基于互信息梯度优化计算的信息判别特征提取

谢文彪[①②]　樊绍胜[①]　费洪晓[②]　樊晓平[②]
[①](长沙理工大学电气与信息工程学院　长沙　410004)
[②](中南大学信息科学与工程学院　长沙　410083)

摘　要：该文将互信息梯度优化引入特征提取矩阵求解，提出一种信息判别分析的特征提取方法。首先，分析了现有线性判别方法的特点和局限，建立了类条件分布参数模型下互信息最大化的信息判别模型。其次，证明了互信息判别的线性变换不变性和贝叶斯一致优化，构造了一个互信息梯度优化计算的特征提取算法。最后通过实际数据上试验验证了该方法的有效性。

关键词：特征提取；信息判别分析；线性变换；互信息梯度
中图分类号：TP391.4；TP274+.3　　**文献标识码**：A　　**文章编号**：1009-5896(2009)12-2975-05

Information Discriminant Feature Extraction Based on Mutual Information Gradient Optimal Computation

Xie Wen-biao[①②]　Fan Shao-sheng[①]　Fei Hong-xiao[②]　Fan Xiao-ping[②]
[①](*School of Electrical & Information Engineering*, *Changsha University of Science and Technology*, *Changsha* 410004, *China*)
[②](*School of Information Science and Engineering*, *Central South University*, *Changsha* 410083, *China*)

Abstract: A linear feature extraction method is present with information discriminant analysis, it is based on a feasible computationally feature extraction matrix used mutual information gradient. Firstly, this paper analyzes the limitation for current linear discriminant, and constructs a information discriminant analysis model which facilitates the maximization of the mutual information under the parametric class-conditional PDF. Then, it is proved that the mutual information is linear transformation invariance and optimal in the sense of Bayes, and the algorithm is present for computing feature extraction matrix with mutual information gradient. Finally, the good performance of the method is proved on real-world data set.

Key words: Feature extraction; Information discriminant analysis; Linear transformation; Mutual information gradient

1　引言

特征提取是模式识别中十分重要的研究内容，它不但能降低数据维数，提高识别速度，更重要的是当训练样本有限、数据维数高时能够克服维数灾难，提高识别精度。在保证分类精度的情况下，寻找最小数目的特征集来表达数据已经成为统计模式识别领域中的一个重要课题[1]。在特征提取领域中，线性判别分析有着重大影响，最典型的线性判别算法由 Fisher 提出，Rao 在此基础上有两类判据扩展到多类判据，之后人们对此方法进行了大量的研究，提出了多个扩展方案。在一定条件下，Fisher 线性判据是一种计算代价小的最优判据。近年来，随着计算机处理技术的发展，Fisher 线性判据广泛应用于高维数据的特征提取和分类判别。Marco[2]等人在 Fisher 线性判别特征提取的基础上提出权重可变的类可分性判据，其后和 Robert[3]合作，针对异方差问题提出一种基于 Chernoff 距离的类可分性判据。

作为高维数据分离度度量的有效工具，互信息建立了特征提取向量和数据分类信息的内在关系，产生了特征提取的信息判据分析方法[4]。文献[5]在分析特征向量和分类判别关系的基础上，在判据目标函数中引入互信息的罚函数机制。文献[6]通过启发式迭代优化进行混合模型的极大似然拟合，一定程度上克服了罚函数的过度拟合。随着信息判别特征提取的广泛研究，互信息作为描述高维数据特征提取下的分类判据成功应用于脑信号分析[7]、文本分类[8]及音频信号识别[9]等统计模式识别问题。

本文从高维数据特性出发，提出一个基于互信息的分类判别算法框架。该框架同时考虑了高斯分布下数据均值和方差的类可分性，并证明了 Fisher 线性判据和文献[3]的算法都是该算法的特例。文章

