架中，用拉普拉斯方法估计 SoftDS-MMP 目标函数的边缘积分，并将最大化拉普拉斯估计值做为高斯混合模型判别选择准则。利用线性搜索策略，同时获得最优高斯混合模型结构和模型参数。该方法主要优点是将判别信息引入到模型选择当中，以判别方式同时学习高斯混合模型的结构和参数，提高了分类器的判别性能。

为了提高判别学习算法的训练速度和优化效果，本文提出了一种基于判别学习目标函数梯度的进化策略优化算法。该方法利用目标函数的梯度信息调整 Cholesky 协方差矩阵自适应进化策略中的参数（Covariance Matrix Adaptation Evolution Strategy based on Cholesky Factorization， Cholesky-CMA-ES），包括加权均值、协方差矩阵和全局步长，提高优化效果和优化效率。该方法主要思想是对每代进化中个体的加权均值用梯度下降算法进行调整，并根据调整后的均值更新协方差矩阵和全局步长。该方法在训练过程中动态调整梯度信息和 Cholesky-CMA-ES 在联合优化方法中所占的比重。在训练初期，Cholesky-CMA-ES 在该联合优化方法中占主导地位，快速确定最优搜索区域；然后，逐渐增加基于梯度优化算法在联合优化方法中所占的比重，以加强联合优化方法的局部探索能力。该联合优化方法实现了 Cholesky-CMA-ES 与梯度下降算法的互补。一方面，利用其多点随机优化策略可以降低陷入局部最优解的概率；另一方面，利用目标函数的梯度信息可以加快收敛速度。

本文将所提出的判别学习准则、模型选择方法和联合优化算法应用于手写体数字识别问题中。在对高斯混合模型进行判别选择和学习的 SoftDS-MMP 方法框架下，利用联合优化算法来优化手写体数字分类器的结构和参数。学习得到的分类器在常用的 CENPAMI 和 MNIST 手写体数字样本库上进行了验证。在 CENPAMI 手写体数字样本库上取得了99.55% 的识别率，在 MNIST 手写体数字样本库上取得了 99.54% 的识别率。实验结果以及与目前最好识别效果的比较证明了我们所提出方法的有效性。

关键词：判别学习；高斯混合模型；图像识别；进化策略；模型选择

Abstract

Gaussian Mixture Model (GMM) is a widely used modeling tool in statistical pattern recognition comuunity. Because of its flexibility and robustness, GMM is being increasingly exploited as a convenient modeling tool in image classification. In the past decade the extent and the potential of the applications of GMM have widened considerably. Fields in which GMM have been successfully applied include document analysis and recognition, image and video retrieval, beiometric identification, object detection and tracking, biomedical image analysis and recognition, intelligent transportation, intelligenet surveillance, and etc.

The learning methods for GMM can be classified into generative learning and discriminative learning. It has been well known that discriminative learning can get better results than generative learning for pattern recognition. GMM has continued to receive increasing attention over the years. This dissertation focuses on the problems of discriminative leaning for GMM of images, including objective function of discrimative learning for Bayesian classifiers, discriminative model selection method, and intelligent optimization method.

This thesis proposes a novel soft target centered learning method for posterior pseudo-probabilities basedy Bayesian classifiers, called SoftDS-MMP for short. Tow adaptive soft targets of posterior pseudo-probabilities are defined for positive samples and negative samples of each class. The empirical loss of a classifier on training data is measured according to these two soft targets. Through minimizing the empirical loss while maximizing the difference between two soft targets, we obtain the optimial parameters in the posterior pseudo-probabilities measure functions of the classes as well as the values of soft targets. We further use the soft targets to dynamically select the training data in the training iterations to redeuce the risk of overfitting and improve the training efficiency. The samples with poster pseudo-probabilities distinctly larger than the corresponding soft target will be temporarily removed from the training set in certain times of iterations. Compared with the hard targets based learning methods, the SoftDS-MMP shows more effectiveness, higher efficiency, and better generalization.

This thesis presents a discriminative method of GMM selection under SoftDS-MMP discriminative learning framework. Acturally, a marginalized Soft-MMP objective function is designed and approximated with Laplace method. Using a line search algorithm to find out the maximum value of approximate marginalized Soft-MMP objective function, the

optimal structure and parameters of the GMM for a class are estimated simultaneously.

This thesis also describes a hybrid method of the Covariance Matrix Adaptation Evolution Strategy based on Cholesky Factorization (Cholesky-CMA-ES) and the gradient decent algorithm for discriminative optimization of Bayesian classifiers. In the hybrid optimization method, the gradient information of objective function is exploited to adjust three crucial factors in Cholesky-CMA-ES, including the weighted mean of parent population, the covariance matrix of distribution and the global step size, to improve the effectiveness and efficiency of Cholesky-CMA-ES. In the first step, the hybrid optimization method is dominated by the Cholesky-CMA-ES to search promising solution regions globally. The influence of gradient information is then enhanced gradually along with training iterations to obtain better local exploitation performance. We also apply the proposed hybrid optimization method to SoftDS-MMP. The hybrid method combines the advantages of Cholesky-CMA-ES and gradient descent algorithm. On one hand, the risk of getting stuck at local optimum is decreased by multi-point stochastic search. On the other hand, the convergence speed is accelerated by exploiting the gradient information of objective function in parameter evolution.

The proposed methods have been applied to handwritten digit recognition. Under SoftDS-MMP discriminative selection and learning framework for GMM, the hybrid optimization method is adopted to estimate the structure and paramters of handwritten digit classifiers. We conduct the experiments on the well-known CENPAMI and MNIST database to evaluate our classifiers. The 99.55% recognition rate on CENPAMI database and 99.54% on MNIST database are reached, respectively. The experiment results demonstrate the effectiveness of our methods.

Key Words: Discriminative Learning; Gaussian Mixture Model; Image Recognition; Evolution Strategy; Model Selection

目录

表格目录

插图目录

算法目录

主要符号表

$\boldsymbol{x}$：任一样本的特征向量

$\hat{\boldsymbol{x}}$：正样本的特征向量

$\bar{\boldsymbol{x}}$：反样本的特征向量

$\boldsymbol{X}$：训练样本集合

$\hat{\boldsymbol{X}}$：某一模式类别的正样本训练集合

$\bar{\boldsymbol{X}}$：某一模式类别的反样本训练集合

C_i：第 i 类模式

κ：后验伪概率公式中的调解参数

β：后验伪概率公式中的调解参数

m：某一模式类的正样本数量

n：某一模式类的反样本数量

d：两个软目标值之间距离

ω：用于控制经验损失和软目标之间距离的权重因子

α_t：梯度下降算法中第 t 次迭代的步长

δ：动态数据选择所用阈值

K：高斯混合模型中高斯成份个数

w_k：第 k 个高斯成份的权重因子

$\boldsymbol{\varphi}_k$：第 k 个高斯成份的均值

$\boldsymbol{\Sigma}_k$：第 k 个高斯成份的协方差矩阵

γ：混合高斯模型的斜方差矩阵中的元素

$p(\omega_i)$：第 i 类模式的先验概率

$p(\boldsymbol{x}|\omega_i)$：第 i 类模式对样本 $\boldsymbol{x}$ 的类条件概率密度

$p(\omega_i|\boldsymbol{x})$：第 i 类模式对样本 $\boldsymbol{x}$ 的后验概率

$\boldsymbol{\Lambda}$：分类器中未知参数集合

ψ：参数集合 $\boldsymbol{\Lambda}$ 中的任意参数

ε：在收敛条件中设置的极小阈值

$\hat{H}$：正样本后验伪概率的软目标值

$\bar{H}$：反样本后验伪概率的软目标值

M：模型的结构
S：模型中的参数个数
ρ： BIC 准则中的惩罚因子

$\zeta(M)$：提出的判别模型选择准则

r：参数空间的维数
μ：进化策略中父代种群的个体数目
λ：即化策略中子代种群的个体数目
g：进化计算中进化代数

$\boldsymbol{y}_k^{(g)}$：第 g 代种群中第 k 个个体

$\langle \boldsymbol{y} \rangle_w^{(g)}$：第 g 代种群的加权均值

σ：Cholesky-CMA-ES 中的全局步长

缩略词含义

GMM：高斯混合模型 (Gaussian Mixture Model)

OGMM：正交高斯混合模型（Orthogonal Gaussian Mixture Model）

EM: 期望最大化 (Expectation Maximization)

MLE：最大似然估计（Maximum Likelihood Estimation）

SVMs: 支持向量机 (Support Vector Machines)

NN: 神经网络 (Neural Networks)

CB：基于分类的判别函数学习方法 (Classification Based objective functions)

SME: 软边缘估计（Soft Margin Estimation）

LM：最大边缘估计（Large Margin Estimation）

PCA：主成分分析 (Principal Component Analysis)

MCE: 最小分类错误 (Minimum Classification Error)

MMI：最大互信息 (Maximum Mutual Information)

MPE：最小语音错误（Minimum Phone Error）

MQDF: 改进二次鉴别函数 (Modified Quadratic Discriminant Function)

MMP：最大最小后验伪概率学习方法 (Max-Min posterior Pseudo-probabilities)

Soft-MMP：基于软目标的最大最小后验伪概率学习方法（Soft target based MMP Learning）

SoftDS-MMP：基于软目标并带有样本选择的最大最小后验伪概率学习方法（Soft target based MMP Learning with Data Selection）

BIC：贝叶斯信息准则（Bayesian Information Criterion）

MDL：最小描述长度（Minimum Description Length）

CMA-ES：协方差自适应进化策略（Covariance Matrix Adaptation Evolution Strategy）

Cholesky-CMA-ES：基于乔里斯基分解的协方差自适应进化策略（Covariance Matrix Adaptation Evolution Strategy based on Cholesky Factorization）

EGS：进化梯度搜索（evolutionary gradient search ）

第 1 章　绪论

1.1 引言

图像统计建模是根据统计学理论和先验知识利用数学模型对图像在特征空间中的分布进行描述。基于统计模型的识别方法是一类广泛应用的图像分类方法[1-3]，具有识别速度快、抗干扰能力强、识别准确率高等优势，并可解决大规模、高维度、形式复杂的图像分类问题。在目前已有的各种统计建模方法中，以高斯混合模型（Gaussian Mixture Model，GMM）为代表的有限混合模型（Finite Mixture Model, FMM）具有形式灵活、建模方便、计算简捷等优点，成为图像识别领域中一种重要的建模方法，在文档分析与识别、图像与视频检索、生物特征提取与识别、目标检测与跟踪、医学图像分析与识别、智能交通、智能监控等领域得到广泛的应用。有限混合模型对数据建模主要包括两个任务：有限混合模型成份个数选择和各成份中具体参数学习[4]。虽然有限混合模型的成份个数和每个成份中的具体参数都是有限混合模型参数，但是二者却存在很大差别。有限混合模型成份个数的选择可归结为函数族的选择问题，而每个成份中具体参数的学习则是在确定的函数族内部选择具体的函数。因此，在用有限混合模型对数据建模过程中，通常对这两部分参数应用不同的学习方法分开学习。为了叙述方便，我们将有限混合模型中每个成份具体参数的学习简称为参数学习，将有有限混合模型成份个数的选择简称为模型选择。

有限混合模型参数学习的传统方法是以最大似然方法（Maximum Likelihood Estimation）为代表的生成学习(Generative Learning)，它以获取各类样本的真实分布为目标。该方法存在两方面问题：第一，最大似然方法取得最小贝叶斯风险（Bayes Risk）的前提条件是所选数据模型与数据的真实分布一致，该条件在实际中很难满足；第二，最大似然方法只有在大量的训练数据上才能学到稳定、有效的数据模型，但实际中很难获得大量的训练数据。随着支持向量机（Support Vector Machine，SVM）等方法的兴起，产生了一类新的面向模式识别的学习方法，即判别学习（Discriminative Learning）。判别学习方法以学习各类样本间的判别信息为重点，以最小化分类器的分类损失为目标，最大化分类器对训练集中正样本和负样本的区分能力[5]。由于支持向量机所取得的成功，判别学习的思想在统计建模领域中也得到了应用。最近十几年提

出了许多面向统计建模的判别学习方法，主要包括最小分类错误学习方法（Minimum Classification Error，MCE）[6-7]、最大互信息学习方法（Maximum Mutual Information，MMI）[8-9]和基于最大边缘的概率密度估计方法（Large Margin Estimation，LME）[10]。大量研究表明，判别学习使模式识别系统的识别性能显著提高，学习效果明显好于传统的生成式学习方法[5-10]。

相对于判别学习方法在有限混合模型参数估计上所取得的成功，有限混合模型判别学习的一个重要问题是估计有限混合模型的成份个数。对有限混合模型参数的学习，通常是先用已有的成份选择方法确定混合模型的成份个数，再用判别学习方法估计模型参数。目前，常用的有限混合模型选择方法包括确定法（Determiinistic Methods）、随机法（Stochastic Mehtods）和再采样方法（Resampling Methods）[11]。

虽然利用判别学习方法学习高斯混合模型受到相关领域的广泛关注，但已有的学习方法都有一定的局限性，如收敛速度、泛化效果等，不能很好的满足各种应用需求。因此，有必要开展图像高斯混合模型的判别学习方法研究。

研究图像混合建模的判别学习方法具有重要的理论意义，主要体现在如下三个方面：首先，通过对判别学习算法的研究，可以更加充分的比较和评估现有机器学习算法，完善和丰富机器学习理论，扩大机器学习应用范围；其次，通过将判别学习应用于图像识别领域，可以提高图像识别的效果，为解决传统的图像识别问题引入新的思想和方法；最后，该课题还涉及到其它相关学科的研究，如人工智能、优化理论、认知科学等。

同时，该选题还具有重要的应用价值。随着计算机科学、模式识别等的飞速发展，图像识别已经被广泛应用到商务、军事、文化等人民生活的各个方面，如文档分析与识别，图像检索，智能交通等等。通过对判别学习理论的研究，将进一步提高现有模式识别系统的性能，扩大模式识别技术的应用范围。

由于判别学习方法对基于统计建模的图像识别方法的重要性和有效性，国内外许多研究机构都开展了该方面的研究工作，如中国科学院自动化研究所、微软研究院、清华大学、香港大学、台湾大学、卡内基-梅隆大学、麻省理工大学、伯克利大学、斯坦福大学、剑桥大学、纽约大学、IBM、Bell 实验室等。在国际知名期刊和会议中每年都有大量的判别学习相关论文发表，如: Journal of Artificial Intelligence、Machine Learning Journal、IEEE Trans. Pattern Analysis and Machine Intelligence、IEEE Trans. Image Processing、Journal of Machine Research、IEEE Trans. Speech and Audio

Processing、IEEE Trans. Neural Networks、 IEEE Trans. On Evolutionary Computation、Pattern Recognition、American Association for AI National Conference、Int. Conf. On Machine Learning、Int. Conf. On Computer Vision and Pattern Recognition、Int. Conf. On Computer Vision 等。

1.2 图像分类中的有限混合建模方法研究现状

根据图像描述方法，将图像统计模型分为三种：描述模型（Descriptive Modeling）、随机模型（Generative Modeling）以及由描述模型和随机模型所构成的混合模型[12]。

描述模型主要包括马尔可夫随机场（Markov Radom Field）和吉布斯（Gibbs）采样，其优点是只需要一个概率模型即可对不同的图像特征进行统计，但是其计算代价很大，尤其对于高维度图像，该问题显得更为突出[13-14]。因此，人们提出了许多改进算法，主要包括：引入一些限定性条件降低计算代价，如随机马尔可夫模型（Causal Markov Model）等[15]；通过对特征向量增加置信度降低计算代价，如均值域估计（Mean Field Approximation）等；伪描述模型（Pseudo-descriptive Model）[16]。

虽然描述模型提出了很多改进方法，但对于高维度和结构复杂的图像识别问题其计算代价仍然很大。另一个解决计算代价的方法是采用随机模型。随机模型假设图像都由一些图像源产生，通过隐藏变量来表述所产生图像之间的相互关系，从而极大的降低了计算代价。随机模型有两种最具代表性的模型：第一种是一层结构的有限混合模型（Finite Mixture Model, FMM）[4]，如高斯混合模型（Gaussian Mixture Model）、有限混合 Dirchlet 模型等；第二种是二层结构的隐马尔可夫模型（Hidden Markov Model）[17-18]、如一维马尔可夫链（Markov Chain）、二维马尔可夫链（2-D Markov Chain）和马尔可夫场（Markov Field）等。

近期的一些文献又提出将描述模型与随机模型相结合的模型，这里我们将其称为混合模型，其中应用比较多的是马尔可夫链蒙特卡罗方法（Markov Chain Monte Carlo）[19]。

1.2.1 有限混合模型

有限混合模型具有形式灵活、建模方便和计算简捷等优点，受到图像识别领域的日益重视，并被应用于很多方面，如文档分析与识别、基于生物特征的图像识别、图像检索等。此外，二层马尔可夫模型也是以有限混合模型为基础进行的扩展，能否有

效的解决有限混合模型中的问题将直接决定着马尔可夫模型的描述效果。所以，本文以有限混合模型作为研究对象，构建最优的图像有限混合模型。

有限混合模型最初由 Pearson 在 1984 年提出，在图像识别领域中的首次应用则是 Chhikara 和 Register 在 1979 年开发出的第一个基于有限混合模型的数字分类系统[4]。有限混合高斯模型具有灵活性好、计算代价小和可以逼近任何复杂的图像分布等优点，成为有限混合模型应用最为广泛的建模方法。然而对于非高斯分布的数据，混合高斯模型对其的描述并不理想。另一种改进的有限混合模型是在扩展贝塔混合模型（Beta Model）基础上提出的有限狄里克莱混合模型（Finite Dirchlet Mixture Model）[20-21]。有限狄里克莱混合模型相对于高斯混合模型具有更大灵活性，因此可以更加准确的描述数据的分布形式，尤其对于非高斯分布数据。除上述的有限高斯混合模型和有限狄里克莱混合模型，有限混合模型在图像识别领域中所采用的概率密度函数形式还包括威尔布分布（Weibull）[22-23]、皮尔森分布（Pearson）[24]、伽玛分布（Gamma）[25-26]、t 分布[27]、范.米塞斯分布（von Mises）[28]、泊松分布（Poisson）、贝塔分布（Beta）和柯西分布（Cauchy）等。

1.2.2 模型选择

有限混合模型假定每个观测数据是由多个数据源产生数据叠加后形成，因此有限混合模型采用多个模型成份对数据进行描述。模型成份个数的确定是有限混合模型中一个关键问题。在用有限混合模型对数据建模时，如果选择的模型成份个数过多，会引起过拟合，而模型成份个数过少，又无法对数据进行准确描述。目前，已经提出了许多用于确定有限混合模型中的成份个数的方法，根据其计算方法的不同，可以将其分为三类：确定方法、随机性方法和再采样方法。

确定方法通过模型选择确定模型成份个数。在确定性方法中，假定一个固定的模型集合中包含最优样本分布模型，然后依据所建立的模型选择准则在模型集合中搜索最优模型。确定性方法中的模型选择准则一般由两部分构成，第一部分是模型对目前样本拟合程度的描述因子，第二部分是随有限混合模型包含成份个数的增加而递增的惩罚因子。常用的模型选择准则可分为如下三类：第一类是贝叶斯准则（Bayesian Criteria），如经验拉普拉斯准则（Laplace-empirical criterion）[4]、贝叶斯推理准则（Bayesian inference criterion, BIC）[29]；第二类是基于信息论/编码理论的模型选择准则，如最小信息长度准则（Minimum Message Length，MML）[30]、最小描述长度（Minimum Description Length，MDL）[31]、Akaike 信息论准则（Akaike information

criterion，AIC）[32]和信息复杂性准则（Information Complexity Criterion，ICC）[4]；第三类是基于全概率的模型选择准则，如分类似然准则（Classification Likelihood Criterion，CLC）、归一熵准则（Normalized Entropy Criterion，NEC）[33]和联合分类似然准则（Integrated Classification Likelihood Criterion，ICL）[34]。该类方法的不足是需要较多的先验知识，且有些准则在模型成份数目发生变化时需要对模型参数再学习。

随机模型选择方法中最常用的是马尔可夫链蒙特卡洛方法（Markov Chain Monte Carlo，MCMC）[35]，可通过两种途径应用于有限混合模型，第一种是利用模型选择准则，第二种是基于贝叶斯框架下在后验概率空间进行采样。

许多文献中提出用再采样方法（Resample-based schemes）[36]和交叉验证方法（Cross-Validation）[37]估计有限混合模型的成份个数。

第二种随机模型选择方法和第三种再采样方法的计算代价大，因此，并不适合参数较多的模式识别系统。

1.3 判别学习方法研究现状

判别学习是一种面向模式识别的学习方法，传统的生成式学习方法是以学习样本的真实分布为目标，而判别学习方法则侧重于学习各类样本间的判别信息，以最小化分类器的分类损失为目标，从而最大化分类器对正样本和负样本的区分能力。相对于传统的生成学习方法，判别学习方法主要有如下三方面的特点：1）判别学习更关注类别之间的差异信息，而不是类别的真实分布，其直接以最小化分类器的分类损失作为学习目标。在学习过程中，每个训练样本的所有类别信息都参加学习，并使其所属类别与竞争类别之间的差异最大化。所以，判别学习更符合模式识别要求；2）判别学习算法更适合小样本学习，大量研究和实验证明在小样本问题中判别学习方法相对于传统的生成式学习方法能够获得更好的实验结果；3）判别学习方法不要求所用模型与真实的模型一致，更适合解决实际问题。在模式识别领域的许多问题中，如语音识别、字符识别、人脸识别等，判别学习方法相对于生成学习方法都显示出了明显的优势。

判别学习方法主要包括：支持向量机（Support Vector Machines，SVM）、最小分类错误学习方法（Minimum Classification Error，MCE）、最大互信息学习方法（Maximum Mutual Information，MMI）和基于贝叶斯分类器的最大边缘学习方法。

有些文献将最大熵方法（Maximum Entropy）和神经网络中的交叉熵（Cross Entropy）也归为判别学习方法。

1.3.1 支持向量机

支持向量机（Support Vector Machines，SVM）是 Vapnik 等人 1992 开始提出的一种统计学习方法 [38-40] 。SVM 是基于统计学习理论中的结构风险最小化原则（Structural Risk Mimimization，SRM），即在降低经验风险的同时，降低 VC 维（Vapnik-Chervonenkis Dimension）以缩小置信范围，控制实际风险上界。SVM 以最大边缘作为学习准则，即所有样本到分类器判决边界的最小距离最大化。对于二分类问题，如果训练样本特征是线性可分的，则 SVM 寻找一条最优分类线，该分类线不但可以将两类训练样本特征无误的分开，并且同时使分类间隔最大，即边缘最大化。对于训练样本特征线性不可分的情况，SVM 利用核函数将样本特征向量映射到一个高维特征空间，并在该特征空间中构造最优分类面。SVM 解决多分类问题的方法主要包括：一对一组和分类法、一对多组和分类法、决策树分类法和全局优化分类法。

标准的支持向量机优化算法可归结为一个受约束的二次型规划问题[41]。SVM 将该问题转换为求解拉格朗日函数的鞍点，并用传统的二次规划方法进行求解，如牛顿法（Newton）、拟牛顿法（Quqsi_Newton）、共轭梯度法（Conjugate Gradient）、原-对偶内点法（Primal-dual Interior Point Methods）等。SVM 优化时需要计算训练样本集中任意两样本间的核函数，生成核函数矩阵，并通过矩阵运算进行优化。当训练集的数据量比较大时，会存在数据存储量大、收敛速度慢等问题，并且使用传统优化算法还存在稳定性问题。为了减少 SVM 的计算复杂性，提高优化效率，人们提出了很多改进的 SVM 优化算法，主要可以分为四类：1）分解算法，如 SMO（sequential Minimal Optimization）[42-43]。分解算法的基本思想是将原二次规划问题分解为一系列二次规划子问题，通过对子问题的迭代求解，使最终结果收敛于原问题的最优解。根据子问题的划分和迭代策略不同，分解算法可以分为“块算法”（Chunking Algorithm）和固定样本集分解法。 2） 多变量更新算法，如对数梯度法（Exponentiated Gradient, EG）[44]。其基本思想是采用梯度下降方法进行优化，与分解算法不同的是在每次迭代时对所有变量进行优化，根据梯度优化策略不同可以分为投影梯度法和对数梯度法。3）序列算法，如增量减量式序列学习算法（IDA）[45]。其基本思想是当出现新的单样本时，根据其与原训练结果的关系对原训练结果进行调整，而不需要重新学习。4）SVM 变异算法。变异优化算法是通过改变 SVM 学习准则降低训练复杂度，提高收敛

效率，如最小二乘支持向量机、基于线形规划的支持向量选择标准等。

以统计学习作为其理论依据的 SVM 方法，其优点主要表现在如下两个方面：第一，SVM 基于结构风险最小化设计学习准则，克服了传统学习方法仅依据经验风险所带来的过学习问题，具有很好的泛化性能；第二，通过核函数将特征映射到高维空间，既解决了样本特征的线性不可分问题，又降低了特征在高维空间中的计算复杂度，克服了维数灾难问题。

贝尔实验室率先将 SVM 应用于手写体数字识别。目前，SVM 已被应用于多种模式分类问题当中，并显示出了显著的效果，如人脸识别、语音识别、字符识别、医学工程等。除模式识别领域外，SVM 还被应用到了其他许多领域中，如路径规划、在线学习、半监督学习、排序、聚类等。已有许多 SVM 工具包发布，常用的有：LIBSVM [46]、SVMlight [47]、SVMTorch2 [48]、HeroSVM [49]。

1.3.2 最小分类错误方法

最小分类错误学习方法（Minimum Classification Error，MCE）是 Juang，Katagiri 和 Lee 等人于 1992-1993 年在语音识别领域中提出的一种判别学习方法 [6-7]。MCE 直接以最小化分类器在训练集上的分类损失作为判别学习准则；对训练集中的样本，基于后验概率定义一个光滑的分类损失函数，使分类器分类正确时分类损失趋近于 1，分类器分类错误时分类损失趋近于 0。通常采用 Sigmoid 函数作为分类损失光滑函数。MCE 以最小化训练集中所有样本的分类损失，即经验分类损失，作为学习目标，通过优化目标函数，获得最优分类器。

在应用 MCE 时，首先用期望最大化算法（Expectation Maximization，EM）学习统计模型，获得模型参数的最大似然估计，并将其作为初始模型参数，然后用随机梯度下降方法（Generalized Probabilistic descent，GPD）优化目标函数。GPD 方法的优点是简单、易于实现，缺点是容易陷入局部最小值，且收敛速度比较慢。针对上述问题，很多文献提出了相应的改进算法，主要分为以下四种：1）半批量优化（Semibatch），该算法是随机优化算法和批量优化算法的折中，每次选取训练集合中的子集进行优化，从而实现在线优化和多模型并行优化；2）基于快速传播的优化算法（Quickprop，QP），该算法是一种基于牛顿定律的二阶快速优化算法，对多个模型可并行优化，最初由 E. McDermott 将其用于优化 MCE 目标函数 [50-51]；3）基于震荡传播的优化算法（Resilient back-propagation，Rprop），该算法只保留目标函数梯度的方向，而梯度的大小则用固定值代替；Rprop 也是一种并行的优化算法，在神经网络中已经被广泛

应用，后被应用到 MCE 算法中[52-53]；4）扩展鲍姆-韦尔奇算法方法（Extended Baum-Welch，EBW）。最近也有一些文献中提出用 EBW 优化 MCE[54]。除上述主要优化算法外，还有一些文献通过在对 MCE 算法增加一些限制条件后，直接解析的求出最优解，而不采用优化的方式，虽然该方法还缺少理论分析，但却取得了较好的实验效果[55-56]。

MCE 学习分类器的泛化性能取决于所定义的损失函数的光滑性，其光滑性越好，则所训练分类器的泛化性能越好[57-59] 。一些文献也提出了改进 MCE 泛化性能的方法，如：限制 GMM 模型协方差矩阵中的方差极小值，对训练集中正样本的赋予较大权重等[60]。还有一些文献给出了 MCE 的光滑损失函数与贝叶斯风险之间关系的理论分析 [61-63]，并推导出其分类损失上限是一致性。MCE 是一种独立于分类器的学习算法，可对多种分类器进行学习。迄今为止，在已发布的文献中被提及可用 MCE 训练的分类器主要有神经网络分类器[64]、基于隐马尔可夫和混合高斯模型的贝叶斯分类器[65-68]和最近邻分类器[69]。

MCE 除在语音领域中发挥重要作用外，还被应用到许多其他领域，如：字符识别[70-72]、特征提取[73-75]、图像处理[76]、文档分析[77]和机器翻译[78]等。

1.3.3 最大互信息方法

最大互信息（Maximum Mutual Information，MMI）是由 Brown 等人提出的一种学习方法，并应用于模式识别领域，初期主要用于语音识别[8-9]，目前在许多研究领域中都得到了应用，如字符识别、特征提取、医学图像处理、目标跟踪、聚类等[84-87]。MMI 的理论基础是信息论。根据信息论，两个变量之间的互信息可作为一种测度，用于度量当一个变量不确定性发生变化时，导致另一个变量不确定性发生变化的大小。在 MMI 方法中，互信息被作为样本与其所属类别之间不确定性的度量，也可描述为样本和所属类别的联合分布与独立分布之间的 KL 距离（Kullback-Leibler Divergence）。MMI 的学习准则是最大化样本与其所属类别之间的互信息。针对 MMI 学习准则，已经提出了很多改进方法，其中最具代表性的是在语音识别领域中广泛应用的最小音素错误学习方法（Minimum Phone Error Training，MPE）和最小语音错误学习方法（Minimum Word Training，MWE），其分别用音素和字作为分类损失的度量单位，并用相应的分类误差函数代替 MMI 中的分类损失函数[79-83]。

MMI 对目标函数的优化算法主要有两种，一种是梯度下降方法，另一种是基于弱性辅助函数（Weak-sense Auxiliary Function）提出来的扩展鲍姆-韦尔奇算法

（Extended Baum-Welch，EBW）。EBW 方法通过对 MMI 目标函数进行变换，并利用辅助函数评价学习效果，对目标函数进行优化。EBW 与梯度下降方法的学习效果相近，但 EBW 的收敛速度明显高于梯度下降方法，因此在 MMI 学习算法中应用较多。

MMI 也是一种独立于分类器的判别学习方法，常用于学习基于隐马尔可夫或混合高斯模型的贝叶斯分类器、神经网络分类器等。对训练集中的任意样本，MMI 不但最大化其所属类别的条件概率，而且最小化竞争类别的条件概率，因此，MMI 的学习效果好于传统的最大似然方法。MMI 的泛化性能与 MCE 一样，也是通过目标函数的光滑性来控制，目标函数的光滑性越好，其泛化性能越好。实验表明 MMI 的学习效果稍差于 MCE，但是由于 MMI 适合用 EBW 算法进行优化，而 MCE 更多用梯度优化。因此，MMI 的优化效率通常高于 MCE，在大规模数据问题中应用更为广泛。

1.3.4 基于边缘最大化的概率密度估计

基于边缘最大化的概率密度估计（Large Margin Estimation，LME）是在语音识别领域中新兴的一种判别学习准则[10]。LME 是贝叶斯学习理论与小样本统计学习理论的融合，是结构风险最小化原则在贝叶斯分类器框架下的实现。对任意样本，贝叶斯分类器的分类边缘被定义为该样本所属类别的后验概率与最大竞争类的后验概率之差。LME 的学习准则是使贝叶斯分类器的边缘最大化。根据最大边缘实现方式不同可分为如下三种：1）基于边缘最大化的判别学习，主要有（Larger Margin Estimation，LME）和最大相对边缘概率密度估计（Large Relative Margin Estimation，LRME）。其仅以贝叶斯分类器的边缘最大化作为学习目标，而不考虑期望分类损失，错误分类的训练样本不参加学习。因为仅用边缘最大化作为学习目标会导致学习过程不收敛，因此在最大边缘概率密度估计方法中对学习准则增加了一些限制性条件，如互信息最大化或限制高斯密度函数中的方差等[87-89]。2）基于最大边缘与最小经验损失联合优化判别学习方法，主要有最大边缘最小分类损失学习方法（Large Margin Minimum Classification Error，LM_MCE）和最大软边缘学习方法（Soft Margin Estimation，SME）。该方法在增加贝叶斯分类器分类边缘的同时降低分类损失，LM_MCE 用 MCE 方法中的分类损失函数度量分类损失，而 SME 则采用 SVM 中常用的铰函数（hinge function）度量分类损失[90-91]。3）基于参数容量的最大边缘贝叶斯分类器判别学习方法，主要有最大边缘训练方法（Large Margin Training，LMT）。该方法是根据 SVM 的思想提出来的，学习目标是在控制分类器边缘的同时最小化模型中参数的范数和松弛变量[92-93]。