2009-01-16 收到，2009-06-18 改回
国家 863 计划项目(2008AA04Z214)资助课题

通过引入互信息梯度给出了一个基于共轭梯度优化求解特征提取矩阵的算法，克服了罚函数和极大似然拟合复杂计算。

2 信息判别分析

由于 Fisher 线性判别分析计算上的简单性，广泛应用于判别特征提取。文献[10]在介绍各种变形的基础上给出了统一的定义：

$$J(\boldsymbol{T})=\left|\boldsymbol{T}\boldsymbol{\Sigma}_B\boldsymbol{T}^{\mathrm{T}}\right|\Big/\left|\boldsymbol{T}\boldsymbol{\Sigma}_W\boldsymbol{T}^{\mathrm{T}}\right| \tag{1}$$

其中 $\boldsymbol{\Sigma}_B=\sum_{i=1}^{c}p_i(\boldsymbol{m}_i-\bar{\boldsymbol{m}})(\boldsymbol{m}_i-\bar{\boldsymbol{m}})^{\mathrm{T}}$，$\boldsymbol{\Sigma}_W=\sum_{i=1}^{c}p_i\Sigma_i$ 分别为类间散度矩阵和类内散度矩阵，$\overline{\boldsymbol{R}}\triangleq\boldsymbol{T}\boldsymbol{R}$，$\boldsymbol{T}:IR^n\to IR^m$ 为满秩特征提取矩阵。但异方差时该判别存在较大误差，因此文献[3]提出了一种基于 Chernoff 距离的异方差判别特征提取算法，但该算法没有考虑均值对类可分性的影响，很难区分重叠数据。

互信息是信息论中的一个基本概念，假设随机向量 $\boldsymbol{R}\in IR^n$，分类参数向量空间 $\boldsymbol{\Omega}$，则它们之间的互信息记为 $\mu(\boldsymbol{R};\boldsymbol{\Omega})$，定义互信息如下：

$$\mu(\boldsymbol{R};\boldsymbol{\Omega})\triangleq H(\boldsymbol{R})-H(\boldsymbol{R}|\boldsymbol{\Omega})=H(\boldsymbol{R})-\sum_{i=1}^{c}H\left(\boldsymbol{R}|\omega_i\right)p_i \tag{2}$$

其中 $H(\boldsymbol{R})\triangleq-\int_R f_R(\boldsymbol{r})\lg\left(f_R(\boldsymbol{r})\right)\mathrm{d}\boldsymbol{r}$，$f_R(\boldsymbol{r})\triangleq\sum_{i=1}^{c}f_{R|\Omega}\left(\boldsymbol{r}|\omega_i\right)p_i$，$f_{R|\Omega}\left(\boldsymbol{r}|\omega_i\right)$ 为随机向量的条件分布，ω_i，p_i 为类 i 的分布参数和先验概率。互信息度量依赖于概率分布，是一种描述分类参数与数据概率关系的有力工具[4]。因此自然定义了一个分类判据：高的互信息量能产生高精度的分类。潜在变量结构理论证明高维数据降维后具有高斯分布特性[11]，依据这一假设，式(3)可以改写为

$$\begin{aligned}\mu(\boldsymbol{R};\boldsymbol{\Omega})&\triangleq H_g(\boldsymbol{R})-H(\boldsymbol{R}|\boldsymbol{\Omega})\\&=H_g(\boldsymbol{R})-\sum_{i=1}^{c}H\left(\boldsymbol{R}|\omega_i\right)p_i\end{aligned} \tag{3}$$

引理 1[12](高斯分布熵) 具有均值 $\boldsymbol{\mu}$ 和方差 $\boldsymbol{\Sigma}$ 多元高斯分布数据集 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$，则

$$H(\boldsymbol{X})=H(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)==\frac{1}{2}\lg\left((2\pi e)^n|\Sigma|\right) \tag{4}$$

由引理 1 有 $H_g(\boldsymbol{R})=\frac{1}{2}\lg\left((2\pi e)^n|\Sigma|\right)$，$H\left(\boldsymbol{R}|\omega_i\right)=\frac{1}{2}\lg\left((2\pi e)^n|\Sigma_i|\right)$，由高斯混合模型定义的性质有数据总体方差：$\Sigma=\sum_{i=1}^{c}\left[\Sigma_i+(\boldsymbol{m}_i-\bar{m})(\boldsymbol{m}_i-\bar{\boldsymbol{m}})^{\mathrm{T}}\right]p_i$。通过观察可以发现系数 $(2\pi e)^n$ 只与分类数相关，在监督聚类中为常数，于是简化得到本文互信息判据计算公式：

$$\mu(\boldsymbol{R};\boldsymbol{\Omega})=\frac{1}{2}\left[\lg\left(|\Sigma|\right)-\sum_{i=1}^{c}\lg\left(|\Sigma_i|\right)p_i\right] \tag{5}$$

类别差异度是类可分性最有效的度量[13]，在概率条件分布下，均值和方差是表征类别差异的两个统计量，下面从同方差和均值近似分别考察均值和方差对类可分性的影响。在同方差条件下有 $\Sigma_1=\Sigma_2=\cdots=\Sigma_c$，则互信息计算公式可以简化为

$$\mu(\boldsymbol{R},\boldsymbol{\Omega})=\frac{1}{2}\lg\frac{\left|\boldsymbol{\Sigma}_B+\boldsymbol{\Sigma}_W\right|}{\left|\boldsymbol{\Sigma}_W\right|} \tag{6}$$

其中 $\boldsymbol{\Sigma}_B\triangleq\sum_{i=1}^{C}\left[(\boldsymbol{m}_i-\bar{\boldsymbol{m}})(\boldsymbol{m}_i-\bar{\boldsymbol{m}})^{\mathrm{T}}\right]p_i$，$\boldsymbol{\Sigma}_W\triangleq\sum_{i=1}^{C}\Sigma_i p_i$ 分别为类间散度和类内散度，这是一个 Fisher 线性判别函数,所以在同方差下该互信息判据等同于 Fisher 线性判据。

在考虑均值相似情况下，两类互信息计算公式可以改写为

$$\mu(\boldsymbol{R};\boldsymbol{\Omega})=\frac{1}{2}\lg\frac{\left|\boldsymbol{\Sigma}_W+(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}\,p_1p_2\right|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}} \tag{7}$$

其中 $\boldsymbol{m}=\sum_{i=1}^{2}\boldsymbol{m}_i p_i$，$\boldsymbol{\Sigma}_W=\sum_{i=1}^{2}\Sigma_i p_i$。式(7)进一步修改为

$$\begin{aligned}\mu(\boldsymbol{R};\boldsymbol{\Omega})=&\frac{1}{2}\lg\frac{\left|\boldsymbol{\Sigma}_W\right|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}}+\frac{1}{2}\log\Big(1+(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}\\&\cdot\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_2)p_1p_2\Big)\end{aligned} \tag{8}$$

由二阶泰勒级数 $\|\boldsymbol{x}\|\approx 0$，$\lg(1+\alpha\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x})\approx\alpha\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}$。于是式(8)最终修改为

$$\begin{aligned}\mu(\boldsymbol{R};\boldsymbol{\Omega})=&\frac{1}{2}\lg\frac{\left|\boldsymbol{\Sigma}_W\right|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}}+\frac{p_1p_2}{2}\\&\cdot(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_2)\end{aligned} \tag{9}$$

这就是文献[3]所提出的 Chernoff 距离类可分性判据。由式(6)和式(9)可知，Fisher 线性判别和基于 Chernoff 距离的异方差判别都是本文式(5)的特例，这说明互信息具有良好的统计特性，是一种通用的类可分性判据。