与 MCE 和 MMI 一样，基于贝叶斯的最大边缘判别学习准则在适当参数控制下是光滑的凸函数，可以用梯度下降对其优化。除梯度下降方法外，目前一些文献中提出了比梯度下降更有效的、更快速的优化方法对目标函数进行优化。H. Jiang 和 L.X.Wei 等人提出用半正定规划方法（Semidefinite Programming，SDP）对 LME 准则进行优化[94]。与传统的梯度下降算法相比，半正定规划方法的优化效率更高，而且是全局优化算法。

目前，基于贝叶斯的分类器的边缘最大化判别学习算法主要应用于语音识别问题，理论和实验表明其学习效果好于 MCE 和 SVM。最大边缘判别学习准则还被用于手写体数字识别问题中，其结果接近于 SVM。

1.3.5 最大最小后验伪概率方法

目前已有判别学习方法主要是面向分类器的判别边缘提出来的，如 SVM、MCE、MMI 等。而最大最小后验伪概率方法（Maximum Minimum Pseudo-Posterior Probability，MMP）则是以样本的分布模型为研究对象[95]。MMP 的学习目标是在后验伪概率框架下，使各模式类的分布模型对其正样本的后验伪概率趋近于 1，而反样本的后验伪概率趋近于 0，从而获得该模式类的最优分布模型。

与 MCE 一样，MMP 先用 EM 方法学习统计模型，获得模型参数的最大似然估计，并将其作为初始模型参数，然后用梯度下降方法优化目标函数。MMP 已成功的应用于字符识别、字符串分割、文本提取以及图像检索等 [96-99]。

1.4 主要问题与难点

目前大量研究人员开展有限混合模型的判别学习的研究工作，提出了许多学习准则和优化方法，但仍有许多有待解决的问题：

（1）如何设计更有效的判别学习准则。设计有效的判别学习准则一直是判别学习研究领域的热点与难点问题，是判别学习方法中最基本和核心的问题，其体现了判别学习方法提出的思想，直接决定着学习的效果。虽然目前图像识别系统通过引入判别学习已经提高了其识别效果，但与理想目标仍有一定的差距。因此，如何设计更有效的判别学习准则，进一步提高模式识别效果，从而使图像识别系统真正的满足我们的需求，仍有待解决。

（2）模型选择方法中的随机方法和再采样方法的计算代价太大。对于基于高斯

混合模型的图像识别问题，通常包含上千的参数，很难应用随机性选择方法或基于马尔可夫链蒙特卡洛方法估计模型结构，因此在本文中不考虑上述两种模型选择方法。

确定方法是基于生成式的模型选择方法，该方法在满足两个前提的条件下，即样本的真实模型在所选模型集合当中和训练样本充足，假设模型在未知数据上的识别错误与其在训练数据上的边缘似然相关，从而根据最大边缘似然准则选择最优模型结构。但该类方法的局限性是其前提条件通常很难满足，我们很难知道数据的真实分布，也很难得到充足的训练样本。并且该类方法忽略了模型的判别能力，用该类模型选择方法有时不能得到满意的学习效果。如何将训练数据的判别信息引入到模型选择当中，从而使模型成份个数估计和模型的参数学习统一到一个判别学习框架当中，提高分类器的判别性能，是判别学习方法中一个有待解决的重要问题。

（3）如何设计有效的优化算法，提高判别学习方法的优化效率。算法复杂度高、学习速度慢是目前已有判别学习方法所共存的一个问题，是判别学习方法的瓶颈，限制了判别学习方法在大数据量、多类别模式识别问题中的应用。如何设计有效的判别学习优化方法，进一步提高优化效率是目前判别学习需要解决的一个重要问题。

本文针对上述问题，确定研究内容，包括以类为中心的贝叶斯分类器判别学习准则、基于判别学习的模型选择方法和判别学习的智能优化算法。

1.5 研究内容

本文的主要研究内容包括以类为中心的贝叶斯分类器判别学习准则、基于判别学习的模型选择方法和判别学习的智能优化算法，并在手写体数字识别应用平台上对所提出的算法进行验证。

（1）以类为中心的贝叶斯分类器判别学习准则

判别学习准则的设计在判别学习方法中起着关键作用，是区别各种判别学习方法的核心因素，其直接决定着判别学习方法的学习效果。根据判别准则关注对象的不同，将判别准则分为两类：以类为中心的判别准则和以数据为中心的判别准则。应用最广泛的以类为中心的判别准则是支持向量机，其以模式类为关注对象，使分类器对两个模式类之间的边缘最大化。而面向统计模型的判别学习准则主要以数据为中心，如最小分类错误、最大互信息和基于最大边缘的概率密度估计，该类判别准则以每一个数据为关注对象，使分类器对每一个数据的可分性最大。本文则是研究面向统计模型的以类为中心的判别准则，使每类模型的判别能力最大化。因此，基于最大最小后验伪

概率方法，通过引入软边缘思想，研究有限混合模型的判别学习准则及在该准则下的样本选择方法。

（2）基于判别学习的有限混合模型选择方法

有限混合模型中一个重要问题是模型成份个数选择，模型成份个数过多，会导致模型的过拟合，而成份个数过少，又无法准确的描述数据分布。虽然针对该问题已经提出了许多解决方法，但已有方法主要都是基于生成学习的，降低了模型的判别性能，且各种方法都存在各自的问题，如确定法需要较多的先验知识、随机法和再采样的计算代价大，不适合复杂、大数据量问题等。判别学习方法虽然在有限混合模型的参数估计中取得了成功，但却没有被引入到模型的成份选择当中。本文基于判别学习的思想，研究新的有限混合模型成份判别选择方法，将模型的参数估计和成份个数选择统一到相同的判别学习框架之下同时学习。

（3）用于判别学习的智能优化算法

判别学习的优化方法是判别学习的关键问题之一。已有的各种判别学习方法普遍存在优化效率低、优化方法中的部分参数需要靠经验设定等问题，很难应用于在超多类别、大规模模式识别问题。如何设计快速的判别学习优化方法，是判别学习中迫切需要解决的问题之一。本文基于智能计算与最优化理论，研究新兴的智能优化与经典的基于梯度优化方法相结合的新的优化方法，使判别学习的优化效率和优化效果同时提高。

1.6 论文结构

第二章为本文工作的应用背景，主要介绍本文所构建手写体数字识别系统，包括手写体数字图像预处理、特征提取、手写体数字特征建模、基于后验伪概率的贝叶斯分类器和本文将使用的手写体数字样本库。本文所提出的算法都将在该平台上进行验证，并与其他算法进行比较。

第三章介绍面向后验伪概率分类器的最大最小后验伪概率学习方法，并在此基础上提出了基于软目标的最大最小后验伪概率学习方法和样本选择方法。

第四章综述了模型选择研究现状和目前存在的问题，进而在基于后验伪概率的贝叶斯分类框架下，提出了一种混合高斯模型的判别选择和学习方法，并将该方法应用于手写体字符识别当中。

第五章详细论述了梯度下降和进化策略各自的优点与局限性，及目前在分类器中

的应用，在此基础上提出了一种 Cholesky-CMA-ES 和梯度下降相结合的联合优化方法，并介绍了基于所提出的联合优化方法的 SoftDS-MMP 在手写体字符识别上的实验结果。

第六章为总结与展望，对全文工作进行总结，概述论文工作的主要内容和意义，提出进一步研究的设想以及对未来应用的展望。

第 2 章　应用背景

2.1 引言

本文选择手写体数字识别问题作为应用背景，所提出方法将在手写体数字识别问题上进行验证，并与已有方法进行比较。本章将介绍作为后续章节中所提出算法应用平台的手写体数字识别系统。图 2.1 是手写体数字识别系统流程框图。下面将分别介绍本文手写体数字识别系统中的预处理、特征提取、后验伪概率分类器和所用的手写体数字样本库。

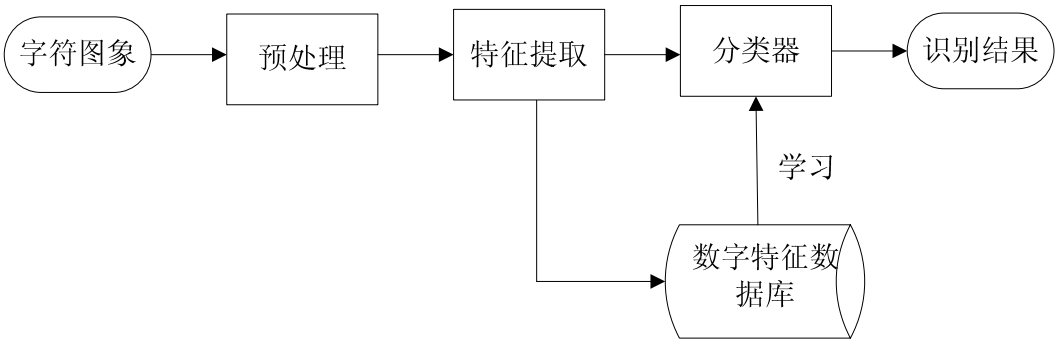


图 2.1 手写体数字识别系统流程框图

手写体数字识别是一种经典的多类模式分类问题，长久以来一直是模式识别领域中最活跃的研究方向之一，并被应用到了许多领域中，如银行票据处理、邮政分拣、个人数字助理、数码摄像机、数字图书馆等。几乎所有经典的机器学习和模式分类方法都曾被应用到手写体数字识别问题当中，如最近邻（1-Nearest-Neighbor）[69]、K-近邻（K-Nearest-Neighbors）[100]、线性分类器（Linear Classifier）[70]、贝叶斯分类器[72]、神经网络[101]、支持向量机[42]、二次判别函数分类器（Quadratic Discriminant Function）[68]、流形学习（Manifold）、提升算法（Boost）[102]等等。因此，手写体数字识别问题是目前公认的一个理想的机器学习和模式分类方法的验证平台[67,103]。许多新方法都会被应用于手写体数字识别，以便与其他算法进行比较，评价该方法的性能。识别正确率是手写体数字识别最重要的性能指标，也是评价机器学习或模式识别方法的重要依据。本文将手写体数字识别的识别正确率作为我们方法的性能评价指标，并根据识别正确率比较各种方法。

2.2 特征提取

本文采用基于灰度图像的 8 方向梯度特征（8-direction gradient feature from gray-scale image，e-grg）来表示数字图像[104-105]。

相对于常用的链码特征，梯度特征采用 Sobel、Robert 或 Kirsch 等算子描述特征，可以对笔划的方向和强度进行更为精确的刻画。该实验中所用的 e-grg 特征提取过程包括四个步骤：

首先，采用宽高比自适应归一化方法对手写体字符图像大小进行归一化操作[104]，其操作可分为两个阶段。

第一阶段是根据原始图像中字符的宽高比，确定归一化后图像中字符的宽度和高度。设归一化操作前，图像中字符的高度和宽度分别为 W_1 和 H_1，则该字符的宽高比为

$$R_1 = \begin{cases} W_1/H_1 & \text{if } W_1 \leqslant H_1 \\ H_1/W_1 & \text{otherwise} \end{cases}. \tag{2-1}$$

设归一化操作后，图像中字符的高度和宽度分别为 W_2 和 H_2，则该字符的宽高比为

$$R_2 = \begin{cases} W_2/H_2 & \text{if } W_2 \leqslant H_2 \\ H_2/W_2 & \text{otherwise} \end{cases}. \tag{2-2}$$

在自适应宽高比归一化方法中，归一化后的字符不需要充满整幅图像。设图像归一化为 $H \times H$，则归一化后的字符满足 $\max(H_2, W_2) = H$，而 $\min(H_2, W_2)$ 根据原图像中字符的宽高比确定。通常采用如下的宽高比影射函数计算

$$\begin{aligned} &F_0 : R_2 = 1, \quad F_1 : R_2 = R_1, \\ &F_2 : R_2 = \sqrt{R_1}, \quad F_3 : R_2 = \sqrt[3]{R_1}. \end{aligned} \tag{2-3}$$

确定归一化后图像中字符的高度和宽度后，在第二阶段，采用重心与中心对齐的归一化方法，将原图像影射到归一化图像中。设原图像的重心为 (x_c, y_c)，归一化后的图像中心 (x'_c, y'_c)，则由原图像到归一化图像的映射为

$$\begin{aligned} x' &= \frac{W_2}{W_1}(x - x_c) + x'_c \\ y' &= \frac{H_2}{H_1}(y - y_c) + y'_c \end{aligned}. \tag{2-4}$$

图 2.2 是归一化方法中的不同映射函数。

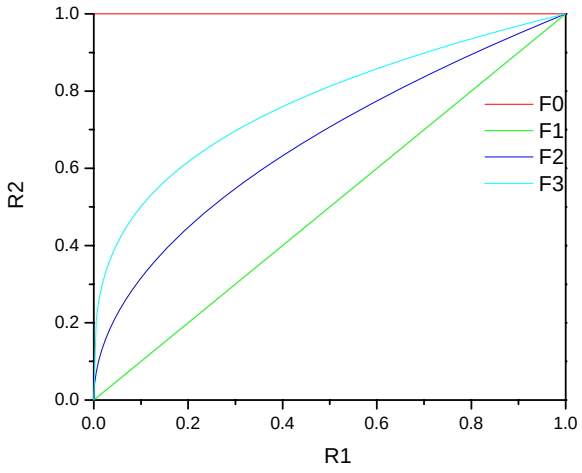


图 2.2 不同大小归一化方法的宽高比映射函数

图 2.3 是用上述四种不同宽高比映射函数所得出的归一化数字图像。文献[104]对各种归一化方法进行比较，$R_2 = \sqrt[3]{R_1}$ 的宽高比映射取得了最好的识别效果。本文的实验研究中字符图像的大小归一化均采用 $R_2 = \sqrt[3]{R_1}$ 的宽高比映射。

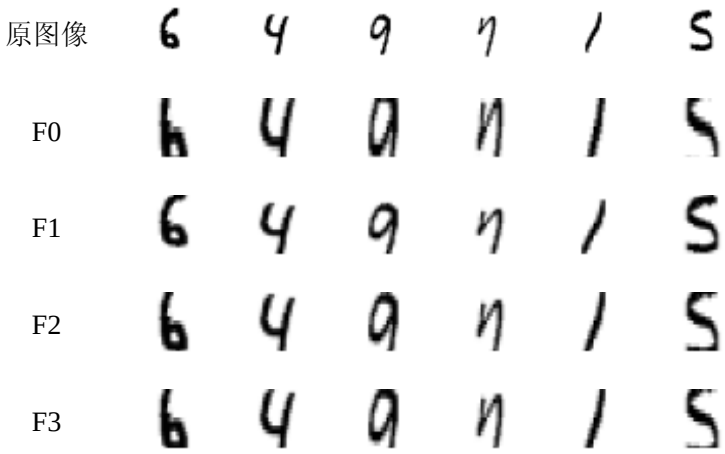


图 2.3 宽高比自适应规一化框图

在第二阶段，用 Sobel 算子计算图像中每个像素点的灰度梯度，数字图像 $f(x,y)$ 在 (x,y) 处水平和垂直两个方向上的灰度梯度分量分别为

$$\begin{aligned} g_x(x,y) &= f(x+1,y-1) + 2f(x+1,y) + f(x+1,y+1) \\ &\quad - f(x-1,y-1) - 2f(x-1,y) - f(x-1,y+1) \\ g_y(x,y) &= f(x-1,y+1) + 2f(x,y+1) + f(x+1,y+1) \\ &\quad - f(x-1,y-1) - 2f(x,y-1) - f(x+1,y-1) \end{aligned} \quad . \tag{2-5}$$

由 $g_x(x,y)$ 和 $g_y(x,y)$ 可以得出该点灰度梯度向量 $\boldsymbol{g}=g_x x+g_y y$ 的幅值和幅角

$$\begin{aligned}\|g\|&=\sqrt{g_x^2+g_y^2}\\ \theta&=arg\boldsymbol{g},\theta\in[0,2\pi]\end{aligned}. \tag{2-6}$$

将幅角 θ 的取值范围 $[0,2\pi]$ 均匀量化为 8 个区间。按照幅角所属的不同区间，可将灰度梯度的幅值图像分解为 8 个方向属性平面。在判定梯度向量的幅角所属范围之后，将梯度向量按照平行四边形法则投影到 $n\theta_0$ 和 $(n+1)\theta_0$ 两个方向上，并根据式（2-7）记录投影长度 g_n 和 g_{n+1}，如图 2.4 所示。图 2.5 是“8”的梯度特征提取中 8 个方向属性平面分解图示。

$$\begin{aligned}g_n&=\frac{\sin(n+1)\theta_0}{\sin\theta_0}g_x-\frac{\cos(n+1)\theta_0}{\sin\theta_0}g_y\\ g_{n+1}&=-\frac{\sin n\theta_0}{\sin\theta_0}g_x+\frac{\cos n\theta_0}{\sin\theta_0}g_y\end{aligned}. \tag{2-7}$$

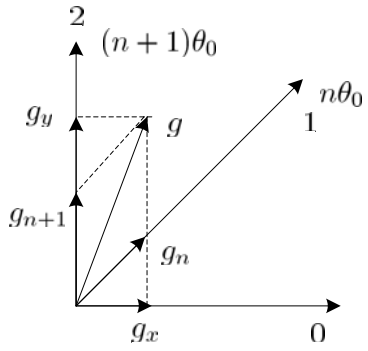


图 2.4 梯度特征抽取图示

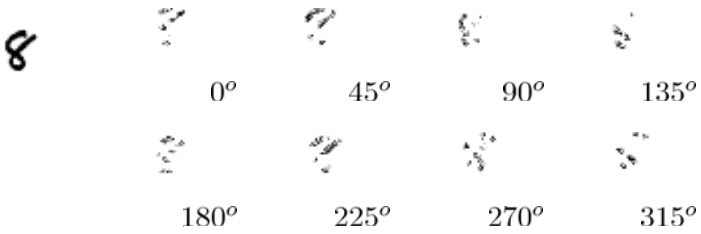


图 2.5 梯度特征的方向属性平面分解

然后，用高斯滤波器对梯度特征的每个方向属性平面进行亚采样，滤除图像中的噪声和字符形变带来的干扰，从而增强特征的鲁棒性。

高斯滤波器的冲激响应函数为

$$h(x,y)=\frac{1}{2\pi\sigma_x^2}\exp\left(-\frac{x^2+y^2}{2\sigma_x^2}\right), \tag{2-8}$$

其中 σ_x 是高斯滤波器的尺度参数

$$\sigma_x = \frac{\sqrt{2}t_x}{\pi}. \tag{2-9}$$

t_x 是采样间隔。

高斯滤波函数为

$$F(x_0, y_0) = \sum_x \sum_y f(x, y)\, h(x - x_0, y - y_0). \tag{2-10}$$

在 20×20 数字图像上以间隔 $t_x = 4$ 进行二维亚采样，即可得到 5×5 个采样值。将 8 个方向属性平面上得到的采样值合并到一起，就得到 $200 = 8 \times 5 \times 5$ 维的梯度特征。

最后，为了使提取的梯度特征更加接近高斯分布，该实验中采用 Box 和 Cox 提出的变换函数族对特征进行整形[106]。Box-Cox 对每一维特征分量的变换形式为

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}. \tag{2-11}$$

为了简化，字符识别中通常对所有的特征分量使用同样的幂方变换 $y = x^\mu$，本文通过实验确定 $\mu = 0.4$。

为了进一步提高学习效率，采用主成份分析（Principal Component Analysis，PCA）方法将 200 维的 e-grg 特征压缩到 120 维。设变换前的特征为 y，变换后的特征为 x，实验中的 PCA 变换方法如下：

计算训练集中所有样本的总散度矩阵为

$$\boldsymbol{S}_t = \frac{1}{n} \sum_{t=1}^{n} (\boldsymbol{y}_i - \boldsymbol{m})(\boldsymbol{y}_i - \boldsymbol{m})^T, \tag{2-12}$$

其中，n 是训练中所有类别的样本总数，m 是样本均值。

计算总散度矩阵 S_t 的特征值 Λ 和特征向量矩阵 Θ

$$\boldsymbol{S}_t \boldsymbol{\Theta} = \boldsymbol{\Theta} \boldsymbol{\Lambda}, \tag{2-13}$$

其中 Λ 对角线上的元素是散度矩阵的特征值 λ_i，Θ 中的列向量 Θ_i 是 λ_i 对应的特征向量。最后，特征变换如下

$$\boldsymbol{x}_i = \boldsymbol{\Theta}_i^T \boldsymbol{y}, \qquad i = 1, 2, \cdots, 120. \tag{2-14}$$

从而得到 120 维的 e-grg 特征。

对于 CENPAMI 手写体数字样本库中的图像，本文将其归一化为 35×35，再用上述方法提取 120 维特征。对于 MNIST 手写体数字样本库中的图像，本文首先从 28×28 的图像中抽取出 20×20 的数字图像，再在该数字图像上用上述方法提取 120 维特征。

2.3 正交高斯混合建模

本文采用有限高斯混合模型（Gaussian Mixture Model GMM）对各数字类的样本特征进行建模。高斯混合模型是目前使用最多的一种建模方法，它采用半参数化的密度估计方法，融合了参数化和半参数化估计方法的优点。高斯混合模型的复杂度决定于所描述问题的复杂度，而与样本集合的大小无关。与单高斯模型相比，高斯混合模型对特征的描述能力更强。理论上，如果模型成份个数选择恰当且训练样本充足，高斯混合模型能够以任意的精度逼近任意的概率分布[107]。

假设 $\boldsymbol{x}$ 是 d 维特征空间中来自 C_i 的特征向量，混合模型定义为 K 个成份密度为 $p_k(\boldsymbol{x}|\theta_k)$ 的线性叠加：

$$p(\boldsymbol{x}|\theta, w) = \sum_{k=1}^{K} \alpha_k p_k(\boldsymbol{x}|w_k), \tag{2-15}$$

其中，M 也被称为混合模型的阶数，w_k 是各成份的非负加权系数，满足

$$\sum_{k}^{M} w_k = 1. \tag{2-16}$$

混合模型将多个简单的密度函数叠加为一个更复杂的密度函数，增加模型的描述能力。

若混合模型中每个成份分布是多元高斯分布，则为高斯混合模型。多元高斯具有一个 $d \times d$ 的对称的、半正定的协方差矩阵 $\boldsymbol{\Sigma}_k$ 和 一个 $d \times 1$ 的均值向量 $\boldsymbol{\varphi}_k$。这些参数决定了此高斯函数的特性，如函数形状的中心点、宽窄和走向。高斯混合模型描述为

$$p\left(\boldsymbol{x}|C_i\right) = \sum_{k=1}^{K} w_k N\left(\boldsymbol{x}|\boldsymbol{\varphi}_k, \boldsymbol{\Sigma}_k\right), \tag{2-17}$$

其中，$N\left(\boldsymbol{x}|\boldsymbol{\varphi}_k, \boldsymbol{\Sigma}_k\right)$ 是单一高斯密度函数

$$N\left(\boldsymbol{x}|\boldsymbol{\varphi}_k,\boldsymbol{\Sigma}_k\right)=\frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_k\right|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\varphi}_k)^T\boldsymbol{\Sigma}_k{}^{-1}\left(\boldsymbol{x}-\boldsymbol{\varphi}_k\right)\Big). \quad (2\text{-}18)$$

在高斯混合模型中，由于基于全方差模型（Full Covariance）的 GMM 存储和计算代价较高，因此在实际应用中需要对高斯混合模型中的协方差矩阵进行约束。在各种方差约束中，对角约束的协方差矩阵应用较多。但由于特征向量之间通常是统计相关的，如果用对角协方差矩阵精确逼近给定的数据分布，需要较多的高斯成份。所以，本文采用正交高斯混合模型（Orthogonal Gaussian Mixture Model, OGMM）[108-109]。正交高斯混合模型利用特征向量矩阵将特征映射到正交变换空间，去除特征间的相关性，在正交变换空间中再用对角协方差矩阵描述特征分布。由于正交高斯混合模型先将信号能量集中于特征分布的各主轴方向上，同对角方差的高斯混合模型相比，可实现特征分布更有效的描述；而同全方差高斯混合模型相比，自由参数的数目大为减少，训练和识别的计算和存储代价大大降低。正交高斯混合模型描述为

$$p\left(\boldsymbol{x}|C_i\right)=\sum_{k=1}^{K}w_kO_k\left(\Omega_i^T\boldsymbol{x}|\boldsymbol{\varphi}_k,\boldsymbol{\Sigma}_k\right) \quad (2\text{-}19)$$

其中 $\boldsymbol{\varphi}_k$ 和 $\boldsymbol{\Sigma}_k$ 分别是正交变换空间中各高斯分量的均值和方差，$\boldsymbol{\Sigma}_k$ 是对角矩阵

$$\boldsymbol{\Sigma}_k=\begin{bmatrix}\gamma_{k1} & & & 0\\ & \gamma_{k2} & & \\ & & \gamma_{k3} & \\ 0 & & & \ddots\end{bmatrix}, \quad (2\text{-}20)$$

γ_{ki} 是第 k 个高斯成份对角线上第 i 个元素。式（2-19）中 Ω_i 是第 i 模式类 C_i 的正交高斯混合模型的变换矩阵，其列向量是第 i 模式类散度矩阵的特征向量。式（2-19）中第 k 个高斯成份描述为

$$O_k\left(\Omega_i^T\boldsymbol{x}\right)=\frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_k\right|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}\left(\Omega_i^T\boldsymbol{x}-\boldsymbol{\varphi}_k\right)^T\boldsymbol{\Sigma}_k{}^{-1}\left(\Omega_i^T\boldsymbol{x}-\boldsymbol{\varphi}_k\right)\Big). \quad (2\text{-}21)$$

当只有一个高斯成份时，正交高斯混合模型就等同于二次鉴别函数（Quadratic Discriminant Function, QDF）。如果进一步对各高斯分量的对角方差进行正则约束

$$\gamma_{ki}=\max\{\gamma_{\min},\gamma_{ki}\}\quad r=1,2\ldots d,\quad k=1,2\ldots K, \quad (2\text{-}22)$$

其中 $\gamma_{\min}$ 为对角线上元素限定阈值，则正交高斯混合模型就等价于改进二次鉴别函数（Modified Quadratic Discriminant Function, MQDF）。

2.4 基于后验伪概率的贝叶斯分类器

本文采用基于后验伪概率的贝叶斯分类器对所得手写体数字特征进行分类[96]。后验伪概率分类器是一种新的贝叶斯分类器，已成功应用于文本提取、字符串分割、以及图像检索等 [96-99]。

假设 $\boldsymbol{x}$ 为输入模式的特征向量，$\{C_1,\cdots,C_n\}$为模式类别。令 $p(C_i)$、$p(\boldsymbol{x}|C_i)$ 和 $p(C_i|\boldsymbol{x})$ 分别是先验概率、类条件概率密度和后验概率。根据贝叶斯理论，输入模式 $\boldsymbol{x}$ 属于具有最大后验概率的类 C^*，即

$$C^* = \arg\max_{C_i} p(C_i|\boldsymbol{x}). \tag{2-23}$$

根据贝叶斯公式

$$p(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i)\,p(C_i)}{\boldsymbol{x}}, \tag{2-24}$$

可以得到

$$\frac{p(C_i|\boldsymbol{x})}{p(C_j|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|C_i)\,p(C_i)}{p(\boldsymbol{x}|C_j)\,p(C_j)}. \tag{2-25}$$

所以，贝叶斯分类器的决策准则为

$$C^* = \arg\max_{C_i} p(\boldsymbol{x}|C_i)\,p(C_i). \tag{2-26}$$

根据式（2-25），可以建立同一类别在不同输入模式下的后验概率之间的关系，即

$$\frac{p(C_i|\boldsymbol{x})}{p(C_i|\boldsymbol{x}')} = \frac{p(\boldsymbol{x}|C_i)\,p(x')}{p(\boldsymbol{x}'|C_i)\,p(x)}. \tag{2-27}$$

式（2-25）与式（2-27）形式类似，但意义不同。式（2-25）中包含两个模式类和一个样本，表示不同模式类别在同一样本下的后验概率之间的相互关系；而式（2-27）包括一个模式类和两个不同样本，表示同一模式类别对不同样本的后验概率之间的相互关系。

由于 $\boldsymbol{x}$ 的取值范围很广，其观测值可认为是以相同概率出现于特征空间中任意一个点。因此，可假设 $\boldsymbol{x}$ 为均匀分布，即 $p(x)=p(x')$。所以，式（2-27）可以简化为

$$\frac{p(C_i|\boldsymbol{x})}{p(C_i|\boldsymbol{x}')} = \frac{p(\boldsymbol{x}|C_i)}{p(\boldsymbol{x}'|C_i)}, \tag{2-28}$$

即有

$$p(C_i|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_i). \tag{2-29}$$

根据式（2-29），可以用一个以 $p(\boldsymbol{x}|C_i)$ 为自变量，光滑、单调递增且值域在 $[0,1]$ 之间的函数来估计 $p(\boldsymbol{x}|C_i)$，将这类函数值称为“后验伪概率（posterior pseudo-probability）”[95]，表示为

$$f\left(p(\boldsymbol{x}|C_i)\right)=1-\exp(-\kappa p^{\beta}(\boldsymbol{x}|C_i)) \tag{2-30}$$

其中 $\kappa\in\mathbb{R}^{+}$ 和 $\beta\in\mathbb{R}^{+}$ 是后验伪概率函数中的控制系数。后验伪概率函数 $f\left(p(\boldsymbol{x}|C_i)\right)$ 是以 $p(\boldsymbol{x}|C)$ 为变量的光滑、单调递增函数，并满足 $f(0)=0$ 和 $f(+\infty)=1$。

对任意给定的输入模式 $\boldsymbol{x}$，我们用式（2-30）计算其对各个模式类别的后验伪概率，取最大后验伪概率对应的类别作为分类结果。因为 $f\left(p(\boldsymbol{x}|C_i)\right)$ 是 $p(\boldsymbol{x}|C)$ 的光滑、单调递增函数，所以基于后验伪概率的分类器与传统的贝叶斯分类器是一致的。但后验伪概率 $f\left(p(\boldsymbol{x}|C_i)\right)$ 的取值范围是 $[0,1]$，所以可作为一种相似度度量，并用于（1）作为拒识标准；（2）进行多分类器联合；（3）作为比传统计数方法更加准确的分类器评估标准。

2.5 实验数据库

本文选择了两个应用广泛的数据库进行实验，分别是 CENPAMI 手写体数字样本库[110]和 MNIST 手写体数字样本库[111]。

（1）CENPAMI 手写体数字样本库

CENPAMI 手写体数字样本库是加拿大肯考迪娅大学（Concordia University）CENPAMI（Centre for Pattern Recognition and Mahine Intelligence）实验室开发的。CENPAMI 数据库中的图像是从美国邮政局所收集到的日常信件以 166DPI 分辨率扫描得到的，共有 6000 幅数字图像（每个数字类有 600 幅图像），其中 4000 幅数字图像用于训练，2000 幅数字图像用于测试。图 2.6 中给出了 CENPAMI 数字图像样本库的测试集中部分手写体数字图像。

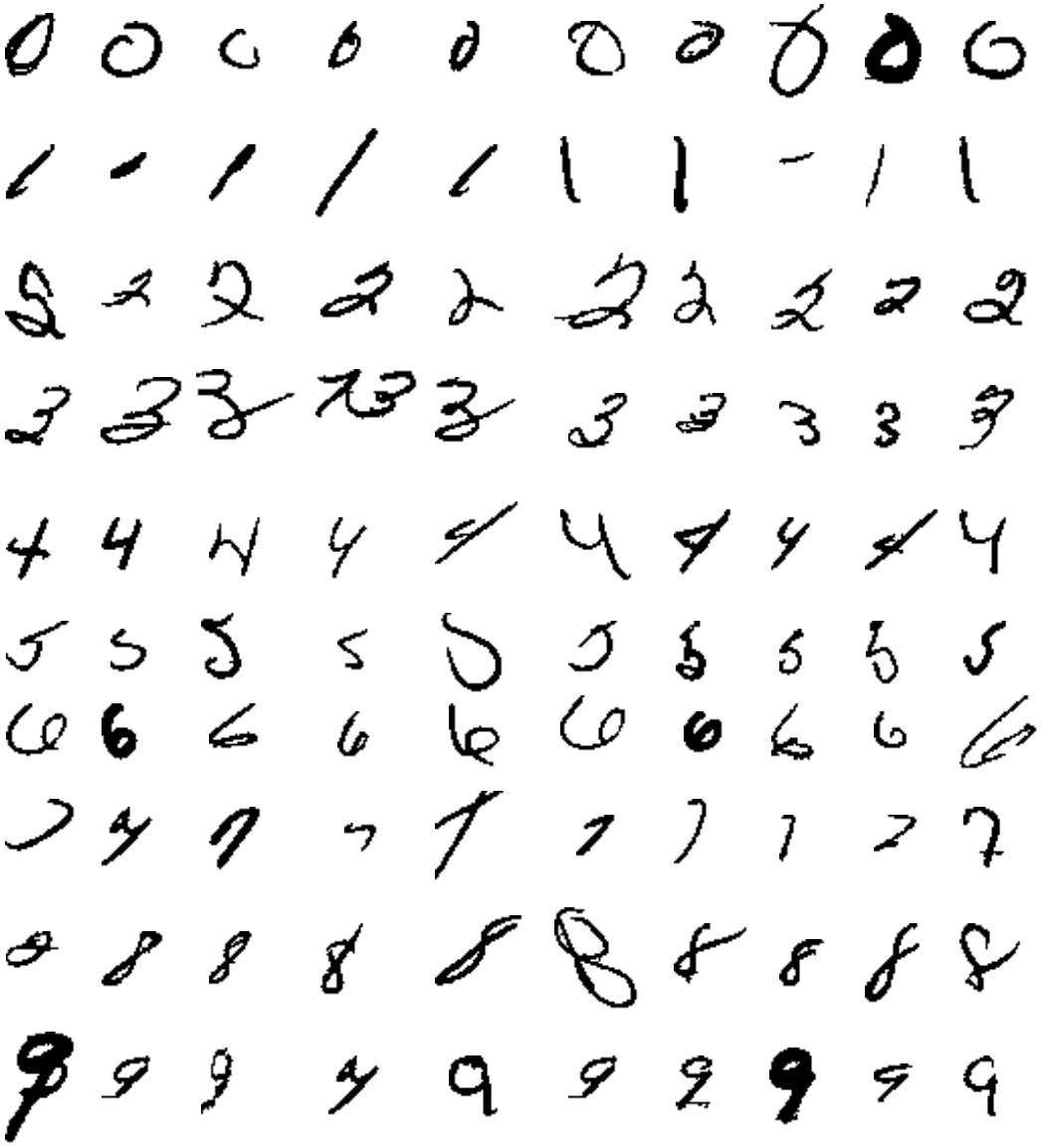


图 2.6 CEMPAMI 样本库测试集中部分手写体数字图像

（2）MNIST 手写体数字样本库

MNIST 数据库是由 AT&T 的 LeCun 将美国标准计量局（American National Institute of Standards and Technology）发布的 NIST-SD19（Special Database 19）手写体数字和英文样本库中的 NIST-SD1 和 NIST-SD3 两个样本集中的手写体数字样本重新整理得到的(http://yann.lecun.com/exdb/mnist/)。该样本库包含 60,000 个训练样本和 10,000 个测试样本，所有样本都被归一化为 20×20 大小，并采用重心与中心重合方式将其置于 28×28 图像中。MNIST 样本库在模式识别研究中已被非常广泛使用。图 2.7 给出了 MNIST 样本库的测试集中部分手写体数字图像。

图 2.7 MNIST 样本库测试集中部分手写体数字图像

2.6 小结

本章作为本文的应用背景，主要介绍本文所构建的手写体数字识别系统。本文选择手写体数字识别问题作为应用平台，所提出方法都将在手写体数字识别问题上进行验证，并与已有方法进行比较。本章构建的手写体数字识别实验系统包括手写体数字图像预处理、特征提取、数据库、图像特征建模和基于后验伪概率的贝叶斯分类器。

第 3 章　基于软目标的后验伪概率判别学习方法

3.1 引言

已有的判别学习方法主要是用预先设定的硬目标或者硬边缘度量分类器在训练集上的分类损失，学习目标或边缘在学习的过程中保持不变[33,46,77, 112-114]。但是让所有的样本都达到预设的硬目标或硬边缘几乎是不可能的，并且还会迫使分类器去继续学习那些已经充分学习的训练样本。因此，采用预设的硬目标或者硬边缘不但容易导致过学习，还会降低分类器的学习效率。

最近有许多文献提出用软目标或软边缘的方法解决由硬目标或者硬边缘所带来的问题，并用于学习各种分类器，如神经网络 (Neural Networks, NN)[115-117]、支持向量机(Support Vector Machines, SVMs) [118-119]、自适应提升算法(Adaboost) [120-122]、贝叶斯分类器(Bayesian Classifiers)[91]等。软目标或者软边缘方法中，学习目标或者边缘并不是预先设定的，而是在学习过程中自适应调整。理论和实验都已证明用软目标或者软边缘代替预设的硬目标或者硬边缘可以提高分类器的泛化性能[115-122]。

基于软目标的方法在神经网络中的应用包括基于排序信息误差平方和准则（sum of squares errors, SSE）的反向传播神经网络（backPROPagation）[115]、面向模式分类的判别函数方法(Classification Based objective functions, CB)[117]等。在 backPROPagation 学习方法中，输入模式依据其输出值进行排序，并根据所有样本的排序结果设定下次迭代的输出目标值。该方法被用于多任务学习，实验证明了其有效性[115-116]。在面向模式分类的判别函数方法中（CB），每个样本根据当前的分类损失动态调整下次训练目标值。实验表明 CB 方法的学习效果明显好于误差平方和最小化准则（SSE）或者交叉熵准则（Cross-Entropy）的学习效果。

利用软边缘方法可以调节支持向量机在训练集上边缘最大化与降低分类损失两个对立的目标[118-119]，Chen 等人[119]还给出基于软边缘支持向量机的误差边界理论分析。除支持向量机外，Li 等人[91]提出了基于软边缘的概率密度估计方法（Soft Margin Estimation, SME），用于学习贝叶斯分类器。SME 将支持向量机中的软边缘思想结合到最小分类错误（MCE）决策反馈方法中，提高分类器的泛化性能，并将该方

法应用到基于隐马尔可夫模型（Hidden Markov Model，HMM）的语音识别当中，实验结果证明了该方法的有效性。

基于软边缘的方法还被用到 Boost 算法中。RÄTSCH 等人[120]通过其提出的三种规则化方法将软边缘的思想推广到 AdaBoost 中，提高了 AdaBoost 的泛化性能。他们将一个根据软边缘思想设计的规则项结合到自适应提升算法目标函数中，降低其对噪声的敏感程度，并且分别在自适应提升算法的线性规划和二次规划的过程中引入松弛因子，使其达到软边缘的泛化效果。Demirize 等人[121-122]通过最小化软边缘误差函数解决该 boosting 问题，用线形规划优化目标函数。

本章提出一种软目标方法学习基于后验伪概率的分类器。目前用于学习后验伪概率分类器的学习方法是最大最小后验伪概率学习方法（Max-Min posterior Pseudo-probabilities，MMP）[95]，该方法使每个模式类别的正样本后验伪概率趋近于 1，反样本的后验伪概率趋近于 0。在 MMP 方法中，学习目标是预先设定的 0 和 1，并在学习过程中保持不变，所以该方法是基于硬目标的学习方法。因此，MMP 学习方法具有前面所提及的硬目标学习方法所具有的局限性，即存在过学习风险和优化速度慢。本文将软目标思想用于 MMP 学习方法中，提高其泛化效果和优化效率。首先，我们对每个模式类的正样本和反样本后验伪概率分别定义相应的自适应软目标，并用该软目标度量分类器在训练集上的分类损失。通过最小化分类损失，同时最大化两个软目标之间的距离，获得贝叶斯分类器的最优参数集合。除分类器的分类损失外，两个目标值之间的距离对学习效果也有很大的影响。所以，既要控制分类器的分类损失，又要控制两个软目标值之间的距离。显然，两个软目标值之间的距离越大，分类器的泛化性能越好；相反，如果目标之间的距离很小，甚至为负数的话，即使分类损失为 0 也不会得到满意的效果。最后，用梯度下降方法对目标函数进行优化。为了降低过学习风险，提高训练速度，本文还利用软目标值对训练数据集进行样本选择，以减少训练样本。在数据选择过程中，对于那些后验伪概率远超过相应目标值的训练样本在设定的训练周期内暂时被移出训练集。

3.2 最大最小后验伪概率判别学习方法

本节将介绍用于学习后验伪概率分类器的最大最小后验伪概率判别学习方法（Max-Min posterior Pseudo-probabilities，MMP）。MMP 方法的核心思想是通过最大化正样本的后验伪概率，使其趋近于 1 ，同时最小化反样本的后验伪概率，使其趋近于 0 ，

获得最佳分类能力。这里，正样本指属于某一类别的样本，反样本指不属于某一类别的样本。为了对该方法进行规则化描述，假设 $f(\boldsymbol{x};\boldsymbol{\Lambda})$ 是某一模式类的后验伪概率度量函数，其中 $\boldsymbol{\Lambda}$ 表示该函数中的未知参数集合。设 $\hat{\boldsymbol{x}}$ 是属于某一模式类的任一正样本的特征向量，$\bar{\boldsymbol{x}}$ 是相对某一模式类的任一反样本的特征向量，m 和 n 分别表示正样本和反样本的数量。根据前面阐述的 MMP 核心思想，MMP 学习方法的目标函数定义为

$$F(\Lambda)=\frac{1}{m}\sum_{i=1}^{m}\left[f(\hat{\boldsymbol{x}};\boldsymbol{\Lambda})-1\right]^2+\frac{1}{n}\sum_{i=1}^{n}\left[f(\bar{\boldsymbol{x}};\boldsymbol{\Lambda})\right]^2. \tag{3-1}$$

由于正样本和反样本的数量级可能存在较大的差别，因此，在目标函数中取正样本和反样本损失函数的均值形式，避免由于正样本和反样本数量不平衡而引起的问题。

显然，$F(\boldsymbol{\Lambda})$ 的值越小，分类效果越好；当 $F(\boldsymbol{\Lambda})=0$ 获得最佳分类效果。通过最小化目标函数 $F(\boldsymbol{\Lambda})$ 获得后验伪概率度量函数的最优参数集合 $\boldsymbol{\Lambda}^*$，即：

$$\boldsymbol{\Lambda}^*=\arg\min_{\boldsymbol{\Lambda}}F(\boldsymbol{\Lambda}). \tag{3-2}$$

用最速梯度下降算法优化目标函数（3-1），得到最优参数集合 $\boldsymbol{\Lambda}^*$。最速梯度下降法沿函数的梯度反方向，迭代更新参数。设 $\boldsymbol{\Lambda}_t$、α_t 分别是第 t 次迭代时的参数集和步长，$\bigtriangledown F(\boldsymbol{\Lambda})$ 表示 $F(\boldsymbol{\Lambda})$ 对 $\boldsymbol{\Lambda}_t$ 的偏导，则

$$\boldsymbol{\Lambda}_{t+1}=\boldsymbol{\Lambda}_t-\alpha_t\bigtriangledown F(\boldsymbol{\Lambda}). \tag{3-3}$$

根据式（3-3），具体学习流程如下：

Step 1. 利用所有样本（正样本和反样本），计算目标函数对各个参数的导数。设 ψ 表示参数集合 $\boldsymbol{\Lambda}$ 中任意参数，则计算公式为

$$\frac{\partial F}{\partial\psi}=\frac{2}{m}\sum_{i=1}^{m}\left(f(\hat{\boldsymbol{x}};\boldsymbol{\Lambda})-1\right)\frac{\partial f(\hat{\boldsymbol{x}};\boldsymbol{\Lambda})}{\partial\psi}+\frac{2}{n}\sum_{i=1}^{n}f(\bar{\boldsymbol{x}};\boldsymbol{\Lambda})\frac{\partial f(\bar{\boldsymbol{x}};\boldsymbol{\Lambda})}{\partial\psi}. \tag{3-4}$$

式（3-4）中的 $\frac{\partial f(\hat{\boldsymbol{x}};\boldsymbol{\Lambda})}{\partial\psi}$ 和 $\frac{\partial f(\bar{\boldsymbol{x}};\boldsymbol{\Lambda})}{\partial\psi}$ 的导数形式由具体应用中的 $f(\boldsymbol{x};\boldsymbol{\Lambda})$ 和参数集合 $\boldsymbol{\Lambda}$ 所决定。

Step 2. 根据目标函数的偏导数计算步长因子。

Step 3. 根据式（3-3）更新参数。

Step 4. 重复以上步骤直到收敛或达到最大迭代次数。设 ε 为极小阈值，则收敛条件定义为

$$\|g_t\| = \left[\sum \left(\frac{\partial F}{\partial \psi}\right)^2\right]^{\frac{1}{2}} \leqslant \varepsilon. \tag{3-5}$$

MMP 对每一模式类别，分别采用上述方法迭代更新类模型中的有关参数，直到收敛或达到预设的最大迭代次数。

3.3 基于软目标的最大最小后验伪概率学习方法

在 3.2 节介绍的 MMP 算法中，我们期望某一模式类的后验伪概率对于属于该类的正样本度量为 1，而对于不属于该类的反样本度量为 0。但如 3.1 节中论述，很难让训练集中的所有样本都达到预定的“0”和“1”硬目标，而且采用硬目标学习方法还会导致分类器对训练集数据过拟合，降低学习效率。本章提出用软目标方法提高 MMP 的泛化性能，加快训练速度。

3.3.1 后验伪概率软目标

本章提出用软目标替代在 MMP 目标函数中对每一个模式类正样本和反样本的后验伪概率所设定的“0”和“1”硬目标值。假设 $\hat{H}$ 和 $\bar{H}$ 分别是正样本和反样本的后验伪概率的软目标值，如图 3.1 所示，图中的圆形和正方形分别代表相对某一模式类的正样本和反样本的后验伪概率，d 是两个软目标值之间距离，$d = \hat{H} - \bar{H}$。

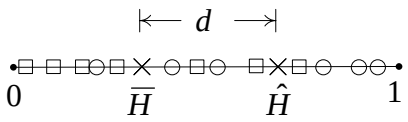


图 3.1 后验伪概率软目标图示

我们提出的软目标与 MMP 方法中的硬目标之间区别是：MMP 学习准则中的“0”和“1”硬目标是提前设定的，并且在学习的过程中保持不变，我们所提出的软目标则在训练过程中自适应调整。

3.3.2 经验损失与目标函数

用每一模式类所对应的软目标度量分类器在训练集上的分类损失。如果某一正样本的后验伪概率值大于相应软目标值，则认为该正样本已经被充分学习，其分类损失为 0。假设 $\hat{\boldsymbol{x}}$ 是第 i 个模式类任意一个正样本的特征向量，则其分类损失度量函数

定义为

$$\hat{l}\left(\hat{\boldsymbol{x}};\boldsymbol{\Lambda}\right)=\begin{cases}0 & f\left(p\left(\hat{\boldsymbol{x}}|C_i\right)\right)>\hat{H}\\ \hat{H}-f(p(\hat{\boldsymbol{x}}|C_i)) & f(p(\hat{\boldsymbol{x}}|C_i))\leqslant\hat{H}\end{cases}. \tag{3-6}$$

同样，第 i 类反样本的分类损失度量函数定义为

$$\bar{l}(\bar{\boldsymbol{x}};\boldsymbol{\Lambda})=\begin{cases}0 & f(p(\hat{\boldsymbol{x}}|C_i))<\bar{H}\\ f(p(\bar{\boldsymbol{x}}|C_i))-\bar{H} & f(p(\bar{\boldsymbol{x}}|C_i))\geqslant\bar{H}\end{cases}, \tag{3-7}$$

其中，$\hat{\boldsymbol{x}}$ 是第 i 类任意一个反样本的特征向量。式（3-6）和（3-7）中，$\boldsymbol{\Lambda}$ 表示经验损失度量函数中的未知参数集合，包括 $\hat{H}$、$\bar{H}$ 和 $f\left(p\left(\boldsymbol{x}|C_i\right)\right)$ 所包含的未知参数。

假设 m 和 n 分别表示训练集中第 i 类正样本和反样本的数量，则第 i 个模式类总的经验损失 $L(\boldsymbol{\Lambda})$ 定义为

$$L(\boldsymbol{\Lambda})=\frac{1}{m}\sum_{i=1}^{m}\hat{l}^2(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda})+\frac{1}{n}\sum_{i=1}^{n}\bar{l}^2(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda}). \tag{3-8}$$

式（3-8）中，我们取正样本和反样本损失函数的均值形式，目的是避免由于正样本和反样本数量差别较大而导致等式不平衡问题。

除前面所论述的经验损失以外，正样本和反样本所对应软目标值之间的距离也很重要。如果正样本和反样本所对应软目标之间的距离很小，甚至为负数，那么即使分类器的经验损失是“0”也不会得到满意的结果。因此，本章提出的最大最小后验伪概率软目标学习方法(SoftDS-MMP)的优化目标是最小化分类器的经验损失 $L(\boldsymbol{\Lambda})$，同时最大化正样本所对应软目标 $\hat{H}$ 和反样本所对应的软目标 $\bar{H}$ 之间的距离。根据以上论述，SoftDS-MMP 的目标函数设计为

$$F(\boldsymbol{\Lambda})=\omega(1-d)^2+(1-\omega)L(\boldsymbol{\Lambda}), \tag{3-9}$$

其中 $\omega\in[0,1]$ 是一个权重因子，用于控制经验损失 $L(\boldsymbol{\Lambda})$ 与软目标之间距离 d 在目标函数中的比例关系。显然，ω 值越大，目标函数越倾向于最大化软目标之间距离 d；相反，ω 值越小，目标函数越倾向于最小化分类损失。通过最小化 $F(\boldsymbol{\Lambda})$ 可以获得最优分类器参数集合 $\boldsymbol{\Lambda}^*$

$$\boldsymbol{\Lambda}^*=\arg\min_{\boldsymbol{\Lambda}}F\left(\boldsymbol{\Lambda}\right). \tag{3-10}$$

3.3.3 优化方法

我们用最速梯度下降方法优化式（3-9）中的 SoftDS-MMP 目标函数，从而获得最优参数集合，具体是用下式对后验伪概率分类器的参数集合迭代更新

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_t - \alpha_t \bigtriangledown F(\mathbf{\Lambda}_t), \tag{3-11}$$

其中，$\mathbf{\Lambda}_t$ 和 α_t 分别代表后验伪概率分类器的参数集合和第 t 次迭代的步长，$\bigtriangledown F(\mathbf{\Lambda}_t)$ 是目标函数 $F(\mathbf{\Lambda}_t)$ 对参数集合 $\mathbf{\Lambda}_t$ 中参数的偏导数。在附录 A 给出了参数集合中每个参数的求导结果。

3.4 基于软目标的训练数据选择方法

大量研究和实验表明判别学习方法的学习效果好于传统的生成学习（Generative Learning），但由于计算复杂度等原因，其优化效率往往并不理想。利用数据选择策略对训练数据集进行压缩是提高判别学习优化效率有效方法之一。

在语音识别领域中已经提出了一些数据选择方法来提高判别学习方法的优化效率[123-125]。Liu 等人[123]提出了三个级别的数据选择策略，分别是语音级、语句级和结构级。Arslan[124]等人用重估公式为每个训练数据动态分配权重，从而控制异常数据的影响。Jiang 等人在文献[125]中提出了一种可以自动从语音数据中发现竞争信息的动态数据选择方法。该方法用 Viterbi 搜索方法对语音数据进行解码，在解码过程中可以自动将每个 HMM 模型的训练数据分成两个数据集，即竞争数据集和真实数据集。在模型的学习过程中，对真实数据集用生成学习方法，而对竞争数据集则采用判别学习方法。

正如前面论述，在 SoftDS-MMP 方法中，当训练样本的后验伪概率超出其相应的软目标值时期望损失为 0。这就意味着这些样本不被用来进行参数更新，这种方式可以看成是一种数据选择过程，使学习的过程中只关注那些经验损失不为 0 的易混淆的数据，降低过学习风险。该思想与软边缘估计（Soft Margin Estimation, SME），最大边缘估计（Large Margin Estimation, LM）和基于分类的判别函数学习方法(Classification Based objective functions)中采用的方式相同。但即使应用这种样本选择方法，计算每个样本的后验伪概率仍然需要较大的计算代价。在本节中，我们提出一种数据选择策略，进一步提高 Soft-MMP 的训练效率，该策略将那些后验伪概率远超出其相应软目标值的样本在一定的周期内移出训练集。

显然，如果一个样本的后验伪概率远超出其相应的软目标值，则其经验损失在一定训练周期内都倾向于保持为 0。因此，一个有效的方法是在一定的训练周期内暂时将这些数据从训练集中移出，减少训练样本数量，压缩训练集，从而提高训练效率。

图 3.2 是对该动态数据选择方法的图示，其中白色的圆形和正方形分别表示被暂时从训练集中移出的正样本和反样本的后验伪概率，黑色圆形和正方形则是仍然保留在训练集中正样本和反样本的后验伪概率，阈值 $\delta \in [0,1]$ 用来判断样本的后验伪概率是否明显超出其相应目标值。

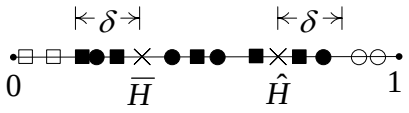


图 3.2 动态数据选择图示

为了对该方法进行形式化描述，假设 $\hat{\boldsymbol{S}}_{t-1}$ 和 $\bar{\boldsymbol{S}}_{t-1}$ 分别是被选择参加第$t-1$ 次训练的正样本和反样本集合，$\hat{\boldsymbol{S}}_t$ 和 $\bar{\boldsymbol{S}}_t$ 则是参加第 t 次训练的样本集合，样本选择策略可以描述为

$$\begin{cases} \hat{\boldsymbol{S}}_t = & \{\hat{\boldsymbol{x}} | f\left(p(\hat{\boldsymbol{x}}|C_t)\right) \leqslant \hat{H} + \delta \ and \ \hat{\boldsymbol{x}} \in \hat{\boldsymbol{S}}_{t-1}\} \\ \bar{\boldsymbol{S}}_t = & \{\bar{\boldsymbol{x}} | f\left(p(\hat{\boldsymbol{x}}|C_t)\right) \geqslant \bar{H} - \delta \ and \ \bar{\boldsymbol{x}} \in \bar{\boldsymbol{S}}_{t-1}\} \end{cases}, \tag{3-12}$$

对于在第 t 次训练时被移出的样本，在第 $t+R$ 次训练时将被重新插入到训练集当中，其中 R 是被移出训练集的训练周期。

在前面介绍的动态数据选择方法中包含两个阈值，分别是用于判断数据移出的阈值 δ 和用于数据插入的阈值 R。假设 t_{max} 是指定的最大训练次数，δ_{max} 和 δ_{min} 分别是阈值 δ 的上界和下界，则在 t 次训练时阈值 δ 描述为

$$\delta = \delta_{\max} - t\left(\delta_{\max} - \delta_{\min}\right)/t_{\max}. \tag{3-13}$$

设 R_0 是将样本暂时移出训练集的训练次数间隔，i 是样本根据式（3-12）被移出训练集的次数，则对用于数据插入的阈值 R 定义为

$$R = iR_0. \tag{3-14}$$

根据式（3-13）和式（3-14）计算阈值 δ 和 R 的主要原因如下。根据式（3-12），用于判断样本移出的阈值 δ 在训练过程中由其上界 $\delta_{\max}$ 逐渐减小到 $\delta_{\min}$。因此，在训练的初始阶段，几乎所有的样本都被保留在训练集中，以保证更多的样本可以被充分学习。随着训练次数的增加，越来越多已经充分学习的样本将被移出训练集。如果样本被移出和插入训练集的次数越多，表示样本被学习的越充分，则逐渐增加该样

本被移出训练集的训练次数间隔。实验表明了式（3-13）和式（3-14）的有效性。

3.5 SoftDS-MMP 算法流程

算法 3.1 给出了本文所提出的基于软目标并带样本选择的最大最小后验伪概率学习方法对每一模式类的具体实现流程。

算法 3.1　学习后验伪概率分类器的 SoftDS-MMP 算法

Input: 训练数据集，后验伪概率分类器的初始参数集合，初始的正样本和反样本的软目标值

优化过程：

Repeat

步骤 1　计算当前分类器在训练数据集上的经验损失；

步骤 2　根据式（3-12）将已经充分学习的数据暂时移出训练集；

步骤 3　计算 SoftDS-MMP 的目标函数 $F(\mathbf{\Lambda})$ 关于参数集合 $\mathbf{\Lambda}$ 中参数在训练数据集上的偏导数；

步骤 4　应用式（3-11）更新未知参数集合；

步骤 5　应用式（3-13）更新阈值 δ；

步骤 6　根据式（3-14）将暂时移出训练集的样本重新插入到训练集中；

Until 算法收敛或者达到最大训练次数 t_{max}。假设 ε 是一个极小值，则该算法的收敛条件是

$$F(\mathbf{\Lambda}_t) - F(\mathbf{\Lambda}_{t+1}) \leqslant \varepsilon$$

Output: 自适应软目标值和后验伪概率分类器的最优参数集合。

SoftDS-MMP 分别对每一模式类别采用上述方法迭代更新类模型中的有关参数，直到收敛或达到预设的最大迭代次数。

3.6 SoftDS-MMP 在手写体数字识别中的应用

我们将 SoftDS-MMP 应用于手写体数字识别问题，在 MNIST 手写体数字样本库上进行实验[111]，并与其他方法进行比较。数字的特征提取和建模方法如本文第二章所述。

本文中所有实验均在计算服务器上完成，其配置为双 2.0GHz Intel 处理器和 2.0 千兆字节的随机存贮器。

后验伪概率分类器中的概率密度形式取正交混合高斯模型，即将式（2-19）代入式（2-30），得到基于后验伪概率的手写体数字分类器，描述为

$$f(\boldsymbol{x};\boldsymbol{\Lambda}) = 1 - \exp\left\{-\kappa\left(\sum_{k=1}^{K} w_k O_k\left(\Omega_i^T \boldsymbol{x}|\boldsymbol{\varphi}_k, \boldsymbol{\Sigma}_k\right)\right)^{\beta}\right\}. \tag{3-15}$$

用 SoftDS-MMP 方法学习该分类器，其中未知参数包括

$$\boldsymbol{\Lambda} = \{\kappa, \beta, w_k, \boldsymbol{\varphi}_k, \boldsymbol{\Sigma}_k, \hat{H}, \bar{H}\}, k = 1, \cdots, K. \tag{3-16}$$

式（3-16）中的部分参数必须满足一定约束条件。为了易于优化，本文将其转换为无约束变量。在表 3.1 中列出了式（3-16）中的有约束变量和变换后的无约束变量。表 3.1 中 τ 是为了防止协方差中元素值过小而引起估计错误，设定的最小阈值。变换后的参数集合为

$$\widetilde{\boldsymbol{\Lambda}} = \{\widetilde{\kappa}, \widetilde{\beta}, \widetilde{w_k}, \boldsymbol{\varphi}_k, \widetilde{\boldsymbol{\Sigma}_k}, h_1, h_2\}, k = 1, \cdots, K. \tag{3-17}$$

表 3.1 数字分类器的 SoftDS-MMP 学习方法中的参数的约束条件及其转换形式

有约束参数	变换后无约束参数
$0 < \hat{H} < 1;\ 0 < \bar{H} < 1$	$\hat{H} = \frac{1}{1+\exp(-h_1)};\ \bar{H} = \frac{1}{1+\exp(-h_2)}$
$\kappa > 0;\ \beta > 0$	$\kappa = \exp(\widetilde{\kappa});\ \beta = \exp(\widetilde{\beta})$
$\gamma_{kj} > \tau$	$\gamma_{kj} = \exp(\widetilde{\gamma}_{kj}) + \tau$
$\sum w_k = 1$	$w_k = \frac{\exp(\widetilde{w}_k)}{\sum \exp(\widetilde{w}_k)}$

应用本文提出的 SoftDS-MMP 学习变换后的参数集合 $\widetilde{\boldsymbol{\Lambda}}$ 中的参数，然后再将这

些参数变换为原始的参数 $\boldsymbol{\Lambda}$。采用梯度下降算法优化目标函数。附录 A 给出了目标函数关于 $\widetilde{\boldsymbol{\Lambda}}$ 中参数的偏导数。

为了验证 SoftDS-MMP 方法的性能，将其应用于手写体数字识别问题中，并在 MNIST 手写体数字样本库上进行实验。

本章提出的手写体数字分类器训练过程可分为三个阶段。第一阶段，采用 AutoClass 算法[126]选择每个数字类 GMM 的成份个数。AutoClass 是一种聚类算法，其利用朴素贝叶斯分类器对数据集合进行分类，通过 EM 算法获得最大后验概率，实现数据的聚类。我们将 AutoClass 所自动获取数据集聚类个数作为对其进行描述的 GMM 模型的成份个数。表 3-3 中列出基于 AutoClass 的每个数字类模型成份选择结果。

表 3.2 各数字类的 GMM 模型成份个数选择结果

数字类	0	1	2	3	4	5	6	7	8	9
模型成份个数	4	8	5	4	5	6	6	9	4	8

第二阶段，用期望最大化算法（Expectation-Maximization，EM）学习得到相应 GMM 模型中参数的最大似然估计（Maximum Likelihood Estimation，MLE），并通过实验选择后验伪概率函数中参数的初始值 $\kappa = 10$ 和 $\beta = 0.02$。

第三阶段，分别应用六种判别学习方法，包括最小分类错误学习方法（MCE）、最大软边缘估计（SME）、基于分类的判别函数学习方法(CB)、最大最小后验伪概率学习方法（MMP）、软目标的学习方法（Soft-MMP）、基于软目标并带样本选择的最大最小后验伪概率学习方法(SoftDS-MMP)，利用每个数字类别的所有正样本和反样本修改由 EM 算法所获得的参数最大似然估计。该实验中所有判别学习均采用批量优化方法（Batch Learning）。各种判别学习中需要经验设置的参数都通过实验用交叉验证的方法确定，具体取值如下：

- Soft-MMP和SoftDS-MMP中用于控制经验损失和软目标之间距离的权重因子取0.05；
- 对于MCE方法，其sigmoid函数中的光滑因子取0.1；
- CB方法中的边缘值取0.02；
- SME中的边缘值取20。

最终通过学习可以得到基于后验伪概率的手写体数字分类器的 7 个参数集，分别对应于 EM、MCE、SME、CB、MMP、Soft-MMP 和 SoftDS-MMP 学习方法。图 3.3

是用七种不同方法学习手写体数字后验伪概率分类器实验的流程框图。用由上述七种不同学习方法所得到的后验伪概率分类器分别对手写体数字识别问题进行开放和封闭测试。根据实验结果，对提出的 SoftDS-MMP 方法学习效果和学习效率分别进行分析。