3 基于互信息的判别特征提取

互信息的引入建立了高维数据和分类信息间内在的关系，大大提高了数据分类的精确度。但在特征提取时极大似然和罚函数的方法一方面需要计算量很大的迭代运算，另一方面他们的引入往往容易陷入局部收敛。因此有必要利用信息论原理，进一步研究线性变换，得到计算简单和物理意义更加明确的特征提取信息判据。

3.1 互信息的线性变换定理

引理 2[12](线性变换熵) 存在系数 a 有

$$H(a\boldsymbol{X}) = H(\boldsymbol{X}) + \lg(|a|) \tag{10}$$

推论 1(矩阵变换熵)　由引理 2 容易得到 $H(\boldsymbol{AX}) = H(\boldsymbol{X}) + \lg(|\boldsymbol{A}|)$。

定理 1(空间不变性)　在可逆线性变换下，信息判别函数值具有不变性。即

$$\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{R};\boldsymbol{\Omega}) \tag{11}$$

其中 $\overline{\boldsymbol{R}} \triangleq \boldsymbol{TR}$，$\boldsymbol{T}: IR^n \to IR^n$ 是一个非奇异变换矩阵。

证明　由推论 1 可得到，对于任意非奇异矩阵 $\boldsymbol{T}$，有 $H(\boldsymbol{TR}) = H(\boldsymbol{TR}) + \lg(|\boldsymbol{T}|)$，根据定义有：$\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = H_g(\boldsymbol{R}) + \lg(|\boldsymbol{T}|) - [H(\boldsymbol{R}|\boldsymbol{\Omega}) + \lg(|\boldsymbol{T}|)] = \mu(\boldsymbol{R};\boldsymbol{\Omega})$。　证毕

定理 2(正交不变性)　对于任意满秩变换矩阵 $\boldsymbol{T}: IR^n \to IR^m$ $(m<n)$，存在一个正交变换矩阵 $\boldsymbol{E} \in IR^{m\times n}$，使得 $\overline{\boldsymbol{R}} \triangleq \boldsymbol{TR}$，$\widetilde{\boldsymbol{R}} \triangleq \boldsymbol{ER}$ 有

$$\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) \tag{12}$$

证明　根据奇异值分解理论，存在正交矩阵 $\boldsymbol{U} \in IR^{m\times m}$ 和 $\boldsymbol{V} \in IR^{n\times n}$ 使得 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{TV} = [\boldsymbol{\Lambda}|0_{m\times d}]$，其中 $\boldsymbol{\Lambda} \in IR^{m\times m}$ 是由变换矩阵 $\boldsymbol{T}$ 的特征值构造的对角矩阵，$d \triangleq n-m$，$\mathrm{rank}(\boldsymbol{T}) = m$。令 $\widetilde{\boldsymbol{R}} \triangleq \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}\overline{\boldsymbol{R}} \triangleq \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{TR}$，由定理 1 有 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega})$。又令 $\boldsymbol{E} \triangleq \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{T} \in IR^{m\times n}$，则 $\boldsymbol{EV} = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{TV} = [\boldsymbol{I}_m|0_{m\times d}]$，有 $\boldsymbol{EV}(\boldsymbol{EV})^{\mathrm{T}} = \boldsymbol{EVV}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}} = \boldsymbol{EE}^{\mathrm{T}} = \boldsymbol{I}_m$，$\boldsymbol{E}$ 是一个正交矩阵。　证毕

定理3(互信息的单调性)　在线性变换下信息判据函数值具有非增性，即

$$\mu(\widehat{\boldsymbol{R}};\boldsymbol{\Omega}) \le \mu(\boldsymbol{R};\boldsymbol{\Omega}) \tag{13}$$

其中 $\widehat{\boldsymbol{R}} \triangleq \boldsymbol{TR}$，$\boldsymbol{T}: IR^n \to IR^m$, $(m<n)$ 是一个满秩矩阵。