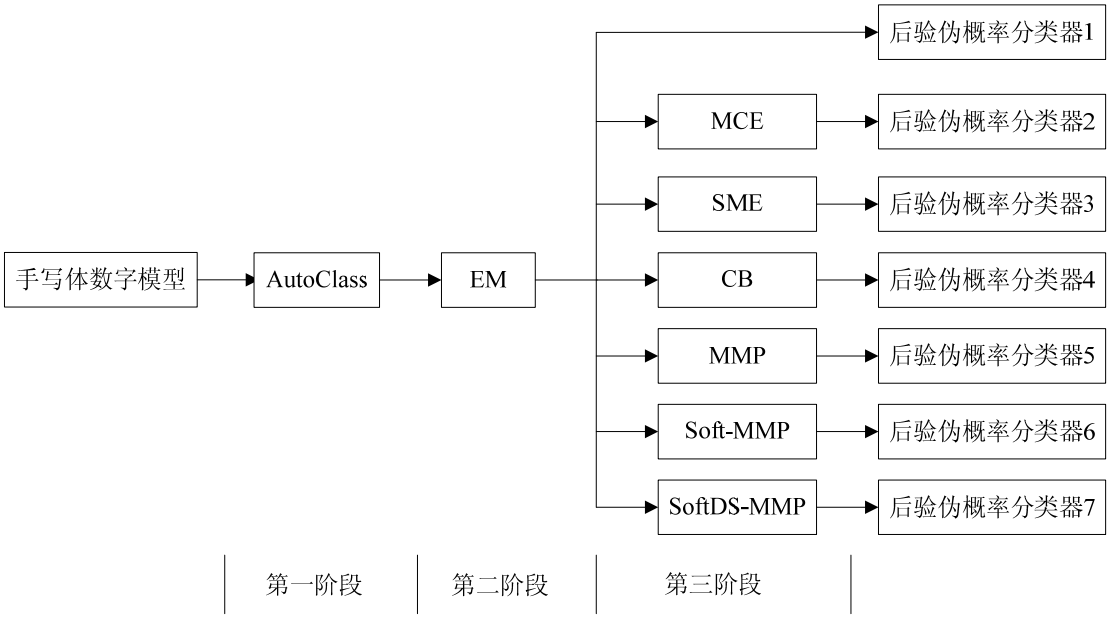


图 3.3 用七种不同方法学习后验伪概率分类器流程框图

3.6.1 学习效果分析

表 3.3 列出了由七种不同的学习方法所得分类器在训练集上的误识率，在测试集的误识率和泛化性能。实验中，我们用分类器在测试集上的识别率与在训练集上的识别率比值评估其泛化性能，该评估方法与 Rimer 和 Martinez 在文献[117]中所用的方法是相同的。显然，该比值越大，说明泛化性能越好；反之，则越差。表 3.3 还列出了应用 SoftDS-MMP 方法与其它学习方法相比所带来的测试集上识别错误下降率。

表 3.3 中实验数据表明，本文所提出的 SoftDS-MMP 学习方法无论识别效果，还是泛化性能都要好于其他的判别学习方法。文献[103]对不同的手写体数字识别技术（包括分类方法和特征提取方法）在 MNIST 手写体数字样本库上进行了全面的研究和比较。他们应用 e-grg 特征在 MNIST 数据库测试集上所取得的最低误识率是用基于 RBF 核函数的支持向量机得到的 0.42％。我们用提出的 SoftDS-MMP 学习方法和相同的 e-grg 特征在测试集上得到了与其接近的误识率。

表 3.3. 不同学习方法的识别率和泛化性能比较

学习方法	训练集 误识率(%)	测试集 误识率(%)	测试集误识 下降率(%)	泛化效果
EM	0.79	0.85	34.92	0.9994
MCE	0.16	0.71	12.70	0.9945
SME	0.18	0.67	6.35	0.9951
CB	0.20	0.66	4.76	0.9954
MMP	0.18	0.69	9.52	0.9950
Soft-MMP	0.22	0.63	_	0.9959
SoftDS-MMP	0.23	0.63	_	0.9960

对获得该实验结果的原因分析如下。MCE、CB 和 SME 学习方法是从观测样本的角度去度量模式类的可分性，而 SoftDS-MMP 则是将每个模式类作为一个整体来度量类的可分性。由于使类最大可分的策略不同，从而降低了 SoftDS-MMP 对数据的敏感性。与 MMP 方法相比，SoftDS-MMP 在训练过程中更加关注那些分布在边界上且与其他样本易混淆的样本，因此 SoftDS-MMP 学习效果好于 MMP 学习方法。

3.6.2. 学习效率分析

为了分析 SoftDS-MMP 方法的优化效率，我们在实验中记录了不同判别学习方法优化分类器所需要的时间，并将记录结果列于表 3.4 中。从表 3.4 中的数据可以看出 SoftDS-MMP 相对于其他的判别学习方法要更有效率。与 MMP 相比，SoftDS-MMP 将训练时间缩短了 63.91%。此外，基于软目标的 MMP 学习方法通过样本选择，将训练时间从最初的 7805 秒缩短为 3549 秒，实验结果说明了所提出的数据选择方法有效性。

表 3.4 不同判别学习方法训练时间比较

学习方法	MCE	SME	CB	MMP	Soft-MMP	SoftDS-MMP
训练时间(秒)	9621	6802	6421	9835	7805	3549

与其他判别学习准则相比，SoftDS-MMP 从两个方面提高了算法的优化效率。一方面，因为 SoftDS-MMP 是从类的角度考虑类的可分性，其所有类的后验伪概率度量函数可以并行优化。相反，MCE、CB 和 MSE 是从样本的角度考虑类的可分性，其各类的后验伪概率函数只能顺序优化。另一方面，SoftDS-MMP 通过样本选择策略压缩了训练数据集，这也是 SoftDS-MMP 相对于 MMP 优化效率更高的主要原因。首先，只有经验损失不为 0 的样本被选择，用于训练分类器，从而降低了计算代价。其次，在本章提出的数据选择策略对提高学习效率起到重要作用。为了分析该策略的有效性，我们分别记录了十个数字类别的前 2000 次训练过程中，训练集保留的样本数量，并将其绘制于图 3.4 中。从图 3.4 可以看出经过样本选择以后训练集中保留的样本数量随着训练次数的增加不断减少。在 200 次训练之后，每个类别的训练集中近一半的训练样本已经被移出训练集，从而降低了计算代价，提高了训练效率。

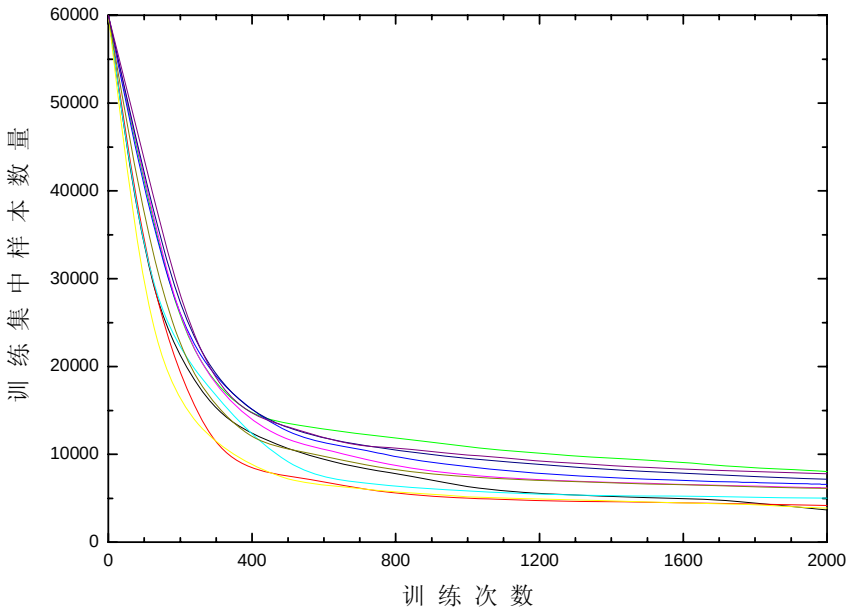


图 3.4 各类别训练过程中被选择训练样本数量变化曲线

3.6.3 用软目标值度量类可分性

在实验中，我们分别记录了应用 SoftDS-MMP 方法训练后各类别的软目标值和训练集中经验损失不为 0 的样本数量 r，如表 3.5 所示。从表 3.5 的实验数据中可以发现正样本和反样本的相应软目标值之间的距离似乎可以用来度量各类别之间的判别能力。如果一个类别容易与其他类别区分，则软目标 $\hat{H}$ 和 $\bar{H}$ 之间的距离较大，

如数字“0”；相反，如果一个类别容易与其他类别混淆，则其对应的两个软目标之间的距离较小，如数字“8”。我将在今后工作中研究软目标与类可分性度量标准之间的关系。

表 3.5　应用 SoftDS-MMP 训练后各类别相应软目标值和训练集中经验损失不为 0 的样本数量

类别	r	$\hat{H}$	$\bar{H}$	d
0	2598	0.85	0.18	0.67
1	1597	0.86	0.15	0.71
2	3733	0.76	0.25	0.51
3	3201	0.76	0.31	0.45
4	3023	0.77	0.25	0.52
5	3798	0.78	0.30	0.48
6	2071	0.82	0.20	0.62
7	3561	0.76	0.22	0.53
8	3848	0.75	0.32	0.43
9	2037	0.76	0.31	0.45

3.7 小结

本章提出了一种新的学习后验伪概率分类器的判别学习方法，即基于软目标的最大最小后验伪概率学习方法（Soft target based MMP Learning with Data Selection，SoftDS-MMP）。与基于硬目标的学习方法相比，SoftDS-MMP 具有更好的学习效果、更快的优化效率和更强的泛化性能。SoftDS-MMP 在学习过程中自适应调整目标值，降低了过学习风险。而且 SoftDS-MMP 还可以利用软目标，在学习过程中动态选择训练样本，压缩训练集，提高优化速度。

为了验证 SoftDS-MMP 方法的有效性，将其应用于手写体数字识别，并与最大最小后验伪概率学习方法（MMP）、最小分类错误学习方法（MCE）、基于分类的判别函数方法（CB）和最大软边缘估计方法（SME）进行比较。我们采用八个方向的梯度特征（e-grg）[104]，并在 MNIST 手写体样本库上进行实验。通过比较，SoftDS-MMP 的学习效率和学习效果都好于其他的学习方法。与 MMP 相比，SoftDS-MMP 使：1）

识别错误率由 0.85% 下降到 0.63%；2）训练时间由 9835 秒缩短到 3549 秒；3）以测试集上的识别率与训练集上识别率的比值作为度量标准的泛化性能由 0.9950 提高到 0.9960。与 MCE、SME 和 CB 三种判别学习方法相比，SoftDS-MMP 使：1）训练集上错误率分别降低了 26.52%、17.80% 和 16.38%；2）训练时间分别缩短了 63.11%、47.82% 和 44.73%；3）泛化性能分别提高了 0.15%、0.09% 和 0.06%。

实验数据显示所提出的软目标还可用于度量类的可分性。与其他类别可以显著区分的类的两个软目标之间的距离要大于易混淆的类的两个软目标之间的距离。我们将在今后的工作中研究软目标与类可分性之间的关系。

第 4 章　高斯混合模型的判别选择方法

4.1 引言

目前，在许多文献和工作中提出了 GMM 的参数估计方法，并发布一些比较满意的实验结果。但如何选择合适的 GMM 结构却仍然是一个具有挑战性的问题。本章将针对 GMM 的模型选择问题展开研究和探讨。

本章将用于学习贝叶斯分类器的基于软目标的最大最小后验伪概率判别学习准则（SOFT target based Max-Min posterior Pseudo-probabilities with dynamic Data Selection，SoftDS-MMP）结合到贝叶斯模型选择框架下，从而提出一种新的 GMM 模型选择判别式方法。该 GMM 模型选择判别式方法用拉普拉斯方法估计 SoftDS-MMP 目标函数的边缘积分，用得到的拉普拉斯估计值评价 GMM 模型对某一模式类的泛化性能，并用线性搜索方法寻找 SoftDS-MMP 边缘积分拉普拉斯估计的最大值。该模型选择判别式方法可同时学习 GMM 的模型结构和模型参数。为了验证所提出的方法，本章将该方法应用于手写体数字识别，并在 CENPARMI 和 MNIST 手写体数字样本库上进行实验。还将该方法与三种常用的模型选择生成式方法进行比较，包括贝叶斯准则（Bayesian Information Criterion, BIC）、最小描述长度（Minimum Description Length，MDL）和 AutoClass。通过实验比较，所提出方法的学习效果不但明显好于人工设定的模型结构，而且还好于由经典的 BIC 模型选择准则、MDL 模型选择准则和 AutoClass 所选择的模型结构，从而说明所提出方法的有效性。

4.2 主流 GMM 选择方法

迄今为止，已发表了许多面向统计建模的模型选择方法。根据各模型选择方法的理论基础不同，可分为模型选择的生成式方法和模型选择的判别式方法。目前，已有的模型选择方法主要是生成式方法。模型选择的生成式方法前提是假设模型集合中包含数据的真实分布模型。在该假设下，选择似然边缘积分最大的模型作为数据的最优模型，而不考虑分类器的分类错误率。近几年，一些文献提出了模型选择的判别式方法，该方法将判别学习相关准则引入到模型选择当中，以分类器的性能为依据选择模型结构。下面分别简单介绍两类模型选择方法中主流方法。

4.2.1 生成式方法

目前，已有的有限混合模型选择方法主要是生成式的，根据各种模型选择方法所依据的理论基础不同，又可将其分为五类[6]：基于贝叶斯准则的模型选择方法（approximate Bayesian criteria）[4,24,128,129]、基于信息准则的模型选择方法[130-132]（informaiton criteria）、基于分类准则的模型选择方法（classification criteria）[133-134]、基于随机采样的模型选择方法（stochatic approaches）[135]和基于交叉验证的模型选择方法（cross-validation approaches）[33-34]。

（1）基于贝叶斯准则的模型选择方法。贝叶斯准则方法来源于贝叶斯理论，其假设模型对未知数据的似然与模型在训练集上的边缘似然紧密相关，即在训练集上具有最大边缘似然的模型为最优模型，描述为

$$M^* = \arg\max_M \int p(\boldsymbol{x}|\boldsymbol{\Lambda}, M)p(\boldsymbol{\Lambda}|M)d\boldsymbol{\Lambda}. \tag{4-1}$$

对模型在训练集上边缘似然的不同估计方法衍变出了不同的贝叶斯模型选择准则，其中常用的贝叶斯模型准则包括 Schwarz 提出的贝叶斯信息准则（Bayesian Information Criterion，BIC）[26,128]、拉普拉斯估计方法（Laplace Approximation）[129-130]、Lewis 和 Raftery 提出的最佳区域修正的拉普拉斯估计方法（Laplace-Metropolis Criterion）[131]、拉普拉斯经验准则（Laplace-Empirical Criterion, LEC）[4]。

（2）基于信息准则的模型选择方法。信息准则是根据信息理论提出的一类模型选择方法。该类方法将模型选择作为寻找数据传输的最优编码过程，数据的概率分布作为编码发生器，信息准则方法假设数据的概率分布 $p(\boldsymbol{X}|M)$ 对于发射器和接收器都是已知的。根据 Shannon 的信息熵理论，编码长度由两部分构成，即数据适应度描述 $-\log p(\boldsymbol{X}|M)$ 和信道代价惩罚因子 $C(M)$。基于信息准则的模型选择标准为

$$M^* = \arg\min_M \{-\log p(\boldsymbol{X}|M) + C(M)\}. \tag{4-2}$$

编码长度中的信道代价惩罚因子 $C(M)$ 与编码发生器的复杂度有关，编码发生器越复杂，则其代价越高。常用的基于信息准则方法有最小描述长度（Minimum Description Length，MDL）[31]、最小信息长度（Minimum Message Length，MML）[30]、Akaike 信息标准（Akaike's Information Criterion，AIC）[32]、基于 Bootstrap 的信息准则（Bootstrap-Based Inforamtion Criterion）[132]和信息复杂度准则（Informational Complexity Criterion，ICOMP）[133]。

（3）基于分类准则的模型选择方法。该方法是基于完备数据的似然值在 EM 框

架下寻找适应当前数据的最优模型。Biernacki 和 Govaert 在 1997 年基于混合模型的似然值和完备数据似然值之间的关系提出了一种选择数据聚类个数的方法，并将其作为模型选择准则。常用的基于分类准则的方法包括分类似然准则（Classification Likelihood Criterion，CLC）[134]、规则熵准则（Normalized Entropy Criterion，NEC）[33]、联合分类似然准则（Integrated Classification Likelihood Criterion，ICL）[34]和 Approximate Weight of Evidence（AEW）[135].

（4）基于随机采样的模型选择方法。另一种常用的模型选择方法是利用随机采样方法确定模型成份个数。其中最常用的是马尔可夫链蒙特卡洛方法（Markov chain Monte Carlo Sampling，MCMC）[35]。基于马尔可夫链的估计方法用参数空间采样点的均值估计贝叶斯证据因子(Bayesian Evidence)，其可描述为

$$\begin{aligned} p(\boldsymbol{X}|M) &= \int p(\boldsymbol{X}|\lambda, M)p(\lambda|M)d\lambda \\ &\approx \frac{1}{N_{mc}}\sum_{i} p(\boldsymbol{X}|\lambda_i, M) \end{aligned} \quad , \qquad (4\text{-}3)$$

其中 λ_i 是模型参数的第 i 个采样，N_{mc} 是采样数量。但简单 MCMC 要求在参数空间中所抽取的采样之间相互独立，该条件在实际当中很难满足，因此又提出了许多改进方法，如基于拒绝采样策略(rejection sampling)的 MCMC、基于重要采样（inportance sampling）策略的 MCMC 等[136]。基于采样方法存在的一个主要问题是其计算代价高，不适合大规模模式识别问题。一个模式识别系统经常包含几千个自由参数，这将导致采样空间维数过高。

（5）基于交叉验证（cross-validation）的模型选择方法[37]。一些文献提出用交叉验证的方法确定模型成份个数。常用的交叉验证方法有 leave-ont-out 交叉验证、v-折交叉验证（v-fold cross-validation）、Monte-Carlo 交叉验证。Smyth 对交叉验证、BIC、bootstrap LRTS 和 Monte-Carlo 四种模型选择方法进行了比较，得到了相近的学习效果[137]。

上述五种模型选择方法都属于模型选择的生成式方法，该方法在满足两个前提的条件下，即样本的真实模型在所选模型集合当中和训练样本充足，假设模型在未知数据上的识别错误与其在训练数据上的边缘似然相关，从而根据最大边缘似然准则选择最优模型结构。但该类方法的局限性是其前提条件通常很难满足，我们很难知道数据的真实分布，通常也很难得到充足的训练样本，且该类方法忽略了模型的判别能力。采用模型选择的生成式方法有时不能得到满意的学习效果。

4.2.2 判别式方法

模型选择的生成式方法，主要来源于经典的最大似然准则，研究重点在于用模型更好的描述正样本的分布，但是却忽略了模型的判别能力。最近，许多判别学习方法，如最小分类错误学习方法（Minimum Classification Error，MCE）[6]、最大互信息（Maximum Mutaual Information, MMI）[8]、最小语音错误（Minimum Phone Error，MPE）[79]、最大最小后验伪概率(Maximum Minimum Pseudo-Posterior Probability，MMP)[95]等，被提出用于模式分类中的 GMM 的参数估计，大量的实验表明其效果要好于传统的生成式学习方法。随着判别学习在模型参数估计中所取得成功，一些文献也提出了模型选择的判别式方法，其可以分为两类：

一类是基于模型分裂策略的模型选择判别式方法。该类方法以模型的判别能力作为模型选择的评价准则，通过对所选择高斯成份的分裂操作增加模型的判别性能。M.padmanabhan 和 L.R.Bahl [138-139]提出在模型选择过程中分别对每个高斯成份进行分裂操作，如果该分裂操作使模型对训练集中正样本的后验概率增加，则保留该分裂操作。该方法被应用于语音识别中，使识别率得到提高。Y.Normandin [140]提出在训练过程中，根据最大互信息准则选择模型进行分裂，如果模型中的某个成份根据最大互信息准则对正样本和反样本的区分能力强，则对该成份进行分裂操作。该类方法的局限性是没有对模型的复杂度进行控制，因此容易引起过学习。

另一类是将已有的判别学习准则融入到模型选择当中。Liu 和 Gales[127]提出了一种基于贝叶斯模型选择框架的 GMM 模型复杂度控制方法。该方法将经典的最小语音错误判别学习准则和最大互信息判别学习准则嵌入到贝叶斯模型选择框架中。该方法被应用到大规模连续语音识别当中，实验结果证明其有效性。

4.3 贝叶斯模型选择方法

贝叶斯模型是一种重要的生成式模型选择方法，本节将简要介绍贝叶斯模型选择方法。关于贝叶斯模型选择框架的详细内容请参考文献[29]。

4.3.1 贝叶斯模型选择准则

贝叶斯模型选择方法的基础是假设未知数据在当前模型下的似然值与模型参数的边缘积分紧密相关。边缘积分值越大，表示模型的泛化性能越好。假设 M 代表模型的结构信息，对于 GMM 模型便是代表其成份个数。$\mathbf{\Lambda}$ 是模型中的未知参数集合，

$p(\boldsymbol{\Lambda}|M)$ 是关于 $\boldsymbol{\Lambda}$ 的先验分布，$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$ 代表训练数据集，则模型参数的边缘概率表示为

$$p(\boldsymbol{X}|M)=\int p(\boldsymbol{X}|\boldsymbol{\Lambda},M)p(\boldsymbol{\Lambda}|M)d\boldsymbol{\Lambda}. \tag{4-4}$$

具有最大似然边缘积分的 M^* 将被选择作为最优模型结构：

$$M^*=\arg\max_M \int p(\boldsymbol{X}|\boldsymbol{\Lambda},M)p(\boldsymbol{\Lambda}|M)d\boldsymbol{\Lambda}. \tag{4-5}$$

在式（4-4）和式（4-5）中，由于缺少模型真实分布和参数个数的先验知识，通常假设模型参数的先验分布 $p(\boldsymbol{\Lambda}|M)$ 是一个无关项。最优模型 M^* 的选择标准仅取决于 $\int p(\boldsymbol{X}|\boldsymbol{\Lambda},M)d\boldsymbol{\Lambda}$，即：

$$M^*=\arg\max_M \int p(\boldsymbol{X}|\boldsymbol{\Lambda},M)d\boldsymbol{\Lambda}. \tag{4-6}$$

$p(\boldsymbol{X}|\boldsymbol{\Lambda},M)$ 通常被称为是贝叶斯证据（Bayesian evidence）。

在贝叶斯模型选择方法中，模型参数作为自由变量，在参数空间中对其进行积分，通过控制模型参数的边缘积分来控制模型的复杂度。如果模型过于简单，则不能很好的描述数据的分布信息；如果模型过于复杂，则会导致过拟和，影响模型的泛化性能。对于贝叶斯方法，过于复杂的模型由于参数空间的自由参数过多，会受到惩罚，影响其边缘积分值。图 4.1 是具有不同模型复杂度的三个模型示意图，11 个黑色的点代表观测数据，绿色、红色和蓝色曲线分别代表具有一个、两个和七个成份的高斯模型。图 4.1 中，只有一个成份的高斯模型过于简单，不能很好的描述数据分布信息；相反，有七个成份的高斯模型虽然可以很好的描述观测数据，但是由于过于复杂，导致泛化性能差。有两个成份的高斯模型在图 4.1 中被标为三个模型中的最佳模型。该模型可以很好的描述一定范围内观测数据和未观测数据的分布。并且，相对于被标为复杂和简单的高斯模型，最佳模型在该范围内贝叶斯证据值最高。因此，在贝叶斯模型选择方法中，能够有效描述观测数据的最简单的模型就是最佳模型。

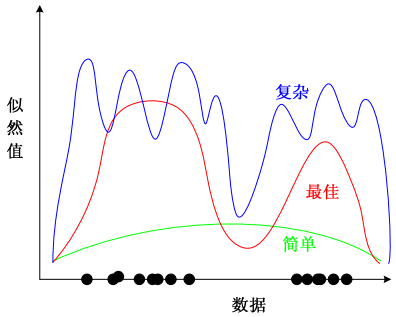


图 4.1 具有不同结构复杂度的三个模型示意图

4.3.2 贝叶斯准则计算方法

直接计算似然的边缘积分是比较困难的，通常采用的方法是利用似然的边缘积分在最优参数集 $\mathbf{\Lambda}^*$ 处的泰勒级数展开式对其做近似估计。目前，最常用的边缘似然估计方法有两种，分别是

$$\log p(\boldsymbol{X}|M) \approx \log p(\boldsymbol{X}|\mathbf{\Lambda}*, M) - \rho\frac{S}{2}\log N, \text{(BIC)} \tag{4-7}$$

和

$$p(\boldsymbol{X}|M) \approx p(\boldsymbol{X}|\mathbf{\Lambda}^*, M)\sqrt{\frac{(2\pi)^S}{|-\bigtriangledown^2_{\Lambda=\Lambda^*}\log p(\boldsymbol{X}|\mathbf{\Lambda}, M)|}}, \text{(Laplace).} \tag{4-8}$$

式（4-7）和式（4-8）中，S 是 GMM 模型中的参数个数。式（4-7）中，ρ 是惩罚因子。式（4-8）中，$|\cdot|$ 代表矩阵的行列式值。

Schwartz 证明当 $\rho = 1$ 时，BIC 是贝叶斯边缘积分的一阶渐近展开[29]。在用 BIC 方法估计似然的边缘积分时存在两个问题[141]：第一，BIC 只是贝叶斯证据的一阶渐近展开，当训练数据较多时，泰勒级数的高阶项可以忽略，但是当训练数据较少时，泰勒级数的高阶项包含较多信息，因此用 BIC 方法估计贝叶斯证据将会引起较大偏差；第二，在 BIC 方法中对模型复杂度的惩罚因子 $\rho\frac{S}{2}\log N$ 仅考虑模型中自由参数的个数 S，而没有考虑到参数的形式。

拉普拉斯估计是贝叶斯证据因子边缘积分的二阶渐近展开，其基本思想是用一个单高斯函数在参数空间拟合模型的最大似然值曲线。高斯函数的均值是模型的最优参数，通常是模型的最大似然估计；高斯函数的方差是目标函数关于最优参数集合的 Hessian 矩阵。通过计算参数空间中高斯函数所涵盖体积，估计目标函数的积分值。相对于 BIC 方法，拉普拉斯估计在 Hessian 矩阵中引入了不同参数的类型信息。Kass 等人在文献[142]中证明在一定的条件下拉普拉斯估计的相对错误是 $O(1/N)$，也可以说当样本足够多的时候，拉普拉斯估计可以以任意的精度逼近其真实值。Mclachlan 和 Ng 进一步通过实验证明拉普拉斯估计的效果要好于 BIC 方法[143]。

图 4.2 是用拉普拉斯方法估计一维函数 $f(x)$ 积分的图示。高斯函数的均值 x^* 是函数 $f(x)$ 的最优解，高斯函数的方差是函数 $f(x)$ 对变量 x 在最优解 x^* 处的二阶导数 $-\bigtriangledown^2_{x=x^*}\log f(x)$。

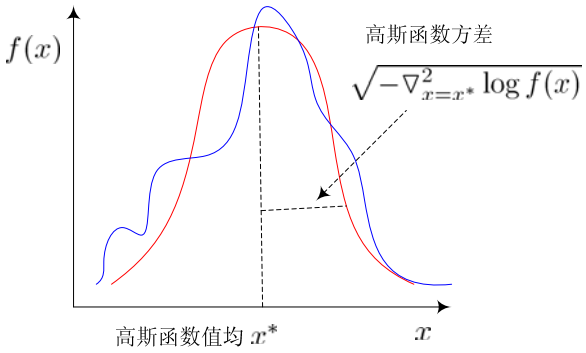


图 4.2 拉普拉斯估计一维函数积分图示

4.4 基于 SoftDS-MMP 的贝叶斯模型选择方法

本节将基于 SoftDS-MMP 目标函数边缘积分的拉普拉斯估计提出一种新的高斯混合模型选择判别式方法。首先，给出模型选择准则和拉普拉斯估计方法；然后，用线性搜索策略同时优化模型结构和模型参数。该工作与 Liu 和 Gales [127]所做的工作相关，但区别在于所采用的模型评价准则不同。我们是基于 SoftDS-MMP 设计判别模型选择准则，而 Liu 和 Gales 是基于 MMI 和 MPE 方法的模型选则准则。

4.4.1 SoftDS-MMP 目标函数的转换

本节根据前面介绍的基于拉普拉斯估计的贝叶斯模型选择策略设计 GMM 模型的评价准则。贝叶斯模型选择策略是基于经典的最大似然生成式学习准则，而本章将采用 SoftDS-MMP 的判别学习准则。令 $F(\boldsymbol{\Lambda}, M)$ 是 SoftDS-MMP 的目标函数，把 $F(\boldsymbol{\Lambda}, M)$ 代入贝叶斯模型选择准则，即得到基于 SoftDS-MMP 的 GMM 模型判别选择准则，为

$$M^* = \arg\max_{M} \int F(\boldsymbol{\Lambda}, M) d\boldsymbol{\Lambda}. \tag{4-9}$$

由于 SoftDS-MMP 学习方法是一个求极小值问题，目标函数的边缘积分不能用拉普拉斯方法估计，因此我们将原始的 SoftDS-MMP 目标函数重新改写为

$$\tilde{F}(\boldsymbol{\Lambda}, M) = C\left(1.0 - \bar{H} + \hat{H}\right) + \left[mn - \frac{n}{2}\sum_{i=1}^{m}\hat{l}^2(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda}) - \frac{m}{2}\sum_{j=1}^{n}\bar{l}^2(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda})\right]. \tag{4-10}$$

（4-10）中的第一项和第二项分别控制各模式类所对应模型的两个软目标之间的距离和其在训练集上的经验损失。参数 $C \in [0,1]$ 用于调节目标函数的两部分的比重。式（4-10）所描述函数与式（3-9）所描述函数的目标相同，都是最大化正样本和反样本

所对应软目标之间的距离，同时最小化训练集上的经验损失。但是，式（4-9）所描述的 SoftDS-MMP 目标函数将最初求极小值问题转变为求极大值问题，适合用拉普拉斯方法估计边缘积分。通过求取变换后的目标函数 $\widetilde{F}(\mathbf{\Lambda}, M)$ 极大值，可以得到最优参数集合 $\mathbf{\Lambda}^*$：

$$\mathbf{\Lambda}^* = \arg\max_{\mathbf{\Lambda}} \widetilde{F}(\mathbf{\Lambda}, M). \tag{4-11}$$

根据式（4-10），用梯度上升方法优化目标函数。设 $\mathbf{\Lambda}_t$ 和 α_t 分别是第 t 次训练时的参数集合和步长因子，$\nabla\widetilde{F}(\Lambda_t, M)$ 是目标函数 $\widetilde{F}(\mathbf{\Lambda}_t, M)$ 关于参数集合 $\mathbf{\Lambda}_t$ 中参数的偏导数，梯度上升方法描述为

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_t + \alpha_t \bigtriangledown \widetilde{F}(\mathbf{\Lambda}_t, M). \tag{4-12}$$

4.4.2 GMM 模型判别评价准则

应用拉普拉斯方法估计目标函数的边缘积分时，如果模型中的参数过多，会导致全 Hessian 矩阵的计算代价过大。为了降低计算代价，假设模型中的变量相互独立，从而将全 Hessian 矩阵简化为对角矩阵，即

$$\nabla^2_{\mathbf{\Lambda}=\mathbf{\Lambda}^*}\widetilde{F}(\mathbf{\Lambda}, M) = \begin{bmatrix} \ddots & & 0 \\ & \nabla^2_{\mathbf{\Lambda}^{(i)}=\mathbf{\Lambda}^{(i)}}\widetilde{F}(\mathbf{\Lambda}, M) & \\ 0 & & \ddots \end{bmatrix}. \tag{4-13}$$

此外，考虑到 $\int \widetilde{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda}$ 值可能很大，存在溢出风险，该工作取 $\int \widetilde{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda}$ 的对数值来评价 GMM 模型，为了进一步简化 $\int \widetilde{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda}$ 的对数计算代价，将目标函数变换为

$$\hat{F}(\mathbf{\Lambda}, M) = \exp\left(2\pi\widetilde{F}(\mathbf{\Lambda}, M)\right). \tag{4-14}$$

目标函数的拉普拉斯估计可以表示为

$$\begin{aligned} \log\int \hat{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda} &\approx \log\left(\hat{F}\left(\mathbf{\Lambda}^*, M\right)\sqrt{\frac{(2\pi)^S}{\left|-\nabla^2_{\Lambda=\mathbf{\Lambda}^*}\log\hat{F}(\mathbf{\Lambda}, M)\right|}}\right) \\ &= 2\pi\widetilde{F}(\mathbf{\Lambda}^*, M) - \frac{1}{2}\sum_{i=1}^{S}\log\left|-\nabla^2_{\mathbf{\Lambda}=\mathbf{\Lambda}^*}\widetilde{F}(\mathbf{\Lambda}, M)\right| \end{aligned}. \tag{4-15}$$

式（4-15）中 $\log\int \hat{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda}$ 的估计值作为模型选择的判别式准则，用于评价 GMM 模型对某一模式类的泛化性能。为了方便论述，在后面的章节中将用符号 $\zeta(M)$ 代表 $\log\int \hat{F}(\mathbf{\Lambda}, M)d\mathbf{\Lambda}$ 的近似估计。使 $\zeta(M)$ 值最大的GMM模型将被选做最优模型结构，

即

$$M^* = \arg\max_{M} \left\{ 2\pi \widetilde{F}(\boldsymbol{\Lambda}^*, M) - \frac{1}{2} \sum_{i=1}^{S} \log \left| -\nabla^2_{\Lambda=\Lambda^*} \widetilde{F}(\boldsymbol{\Lambda}, M) \right| \right\}. \tag{4-16}$$

附录B中给出了基于SoftDS-MMP模型判别选择和学习方法中所用的模型选择的判别式准则 $\log \int \hat{F}(\boldsymbol{\Lambda}, M) d\boldsymbol{\Lambda}$ 关于参数集合 $\boldsymbol{\Lambda}$ 中参数的一阶和二阶导数求导结果。

4.4.3 线性搜索策略

现有的模型选择方法中大部分方法采用穷举搜索策略寻找最优模型结构，并根据先验知识确定搜索间隔。本章所提出的判别模型选择方法采用线性搜索策略寻找最优模型结构。

SoftDS-MMP目标函数边缘积分的拉普拉斯估计 $\log \int \hat{F}(\boldsymbol{\Lambda}, M) d\boldsymbol{\Lambda}$ 是以GMM模型成份个数为变量的凸函数，因此可以用线形搜索策略来确定包含最优解的搜索区域。线形搜索的具体方法是采用进退法[144]确定最初的搜索区域。进退法的基本思想描述如下：首先，随机选择一点，并将其作为初始点。从该点出发，沿着固定的方向，按照一定的步长，试图找到函数值呈现“低-高-低”的三个点。若该方向搜索不成功，则退回到出发点，并沿相反的方向搜索。

在确定了搜索区域之后，我们采用黄金分割算法[145]不断的缩小搜索区域，直到找到具有最大目标函数边缘积分的模型结构。该方法通过取试探点和函数值进行比较，使包含极小点的搜索区间不断减小，直到搜索区间内没有可选择的点为止，并将搜索区间中最终具有极大函数值的点作为最优解输出。

综上所述，本章提出的高斯混合模型选择判别式方法由SoftDS-MMP参数估计方法、GMM模型判别选择准则和线性搜索策略构成，详细算法见算法4.1。

算法 4.1 GMM 模型的判别选择与学习算法

输入: g_0：初始化 GMM 模型结构；$[a,b]$：搜索区间；η：搜索步长；$v>1$：加速因子；$\lambda=2$：搜索间隔终止长度。

模型选择与学习:

步骤 1　根据进退初始化搜索区间

步骤 1.1　设置迭代次数 $t=0$。计算面向 g_t 个成份 GMM 模型的 Soft-MMP 目标函数值边缘拉普拉斯估计 $\zeta(g_t)$。

步骤 1.2　令 $g_t=g_t+\eta$。计算拉普拉斯估计值 $\zeta(g_{t+1})$。如果 $\zeta(g_{t+1})>\zeta(g_t)$，则转到步骤 1.3，否则转到步骤 1.4。

步骤 1.3　令 $\eta=v\eta$，$g=g_t$，$g_t=g_{t+1}$，$t=t+1$，转到步骤 1.2。

步骤 1.4　如果 $t=0$，则令 $g_t=g_{t+1}$，$\eta=-\eta$，并转到步骤 1.2，否则令 $a=\min\{g,g_t\}$，$b=\max\{g,g_t\}$。

步骤 2　利用黄金分割方法缩小搜索区域

步骤 2.1　令$a_1=a$，$b_1=b$，$t=1$。在搜索区间 $[a,b]$ 确定分割点 p 和 q值: $p_1=\lfloor a_1+0.382(b_1-a_1)\rfloor$, $q_1=\lceil a_1+0.618(b_1-a_1)\rceil$, 计算$\zeta(p_1)$和$\zeta(q_1)$。

步骤 2.2　如果 $\zeta(p_t)<\zeta(q_t)$, 转到步骤 2.3，否则转到步骤 2.4。

步骤 2.3　如果$b_t-p_t\leqslant\lambda$，则结束搜索过程，并且输出具有 q_t 个成份的最优 GMM 模型；否则，令 $a_t=p_t$，$b_{t+1}=b_t$，$p_{t+1}=q_t$，$q_{t+1}=\lceil a_{t+1}+0.618(b_{t+1}-a_{t+1})\rceil$，计算$\zeta(q_{t+1})$，转步骤 2.5。

步骤 2.4　如果$q_t-a_t\leqslant\lambda$，则结束搜索过程，并且输出具有 p_t 个成份的最优 GMM 模型；否则，令$a_{t+1}=a_t$，$b_{t+1}=q_t$，$q_{t+1}=p_t$，$p_{t+1}=\lfloor a_{t+1}+0.382(b_{t+1}-a_{t+1})\rfloor$，计算$\zeta(p_{t+1})$，转步骤 2.5。

步骤 2.5　令 $t=t+1$，且转到步骤 2.2。

输出: 最优 GMM 模型

4.5 实验结果

将提出的高斯混合模型选择判别式方法应用于手写体数字识别中，评价该方法的性能，并在广泛使用的 CENPAMI[110]和 MNIST[111]手写体数字样本库上进行实验。手写体数字图像特征提取与建模方法见第二章。

4.5.1 模型选择方法比较

在实验中，将提出的模型选择方法与经典的贝叶斯模型选择准则(BIC)、最小描述长度模型选择准则(MDL)、AutoClass 模型选择方法和人工设定的模型结构进行比较。在 CENPAMI 手写体数字样本库上，应用上述四种自动模型选择方法，即 BIC、MDL、AutoClass 和模型选择判别式方法，获得的所有数字类别最优模型成份个数均值约为 4，所以在人工设定的模型结构中，分别将高斯混合模型的成份个数设置成 3 至 5 个。在 MNIST 手写体数字样本库上，应用四种自动模型选择方法获得的最优模型成份个数均值约为 7，在人工设定的模型结构中，将高斯混合模型的成份个数设置为 6 至 8 个。为了对自动模型选择方法进行比较，本文对每一数字类分别应用上述四种不同的自动模型选择方法优化高斯混合模型结构。表 4.1 和表 4.2 中分别列出了各种不同模型选择方法在 CENPAMI 和 MNIST 手写体数字样本库上的选择结果，其中，贝叶斯模型选择方法中的参数 ρ 根据实验选择为 0.3。

除模型选择外，模型的参数估计是 GMM 建模方法的另一个重要问题。本章提出的模型选择判别式方法在判别学习框架下同时完成模型选择和参数估计。BIC、MDL 和 AutoClass 三种方法的模型选择和参数估计则需要分开进行，并需要采用生成式方法估计模型参数。为了公平比较 BIC、MDL 和 AutoClass 三种自动模型选择方法与本章提出的模型选择判别式方法，分别对由 BIC、MDL 和 AutoClass 三种模型选择方法获得的参数也用 SoftDS-MMP 判别方法进行调整。

表 4.1 CENPAMI 上采用四种不同模型选择方法所得最优 GMM 模型成份个数

数字类 / 方法	0	1	2	3	4	5	6	7	8	9	平均
BIC	4	4	3	4	5	4	4	5	5	4	4.2
MDL	3	3	2	2	4	3	2	4	3	2	2.8
AutoClass	4	3	4	4	4	4	3	5	5	5	4.1
Our	4	3	3	5	3	3	5	4	6	5	4.1

表 4.2 MNIST 上采用四种不同模型选择方法所得最优 GMM 模型成份个数

数字类 / 方法	0	1	2	3	4	5	6	7	8	9	平均
BIC	6	12	6	7	4	7	6	9	6	7	7.0
MDL	3	5	3	3	5	4	3	5	3	4	3.8
AutoClass	4	8	5	4	5	6	6	9	4	8	5.9
Our	5	10	4	4	6	5	4	7	9	10	6.4

综上所述，共获得七种数字分类器，分别是三种人工设定的 GMM 模型结构和四种自动获取的 GMM 模型结构。用上述七种分类器分别在手写体数字识别问题上进行封闭和开放的测试，并将其在 CENPAMI 和 MNIST 上的实验结果分别列于表 4.3 和表 4.4 中。该章采用分类器在测试集上的识别率和在训练集上的识别率的比值作为分类器泛化性能的度量标准。显然，该比值越大表示其泛化性能越好；相反，则泛化性能越差。表 4.3 和表 4.4 还列出了本章所提出的模型选择方法与采用其他模型选择方法相比所带来的错误率下降率。

表 4.3 CENPAMI 上人工设定和自动模型选择方法的误识率

模型选择方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)	泛化性能
3 个成份	0.300	1.35	51.85	0.9895
4 个成份	0.225	1.15	43.48	0.9907
5 个成份	0.200	1.15	43.48	0.9904
BIC	0.125	0.90	27.78	0.9922
MDL	0.225	1.10	40.91	0.9912
AutoClass	0.200	1.05	38.10	0.9915
Our	**0.050**	**0.65**	-	**0.9940**

表 4.4 MNIST 上人工设定和自动模型选择方法的误识率

模型选择方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)	泛化性能
6 个成份	0.31	0.80	33.75	0.9951
7 个成份	0.25	0.76	30.26	0.9949
8 个成份	0.21	0.67	20.90	0.9954
BIC	0.29	0.73	27.40	0.9956
MDL	0.37	0.79	32.91	0.9958
AutoClass	0.23	0.63	15.87	0.9960
Our	**0.19**	**0.53**	**-**	**0.9966**

如表 4.3 和表 4.4 所示，采用提出的模型选择方法所获得的识别效果不但好于人工设定的方法，而且还好于模型选择的生成式方法。在 CENPAMI 数据库上，与 BIC、MDL 和 AutoClass 模型选择方法相比，本章所提出的方法使测试集上的错误率分别降低 27.78%、40.91%和 38.10%，并且使分类器的泛化性能分别由 0.9922（BIC 方法）、0.99912（MDL 方法）和 0.9915（AutoClass 方法）提高到 0.9940（本章提出的方法）。在 MNIST 数据库上，与 BIC、MDL 和 AutoClass 模型选择方法相比，所提出的方法使测试集上的错误率分别降低 27.40%、32.91%和 15.87%，并且使分类器的泛化性能分别由 0.9956（BIC 方法）、0.9958（MDL 方法）和 0.9960（AutoClass 方法）提高到 0.9966（本章提出的方法）。

4.5.2 数字分类器比较

我们还根据已查阅文献进一步收集了一些不同的识别方法在 CENPAMI 和 MNIST 手写体数字样本库上所取得的识别结果，与本章提出的算法所取得的识别效果进行比较。

（1）CENPARMI样本库

文献[103]对目前各种手写体数字识别技术，包括特征和分类器，在 CENPAMI 手写体数字样本库上进行了研究和比较，其发布的最低误识率为 0.95%，采用的特征为八个方向的梯度特征（e-grg），分类器为基于 RBF 核的 SVM 和判别学习二次判别函数（Discriminative Learning Quadratic Discriminant Function，DLQDF）[103-104]。应

用相同的数字特征（e-grg），本章提出的分类器取得了更好的识别效果。此外，本章还收集了目前为止在 CENPAMI 样本库上所报导的较好的识别效果，并在表 4.5 中与本文所提出的方法进行比较。从表 4.5 中可以看出应用本文所提出的 GMM 模型选择和学习方法获得的分类器在手写体数字识别实验中所表现出来的性能要明显好于其它分类器。

表 4.5 CENPARMI 数据库上各种分类器所获得的误识率

分类方法	特征	测试集上误识率(%)
Modular Neural Network [146]	Class dependent features	2.15
Local Learning Framework[147]	32 direction gradient features	1.90
Neural Network[148]	Random features	1.70
Virtual SVM [42]	32 direction gradient features	1.30
SVC-rbf[104]	8 direction deslant chain code features	0.85
Our Method	**e-grg**	**0.65**

（2）MNIST样本库

文献[103]还在 MNIST 手写体数字样本库上对各种手写体数字识别技术进行了研究和比较，其发布的最低误识率为 0.42%，采用的特征为八个方向的梯度特征（e-grg），分类器为基于 RBF 核的 SVM。本章提出的分类器取得了更好的识别效果。此外，本章收集了在 MNIST 样本库上所报导的较好的识别效果，并在表 4.6 中与本章所提出的方法进行比较。从表 4.6 中可以看出应用本章所提出的 GMM 模型选择和学习方法获得的分类器在手写体数字识别实验中取得的识别率接近目前已有最好识别效果。

表 4.6 MNIST 数据库上各种分类器所获得的误识率

方法	特征	误识率(%)
Convolutional Net LeNet-1[111]	Subsampling	1.7
Polynomial SVM [149]	32 direction gradient features	1.4
Boosted LeNet4 [70]	Subsampling	0.7
Large Convolutional Net[150]	Unsup features	0.62
SVM[151]	Vision-based feature	0.59
SVMs[152]	Trainable feature	0.54
K-NN[100]	Shiftable edges	0.52
VSVM [42]	32 direction gradient features	0.44
SVC-rbf [103]	e-grg	0.42
Large Convolutional Net[101]	Trainable feature	**0.39**
Our Method	e-grg	**0.53**

4.6 小结

本章在面向贝叶斯分类器的最大最小后验伪概率软目标学习框架（Soft-MMP）下，提出了一种新的 GMM 模型判别选择方法。该章根据模型选择的要求调整了 Soft-MMP 判别学习方法的目标函数，并将其结合到基于拉普拉斯估计的贝叶斯模型选择框架下。模型选择标准是选择使 Soft-MMP 目标函数边缘积分的拉普拉斯估计值最大的 GMM 模型。利用线性搜索策略，我们可以同时获得 GMM 的最优模型结构和模型参数。

在手写体数字识别问题上对提出的 GMM 模型判别选择方法进行评估，分别在使用广泛的 CENPAMI 和 MNIST 数据库上进行实验。通过实验，提出的 GMM 模型判别选择方法无论在识别准确率还是泛化性能上都要好于经典的贝叶斯模型选择准则

(BIC)、最小描述长度模型选择准则(MDL)和 AutoClass 模型选择方法。在 CENPAMI 数据库上，提出的判别模型选择方法与固定模型结构、BIC、MDL 和 AutoClass 相比，使：1）测试集上的错误率下降 27.78% 至 51.85%；2）泛化性能提高 0.18% 至 0.45%。在 MNIST 数据库上，提出的判别模型选择方法与固定模型结构、BIC、MDL 和 AutoClass 相比，使：1）测试集上的错误率下降 15.81% 至 33.75%；2）泛化性能提高 0.06% 至 0.17%。此外，根据我们现有知识，该章提出的基于 SoftDS-MMP 模型选择判别式方法所在 CENPAMI 数据库上取得的 0.65% 误识率是目前取得的最好结果。在 MNIST 数据库上所取得的 0.53% 误识率接近目前已有最好识别结果。

第 5 章　结合目标函数梯度的进化策略优化算法

5.1 引言

面向统计模式识别的判别学习研究主要包括判别学习准则设计和判别学习准则的优化方法。目前, 关于判别学习的研究主要集中于设计最优的判别学习准则，并提出了许多相应方法，如最大互信息(Maximum Mutal Information, MMI）[8-9]、最小分类错误（Minimum Classification Error, MCE）[6-7]、最大最小后验伪概率方法（Max-Min Posterior pseudo-probability, MMP）[95-99]等。与判别学习准则相比，判别学习的优化方法还没有得到研究人员足够重视。现有判别学习方法主要采用传统的基于梯度的优化方法，如最速梯度法、共轭梯度法等[10,50-53,87-94]。但由于梯度下降方法易陷入局部最优解，因此判别学习方法有时不能取得满意的学习效果。

与传统的基于梯度的优化算法相比，进化计算不但能在高维、复杂的搜索空间中快速的找到一个较优的搜索区域，而且可以以一定的概率跳出局部最优解，找到全局最优解。最近，应用进化计算方法学习统计分类器受到该研究领域的广泛关注，每年在相关期刊和会议都有大量论文发表，如用进化计算方法优化神经网络[153-160]、支持向量机 [161-166]、Boost[167-168]、贝叶斯网络[169]等。在用进化计算方法学习统计分类器过程中，根据优化目标不同，可以分为以下四类:

（1）优化分类器参数。如优化神经网络中连接权值[153-154]、优化支持向量机中的核参数和超参数[162-163]等。

（2）优化分类器结构。由于分类器中结构变量是非连续的，不能用传统的基于梯度的方法进行优化。可用于优化非连续函数的进化计算在该方向表现出了优势。进化计算可用于优化神经网络结构变量中的神经原节点个数、神经网络层数、节点之间是否存在连接、以及信息传输函数（transfer function）[155-157]。对于支持向量机，进化计算可用于选择支持向量机中核函数的形式。Tom 和 Michael[164]提出了基于进化核的支持向量机（Genetic Kernel SVM），该方法采用进化规则优化支持向量机的核函数。他们在一些数据库上对该方法进行验证，并与常用的标准核函数进行比较，包括多项式核（Ploynomial）、径向基函数核(radial basis function, RBF)、Sigmoid 核，实验结果表明基于进化核的支持向量机性能明显好于其他常用的标准核函数。除神经网

络和支持向量机外，进化计算还被用于优化贝叶斯网络的网络结构，并在市场预测问题中表现出了有效性。

（3）输入特征优化。进化计算还可用于特征选择[170]、特征空间变换[171]、生成虚拟特征[172]。

（4）多目标优化。相对于传统优化方法，进化计算一个重要优势是可以同时优化多个目标函数，即使这些目标函数之间相互冲突。进化计算的多目标优化方法在优化分类器上表现出了明显优势，如同时最大化分类器的容量和最小化分类器的结构复杂度。对于神经网络，进化计算通常被用来同时优化神经网络的连接权值和网络结构，在最大化神经网络容量的同时控制网络复杂度[158-160]。最近，进化计算的多目标优化方法还被用于优化支持向量机，在训练错误与模型复杂度两个相互冲突的目标之间寻求最优解[165-166]。

进化计算的主要优点是可利用多点随机搜索策略跳出局部最优解，但其搜索效率通常不能达到期望目标。目前，在判别学习领域，尤其在神经网络的研究中，提出将梯度信息结合到进化计算中，加快进化计算的收敛速度。由于该联合优化方法取得了较好的实验结果，因而受到到相关研究领域的广泛关注[172-181]。梯度下降和进化计算的联合优化方法可以分为两类：1）进化计算和梯度下降分别优化神经网络的不同部分，如用进化计算优化神经网络的网络结构，用梯度下降算法优化神经网络的连接权值[172-174]；或者用进化计算和梯度下降分别优化神经网络中不同层的连接权值[175]。2）进化计算和梯度下降分别用在优化过程中的不用阶段[176]。该类联合优化方法通常是在梯度下降陷入局部最优解停滞时，改用进化计算进行优化，从而使其跳出局部最优解；或者先用进化计算找到最优搜索区域后，改用梯度下降在该搜索区域内寻求最优解，加快优化速度。

但在上述两类联合策略中，进化计算和梯度下降本质上却是独立的，两类算法只是实现协同优化，却没有实现内在互补。本章将在第三章提出的基于软目标的最大最小后验伪概率判别学习框架下（SoftDS-MMP），探讨一种面向贝叶斯分类器的进化策略与梯度下降内在结合方法，实现梯度下降与进化策略的互补。在前面章节中，SoftDS-MMP 已经应用于字符识别问题，并采用梯度下降算法进行优化。本章将梯度优化算法和基于 Cholesky 分解的协方差矩阵自适应进化策略(Covariance Matrix Adaptation Evolution Strategy based on Cholesky Factorization, Cholesky-CMA-ES)相结合，提出一种基于目标函数梯度的进化策略优化算法，提高优化效果和优化效率，为

了便于叙述，本文将该方法简称为联合优化算法。Cholesky-CMA-ES 是一种新的进化策略，该方法利用个体的进化路径信息自适应更新协方差矩阵和全局步长因子[182-183]。

本章提出的联合优化算法利用目标函数的梯度信息调整 Cholesky-CMA-ES 中三个重要参数，即加权均值、协方差矩阵和全局步长，提高优化效果和效率。该联合优化算法的主要思想是对每代个体的加权均值用梯度下降算法进行调整，并根据调整后的均值调整协方差矩阵和全局步长。该联合优化算法还在训练过程中动态调整梯度下降与 Cholesky-CMA-ES 所占的比重。在训练初期，为了快速找到最优搜索区域，Cholesky-CMA-ES 在联合优化算法中占主导地位。然后，逐渐增加梯度下降算法在联合优化算法中所占的比重，以加强联合优化算法的局部探索能力。通过该联合优化算法可实现 Cholesky-CMA-ES 和梯度下降算法的互补。一方面，利用 Cholesky-CMA-ES 的多点随机搜索策略可以以一定的概率跳出局部最优解；另一方面，利用目标函数的梯度信息可以加快优化算法的收敛速度和增强局部探索性能。

5.2 相关工作

已有一些文献提出了 CMA-ES 和梯度下降的联合优化方法[178-181]。Auger 等人[178]提出一种面向进化策略协方差矩阵自适应更新的二阶算法（LS-CMA-ES），该方法通过学习适应函数（fitness function）的曲率更新协方差矩阵。LS-CMA-ES 对于椭圆函数，采用最小二乘法估计目标函数的梯度方向和 Hessian 矩阵；对于非椭圆函数，由于用最小二乘法无法获得准确的估计，LS-CMA-ES 改用 CMA-ES 优化目标函数。在优化过程中，LS-CMA-ES 通过一个动态标准评估对目标函数逼近的准确程度，确定当前函数是否是椭圆函数，以选择相映的优化方法。对于椭圆函数，LS-CMA-ES 的优化效果明显好于 CMA-ES；对于非椭圆函数，该方法优化效果接近于 CMA-ES。LS-CMA-ES 的局限性是计算代价高，很难用于大规模模式识别问题中。LS-CMA-ES 优化算法是在优化的过程中动态选择用梯度信息还是 CMA-ES 优化目标函数，没有对两种算法的内在优化机制进行改进，实现两种算法的内在结合与互补。

Salmon[179]也提出了一种进化策略与梯度下降相结合的优化算法，并将该算法命名为“进化梯度搜索（Evolutionary Gradient Search , EGS）”。在每代进化中，进化梯度搜索算法利用高斯函数对参数空间中一指定的点进行变异操作，生成一些采样点，并根据采样点所提供的信息估计目标函数的梯度方向，沿最速梯度的反方向更新目标

函数中的参数，并自适应调整进化策略的步长因子。在多数测试函数上，EGS 取得了理想的优化效果。对于特征值差异较大或者陡峭的函数，EGS 的优化效果不十分理想。对于多峰函数，EGS 优化效果与初试点选择相关。Arnold 和 Salomon 在文献[180]中对 EGS 优化方法进行了改进，通过尺度变换方法增加 EGS 的鲁棒性。他们将改进后的 EGS 优化算法与 CMA-ES 相结合，命名为 CMA-EGS 优化算法。CMA-EGS 首先利用 CMA-AS 对参数空间中所选择的初始点进行变异操作，生成一些采样点，再用 EGS 算法根据采样点所提供的信息估计目标函数的梯度方向。利用该梯度方向调整 CMA-ES 中高斯突变函数的均值和所构建的搜索路径。根据调整后的搜索路径更新进化策略的协方差矩阵和步长因子。他们在球函数上对 CMA-EGS 和 CMA-ES 的优化效果进行比较，CMA-EGS 的优化效果好于 CMA-ES。

与 CMA-EGS 方法相似，Wierstra 等人[181]提出了一种自然进化策略（Natural Evolution Strategies，NES）用于优化“黑箱”函数，即优化函数形式未知的目标。该方法利用蒙特卡罗算法估计“黑箱”函数的自然梯度方向，并用该梯度调整父辈个体中的参数和进化策略中的突变矩阵。他们用自然进化策略优化一些基准函数。对于单峰函数，自然进化策略的优化效果接近 CMA-ES 算法。对于多峰函数，自然进化策略的优化效果要好于 CMA-ES 算法。

前面所介绍的联合优化算法均是凭借采样点所提供的信息估计目标函数的梯度方向，而本章所提出的联合优化算法则是在 CMA-ES 进化机制下，计算目标函数的精确梯度方向，用该梯度方向提高联合优化算法的收敛速度。

5.3 Cholesky-CMA-ES

本章将协方差矩阵自适应进化策略（Covariance Matrix Adaptation Evolution Strategy，CMA-ES）[184-186]引入到基于 SoftDS-MMP 的参数估计和模型选择方法中。与传统的进化策略不同，CMA-ES 利用进化过程中成功突变所传递的信息自动更新进化策略中的参数。大量实验表明，CMA-ES 不但提高了进化策略的收敛速度，而且减少了对个体数量的需求。为了用正态分布函数在参数空间中采样，CMA-ES 需要对协方差矩阵进行奇异值分解。令 r 是参数空间的维数，则协方差矩阵奇异值分解所需要的计算复杂度是 $O\left(r^3\right)$。为了提高 CMA-ES 在高维空间的搜索效率，Thorsten Suttorp 等人[182]提出用 Cholesky 分解代替协方差矩阵的奇异值分解，用更新 Cholesky 因子代替更新协方差矩阵，从而避免了进化过程中计算复杂度较高的奇异值分解操作，使

计算代价由 $O\left(r^3\right)$ 下降到 $O\left(r^2\right)$。为论述方便，本文将基于 Cholesky 分解的协方差矩阵自适应进化策略简称为 Cholesky-CMA-ES。

5.3.1 协方差矩阵自适应进化策略

协方差矩阵自适应进化策略（CMA-ES）本质上是一种 (μ,λ) 进化策略，每次生成 λ 个子代个体，并从中选择 μ 个最优个体作为下次优化的父代种群。种群中的每个个体代表目标函数的一个解向量。与传统的进化策略相比，CMA-ES 生成子代个体包括两个操作，即所选个体的加权重组和高斯突变。假设 $\boldsymbol{y}_k^{(g)} \in \mathbb{R}^s$ 是第 g 代种群中第 k 个个体， CMA-ES 生成子代个体的函数为

$$\boldsymbol{y}_k^{(g+1)} = \langle \boldsymbol{y} \rangle_w^{(g)} + \sigma^{(g)} v_k^{(g)}, \tag{5-1}$$

$\sigma^{(g)}$ 是全局步长因子，$\langle \boldsymbol{y} \rangle_w^{(g)}$ 是 μ 个最优子代个体的加权均值，即

$$\langle \boldsymbol{y} \rangle_w^g = \sum_{i=1}^{\mu} u_i \boldsymbol{y}_{i:\lambda}^{(g)}, \tag{5-2}$$

其中$v_k^{(g)}$ 是由均值为零且方差为 $\boldsymbol{C}_k^{(g)} \in \mathbb{R}^{n\times n}$ 的高斯函数生成的 r 维随机向量，即

$$v_k^{(g)} \sim \mathcal{N}\left(0, \boldsymbol{C}_k^{(g)}\right). \tag{5-3}$$

5.3.2 基于 Cholesky 分解的协方差矩阵更新

为了避免 CMA-ES 算法中的协方差奇异值分解，Cholesky-CMA-ES 将高斯突变函数中的对称正定协方差矩阵 $\boldsymbol{C}_i^{(g)}$ 分解为 Cholesky 因子，即

$$\boldsymbol{C}_i^{(g)} = \boldsymbol{A}_i^{(g)} \left(\boldsymbol{A}_i^{(g)}\right)^T, \tag{5-4}$$

其中 $\boldsymbol{A}_i^{(g)} \in \mathbb{R}^{n\times n}$ 是协方差矩阵 $\boldsymbol{C}_i^{(g)}$ 的 Cholesky 分解因子。式(5-1)中的采样点 $\boldsymbol{v}_k^{(g)}$ 可以用下式生成

$$\boldsymbol{v}_k^{(g)} = \boldsymbol{A}_i^{(g)} \boldsymbol{z}_i^{(g)}, \tag{5-5}$$

其中 $\boldsymbol{z}_i^{(g)} \sim \mathcal{N}(0, \boldsymbol{I}) \in \mathbb{R}^n$ 是由 $(0, \boldsymbol{I})$ 正态分布所生成的随机向量，$\boldsymbol{I}$ 是单位矩阵。把式（5-5）带入式（5-1）中，得到

$$\boldsymbol{y}_k^{(g+1)} = \langle \boldsymbol{y} \rangle_w^{(g)} + \sigma^{(g)} \boldsymbol{A}_i^{(g)} \boldsymbol{z}_i^{(g)}. \tag{5-6}$$

Cholesky-CMA-ES 算法在每代进化过程中并不直接更新 $\boldsymbol{C}^{(g)}$， 而是更新 Cholesky 因子 $\boldsymbol{A}^{(g)}$，从而避免了协方差矩阵的奇异值分解操作。Cholesky 因子 $\boldsymbol{A}^{(g)}$ 的迭代更新公式为

$$\boldsymbol{A}^{(g+1)}=\sqrt{1-c_{cov}}\boldsymbol{A}^{(g)}+\frac{\sqrt{1-c_{cov}}}{\|v\|^2}\left(\sqrt{1+\frac{c_{cov}}{1-c_{cov}}\|v\|^2}-1\right)\boldsymbol{p}_c\boldsymbol{v}^T, \quad (5\text{-}7)$$

其中 $c_{cov}\in\mathbb{R}^+$ 是学习速度控制因子，$\boldsymbol{p}_c^{(g)}$ 是根据进化过程中成功的进化步骤所构造的进化路径，定义为

$$\boldsymbol{p}_c^{(g+1)}=(1-c_c)\,\boldsymbol{p}_c^{(g)}+\sqrt{c_c\,(2-c_c)\,\mu_{eff}}\boldsymbol{A}\boldsymbol{z}_w^{(g)}, \quad (5\text{-}8)$$

其中 $c_c\in\mathbb{R}^+$ 用于均衡先前最优进化路径和当前最优路径，$\mu_{eff}=\left(\sum_{i=1}^{\mu}w_i^2\right)^{-1}$是选择质量方差，$\boldsymbol{z}_w^{(g)}=\sum_{i=1}^{\mu}w_i\boldsymbol{z}_{i:\mu}^{(g)}$ 是可产生 μ 个最优子代个体的由标准正态分布函数生成的μ 个随机变量加权均值。式（5-7）中，向量 $\boldsymbol{v}$ 是一个假定服从 $\mathcal{N}(0,\boldsymbol{I})$ 正态分布的虚拟向量，计算方法为

$$\boldsymbol{v}=\boldsymbol{A}^{-1}\cdot\boldsymbol{p}_c. \quad (5\text{-}9)$$

$\boldsymbol{A}^{-1}$ 在 Cholesky-CMA-ES 优化算法中不需要用逆计算求得，而是直接用迭代公式进行更新，更新公式为

$$\boldsymbol{A}^{(g+1)-1}=\frac{1}{\sqrt{1-c_{cov}}}\boldsymbol{A}^{(g)-1}-\frac{1}{\sqrt{1-c_{cov}}\|\boldsymbol{v}\|^2}\left(1-\frac{1}{\sqrt{1+\frac{c_{cov}}{1-c_{cov}}\|\boldsymbol{v}\|^2}}\right)\boldsymbol{v}\left[\boldsymbol{v}^T\boldsymbol{A}^{-1}\right]. \quad (5\text{-}10)$$