证明　假设 $\boldsymbol{T}^{\perp} \in IR^{(n-m)\times n}$ 为矩阵 $\boldsymbol{T}$ 的行生成零矩阵，定义变换矩阵 $\widetilde{\boldsymbol{T}}$ 和随机向量 $\widetilde{\boldsymbol{R}}$：

$$\widetilde{\boldsymbol{T}} \triangleq \begin{bmatrix}\boldsymbol{T}\\ \boldsymbol{T}^{\perp}\end{bmatrix} \in IR^{n\times n},\quad \widetilde{\boldsymbol{R}} \triangleq \widetilde{\boldsymbol{T}}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{TR}\\ \boldsymbol{T}^{\perp}\boldsymbol{R}\end{bmatrix} \triangleq \begin{bmatrix}\widehat{\boldsymbol{R}}\\ \widehat{\boldsymbol{N}}\end{bmatrix}$$

显然 $\widetilde{\boldsymbol{T}}$ 是满秩矩阵，有 $\mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{R};\boldsymbol{\Omega})$，且：$\mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\widehat{\boldsymbol{R}},\widehat{\boldsymbol{N}};\boldsymbol{\Omega}) = H_g(\widehat{\boldsymbol{R}},\widehat{\boldsymbol{N}}) - H(\widehat{\boldsymbol{R}},\widehat{\boldsymbol{N}}|\boldsymbol{\Omega})$　根据熵的链式法则和高斯条件熵有 $H_g(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}}) \ge H(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}}) \ge H(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}},\boldsymbol{\Omega})$，则 $\mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) = H_g(\widehat{\boldsymbol{R}}) + H_g(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}}) - [H(\widehat{\boldsymbol{R}}|\boldsymbol{\Omega}) + H(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}},\boldsymbol{\Omega})] = \mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) + H_g(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}}) - H(\widehat{\boldsymbol{N}}|\widehat{\boldsymbol{R}},\boldsymbol{\Omega}) \ge \mu(\widehat{\boldsymbol{R}};\boldsymbol{\Omega})$ 有 $\mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{R};\boldsymbol{\Omega}) \ge \mu(\widehat{\boldsymbol{R}};\boldsymbol{\Omega})$。　证毕

以上定理解释了数据线性变换中互信息不变性，为基于互信息判据函数在数据特征提取中的应用提供了严格的数学准则，能体现数据处理上更好的物理意义。

3.2 贝叶斯一致优化

贝叶斯分类器作为判别特征提取性能评估工具，已成为事实上的标准。本节首先介绍贝叶斯分类误差的计算：

$$P_R(\varepsilon) = 1 - \sum_{i=1}^{c}\int_{R_i} f_{R|\Omega}(\boldsymbol{r}|\omega_i)\,p_i\mathrm{d}\boldsymbol{r} \tag{14}$$

其中 $f_{R|\Omega}(\boldsymbol{r}|\omega_i)$ 为类条件分布，$\boldsymbol{R} \in IR^n$ 和 $\boldsymbol{\Omega} = \{\omega_1, \omega_2, \cdots, \omega_c\}$ 分别为数据分布接受域和分布参数向量，p_i 分类先验概率。因此有分类判据：

$$i^* = \arg\max_{1\le i\le c} P(\omega_i|\boldsymbol{r}_0) \tag{15}$$

其中 $\boldsymbol{r}_0$ 为无标记的观测数据，$P(\omega_i|\boldsymbol{r}_0)$ 为分类后验概率。于是定义 $\varepsilon_R \triangleq \min\ P_R(\varepsilon)$ 贝叶斯误差，且有可逆线性变换不变性。