5.3.3 全局步长自适应更新

Cholesky-CMA-ES 的全局步长更新方法与 CMA-ES 方法类似，也是利用共轭进化路径。Cholesky-CMA-ES 方法中的共轭进化路径描述为

$$\boldsymbol{p}_\sigma^{(g+1)}=(1-c_\sigma)\,\boldsymbol{p}_\sigma^{(g)}+\sqrt{c_\sigma\,(2-c_\sigma)\,\mu_{eff}}\boldsymbol{z}_w^{(g)}, \quad (5\text{-}11)$$

其中 $c_\sigma<1$ 是用于控制累加步长的学习因子。与式（5-8）中定义的进化路径相比，该进化路径的期望长度与方向无关，因此称为共轭进化路径。Cholesky-CMA-ES 通过比较共轭进化路径长度与正态随机分布的期望长度，判断增加还是减少当前进化策略中的全局步长因子 $\sigma^{(g)}$。如果当前共轭路径的长度小于随机正态分布的期望长度，则缩短当前进化策略中的全局步长因子；否则，增加当前全局步长因子。Cholesky-CM

A-ES 的全局步长因子自适应更新函数描述为

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma^{(g+1)}\|}{\chi} - 1\right)\right), \tag{5-12}$$

其中参数 d_σ 用于减弱共轭路径长度变化对全局步长的影响，$\chi \approx \sqrt{s}(1 - \frac{1}{4s} + \frac{1}{21s^2})$ 是符合正态分布 $N(0, \boldsymbol{I})$ 的随机变量的期望长度。

5.3.4 Cholesky-CMA-ES 算法

算法 5.1 给出了 Cholesky-CMA-ES 优化算法的优化流程。

算法 5.1 Cholesky-CMA-ES 优化算法

Initial

步骤 1. 随机生成 λ 个变量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda$，构建成初始种群；

步骤 2. 初始化全局步长 $\sigma \in \mathbb{R}^+$；

步骤 3. 分别初始化 Cholesky 因子 $\boldsymbol{A}$ 和 $\boldsymbol{A}^{-1}$ 为单位矩阵；

步骤 4. 分别初始化进化路径 $\boldsymbol{p}_c$ 和共轭进化路径 $\boldsymbol{p}_\sigma$ 为零向量；

Repeat

步骤 5. 根据目标函数值从采样点 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda$ 中选择 μ 个最优点，利用式（5-2）重组已选择的 μ 个采样点，计算加权均值 $\boldsymbol{m}^{(g)}$；

步骤 6. 根据式（5-6）生成新的 λ 个采样点 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda$；

步骤 7. 基于式（5-8）度量 Cholesky-CMA-ES 的进化路径长度；

步骤 8. 分别用式（5-7）和式（5-10）更新 Cholesky 因子 $\boldsymbol{A}$ 及其逆 $\boldsymbol{A}^{-1}$。

步骤 9. 利用式（5-11）度量共轭进化路径 $\boldsymbol{p}_\sigma$长度；

步骤 10. 根据式（5-12）调整全局步长 σ；

Untill 满足收敛条件或者达到最大迭代次数。令 ε 是一个预先设置的极小阈值，则其收敛条件定义为

$$\max\{F(\boldsymbol{y}_i)\} - \min\{F(\boldsymbol{y}_i)\} \leqslant \varepsilon, i = 1 \cdots \lambda。$$

即采样点对应的最大目标函数值与最小目标函数值之差小于预设的阈值。

5.4 Cholesky-CMA-ES 与梯度优化的结合

进化计算主要优势是利用其多点随机搜索策略可以跳出局部最小值，但进化计算的优化效率通常不能满足要求。梯度向量的一个重要特点是对搜索空间内任意一点，始终指向最快增加目标函数值的方向。所以，本章利用目标函数的梯度信息，加快进化策略的收敛速度。该方法实现了进化策略与梯度下降算法的互补。一方面，利用进化策略的多点随机搜索可以以一定概率跳出局部最小值；另一方面，利用目标函数的梯度信息可以加快进化策略的收敛速度。基于该思想，本章提出了一种面向 SofDS-MMP 的进化策略与梯度下降联合优化算法。

如前所述，在面向 Soft-MMP 的梯度下降算法中，SoftDS-MMP 中参数迭代更新公式为

$$\boldsymbol{\Lambda}_{t+1} = \boldsymbol{\Lambda}_t - \alpha_t \nabla F(\boldsymbol{\Lambda}_t), \tag{5-13}$$

其中 $\boldsymbol{\Lambda_t}$ 和 α_t 分别表示参数集合和第 t 次迭代时的步长因子，$\nabla F(\boldsymbol{\Lambda_t})$ 是目标函数关于参数集合 $\boldsymbol{\Lambda_t}$ 中参数的偏导数。

根据式（5-1）和式（5-3），Cholesky-CMA-ES的优化效果和效率主要取决于三个重要变量，即：（1）所选最优采样点的加权均值 $\langle \boldsymbol{y} \rangle_w^{(g)} \in \mathbb{R}^n$，（2）协方差矩阵 $\boldsymbol{C}^{(g)} \in \mathbb{R}^{n \times n}$，（3）全局步长因子 $\sigma^{(g)} \in \mathbb{R}^+$。Cholesky-CMA-ES优化算法仅利用进化信息对这三个变量进行更新，但本章所提出的联合优化方法将结合进化信息和梯度信息一起更新上述三个变量。

5.4.1 加权均值更新

$\langle \boldsymbol{y} \rangle_w^{(g)}$ 是从采样点 $\boldsymbol{x}_1^{(g)}, \cdots, \boldsymbol{x}_\lambda^{(g)}$ 中选择的 μ 个最优采样点的加权均值，并被作为中心产生下一代采样点。$\langle \boldsymbol{y} \rangle_w^{(g)}$ 作为择优采样点的重组，引导新的采样点进入一个更加有效的搜索区域。显然，如果能够调整重组后的加权均值 $\langle \boldsymbol{y} \rangle_w^{(g)}$，使其越靠近最优解，则以 $\langle \boldsymbol{y} \rangle_w^{(g)}$ 为中心生成的新的采样点越有效，从而加快优化算法的收敛速度。图5.1是用进化策略优化目标函数 $f(x, y) = x^2 + y^2$ 的图示，图中的虚线是目标函数的等值线，绿色的点和蓝色的点分别是由相同协方差矩阵（单位矩阵）不同均值的高斯函数生成的采样点，红色的点分别代表两个高斯函数的均值，红色的星代表目标函数最优解。从图5.1中可以看出生成蓝色采样点的高斯函数的均值相对于生成绿色采样点的高斯函数的均值更靠近最优解。因此，以其为中心所生成的蓝色采样点更靠近最优解。

由于目标函数的梯度总是指向使目标函数值增加最快的方向，所以本章利用目标函数的梯度信息调整加权均值，使其更靠近最优解，加快算法的收敛速度。具体是在Cholesky-CMA-ES的每代进化中，沿目标函数梯度的反方向调整加权均值，即

$$\langle \boldsymbol{q} \rangle_w^{(g)} = \langle \boldsymbol{y} \rangle_w^{(g)} - \alpha \nabla F\left(\langle \boldsymbol{y} \rangle_w^{(g)} \right), \tag{5-14}$$

其中 α 是调整的步长因子。步长因子 α 的大小将影响梯度下降算法和进化策略在该联合优化算法中所占的比值关系。显然，步长因子 α 越大，则联合优化算法越倾向于梯度下降；相反，步长因子 α 越小，则联合优化算法越倾向于Cholesky-CMA-ES。本文采用Armijo-Goldstein不精确搜索方法[187]确定联合优化算法中梯度下降的最优步长因子。算法5.2中给出了基于Armijo-Goldstein不精确搜索的步长因子α 选择流程。

算法 5.2 基于 Armijo-Goldstein 不确定搜索步长因子选择算法

步 1：选定初始点 $\alpha_0 = 1$，给出 $\rho \in \left(0, \frac{1}{2}\right)$，$l > 1$，令 $b_0 = 0$；$c_0 = +\infty$；$k = 0$

步 2：令 $\varphi(\alpha) = F(\Lambda - \alpha g)$，如果

$$\varphi(\alpha_k) \leqslant \varphi(0) - \rho \alpha_k \|g\|^2.$$

转步 3，否则，令 $b_{k+1} = b$，$c_{k+1} = \alpha_k$，转步 4;

步 3：如果

$$\varphi(\alpha_k) \geqslant \varphi(0) - (1-\rho)\alpha_k \|g\|^2.$$

停止迭代，输出 α_k；否则，令

$$b_{k+1} = \alpha_k, \quad c_{k+1} = c_k.$$

若 $c_{k+1} < +\infty$，转步 4；否则，令 $\alpha_{k+1} = l\alpha_{k+1}$，$k = k+1$，转步 2;

步 4：取

$$\alpha_{k+1} = \frac{b_{k+1} + c_{k+1}}{2}$$

令 $k = k+1$，转步 2;

基于式（5-14）用梯度调整加权均值 $\langle \boldsymbol{y} \rangle_w^{(g)}$ 后，以 $\langle \boldsymbol{q} \rangle_w^{(g)}$ 作为高斯函数的均值，生成新的采样点 $\boldsymbol{y}^{(g+1)}$。

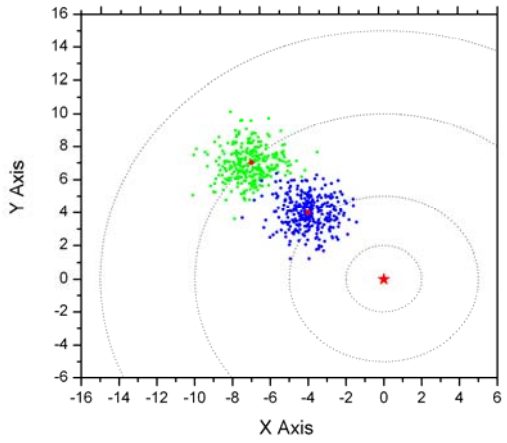


图 5.1 不同均值对进化策略优化效果影响图示

5.4.2 协方差矩阵更新

Cholesky-CMA-ES方法中的协方差矩阵决定着以均值为中心的各采样点出现的概率。在Cholesky-CMA-ES中，将采样点与加权均值之间的距离 $\boldsymbol{y}_k^{(g+1)} - \langle \boldsymbol{y} \rangle_w^{(g)}$ 定义为采样步长（Sample Step）。Cholesky-CMA-ES方法中的协方差矩阵控制着不同采样步长在下次迭代中出现的概率。Cholesky-CMA-ES更新协方差矩阵主要包括两个步骤：（1）估计下代能够带来最大目标函数值的最优采样步长；（2）利用rank-one方法调整Cholesky-CMA-ES的协方差矩阵，以增加该采样步长在下代优化中出现的概率。显然，对最优采样步长估计越准确，则算法的收敛速度越快。

图5.2是根据不同采样步长调整协方差矩阵后，Cholesky-CMA-ES生成采样点的图示，其优化目标函数为$f(x,y) = x^2 + y^2$。图5.2中的红色箭头代表估计的当前最优采样步长，并根据该步长调整协方差矩阵，绿色和蓝色的点分别是用图中所示最优采样步长更新协方差矩阵后所得的新的采样点。从图5.2中可以看出，右图中估计的最优采样步长比左图准确，用该采样步长调整后的协方差矩阵所生成的随机采样点更接近最优解。

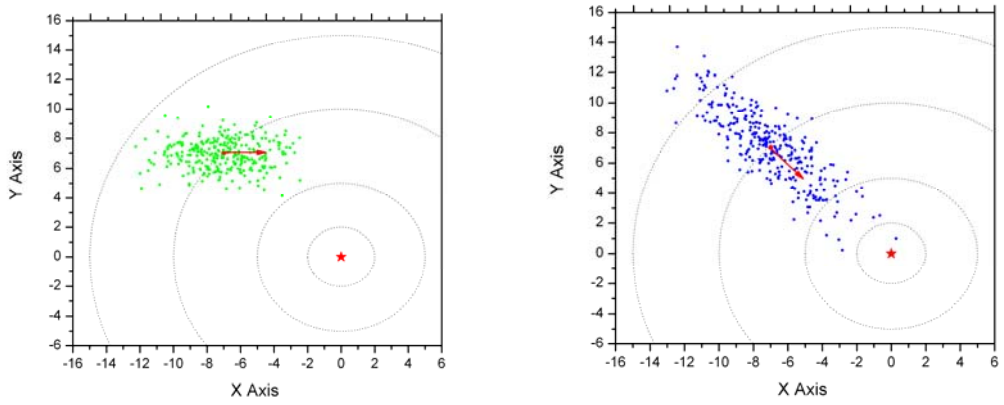


图 5.2 基于不同采样步长调整协方差矩阵生成采样点图示

Cholesky-CMA-ES由于仅利用采样点的进化信息估计最优采样步长，当采样点的数目较少，或者搜索空间过于复杂时，Cholesky-CMA-ES很难准确估计出最优采样步长。本章提出的结合目标函数梯度的进化策略优化算法不仅利用采样点的进化信息，还利用目标函数的梯度信息来估计下次进化的最优采样步长。该方法用采样点所估算出的最优采样步长 $\langle \boldsymbol{y} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}$ 和梯度下降步长 $\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{y} \rangle_w^{(g+1)}$ 构建联合步长，再利用Cholesky-CMA-ES的进化路径方法将构建的联合步长与先前累加的最优步长相结合（如式（5-8）），根据结合后的进化路径更新协方差矩阵。图5.3是用由采样点所估算出的最优采样步长和梯度下降步长构建联合步长的示意图，其中红色点代表Cholesky-CMA-ES中的第 g 代高斯函数的均值 $\langle \boldsymbol{q} \rangle_w^{(g)}$，蓝色点代表第 $g+1$代中被选择的采样点加权重组后的均值 $\langle \boldsymbol{y} \rangle_w^{(g+1)}$，绿色点代表利用梯度下降算法调整后的加权均值 $\langle \boldsymbol{q} \rangle_w^{(g+1)}$。蓝色箭头表示根据采样点计算出的最优采样步长 $\langle \boldsymbol{y} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}$，绿色箭头表示梯度下降步长 $\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{y} \rangle_w^{(g+1)}$，红色箭头即为联合步长。为了消除Cholesky-CMA-ES中全局步长 σ 的影响，最终采用的联合步长形式为 $\frac{\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}}{\sigma}$，并用该联合步长构建进化路径为

$$\boldsymbol{p}_c^{(g+1)} = (1-c_c)\,\boldsymbol{p}_c^{(g)} + \sqrt{c_c\,(2-c_c)\,\mu_{eff}}\left(\frac{\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}}{\sigma}\right). \tag{5-15}$$

用联合步长更新协方差矩阵有两个优点：第一，联合步长利用其所包含的随机信息可以跳出局部最优解；第二，通过引入梯度信息可以减少仅利用采样点所估计的最优采样步长与真实最优采样步长之间的差异，使对最优采样步长的估计更为准确，提高算法的收敛速度。

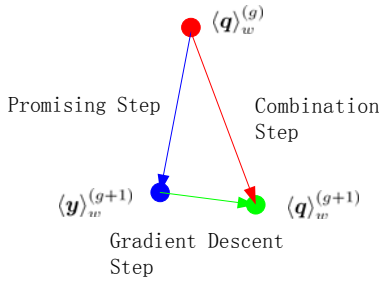


图 5.3 构建联合步长示意图

5.4.3 全局步长控制

Cholesky-CMA-ES中的全局步长因子 $\sigma \in \mathbb{R}^+$ 可根据被优化目标函数的特点控制全局优化的尺度。Cholesky-CMA-ES利用累加长度控制技术（Cumulative step size adaptation）自适应调整全局步长，通过比较共轭进化路径长度与正态分布的期望长度，决定增加或者减少当前全局步长因子。Cholesky-CMA-ES的共轭进化路径构建如式（5-11）所示。如前所述，本章提出的联合优化算法在每代优化过程中用梯度下降将加权均值从 $\langle \boldsymbol{y} \rangle_w$ 调整到 $\langle \boldsymbol{q} \rangle_w$，所以要用调整后的加权均值 $\langle \boldsymbol{q} \rangle_w$ 重新构造共轭进化路径，调整全局步长。

由于该联合优化算法利用向量 $\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}$ 更新协方差矩阵，可以假设 $\langle \boldsymbol{q} \rangle_w$ 不是利用梯度下降方法调整得到，而是由当前分布所生成的采样点经选择和重组后所得的加权均值，即

$$\langle \boldsymbol{q} \rangle_w^{(g+1)} = \langle \boldsymbol{q} \rangle_w^{(g)} + \sigma^{(g)} \boldsymbol{A}_i^{(g)} \langle \boldsymbol{z} \rangle_w, \tag{5-16}$$

其中 $\langle \boldsymbol{z} \rangle_w = \sum_{i=1}^{\mu} w_i \boldsymbol{z}_{i:\lambda}$ 是可产生 μ 个最优子代个体的由标准正态分布函数生成的随机变量的加权均值。由式（5-16）可得

$$\langle \boldsymbol{z} \rangle_w = \frac{\langle \boldsymbol{q} \rangle_w^{(g+1)} - \langle \boldsymbol{q} \rangle_w^{(g)}}{\sigma} \boldsymbol{A}^{-1}. \tag{5-17}$$

式（5-17）中的 $\boldsymbol{A}^{-1}$ 可由式（5-10）通过迭代方法获得。将式（5-17）所得的加权向量 $\langle \boldsymbol{z} \rangle_w$ 带入式（5-11）中，重新构造共轭进化路径，并根据式（5-12）调整联合优化算法中 Cholesky-CMA-ES 的全局步长因子σ。

5.5 Cholesky-CMA-ES 与梯度优化的均衡控制

在联合优化算法中，合理调整Cholesky-CMA-ES和梯度优化两部分所占的比例十分重要。进化计算一个重要优点是其全局搜索能力比较强，利用多点随机搜索策略可以在复杂、高维的参数空间中快速找到最优搜索区域。但是进化计算的局部探测能力却稍弱。相反，基于梯度的优化算法是一类常用的局部搜索方法，局部探测能力较强。联合优化算法动态调整Cholesky-CMA-ES和梯度下降两种算法所占的比例，使该联合优化算法可以同时获得较好的全局搜索能力和局部探测能力。联合优化算法通过控制每代进化中梯度下降算法的迭代次数调整两种优化算法在联合优化算法中所占的比例关系。显然，在每代的进化中，梯度下降运行的次数越多，则联合优化算法越倾向于梯度下降部分；否则，该联合优化算法倾向于Cholesky-CMA-ES部分。

为了在能够在搜索空间中快速锁定一个最优搜索区域，在优化初始阶段，联合优化算由Cholesky-CMA-ES主导其优化过程。然后，逐步由梯度优化取代Cholesky-CMA-ES，主导该联合优化算法，加强优化算法的局部探测能力，使其能够在一个较小的搜索区域内快速的找到最优解。在优化的初始阶段，联合优化算法的每代进化中只用梯度下降算法对参数进行一次迭代更新。随后，逐步增加每代进化中梯度下降迭代次数，从而使联合优化算法由倾向于Cholesky-CMA-ES转变为倾向梯度下降。联合优化算法的性能也由初始的以全局搜索能力为主，逐渐转变为以局部探测能力为主。令 N 代表联合优化算法每代进化中用梯度下降对参数进行迭代更新的次数，则联合优化算法对Cholesky-CMA-ES和梯度下降两部分控制策略描述为

$$N = \left\lceil \frac{g}{T} \right\rceil, \tag{5-18}$$

其中 g 是联合优化算法进化的代数，T 是预先设置的阈值，其控制着在联合优化算法每代进化中目标函数梯度信息的增加程度。随着联合优化算法进化代数 g 的增加，在每代梯度下降的迭代次数 N 也在增加，使联合优化算法的重点由最初的Cholesky-CMA-ES逐步转移到梯度下降算法。

5.6 联合优化算法在 SoftDS-MMP 中的应用

将本章所提出的联合优化算法用于优化SoftDS-MMP，并将其流程列于算法5.3中。SoftDS-MMP的全部学习过程即是分别对每一模式类别执行算法5.3中的所列操作。

算法 5.3 基于联合优化算法的 SoftDS-MMP 判别选择与学习方法

Initial

对给定的 SoftDS-MMP 初始参数集合，按固定顺序排列其参数，形成初始基因序列。

Repeat

步骤 1. 根据式（5-6）生成 λ 个子代个体；

步骤 2. 计算子代个体中每个个体的目标函数值；

步骤 3. 依据每个个体的目标函数值选择 μ 个最优个体；

步骤 4. 根据式（5-2）计算所选个体的加权均值 $\langle \boldsymbol{y} \rangle_k^{(g)}$；

步骤 5. 利用梯度下降算法，根据式（5-14）调整加权均值；

步骤 6. 根据式（5-15），计算联合优化算法的进化路径长度；

步骤 7. 分别用式（5-7）和式（5-10）更新 Cholesky 因子 $\boldsymbol{A}$ 及其逆 $\boldsymbol{A}^{-1}$；

步骤 8. 根据式（5-17）计算加权向量 $\langle \boldsymbol{z}_k \rangle$；

步骤 9. 用求出的加权向量 $\langle \boldsymbol{z}_k \rangle$，由式（5-11）计算共轭进化路径 $\boldsymbol{p}_\sigma$的路径长度；

步骤 10. 用所得的共轭进化路径 $\boldsymbol{p}_\sigma$的路径长度，利用式（5-12）调整全局步长因子 σ

步骤 11. 根据式（5-18）更新每代进化过程中的梯度下降算法的迭代次数。

Until 学习过程收敛或者达到最大迭代次数。令 ε 是一个预先设置的极小值，则其收敛条件定义为

$$\max\{F(\boldsymbol{y}_i)\} - \min\{F(\boldsymbol{y}_i)\} \leqslant \varepsilon, i = 1 \cdots \lambda。$$

即采样点对应的最大目标函数值与最小目标函数值之差小于预设阈值。

5.7 实验结果

为了评估本章所提出的联合优化算法性能,将该算法用于手写体数字识别问题中,并在CENPAMI和MNIST手写体数字样本库上进行实验，数字特征提取和建模过程如本文第 2 章所述。

在本章的手写体数字识别实验中，分别用第4章中提出的模型选择判别式方法和AutoClass两种方法学习数字模型结构。

对用AutoClass方法学习模型结构的数字分类器，其训练过程包括三个阶段：

第一阶段，用AutoClass方法[126]估计每个数字类GMM模型的成份个数。

第二阶段，在每个数字类的正样本上用期望最大化方法(Expectation Maximization，EM)获得相应GMM模型中参数的最大似然估计(Maximum Likelihood Estimation，MLE)。根据实验，设置后验伪概率中的参数 $\kappa = 10$ 和 $\beta = 0.01$，设置SoftDS-MMP中的参数为 $\omega = 0.01$。

第三阶段，在SoftDS-MMP学习准则下，用三种不同优化算法的学习分类器中的未知参数，三种优化方法分别是梯度下降优化算法、Cholesky-CMA-ES优化算法和本章所提出结合目标函数梯度信息的优化策略，从而得到三种不同的数字识别分类器。图5.4是用三种不同优化算法优化手写体数字分类器实验的流程框图。

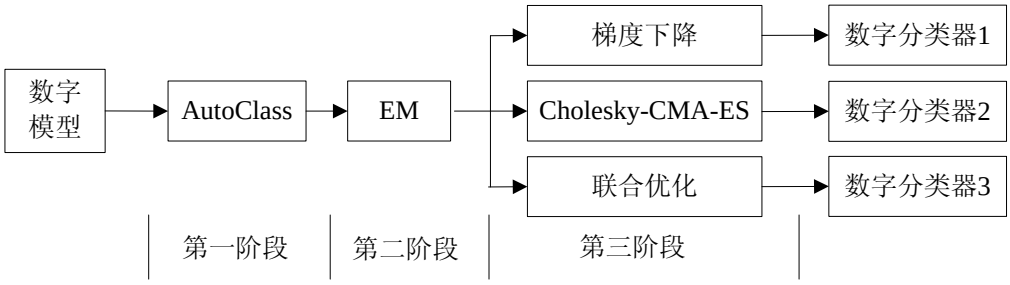


图 5.4 是用三种不同优化算法优化后验伪概率数字分类器实验流程框图

由于模型选择的判别式方法是求取目标函数的极大值，因此在联合优化策略中用梯度上升代替前面所阐述的梯度下降算法。梯度上升和Cholesky-CMA-ES的联合优化方法与在本章中论述的梯度下降和Cholesky-CMA-ES的联合优化方法相同。在基于SoftDS-MMP的判别模型选择框下，用三种不同的优化算法同时优化数字模型的结构和参数。三种优化方法分别是梯度上升优化算法，Cholesky-CMA-ES优化算法和结合目标函数梯度信息的优化策略，获得三种手写体数字分类器。

对上述两种模型选择方法中的三种优化算法的最大迭代次数均设置为50000次。其中，Cholesky-CMA-ES中的参数设置为文献[182]所给出的默认值，并将其列于表5.1中。对Cholesky-CMA-ES和本章所提出的联合优化算法，分别进行10次实验，并取最

优识别效果。最终，得到六种不同的数字识别分类器。应用六种分类器分别对手写体数字识别问题进行开放和封闭测试。

表 5.1　Cholesky-CMA-ES 参数设置

重组和突变操作中所用参数:

$$\lambda = 4 + \lfloor 3\ln(s) \rfloor,\ \mu = \lceil \lambda/2 \rceil,\ u_i = u_i' / \textstyle\sum_{i=1}^{\mu} u_i',\ u_i' = \ln(\mu+1) - \ln i,\ \boldsymbol{C}^0 = \boldsymbol{I}$$

协方差矩阵自适应更新时所用参数:

$$\boldsymbol{p}_c^{(0)} = 0,\ c_c = \frac{4}{s+4},\ \mu_{cov} = \mu_{eff},\ c_{cov} = \frac{2}{\left(s+\sqrt{2}\right)^2}$$

全局步长自适应更新时所用参数:

$$d_\sigma = 1 + 2 \cdot \max\left(0, \sqrt{\frac{\mu_{eff}-1}{s+1}} - 1\right) + c_\sigma,\ c_\sigma = \frac{\sqrt{\mu_{eff}}}{\sqrt{s}+\mu_{eff}}$$

5.7.1 优化方法效果比较

表5.2和表5.3中分别列出了用AutoClass所学得的GMM模型结构在CENPAMI和MNIST手写体数字样本集上三种不同优化算法的实验结果。

表 5.2 用 AutoClass 所得模型结构在 CENPAMI 上三种优化算法取得误识率比较

优化方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)
梯度下降	0.200	1.05	28.57
Cholesky-CMA-ES	0.120	0.80	12.50
联合优化算法	**0.075**	**0.70**	-

表 5.3 用 AutoClass 所得模型结构在 MNIST 上三种优化算法取得误识率比较

优化方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)
梯度下降	0.230	0.63	22.22
Cholesky-CMA-ES	0.201	0.55	10.91
联合优化算法	**0.175**	**0.49**	-

表5.4和表5.5中分别列出了用模型选择的判别式方法所学得的GMM模型结构在CENPAMI和MNIST手写体数字样本库上三种不同优化算法的实验结果。

表 5.4 用判别模型选择方法所得模型结构在 CENPAMI 上三种优化算法取得错误率比较

优化方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)
梯度上升	0.180	0.85	47.01
Cholesky-CMA-ES	0.115	0.70	35.71
联合优化算法	**0**	**0.45**	-

表 5.5 用判别模型选择方法所得模型结构在 MNIST 上三种优化算法取得误识率比较

优化方法	训练集误识率 (%)	测试集误识率(%)	错误降低率(%)
梯度上升	0.215	0.56	23.21
Cholesky-CMA-ES	0.183	0.51	9.80
联合优化算法	**0.167**	**0.46**	-

如表5. 2 至表5. 5 所示，与传统的梯度下降算法（梯度上升）和Cholesky-CMA-ES优化算法相比，本章所提出结合目标函数梯度的进化策略优化算法在CENPAMI和

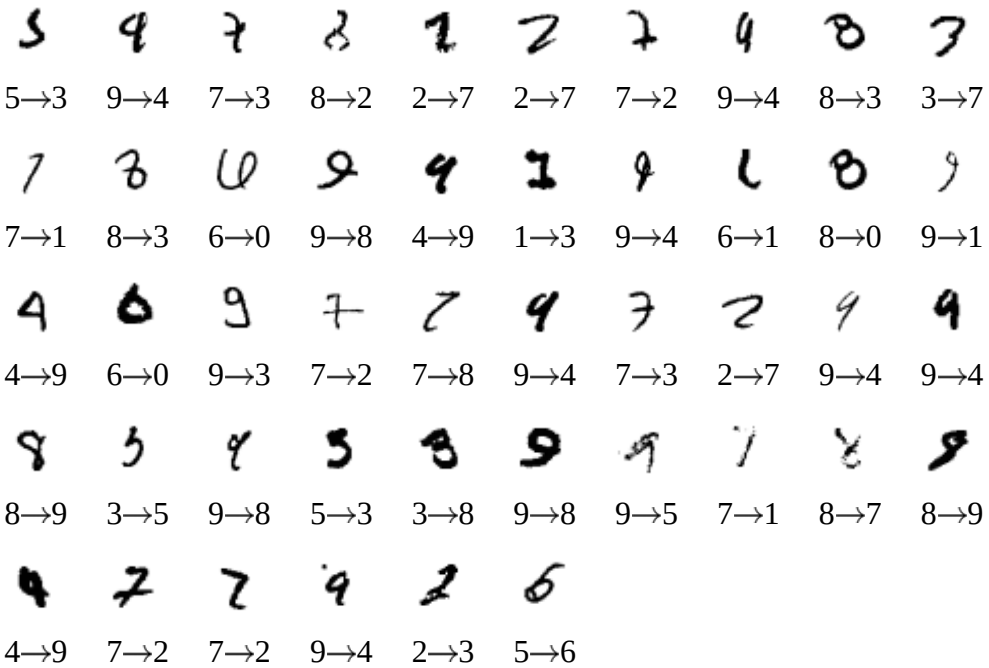
MNIST手写体数字样本库上均取得了最好的优化效果。实验结果表明，结合目标函数梯度的进化策略优化算法是一种有效的优化方法。

图5.5和图5.6中分别列出了由SoftDS-MMP判别学习准则、模型选择的判别式方法和结合目标函数梯度的进化策略优化算法所构成的手写体数字分类器在CENPAMI和MNIST手写体数字样本库上的所有误识字符样本图像，其中部分图像人眼也难以正确识别。

2→3　3→9　4→9　4→9　5→8

5→3　7→4　7→2　7→1

图 5.5 CENPAMI 测试集上所有误识数字样本



5→3　9→4　7→3　8→2　2→7　2→7　7→2　9→4　8→3　3→7

7→1　8→3　6→0　9→8　4→9　1→3　9→4　6→1　8→0　9→1

4→9　6→0　9→3　7→2　7→8　9→4　7→3　2→7　9→4　9→4

8→9　3→5　9→8　5→3　3→8　9→8　9→5　7→1　8→7　8→9

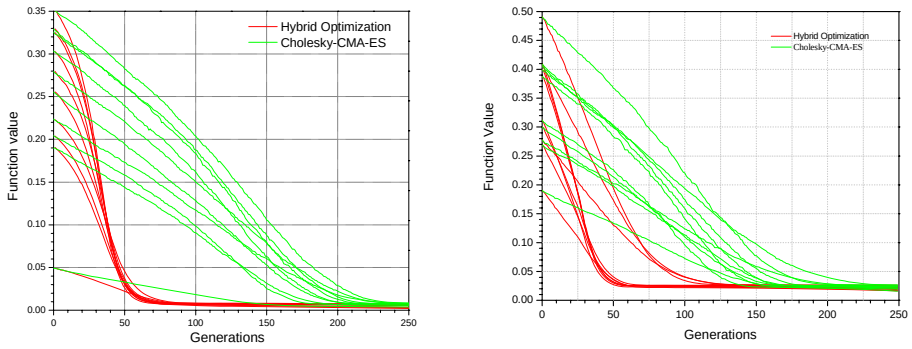
4→9　7→2　7→2　9→4　2→3　5→6

图 5.6 MNIST 测试集上所有误识数字样本

5.7.2 优化方法效率比较

为了分析本章所提出的联合优化算法在收敛速度上所带来的提高，在用AutoClass方法学得的模型结构后，分别记录了Cholesky-CMA-ES和联合优化算法对各数字类在CENPAMI和MNIST手写体数字样本库上前250次迭代的优化结果，其中每次迭代的优化结果是指当前最优目标函数值。从图5.7的优化结果中可以明显的看出联合优化算法的收敛速度要快于Cholesky-CMA-ES算法。在CENPAMI手写体数字样本库上，采用联合优化算法在约60次迭代后，目标函数值可以达到0.01，但Cholesky-CMA-ES则需要约200次迭代才能达到相同的目标值。在MNIST手写体数字样本库上，采用联合优化算法在约120次迭代后，所有数字类的目标函数值可以达到0.03，但Cholesky-CMA-ES则要在230次迭代后，所有数字类的目标函数值达到0.03。

本章还记录了三种优化方法的优化时间。该实验的三种优化方法均是在3.4GHz计算处理器和1.0G内部存储器的个人计算机上完成。在CENPAMI手写体数字样本库上，梯度下降、Cholesky-CMA-ES和联合优化算法的优化时间分别为2133秒，2036秒和1659秒。相对于梯度下降和Cholesky-CMA-ES算法，联合优化算法将训练时间分别缩短了22.22%和18.52%。在MNIST手写体数字样本库上，梯度下降，Cholesky-CMA-ES和联合优化算法的优化时间分别为3549秒、3735秒和3018秒。相对于梯度下降和Cholesky-CMA-ES方法，联合优化方法分别将训练时间缩短了14.96%和19.20%。



CENPAMI数据库　　MNIST数据库

图 5.7 Cholesky-CMA-ES 和联合优化算法的实验结果

5.7.3 数字分类识别结果比较

用于SoftDS-MMP的联合优化算法在CENPAMI和在MNIST手写体数字样本库上均取得了满意的识别效果。我们还根据已阅文献收集了目前在CENPAMI和MNIST手写

体数字样本库上所取得的最好识别效果，与联合优化算法所取得的识别效果进行比较，从而说明结合目标函数梯度的进化策略优化算法一种有效的判别学习优化方法。

（1）CENPARMI样本库

文献[103]对各种手写体数字识别技术在CENPAMI样本库上进行了比较，包括手写体数字特征和分类器，所发布的测试集上最低误识率是 0.95%，所用的特征是e-grg，分类器是基于RBF核的SVM和判别学习二次判别函数(Discriminative Learning Quadratic Discriminant Function，DLQDF)。在相同的特征上，用Cholesky-CMA-ES和联合优化算法取得了更好的识别效果。表5.6中列出了目前在CENPAMI手写体数字样本库上发布的较好识别效果与在本文中所取得的最好识别效果的比较。根据表5.6中所列结果，基于联合优化算法的SoftDS-MMP在CENPAMI样本库上取得识别效果好于其它已有识别方法。

表 5.6 不同识别方法在 CENPARMI 数据库上所取得的误识率

方法	特征	误识率(%)
Modular Neural Network [143]	Class dependent features	2.15
Local Learning Framework[147]	32 direction gradient features	1.90
Neural Network[148]	Random features	1.70
Virtual SVM [42]	32 direction gradient features	1.30
SVC-rbf [104]	8 direction deslant chaincode features	0.85
Cholesky-CMA-ES	e-grg	0.70
Our Method	**e-grg**	**0.45**

（2） MNIST样本库

文献[103]还在MNIST数据库上比较了各种数字识别技术，其所报道的最低误识率是0.42%，所用特征是e-grg，分类器是基于RBF核的SVM。在相同的特征上，基于联合优化算法的SoftDS-MMP取得了0.46%的识别效果。本章还收集了目前为止在MNIST手写体数字样本库上所发布的较好的识别效果，并将其列于表5.7中。从表5.7所列的

识别结果可以看出基于联合优化算法的Soft-MMP所取得识别效果要好于目前大部分已有识别技术，并接近现有最好识别率。文献[42]，[103]和[101]发布的识别效果好于本文所取得的最好识别效果。但文献[37]和[103]的所用的数字分类器是SVM，而本文所用的分类器是基于后验伪概率的贝叶斯分类器。贝叶斯分类器的分类速度通常比SVM快很多，而且是一种多类分类器，更适合大规模模式识别问题。在本章实验中，用基于后验伪概率的贝叶斯分类器在中央处理器为3.4GHz的个人电脑上完成MNIST手写体数字样本库测试集分类任务所需时间是20秒，而文献[103]所发布在中央处理器为1.5GHz的个人电脑上完成MNIST手写体数字样本库测试集分类任务所需时间是219秒。目前，在MNIST上发布的最低误识率是0.39%，由文献[101]给出，但文献[42]和文献[101]在训练过程中都使用了大量的虚拟样本，训练的时空复杂度都很高，而本文所提出的算法显然要简单得多。

表 5.7 不同字符识别技术在 MNIST 数据库上所取得的误识率

方法	特征	误识率(%)
Convolutional net LeNet-1[111]	Subsampling	1.7
Polynomial SVM [149]	32 direction gradient features	1.4
Boosted LeNet4[70]	Subsampling	0.7
Large Convolutional Net[150]	Unsup features	0.62
SVM[151]	Vision-based feature	0.59
SVMs[152]	Trainable feature	0.54
K-NN[100]	Shiftable edges	0.52
VSVM [42]	32 direction gradient features	0.44
SVC-rbf[103]	e-grg	0.42
Large Convolutional Net[101]	Trainable feature	**0.39**
Our Method	e-grg	**0.46**

5.8 小结

目前为止，基于梯度的优化算法是判别学习领域的主要优化方法。但是基于梯度的优化算法存在一个重要问题是易陷入局部最优解，优化效果有时并不理想。本章将进化策略引入到本文所提出的SoftDS-MMP判别学习框架中，提出一种结合目标函数梯度信息的进化策略优化算法。利用该算法实现了梯度下降与Cholesky分解的协方差自适应进化策略的互补(Covariance Matrix Adaptation Evolution Strategy based on Cholesky factorization，Cholesky-CMA-ES)。一方面，利用多点随机搜索策略可以降低陷入局部最优解的概率；另一方面，利用目标函数的梯度信息可以加快收敛速度。我们用联合优化算法优化SoftDS-MMP，并将其用于手写体数字识别问题，在广泛使用的CEMPAMI和MNIST手写体数字样本库上进行实验。与梯度下降（梯度上升）和Cholesky-CMA-ES优化算法相比，所提出的联合优化算法具有更好的优化效果和效率。本文在CENPAMI数据库上所取得了0.45%的误识率，根据我们现有知识是目前在CENPAMI数据库上报道的最底误识率。在MNIST数据库上，本文取得的了0.46%误识率，接近于目前在MNIST数据库上所报道的最底误识率。

需要说明的是在该实验中由于对进化策略采用的是串行实现，所以并没有达到最快优化速度。在今后的工作中，将通过并行实现进一步提高所提出联合优化算法的优化速度。今后工作的另一个重要研究方向是探讨更加适合的策略控制联合优化算法中Cholesky-CMA-ES和梯度下降两部分间的比例关系。目前，采用的策略是在学习过程中不断增加梯度对该联合优化算法的影响。今后将根据采样点的分布来调整两部分之间的关系。此外，在下一步研究工作当中还将把联合优化算法应用于其他优化问题，尤其是应用到其他的判别学习问题当中。

第 6 章 总结与展望

6.1 本文研究工作总结

基于有限混合模型的图像识别方法，具有形式灵活、识别速度快、抗干扰能力强、识别准确率高等优点、成为一类重要的图像分类方法，受到广泛关注。有限混合模型对数据建模主要包括两个任务：有限混合模型成份个数选择和各成份具体参数学习。有限混合模型学习方法按照学习目标不同可分为传统的生成学习和判别学习。大量研究表明，判别学习使模式识别系统的识别性能明显提高，其学习效果明显好于传统的生成式学习。很多研究机构都对判别学习展开了研究，并提出了多种判别学习方法，如支持向量机(Support Vector Machine，SVM)、最大互信息(Maximum Mutal Information，MMI)、最小分类错误(Minimum Classification Error，MCE)等等，但目前已有方法的性能与人们的期望还有相当大差距，并且已有判别学习方法主要集中于学习有限混合模型的参数，忽略了学习有限混合模型成份个数。本文的研究目标为面向图像分类问题，研究图像有限混合模型的判别学习方法，用于从训练样本中获得判别能力强的有限混合模型，实现有效分类；主要研究了以类为中心的贝叶斯分类器判别学习准则、基于判别学习的模型选择方法和判别学习智能优化算法，以及在文档分析与识别领域方面的应用。本文主要工作和创新点有：

（1）提出了一种新的学习贝叶斯分类器的判别学习方法，称为基于软目标的最大最小后验伪概率学习方法（SoftDS-MMP）。SoftDS-MMP 对每个模式类正样本和反样本的后验伪概率分别定义相应的自适应软目标，进而用该软目标度量分类器在训练集上期望分类损失，通过最小化期望分类损失，同时最大化两个软目标之间的距离，获得最优分类器参数集合。SoftDS-MMP 还进一步利用软目标在训练过程中动态选择训练数据，从训练集中移出和插入样本，减少训练样本数量，压缩训练数据集，加快训练速度。在数据选择过程中，对于那些后验伪概率远超过其相应目标值的样本，在一定的训练周期内暂时将其移出训练集。与基于硬目标的学习方法相比，SoftDS-MMP 判别学习方法在学习过程中根据训练样本自适应的调节训练目标，从而降低了过学习风险，并利用软目标值自适应进行动态样本选择，加快训练速度。

（2）在 SoftDS-MMP 判别学习框架下，提出了一种新的高斯混合模型选择判别

式方法。该方法根据模型选择的要求调整 Soft-MMP 判别学习方法的目标函数，并将其结合到基于拉普拉斯估计的贝叶斯模型选择框架下。相应的模型选择标准是选择使 SoftDS-MMP 目标函数边缘积分的拉普拉斯估计值最大的高斯混合模型。利用线性搜索策略同时获得最优高斯混合模型结构和模型参数。该方法将判别信息引入到模型选择当中，从而提高了分类器的判别性能，并使模型选择和参数学习同时进行。

（3）本文将进化策略引入到 SoftDS-MMP 判别学习框架之中，该方法利用目标函数的梯度信息调整 Cholesky 协方差矩阵自适应进化策略中的参数（Covariance Matrix Adaptation Evolution Strategy based on Cholesky Factorization，Cholesky-CMA-ES），包括加权均值、协方差矩阵和全局步长，提高优化效果和优化效率。该方法主要思想是对每代个体的加权均值都用梯度优化进行调整，并根据其调整的均值调整相应协方差矩阵和全局步长。此外，本文还进一步在训练过程中动态调整梯度下降与 Cholesky-CMA-ES 在联合优化方法中所占的比重。在训练初期，为了快速找到最优搜索区域，Cholesky-CMA-ES 在联合优化方法中占主导地位。然后，逐渐增加梯度下降算法在联合优化方法中所占的比重，以加强联合优化方法的局部探索能力。该联合方法实现了 Cholesky-CMA-ES 与梯度下降方法的互补，一方面，利用其多点随机优化策略可以降低其陷入局部最优解的概率；另一方面，利用目标函数的梯度信息可以加快其收敛速度。

（4）本文将上述判别学习准则、模型选择方法和联合优化算法应用于手写体数字识别问题当中，通过用联合优化算法优化基于SoftDS-MMP的判别模型选择准则，学习手写体数字分类器的结构和参数，在广泛使用的CENPAMI和MNIST手写体数字样本库上进行实验。在CENPAMI手写体数字样本库上我们取得了 99.55% 的识别率，根据我们已有知识，这是目前最好的识别效果。在MNIST手写体数字样本库上取得了 99.54% 的识别率，接近目前已有最好识别效果。实验结果证明了所提出方法的有效性。

6.2 进一步研究工作展望

本文在前人工作基础上，针对判别学习中的学习准则、优化算法和模型选择的判别式方法三个方面的内容展开了研究，取得一定的成果，同时也还有一些问题值得进一步的研究和探讨。下一步将在目前已有工作的基础上，在以下几个方面上展开进一步的研究。

（1）目前，所提出的SoftDS-MMP是一种面向类的判别学习方法，在下一步工作中，将把样本的判别信息也引入到所提出的判别学习准则当中，即在SoftDS-MMP中引入关于分类器判决边界的信息。同时，SoftDS-MMP的实验数据显示软目标还具有一个值得关注的特征。训练后的软目标可以度量不同类别的判别能力。与其他类别可以显著区分的类的两个软目标之间的距离要大于易混淆的类的两个软目标之间的距离。在今后的工作中将对该特征进一步研究。

（2）目前，本文对所提出的组合进化策略采用的是串行实现，并没有发挥其最佳性能。在今后的工作中，将通过并行实现进一步提高所提出的组合优化方法的优化速度。今后工作的另一个重要研究方向是探讨更加适合的策略控制组合优化方法中Cholesky-CMA-ES和梯度下降两部分间的比例关系。目前，本文采用的策略是在学习过程中不断增加梯度对该联合优化方法的影响。今后将根据采样点的分布来调整两部分之间的关系。此外，将提出的联合优化方法应用于其他优化问题，尤其是应用到其他的判别学习问题当中。

（3）本文主要面向贝叶斯分类器提出相应判别学习方法，而没有考虑图像的特征表示。在今后的工作中，我们将进一步研究探讨判别学习在特征选择中的应用。从特征向量中选择一些判别能力强的特征，压缩特征向量，增加特征的判别能力。将特征选择与分类器学习统一到判别学习框架之中，构成统一整体，提高模式识别系统的识别性能。

（4）在更多的实际应用，数据库和除高斯混合模型以外的其它有限混合模型中评价该判别模型选择方法的性能。此外，本文对模型准则是采用线形搜索的策略，在今后的工作中将研究一些更自动的模型选择方法。

参考文献

[1] Mario Castelan, W.A.P. Smith, E.R. Hancoc. A Coupled Statistical Model for Face Shape Recovery from Brightness Images [J]. IEEE Trans. Image Processing, 2007, 16(4): 1139-1151.

[2] W.A.P. Smith, E. R. Hancock. Recovering Facial Shape Using a Statistical Model of Surface Normal Direction [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2006, 28(12): 1914-1930.

[3] E.G. Learned-Miller. Data Driven Image Models through Continuous Joint Alignment [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2006, 28(2): 236-250

[4] G. Mclachlan, K. Basford. Finite Mixture Models [J]. New York: John Wiley Sons, 2000.

[5] X.C. Wang, K.K. Paliwal. Discriminative Learning and Informative Learning In Pattern Recognition [C]. Proceeding of 9th international Conference on Neural Information Processing (ICONIP), 2000(2): 862-865.

[6] S. Katagiri, C.H. Lee, B.H. Juang. New discriminative training algorithms based on the generalized descent method [C]. Proc. IEEE worshop Neural Networks for signal Processing, 1991: 299-308.

[7] B.H. Juang, S. Katagiri. Discriminative Learning for Minimum Error Classification [J]. IEEE Trans. Acoust, Speech, Signal Process, 1992 40(12): 3043-3054.

[8] Brown, The Acoustic-Modeling Problem in Automatic Speech Recognition [D]. 1987, Ph.D. thesis, Carnegie-Mellon University.

[9] L.R.Bahl, P.F. Brown, P.V. Souza, R.L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition [C]. ICASSP, 1986: 49-52.

[10] H. Jiang, X. Li. Incorporating training errors for large margin HMMs under semi-definite programming framework [C]. ICASSP, 2007:629-632.

[11] Mario A.T. Figueiredo, Anil K. Jain. Unsupervised Learning of Finite Mixture Models [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002, 24(3):381-396.

[12] Song-Chun Zhu. Statistical Modeling and Conceptualization of Visual Patterns [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2003, 25(6):691-711.

[13] Fabrice Heitz, Patrick Bouthemy. Multimodal Estimation of Discontinuous Optical Flow Using Markov Random Fields [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1993, 15(12):1217-1232.

[14] S. Geman, D. Geman, Stochastic Relaxation. Gibbs Distributions and the Bayesian Restoration of Images [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1984, 6: 721-741.

[15] A. Efros, T. Leung. Image Quilting for Texture Synthesis and Transfer [C], Proc. SIGGRAPH, 2001.

[16] J.S. De Bonet, P. Viola. A Non-parametrid Multi-Scale Statistical Model for Natural Images [J], Advances in Neural Information Processing, 1997.

[17] J. J. Hull. Incorporating Language Syntax in Visual Text Recognition with a Statistical Model [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1996, 18(12):1251-1256.

[18] S.S. Kuo, E. O. Agazzi. Keyword Spotting in Poorly Printed Documents Using Pesudo 2-D Hidden Markov Models [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1994, 16(8):842-848.

[19] Z.W. Tu and S. C. Zhu. Image Segmentation by Data Driven Markov Chain Monte Carlo [J]. IEEE Trans. Pattern Analysis and Machine Intelligence,2002, 24(5):657-673.

[20] Nizar Bouguila, Djemel Ziou. Unsupervised Selection of a Finite Dirichlet Mixture Model: An MML-Based Approach [J]. IEEE Trans. Knowledge and Engineering, 2006, 18(8):993-1009.

[21] Nizar Bouguila, Djemel Ziou. High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2007, 29(10):1716-1731.

[22] J.A. Al-Saleh, S.K. Agarwal. Extended Weibull type distribution and finite mixture of distributions [J]. Ststistical Methodogy, 2006, 3:224-233.

[23] Lianjun Zhang, Chuangmin Liu. Fitting irregular diameter distributions of forest stands by Weibull, modified Weibull, and mixture Weibull models [J]. The Japanese Forest Society and Springer, 2006, 11:369-372.

[24] S. Medasani, R. Krishnapuram. A comparison of Gaussian and Pearson mixture modeling for pattern recognition and computer vision applications [J]. Pattern Recognition Letter. 1999, 20:305-313.

[25] S.K. Agarwal and S.L. Kalla. A generalized gamma distribution and its application in reliability [J]. Commun. Statist. Theory Methods, 1996, 25(1):201-210.

[26] A.J. Al-salen, S.K. Agarwal. Finite mixture of gamma distributions: A conjugate prior [J]. Computational Statistics & Data Analysis, 2007,51: 4369-4378.

[27] Y. Agusta and D.L. Dows. MML Clustering of Continuous Valued Data Using Gaussian and t Distributions [C]. Proc. Australian Joint Conf. Artifical Intelligence, 2002:143-154.

[28] C.S. Wallace and D.L. Dowe. MML Clustering of Multi-State and Poisson, von Mises Circular and Gaussian Distributions [J]. Statistics and Computing, 2000, 10(1):73-83.

[29] G. Schwarz. Estimating the Dimension of a Model [J]. The Annals of Statistics, 1978, 6(2):461-464.

[30] J. Oliver, R. Baxter, and C. Wallace. Unsupervised Learning Using MML [C], Porc. 13th Int. Conf. Machine Learning, 1996:364- 372.

[31] J. Rissanen. Universal Coding, Information, Prediction and Estimation [J]. IEEE Trans. Information Theory, 1984, 30(4):629~636.

[32] H. Akaike. A New Look at the Statistical Model Identification [J]. IEEE Trans. Automatic Control, 1974, 9(6):716-723.

[33] C. Biernacki, G. Celeux, and G.Govaert. An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model [J]. Pattern Recognition Letters, 1999, 20:267-272.

[34] C. Biernacki, G. Celeux, and G. Govaert. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2000,22(7):719-725.

[35] H. Bensmail, G. Celeus. Inference in Model-Based Cluster Analysis [J], statistics and Computing, 1997, 7:1-10.

[36] G. Mclachlan. On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture [J]. J.Royal Statistical Soc, 1987(36):318-324.

[37] P. Smyth. Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood [J]. Statistics and Computing, 2000, 10(1):63-72.

[38] V.N. Vapnik, The nature of statistical learning theory [J]. New York: Wiley, 1998.

[39] C.Burges. A tutorial on support vector macines for pattern recognition [J]. Data Mining Knowl. Dis., 1998, 2(2): 121-167.

[40] 张学工. 关于统计学习理论与支持向量机 [J] 自动化学报, 2000, 26(1): 32-42.

[41] 张铃. 支持向量机理论与基于规划的神经网络学习算法 [J]. 计算机学报, 2001, 24(2);113-118.

[42] Jian-xiong Dong, Adam Krzyzak, and C.Y. Suen. Fast SVM Training Algorithm with Decomposition on Very Large Data sets [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2005, 27(4): 603-618.

[43] S.S. Keerthi, S.K. Shevade, and C.Bhattachayya. Improvments to Platt's SMO Algorithm for SVM Classifier Design [J]. Neural Computation, 2001(13): 637-649.

[44] 李健民, 张钹林, 富宗. 支持向量机的训练方法 [J] 清华大学学报, 2003, 43(1) 120-124.

[45] J.C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization [J]. Advances in Kernel methods supoort Vector Machines, MIT Press, 1998.

[46] http://svmlight.joachims.org/

[47] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[48] http://www.idiap.ch/learning/SVMTorch.html

[49] http://www.cenparmi.concordia.ca/~jdong/HeroSvm.html

[50] E. McDermott and S. Katagiri. Strinig-level MCE for continuous phoneme recognition [J]. Eurospeech, 1997:123-126.

[51] J.Leroux and E.Mcdermott. Optimization methods for discriminative training [J]. Eurospeech, 2005: 3341-3344.

[52] L. Nguyen and B. Xiang. Light supervision in acoustic model training [C]. IEEE ICASSP, 2004(1):209-212.

[53] G. Evermann, H.Chan. M.Gales, et al. Training LVCSR systems on thousands of hours of data [C]. IEEE, ICASSP, 2005(1):209-212.

[54] W. Macherey, L. Halferkamp, et al. Investigations on error minimizing training criteria for discriminative training in automatic speech recognition [J]. Eurospeech, 2005: 2133-2136.

[55] Qi Li and B.H. Juang. A New Algorithm for Fast Discriminative Training [J]. IEEE ICASSP, 2002(1):97-100.

[56] Qi Li and B.H. Juang. Study of a Fast Discriminative Training Algorithm for Pattern Recognition [J]. IEEE Trans. Neural NetWorks, 2006, 17(5):1212-1222.

[57] S. KataGiri, B.H. Juang and C.H. Lee. Pattern Recognition Using a Family of Design Algorithms Based Upon the Generalized Probabilistic Descent Method [C]. Proceedings of IEEE, 1998, 86(11): 2345-2373.

[58] 袁伟, 高剑峰, 步丰林. 语言建模中最小化样本风险算法的研究与改进 [J]. 软件学报, 2007, 18(2): 196-204.

[59] 于浩, 高剑峰, 步丰林. 一种新的语言模型判别训练方法[J]. 计算机学报, 2005, 28(10):1708-1715.

[60] D. W. Purnell and E.C. Botha. Improved Generalization of MCE Parameter Estimation With Application to Speech Recognition [J]. IEEE Trans. Speech and Processing, 2002, 10(4): 232-239.

[61] M. Afify, X.W. Li and H. Jiang. Statistical Analysis of Minimum classification Error Learning for Gaussian and Hidden Markov Model Classifiers [J]. IEEE Trans. Audio, Speech, And Language Processing, 2007, 15(8): 2405~2417.

[62] E.MCDermott and S. Katagiri. A derivation of minimum classification error from the theoretical classification risk using Pazen estimation [J]. Compute, Speech Lang, 2004(18): 107-122.

[63] Ralf Schluter. Investigations on Discriminative Training Criteria [D]. PHD, 2000.

[64] H. Gish. A minimum classification error, maximum likelihood, neural network [C]. IEEE ICASSP, 1992(2): 289~292.

[65] W. Chou, B.H. Juang and C.H. Lee. Segmental GPD training of HMM based speech recognization [C]. Proc. IEEE ICASSP, 1992(1): 473-476.

[66] Biing-Hwang Juang, Wu Chou and Chin-Hui Lee. Minimum Classification Error Rate Methods for Speech Recognition [C]. IEEE Transaction on Speech And Audio Processing, 1997, 15(3):257-265.

[67] E. McDermott, T.J. Hazen, J.L. Roux, A. Nakamura and S. Katagiri. Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error [J]. IEEE Trans. Audio, Speech, And Language Processing, 2007, 15(1):203-223.

[68] C.L. Liu and H. Sako, Hiromichi Fujisawa. Discriminative Learning Quadratic Discriminant Function for Handwriting Recognition [J]. IEEE Trans. Neural Networks, 2004, 15(2):430-444.

[69] Cheng-Lin Liu and Masaki Nakagawa. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition [J]. Pattern Recognition, 2001(34):601-615.

[70] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. Gradient-Based Learning Applied to Document Recognition [C]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[71] Yaodong Zhang, Peng Liu, and Frank K.Soong. Minimum Error Discriminative Training for Radical-based Online Chinese Handwriting Recognition [C], Proceedings of the 10th International Conference on Document Analysis and Recognition, 2007:53-57.

[72] Rui Zhang and Xiaoqing Ding. Minimum Classification Error Training for Handwritten Character Recognition [C]. Proceedings of the 10th International Conference on Document Analysis and Recognition, 2007.

[73] C.L. Cheng. High Accuracy Handwritten Chinese Character Recognition Using Quadratic Classifiers with Discriminative Feature Extraction [C]. The 18th International Conference on Pattern Recognition, 2006.

[74] R. Chengalvarayan and L.Deng. HMM-based speech recognition using state-dependent, discriminatively derived trandforms on Mel-warped DFT features [J]. IEEE Trans. Speech Andio Processing, 1997, 5(3):243-256.

[75] Li Deng, K. Wang, and W. Chou. Guest editorial [J]. IEEE Signal Process, 2005, 22(5):12-14.

[76] C. Yen, S.S. Kuo, and C.H. Lee. Minimum error rate training for PHMM-based text recognition [J]. IEEE Trans. Image Processing, 1999, 8(8):1120-1124.

[77] S. Gao, W.Wu, C.H.Lee, and T.S. Chua. A maximal figure-of-merit Learning Approach Text Categorization [C]. Poc. ACM SIGIR Conf. Research and Development in Information Retrieval, 2003:174-181.

[78] F.J.Och. Minimum Error Rate Training in Statistical Machine Translation [C]. ACL: Proc. Assoc. Comput. Ling., 2003:160-170.

[79] D.Povey, P.C. Woodland. Minimum Phone Error and I-Smoothing for Improved Discriminative Training [J]. IEEE 2002.

[80] K. Yu and M.J.F.Gales. Discriminative cluster adaptive training [J]. IEEE Trans. Audio, Speech, and Language Processing, 2006, 14(5):1694-1703.

[81] Khe Chai Sim, Mark J.F. Gales. Minimum Phone Error Training of Precision Matrix Models [J]. IEEE Transactions On Audio, Speech, And Language Processing, vol. 14, no. 3, May 2006.

[82] P.C. Woodland and D.Povey. Large Scale Discriminative Training for Speech Recognition [D].

[83] J. Du, P. Liu, F.K. Soong, J.L. Zhou and R.H. Wang. Minimum Divergence Based Discriminative Training [C]. Ninth International Conference on Spoken Language Processing, 2006.

[84] Roongroj Nopsuwanchai. Discriminative training methods and their applications to handwriting recognition [D]. PHD, 2005.

[85] Roongroj Nopsuwanchai, Alain Biem, and William F.Clocksin. Maximization of Mutual Information for Offline Thai Handwriting Recognition [J]. IEEE Trans. Pattern Analysis and Macine Intelligence, vol. 28, no.8, August 2006, pp: 1347-1351.

[86] Kenneth E. Hild, Deniz Erdogmus, Kari Torkkola, and Jose Principe, Feature Extraction Using Information-Theoretic Learning [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.28, No.9, 2006, pp.1385-1392.

[87] 杨国亮，王志良，刘翼伟，王国江，陈锋军. 基于改进 MMI 的 HMM 训练算法及其在面部表情识别中的应用 [J]. 北京科技大学学报, 2007, 29(4):432-437.

[88] C. Liu, H. Jiang, and L. Rigazio. Recent improvement on maximum relative margin estimation of HMMs for speech recognition [C]. ICASSP, 2006: 269-272.

[89] H. Jiang, X.W. Li, and C.J. Liu. Large Margin Hidden Markov Models for Speech Recognition [J]. IEEE Trans. Audio, Speech, and Language Process, 2007, 14(5):1584-1595.

[90] D. Yu, L. Deng, X. He, and A. Acero. Large-margin Minimum Classification Error Training for Large-Scale Speech Recognition Tasks [C]. Proc. ICASSP, 2007(1): 1137-1140.

[91] Jinyu Li, M. Yuan, and C.H. Lee. Approximate Test Risk Bound Minimization through Soft Margin Estimation [J]. IEEE Trans. Audio, Speech, and Language Process, 2007, 15(8):2293-2404.

[92] F. Sha and L. Saul. Large margin Gaussian Mixture Modeling for Phonetic Classification and Recognition [C]. Proc. ICASSP, 2006(1): 265-268.

[93] F. Sha. Large margin training of acoustic models for speech recognition [D]. University of Pennsylvania, 2007.

[94] X.W. Li and H. Jiang. Solving Large-Marin Hidden Markov Model Estimation via Semidefinite Programming [J]. IEEE Trans. Audio, Speech, and Language Process, 2007, 15(8): 2383-2392.

[95] Xiabi Liu, Yunde Jia, Xuefeng Chen, Yuan Deng, and Hui Fu. Max-Min Posterior Pseudo-Probabilities Estimation of Posterior Pseudo-Probabilities Estimation of Posterior Class [R]. 2006，http://www.mcislab.org.cn/technicalreports/MMP.PDF.

[96] Xiabi Liu, HuiFu, and Yunde Jia. Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images [J]. Pattern Recognition, Vol. 41, 2008, pp. 484-493.

[97] 付慧，刘峡壁，贾云得. 基于最大最小相识度学习方法的文本提取[J]. 软件学报，2008,19(3):621-629.

[98] Yuan Deng, Xiabi Liu, and Yunde Jia. Learning Semantic Concepts for Image Retrieval Using the Max-min Posterior Pseudo-Probabilities Method [C]. IEEE International Conference on Multimedia and Expo, 2007, pp. 1970-1973.

[99] Xuefeng Chen, Xiabi Liu, and Yunde Jia. Learning Handwritten Digit Recognition by the Max-Min Posterior Pseudo-Probabilities Method [C]. Ninth International Conference on Document Analysis and Recognition, 2007, pp. 342-346.

[100] Daniel Keysers, Thomas Deselaers, Christian Gollan, and Hermann Ney. Deformation Models for Image Recognition [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2007, 29(8): 1422-1435.

[101] Ranzato Marc'Aurelio, Christopher Poultney, Sumit Chopra and Yann LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model [J]. Advances in Neural Information Processing Systems, MIT Press, 2006.

[102] B. Kegl and R. Busa-Fekete. Boosting products of base classifiers [C]. Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.

[103] C.L. Liu, Kazuki Nakashima, Hiroshi Sako, and Hiromichi Fujisawa. Handwritten digit recognition: benchmarking of state-of-art techniques [J]. Pattern Recognition, 2003, 36: 2271-2285.

[104] C.L., Liu, K. Nakashima, H. Sako, and H. Fujisawa. Handwritten digit recognition: investigation of normalization and feature extraction techniques [J]. Pattern Recognition, 2004, 37: 265-279.

[105] 刘海龙. 基于描述模型和鉴别学习的脱机手写字符识别研究[D]. 北京：清华大学，2006.

[106] Box GEP, Cox DR. An analysis of transformation [J]. J.R. Statistical Society. 1964, 26(Series B): 211-243.

[107] P. Moerland. A comparison of mixture models for density estimation [C]. Proceedings of the 9th International Conference on Artificial Neural Networks, 1999:25~30.

[108] L. Liu and J.L. He. On the use of Orthogonal GMM in Speaker Recognition [C]. Proc. Int. Conf. Acoust., Speech, Signal Processing, 1999: 845-848.

[109] R. Zhang and X.D. Ding. Offline Handwritten Numeral Recognition Using Orthogonal Gaussian Mixture Model [C]. Proc. 6th Int. Conf. Document Analysis and Recognition, 2001:1126-1129.

[110] C.Y. Suen, et al. Computer recognition of unconstrained handwritten numerals [C]. Proc. IEEE, 1992, 80(7): 1162-1180.

[111] Y. LeCun, et al. Comparison of Learning Algorithms for Handwritten Digit Recognition [C]. Nanterre, France, Proc. Int. Conf. Artificial Neural Networks, 1995:53-60.

[112] Yun Lei, Xiaoqing Ding, and Shenjin Wang. Visual Tracker Using Sequential Bayesian Learning: Discriminative, Generative, and Hybrid [J]. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics, 2008, 38(6): 1578-1591.

[113] Yanmin Sun, Andrew K. C. Wong, and Yang Wang. Generative and Discriminative Learning by CL-net [J]. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics, 2007, 37(4): 1022-1029.

[114] Georg Heigold, Thomas Deselaers, and Ralf Schlüter, Hermann Ney. Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition [C]. The 25th International Conference on Machine Learning, 2008.

[115] Caruana R., Baluja S., and T. Mitchell. Using the future to ‘sort out’ the present: rankprop and multitask learning for medical risk evaluation [C]. Advances in Neural Information Processing Systems, 1996.

[116] R. Caruana. Multitask Learning [J]. Journal of machine learning, 1997, 28: 41-75.

[117] Michael Rimer, Tony Martinez. Classification-based objective functions [J]. Journal of Machine Leaning. 2006, 63(2): 183-205.

[118] Corinna Cortes, Vladimir Vapnik. Support-Vector Networks [J]. Journal of Machine Learning, 1995, 20(3): 273-297.

[119] Di-Rong Chen, Qiang Wu, Yiming Ying, Ding-Xuan Zhou. Support Vector Machine Soft Margin Classifiers: Error analysis [J]. Journal of Machine Learning Research, 2004: 1143-1175.

[120] G.RÄTSCH, T.ONODA, and K.R. MÜLLER. Soft Margins for AdaBoost [J]. Journal of Machine Learning, 2001, 42(3): 287-320.

[121] A. Demiriz, K.P. Bennett, and J. Shawe-Taylor. A column generation algorithm for boosting [J]. Journal of Machine Learning, 2002, 46(1-3): 225-254.

[122] John Shawe-Taylor and Nello Cristianini. On the Generalization of Soft Margin Algorithms [J]. IEEE Trans. Information Theory, 2002, 48(10): 2721-2735.

[123] Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin, Hung-Shin Lee, Berlin Chen. Training Data Selection for Improving Discriminative Training Of Acoustic Models [C]. IEEE Workshop on Automatic Speech Recognition and Understanding, 2007.

[124] L.M. Arslan, J.H. Hansen. Selective Training for Hidden Markov Models with Applications to Speech Classification [J]. IEEE Trans. Speech and Audio Processing, 1999, 7(1): 46-54.

[125] H. Jiang, F.K. Soong, C.H. Lee. A Dynamic In-Search Data Selection Method with Its Applications to Acoustic Modeling and Utterance Verification [J]. IEEE Trans. Speech and Audio Processing, 2005, 13(5): 945-955.

[126] Peter Cheeseman and John Stutz. Bayesian classification (AutoClass): theory and results, Advances in knowledge discovery and data mining [J], Menlo Park, CA. USA, American Association for Artificial Intelligence Press, 1996: 153-180.

[127] X.Y. Liu and M. Gales. Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions [J]. IEEE Trans. Audio, Speech, and Language Processing, 2007, 12(4): 1414-1424.

[128] A. Dasgupta and A. Raftery. Detecting Features in Spatial Point Patterns with Clutter via Model-Based Clustering [J]. J. Am. Statistical Assoc., 1998, 93:294-32.

[129] D.J.C. Mackay. Choices of Basis for Laplace Approximation [J]. Machine Learning, 1998, 33(1).

[130] A.H. Welsh. Aspects of Statistical Inference [J]. John Wiley & Sons, Inc., 1996.

[131] A.E. Raftery. Hypothesis testing and model selection [J]. Lodon, Markov Chain Monte Carlo in Practice, 1996:115-13.

[132] M. Ishiguro, Y. Sakamoto and G. Kitagawa. Boostrapping log-likelihood and EIC, an extension of AIC [J]. Annals of the institute of Statistical Mathematics, 1997, 49: 411-434.

[133] H. Bozdogan. Choosing the Number of Component Clusters in the Mixture Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix [J]. Information and Classification, 1993:40-54.

[134] C. Biernacki and G. Govaert. Using the Classification Likelihood to Choose the Number of Clusters [J]. Computing Science and Statistics, 1997, 29: 451-457.

[135] J. Banfield and A. Raftery. Model-Based Gaussian and Non Gaussian Clustering. Biometrics, 1993, 49: 803-821.

[136] D.J.C. Mackay. Introduction to Monte Carlo Methods [J]. Learning in Graphical Models, NOTA Science Series, Kluwer Academic Press, 1998:175 – 204.

[137] P.Smyth. Model selection for probabilistic clustering using cross-validated likelihood [J]. Statistics and Computing, 2000, 10: 63-72.

[138] M.Padmanabhan and L.R. Bahl. Model Complexity Adaptation Using a Discriminant Measure [J]. IEEE Transactions on Speech and Audio Processing, March 2000, 8(2):205-208.

[139] L. R. Bahl and M. Padmanabhan. A Discriminant Measure for Model Complexity Adaptation [C]. Proc. ICASSP, 1998, 1:453-457.

[140] Y. Normandin. Optimal Splitting of HMM Gaussian Mixture Components with MMIE Training [J]. IEEE Transactions on Speech and Audio Processing, 1995, 5: 449-453.

[141] Xunying Liu. Discriminative Complexity Control and Linear Projections for Large Vocabulary Speech Recognition [D]. U.K.: Cambridge, 2005.

[142] R. Kass, L. Tierney and J. Kadane. Asymptotic in Bayesian computation. Bayesian statistics [J]. 1988:773—795.

[143] G.J. McLachlan and S.K. Ng, A comparison of some information criteria for the number of components in a mixture model [R]. Department of Mathematics, University of Queensland, 2000.

[144] E. Polak. Optimization: Algorithms and Consistent Approximations [J]. New York, USA, Springer-Verlag, 1997.

[145] G.E. Forsythe, M.A. Malcolm and C.B. Moler. Computer Methods for Mathematical Computations [J]. Prentice Hall, Englewood Cliffs, 1977.

[146] I.S. Oh, J.S. Lee and C.Y. Suen. Analysis of class separation and combination of class-dependent features for handwriting recognition [J]. IEEE Trans. Pattern Analysis and Machine Intelligence 1999, 12(10): 1089-1094.

[147] J.X. Dong, A. Krzyzak, C.Y. Suen. Local Learning framework for handwritten character recognition [J]. Engineering Applications of Artificial Intelligence, 2002, 15: 151-159.

[148] P.D. Gader, M.A. Khabou. Automatic feature generation for handwritten digit recognition [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1996, 18(12): 1256-1261.

[149] C.J.C. Burges and B. Scholkopf. Improving the accuracy and speed of support vector learning machines [J]. Advances in Neural information Processing Systems, MIT Press, 1997.

[150] M.A. Ranzato, Fu-Jie Huang, Y.L. Boureau, and Yann LeCun. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition [C]. Proc. Computer Vision and Pattern Recognition Conferenc, 2007.

[151] L.N. Teow and K.F. Loe. Robust vision-based features and classification schemes for off-line handwritten digit recognition [J]. Pattern Recognition, 2002, 35: 2355-2364.

[152] F. Lauer, C.Y. Suen and G. Bloch. A trainable feature extractor for handwritten digit recognition [J]. Pattern Recognition, 2007, 40: 1816-1824.

[153] K. Chellapilla and D.B. Fogel. Evolving an Expert Checkers Playing Program without Using Human Expertise [J]. IEEE Trans. Evolutionary Computation, 2001, 5(4): 422-428.

[154] X. Yao. Evolving Artificial Neural Networks [C]. Proceedings of the IEEE, 1999, 87(9): 1423-1447.

[155] J.Y. Jung and J.A. Reggia. Evolutionary Design of Neural Network Architectures Using a Descriptive Encoding Language [J]. IEEE Trans. Evolutionary Computation, Dec. 2006, 10(6): 676-688.

[156] F. Leung, H. Lam, S. Ling, and P. Tam. Tuning of the structure and parameters of a neural network using an improved genetic algorithm [J]. IEEE Trans. Neural Network, 2003, 14, (1): 79-88.

[157] N. Nikolaerv and H. lba. Learning polynomial feedforward neural networks by genetic programming and backpropagation [J]. IEEE Trans. Neural Network, Mar. 2003, 14 (2): 337-350.

[158] G. Cochenour, J. Simon, S. Das, A. Pahwa, and S. Nag. Evolutionary strategy based radial basis function neural network training algorithm for failure rate prediction in overhead feeders. Proceedings of the Genetic and Evolutionary Computation Conference [C]. ACM Press, New York, NY, 2005: 2127-2132.

[159] C.K. Goh, E.J. Teoh, and K.C. Tan. Hybrid Multiobjective Evolutionary Design for Aritificial Neural Networks [J]. IEEE Trans. Neural Networks, Sep. 2008, 19(9): 1531-1548.

[160] P.P. Palmes, T. Hayasaka, and S. Usui. Mutation-Based Genetic Neural Network [J]. IEEE Trans. Neural Networks, May, 2005, 16(3): 587-600.

[161] H. Frőhlich, O. Chapelle, and B. Schőlkopf. Feature selection for support vector machines by means of genetic algorithms. Proceedings of the 15th IEEE International Conference on Tools with AI [C]. Sacramento, CA, United states, Nov. 03-05, 2003: 142-148.

[162] T. Phienthrakul and B. Kijsirikul. Evolutionary Strategies for Multi-Scale Radial Basis Function Kernels in Support Vector Machines. In Proceedings of the Genetic and Evolutionary Computation Conference [C]. ACM Press, New York, NY, 2005: 905-911.

[163] C.H. Wu, G.H. Tzeng, Y.J. Goo and W.C. Fang. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy [J]. Expert Systems with Applications, 2007, 32: 397-408.

[164] T. Howley and M. G. Madden. The Genetic Kernel Support Vector Machine: Description and Evaluation [J]. Artificial Intelligence Review, 2005, 24:379–395.

[165] C. Igel. Multiobjective Model Selection for Support Vector Machines. Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization [C]. Springer, Berlin, March 9-11, 2005:534-546.

[166] I. Mierswa. Controlling Overfitting with Multi-Objective Support Vector Machines. Proceedings of the Genetic and Evolutionary Computation Conference [C]. ACM Press, New York, NY, 2007: 1830-1837.

[167] M.J. Jesus, F. Hoffmann, L.J. Navascués, and L. Sánchez. Induction of Fuzzy-Rule-Based Classifiers with Evolutionary Boosting Algorithms [J]. IEEE Trans. Fuzzy Systems, 2004, 12(3): 296-308.

[168] A. Treptow and A. Zell. Combining Adaboost learning and evolutionary search to select features for real-time object detection. Proceedings of the 2004 Congress on Evolutionary Computation [C]. 2004, 2: 2107-2113.

[169] M.L. Wong and K. S. Leung. An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm-Based Hybrid Approach [J]. IEEE Trans. Evolutionary Computation, Aug. 2004, 8(4): 378-404.

[170] C.R. Reeves and S.J. Taylor. Selection of training data for neural networks by a genetic algorithm [J]. Lecture Notes in Computer Science, Springer, Berlin, 1998.

[171] B.T. Zhang and G. Veenker. Neural networks that teach themselves through genetic discovery of novel examples. Poc. IEEE Int. Joint Conf. Neural Networks[C]. 1991, 1: 690–695.

[172] S. Cho and K. Cha. Evolution of neural network training set through addition of virtual samples. Proc. IEEE Int.Conf. Evolutionary Computation [C]. 1996: 685–688.

[173] G. P. Nicolás, O.B. Domingo, and H.M.C. Martínez. An alternative approach for neural network evolution with a genetic algorithm: Crossover by combinatorial optimization [J]. Neural Networks 2006(19): 514–528.

[174] C. MACLEOD and G. M. MAXWELL. Incremental Evolution in ANNs: Neural Nets which Grow [J]. Artificial Intelligence Review 2001, 16: 201–224.

[175] M. Rochaa, P. Cortezb, and J. Nevesa. Evolution of neural networks for classification and regression [J]. Neurocomputing, 2007(70): 2809-2816.

[176] Pavol Malinak and Rudolf Jaksa. Simultaneous Gradient and Evolutionary Neural Network Weights Adaptation Methods. IEEE Congress on Evolutionary Computation [C]. 2007: 2665-2671.

[177] E. J. Teoh and C. Xiang. A Global-Local Hybrid Evolutionary Strategy (ES) for Recurrent Neural Networks (RNNs) in System Identification. IEEE Congress on Evolutionary Computation [C], 2007: 1628-1635.

[178] A. Auger, M. Schoenauer, and N. Vanhhaecke. LS-CMA-ES: A Second-Order Algorithm for Covariance Matrix Adaptation. In Proceedings of Eighth International Conference on Parallel Problem Solving from Nature PPSN VIII [C]. Springer, Berlin, 2004:1611-3349.

[179] R. Salomon. Evolutionary Algorithms and Gradient Search: Similarities and Differences [J]. IEEE Trans. Evolutionary Computation, 1998, 2(2): 45-55.

[180] D.V. Arnold and R. Salomon. Evolutionary Gradient Search Revisited [J]. IEEE Trans. Evolutionary Computation, 207, 11(4): 480-495.

[181] D. Wierstra, T. Schaul, J. Peters and J. Schmidhuber. Natural Evolution Strategies. IEEE Congress on Evolutionary Computation [C]. 2008: 3381-3387.

[182] T. Nikolaus Hansen and C. Igel. Efficient covariance matrix update for variable metric evolution strategies [J]. Journal of Machine Learning, 2009, 75: 167–197.

[183] C. Igel, T. Suttorp and N. Hansen. A Computational Efficient Covariance Matrix Update and a (1+1)-CMA for Evolution Strategies. In Proceedings of the Genetic and Evolutionary Computation Conference, ACM Press, 2006: 453-460.

[184] N. Hansen. An analysis of mutative σ-self-Adaptation on linear fitness functions [J]. Evolutionary Computation, 2006, 14(3), 255-275.

[185] N. Hansen, A.S.P. Niederberger, L. Guzzella, and P. Koumoutsakos. A Method for Handling Uncertainty in Evolutionary Optimization with an Application to Feedback Control of Combustion [J]. IEEE Trans. Evolutionary Computation, 2009, 13(1): 180-197.

[186] T. Suttorp, N. Hansen, and C. Igel. Efficient covariance matrix update for variable metric evolution strategies [J]. Machine Learning, 2009(75): 167-197.

[187] 袁亚湘，孙文瑜. 最优化理论与方法[M]. 北京：科学出版社，2003:108-109.

附录 A

该附录提供了用梯度下降算法优化 SoftDS-MMP 所需的导数 $\bigtriangledown F(\mathbf{\Lambda}_t)$。该附录将式（A-1）至式 (A-15)中的 $N(\boldsymbol{x}|\boldsymbol{\varphi}_k, \boldsymbol{\Sigma}_k)$ 简写为 $N_k(\boldsymbol{X})$，其中 k 表示高斯成份在有限混合高斯模型中的顺序，j 表示该元素在手写体数字特征向量中的顺序:

$$\frac{\partial F}{\partial h_1} = 2\left(\omega(1-d) + \frac{1-w}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\right)\hat{H}^2 e^{-h_1}, \tag{A-1}$$

$$\frac{\partial F}{\partial h_2} = -2\left(\omega(1-d) + \frac{1-w}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\right)\bar{H}^2 e^{-h_2} \tag{A-2}$$

$$\frac{\partial F}{\partial \widetilde{\kappa}} = 2(1-\omega)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\kappa}} - \frac{1}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\kappa}}\right) \tag{A-3}$$

$$\frac{\partial F}{\partial \widetilde{\beta}} = 2(1-\omega)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\beta}} - \frac{1}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\beta}}\right) \tag{A-4}$$

$$\frac{\partial F}{\partial \widetilde{w}_k} = 2(1-\omega)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{w}_k} - \frac{1}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{w}_k}\right) \tag{A-5}$$

$$\frac{\partial F}{\partial \boldsymbol{\varphi}_{kj}} = 2(1-\omega)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}} - \frac{1}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}}\right) \tag{A-8}$$

$$\frac{\partial F}{\partial \widetilde{\gamma}_{kj}} = 2(1-\omega)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{l}(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\gamma}_{kj}} - \frac{1}{m}\sum_{i=1}^{m}\hat{l}(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda})\frac{\partial f}{\partial \widetilde{\gamma}_{kj}}\right) \tag{A-9}$$

在式（A-1）到式（A-9）中,

$$\frac{\partial f}{\partial \widetilde{\kappa}} = \left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} + \widetilde{\kappa}\right)} \tag{A-10}$$

$$\frac{\partial f}{\partial \widetilde{\beta}} = \kappa\ln(\beta)\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} + \widetilde{\beta}\right)} \tag{A-11}$$

$$\frac{\partial f}{\partial \widetilde{\beta}} = \kappa\ln(\beta)\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta} + \widetilde{\beta}\right)} \tag{A-12}$$

$$\frac{\partial f}{\partial \widetilde{w}_k} = \kappa\beta w_k N_k(\boldsymbol{X})(1-w_k) \left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta-1} e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)\right)^{\beta}} \tag{A-13}$$

$$\frac{\partial f}{\partial \varphi_{kj}} = \kappa\beta w_k N_k(\boldsymbol{X}) \left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta-1} \left(\frac{x_j-\varphi_{kj}}{\gamma_{kj}}\right) e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)\right)^{\beta}} \tag{A-14}$$

$$\frac{\partial f}{\partial \widetilde{\gamma}_{kj}} = \kappa\beta w_k N_k(\boldsymbol{X}) \left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta-1} \left(\frac{(\boldsymbol{x}_j-\varphi_{kj})^2}{2\gamma_{kj}^2}\right) e^{\left(-\kappa\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}+\widetilde{\gamma}_{kj}\right)} \tag{A-15}$$

附录 B

该附录提供了第四章模型选择判别式方法所需的一阶导数 $\bigtriangledown\widetilde{F}\left(\mathbf{\Lambda}_t\right)$ 和二阶导数 $\bigtriangledown^2\widetilde{F}\left(\mathbf{\Lambda}_t\right)$。该附录将式（B-1）至式（B-22）中的 $N\left(\boldsymbol{x}|\boldsymbol{\varphi}_k,\boldsymbol{\Sigma}_k\right)$ 简写为 $N_k(\boldsymbol{x})$，其中 k 表示高斯成份在有限混合高斯模型中的顺序，j 表示该元素在手写体数字特征向量中的顺序：

$$\frac{\partial\widetilde{F}}{\partial h_1}=2e^{-h_1}\hat{H}^2\left[\omega\left(\hat{H}-\bar{H}\right)-\frac{1-\omega}{m}\sum_{f<\bar{H}}\left(1.0-\hat{H}+f\right)\right] \tag{B-1}$$

$$\frac{\partial^2\widetilde{F}}{\partial h_1^2}=\left[2\frac{\partial\widetilde{F}}{\partial h_1}-n\hat{H}^3e^{-h_1}\sum_{f(\widetilde{x})_i<\hat{H}}1\right]\hat{H}e^{-h_1}-\frac{\partial\widetilde{F}}{\partial h_1} \tag{B-2}$$

$$\frac{\partial\widetilde{F}}{\partial h_2}=2e^{-h_2}\bar{H}^2\left[-\omega\left(\hat{H}-\bar{H}\right)+\frac{1-\omega}{n}\sum_{f>\bar{H}}\left(1.0-f+\bar{H}\right)\right] \tag{B-3}$$

$$\frac{\partial^2\widetilde{F}}{\partial h_2^2}=\left[2\frac{\partial\widetilde{F}}{\partial h_2}-m\bar{H}^3e^{-h_2}\sum_{f(\widetilde{x})_i<\bar{H}}1\right]\bar{H}e^{-h_2}-\frac{\partial\widetilde{F}}{\partial h_2} \tag{B-4}$$

$$\frac{\partial\widetilde{F}}{\partial\widetilde{\kappa}}=n\sum_{i=1}^{m}\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{\kappa}}-m\sum_{i=1}^{n}\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{\kappa}} \tag{B-5}$$

$$\frac{\partial^2\widetilde{F}}{\partial\widetilde{\kappa}^2}=n\sum_{i=1}^{m}\left[\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{\kappa}^2}-\frac{\partial f}{\partial\widetilde{\kappa}}\right]-m\sum_{i=1}^{n}\left[\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{\kappa}^2}+\frac{\partial f}{\partial\widetilde{\kappa}}\right] \tag{B-6}$$

$$\frac{\partial\widetilde{F}}{\partial\widetilde{\beta}}=n\sum_{i=1}^{m}\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{\beta}}-m\sum_{i=1}^{n}\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{\beta}} \tag{B-7}$$

$$\frac{\partial^2\widetilde{F}}{\partial\widetilde{\beta}^2}=n\sum_{i=1}^{m}\left[\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{\beta}^2}-\frac{\partial f}{\partial\widetilde{\beta}}\right]-m\sum_{i=1}^{n}\left[\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{\beta}^2}+\frac{\partial f}{\partial\widetilde{\beta}}\right] \tag{B-8}$$

$$\frac{\partial\widetilde{F}}{\partial\widetilde{w}_k}=n\sum_{i=1}^{m}\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{w}_k}-m\sum_{i=1}^{n}\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\widetilde{w}_k} \tag{B-9}$$

$$\frac{\partial^2\widetilde{F}}{\partial\widetilde{w}_k^2}=n\sum_{i=1}^{m}\left[\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{w}_k^2}-\frac{\partial f}{\partial\widetilde{w}_k}\right]-m\sum_{i=1}^{n}\left[\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial^2 f}{\partial\widetilde{w}_k^2}+\frac{\partial f}{\partial\widetilde{w}_k}\right] \tag{B-10}$$

$$\frac{\partial\widetilde{F}}{\partial\boldsymbol{\varphi}_{kj}}=n\sum_{i=1}^{m}\hat{l}\left(\hat{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\boldsymbol{\varphi}_{kj}}-m\sum_{i=1}^{n}\bar{l}\left(\bar{\boldsymbol{x}}_i;\mathbf{\Lambda}\right)\frac{\partial f}{\partial\boldsymbol{\varphi}_{kj}} \tag{B-11}$$

$$\frac{\partial^2 \widetilde{F}}{\partial \boldsymbol{\varphi}_{kj}^2} = n\sum_{i=1}^{m}\left[\hat{l}\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial^2 f}{\partial \boldsymbol{\varphi}_{kj}^2} - \frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}}\right] - m\sum_{i=1}^{n}\left[\bar{l}\left(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial^2 f}{\partial \boldsymbol{\varphi}_{kj}^2} + \frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}}\right] \tag{B-12}$$

$$\frac{\partial \widetilde{F}}{\partial \widetilde{\gamma}_{kj}} = n\sum_{i=1}^{m}\hat{l}\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial f}{\partial \widetilde{\gamma}_{kj}} - m\sum_{i=1}^{n}\bar{l}\left(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial f}{\partial \widetilde{\gamma}_{kj}} \tag{B-13}$$

$$\frac{\partial^2 \widetilde{F}}{\partial \widetilde{\gamma}_{k}^2} = n\sum_{i=1}^{m}\left[\hat{l}\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial^2 f}{\partial \widetilde{\gamma}_{kj}^2} - \frac{\partial f}{\partial \widetilde{\gamma}_{kj}}\right] - m\sum_{i=1}^{n}\left[\bar{l}\left(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\frac{\partial^2 f}{\partial \widetilde{\gamma}_{kj}^2} + \frac{\partial f}{\partial \widetilde{\gamma}_{kj}}\right] \tag{B-14}$$

在式（B-1）到式（B-14）中

$$\frac{\partial f}{\partial \widetilde{\kappa}} = \kappa p^{\beta} e^{-K\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}} \tag{B-15}$$

$$\frac{\partial^2 f}{\partial \widetilde{\kappa}^2} = \left[1 - K\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta}\right]\frac{\partial f}{\partial \widetilde{\kappa}} \tag{B-16}$$

$$\frac{\partial f}{\partial \widetilde{\beta}} = \kappa\beta\log\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta} e^{-K\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}} \tag{B-17}$$

$$\frac{\partial^2 f}{\partial \widetilde{\beta}^2} = \left\{1 + \beta\log\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)\left[1 - \kappa\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta}\right]\right\}\frac{\partial f}{\partial \widetilde{\beta}} \tag{B-18}$$

$$\frac{\partial f}{\partial \widetilde{w}_k} = \kappa\beta N_k\left(\boldsymbol{X}\right) w_k\left(1 - w_k\right)\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta-1} e^{-K\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}} \tag{B-19}$$

$$\begin{aligned}\frac{\partial^2 f}{\partial \widetilde{w}_k^2} =& \frac{\partial f}{\partial \widetilde{w}_k}\left\{1 - 2w_k + \frac{w_k\left(1 - w_k\right)N_k\left(\boldsymbol{X}\right)}{\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)}\right.\\ &\left.\left[\beta - 1 - \kappa\beta\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)^{\beta}\right)\right]\right\}\end{aligned} \tag{B-20}$$

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}} =& \kappa\beta N_k\left(\boldsymbol{X}\right) w_k\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta-1}\\ &\left(\frac{\boldsymbol{x}_j - \boldsymbol{\varphi}_{kj}}{\gamma_{kj}}\right) e^{-K\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}}\end{aligned} \tag{B-21}$$

$$\begin{aligned}\frac{\partial^2 f}{\partial \boldsymbol{\varphi}_{kj}^2} =& \frac{\partial f}{\partial \boldsymbol{\varphi}_{kj}}\left\{\left(\frac{\boldsymbol{x}_j - \boldsymbol{\varphi}_{kj}}{\gamma_{kj}}\right)\left[1 + \frac{\left(\beta - 1\right)w_k N_k\left(\boldsymbol{X}\right)}{\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)}\right.\right.\\ &\left.\left. - \kappa w_k\beta N_k\left(\boldsymbol{X}\right)\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta-1}\right] - \frac{1}{\boldsymbol{x}_j - \boldsymbol{\varphi}_{kj}}\right\}\end{aligned} \tag{B-20}$$

$$\begin{aligned}\frac{\partial f}{\partial \widetilde{\gamma}_{kj}} =& \frac{1}{2}\kappa\beta N_k\left(\boldsymbol{X}\right) w_k \left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta-1} \\ & \left(\frac{\left(\boldsymbol{x}_j-\boldsymbol{\varphi}_{kj}\right)^2}{\gamma_{kj}}-1\right) e^{-K\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{X})\right)^{\beta}}\end{aligned} \tag{B-21}$$

$$\begin{aligned}\frac{\partial^2 f}{\partial \widetilde{\gamma}_{kj}^2} =& \frac{\partial f}{\partial \widetilde{\gamma}_{kj}}\left\{\left(\frac{\left(\boldsymbol{x}_j-\boldsymbol{\varphi}_{kj}\right)^2}{2\gamma_{kj}}-2\right)\left[1+\frac{w_k N_k\left(\boldsymbol{X}\right)}{\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)}\right.\right. \\ & \left.\left.\left(\beta-1-\kappa\beta\left(\sum_{k=1}^{K} w_k N_k\left(\boldsymbol{X}\right)\right)^{\beta}\right)\right]-\frac{1}{\left(\boldsymbol{x}_j-\boldsymbol{\varphi}_{kj}\right)^2-\gamma_{kj}}-\frac{1}{\gamma_{kj}^2}\right\}\end{aligned} \tag{B-22}$$

攻读学位期间发表论文与研究成果清单

[1] Xuefeng Chen, Xiabi Liu, Yunde Jia. Combining Evolution Strategy and Gradient Descent Method for Discriminative Learning of Bayesian Classifiers, in Proc. of the 11th ACM SIGEVO Annual Conference on Genetic and Evolutionary Computation(GECCO2009), pp. 507-514, 2009.7.8-12, Montréal, Québec, Canada.(EI)

[2] Xuefeng Chen, Xiabi Liu, Yunde Jia. Unsupervised Selection and Discriminative Estimation of Orthogonal Gaussian Mixture Models for Handwritten Digit Recognition, in Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), pp. 1151-1155, 2009.7.26-29, Barcelona, Spain . (EI)

[3] Xuefeng Chen, Xiabi Liu, Yunde Jia. A Soft Target Method of Learning Posterior Pseudo-probabilities based Classifiers with its Application to Handwritten Digit Recognition, The 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008), Aug. 19-21, 2008, Montréal, Québec, Canada.

[4] Xuefeng Chen, Xiabi Liu, Yunde Jia. Learning Handwritten Digit Recognition by the Max-Min Posterior Pseudo-Probabilities Method, in Proc. of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 342-346, 2007.9.23-26, Curitiba, Brazil . (EI)

[5] Xuefeng Chen, Xiabi Liu, Yunde Jia. Soft Learning Method of Posterior Pseudo-probabilities Based Classifiers, Submitted to Patten Recognition.

[6] Xuefeng Chen, Xiabi Liu, Yunde Jia. A Hybrid Optimization Method of Evolution Strategy and Gradient Descent for Discriminative Learning of Bayesian Classifiers, Submitted to Information Sciences.

[7] Xuefeng Chen, Xiabi Liu, Yunde Jia. Discriminative Selection and Learning of Gaussian Mixture Models for Pattern Classification, Submitted to Neurocomputing

[8] Xiabi Liu, Yunde Jia, Xuefeng Chen. Image Classification Using the Max-Min Posterior Pseudo-Probabilities Method, Submitted to Pattern Recognition.

攻读博士学位期间参加的科研项目

国家自然科学基金项目（No. 60973059）

国家“九七三”重点基础研究发展规划项目(No. 2006CB303103)

国家 863 高技术项目(No.2006AA01Z120)

微软研究院国际合作项目(No.FY08-RES-THEME-158)

致谢

时光荏苒，艰辛而又有收获的博士生活即将结束。在这段时光中，许多人的支持和帮助成就了我研究工作的完成，我在此谨向他们表示衷心的感谢。

首先要感谢贾云得教授几年来对我的悉心指导。在读博期间，贾老师从学业、生活等各个方面给予我莫大的鼓励和帮助。从论文的选题到研究，直到论文的攥写、修改，我的每一个进步都凝聚着贾老师的无数心血。贾老师刻苦认真地工作作风、严谨求实的治学态度、平易近人的处事风范深深的影响着我，在今后的人生道路上将用永远激励着我不断前进。

衷心感谢刘峡壁博士。刘老师不仅在我课题研究中给予我无私的指导和帮助，他还教给我许多为人处事的道理，使我受益终生，在此向刘老师表示最诚挚的谢意！

感谢媒体计算与智能系统实验室为我们提供了宽松的学习环境。感谢实验室的各位老师和同学共同营造了良好的学习氛围，并在生活和工作上给予我很大的帮助和支持。特别感谢同组的王彦杰和邓元同学，和他（她）们在工作和生活方面的讨论中使我受益非浅。

最后要感谢我的家人，他们始终如一的支持让我能够安心的完成学业，我只能用更多更好的成绩来回报他们。

作者简介

1979 年 11 月 21 日出生于吉林省。

1998 年 9 月考入长春理工大学计算机科学与技术专业，2002 年 7 月本科毕业并获得工学学士学位。

2002 年 9 月考入长春理工大学计算机科学与技术专业攻读工硕士，2005 年 7 月硕士毕业并获得工学硕士学位。

2005 年 9 月考入北京理工大学计算机科学与技术专业攻读工学博士至今，师从贾云得教授。