为分析互信息线性变换的贝叶斯优化，依据数据高斯分布，将数据分解成独立高斯信号空间和噪声空间，即 $\boldsymbol{S}|\boldsymbol{\Omega} \in IR^m$ 和 $\boldsymbol{N}|\boldsymbol{\Omega} \in IR^d$ 且 $f_{N|S,\Omega}(\boldsymbol{n}|\boldsymbol{s}, \omega_i) = f_{N|S}(\boldsymbol{n}|\boldsymbol{s})$，$\forall i=\{1,2,\cdots,c\}$，$\forall \boldsymbol{s} \in IR^m$，$\forall \boldsymbol{n} \in IR^d$。则有

$$\boldsymbol{R}|\omega_i = \boldsymbol{M}\begin{bmatrix}\boldsymbol{S}|\omega_i\\ \boldsymbol{N}|\omega_i\end{bmatrix},\quad \forall i=\{1,2,\cdots,c\} \tag{16}$$

其中 $\boldsymbol{M} \in IR^{n\times n}$ 非奇异矩阵。令 $\overline{\boldsymbol{R}} \triangleq \boldsymbol{TR}$，$\boldsymbol{T} \in IR^{m\times n}$ 是使 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega})$ 值最大的满秩矩阵，由定理 1 有 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{R};\boldsymbol{\Omega})$，$\varepsilon_R = \varepsilon_{\overline{R}}$。令 $\boldsymbol{T}^{\perp} \in IR^{(n-m)\times n}$ 为矩阵 $\boldsymbol{T}$ 的零行生成矩阵，则有

$$\widetilde{\boldsymbol{T}} \triangleq \begin{bmatrix}\boldsymbol{T}\\ \boldsymbol{T}^{\perp}\end{bmatrix} = \boldsymbol{M}^{-1} \in IR^{n\times n},\ \widetilde{\boldsymbol{R}} \triangleq \widetilde{\boldsymbol{T}}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{TR}\\ \boldsymbol{T}^{\perp}\boldsymbol{R}\end{bmatrix} \triangleq \begin{bmatrix}\boldsymbol{S}\\ \boldsymbol{N}\end{bmatrix}$$

由定理 1 可知 $\mu(\widetilde{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{R};\boldsymbol{\Omega})$，且 $\varepsilon_{\widetilde{R}} = \varepsilon_R$。依据链式熵法则有

$$\begin{aligned}\mu(\boldsymbol{R};\boldsymbol{\Omega}) &= \mu(\boldsymbol{S},\boldsymbol{N};\boldsymbol{\Omega})\\ &= \mu(\boldsymbol{S};\boldsymbol{\Omega}) + H_g(\boldsymbol{N}|\boldsymbol{S}) - H(\boldsymbol{N}|\boldsymbol{S},\boldsymbol{\Omega})\end{aligned}$$

又由 $f_{N|S}(\boldsymbol{n}|\boldsymbol{s}) = f_{N|S,\Omega}(\boldsymbol{n}|\boldsymbol{s},\omega_i)$ 和条件熵原理，有 $\boldsymbol{N}$ 独立于 $\boldsymbol{\Omega}$，即给定 $\boldsymbol{S}$，有 $H(\boldsymbol{N}|\boldsymbol{S},\boldsymbol{\Omega}) = H(\boldsymbol{N}|\boldsymbol{S})$，得 $\mu(\boldsymbol{R};\boldsymbol{\Omega}) = \mu(\boldsymbol{S};\boldsymbol{\Omega})$，则 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{S};\boldsymbol{\Omega})$。依据贝叶斯原理和 $f_{N|S}(\boldsymbol{n}|\boldsymbol{s}) = f_{N|S,\Omega}(\boldsymbol{n}|\boldsymbol{s},\omega_i)$ 有

$$\begin{aligned}P(\omega_i|\boldsymbol{r}) &= \frac{f_{R|\Omega}(\boldsymbol{r}|\omega_i)}{f_R(\boldsymbol{r})} = \frac{f_{N|S,\Omega}(\boldsymbol{n}|\boldsymbol{s},\omega_i)\,f_{S|\Omega}(\boldsymbol{s}|\omega_i)\,p_i}{f_R(\boldsymbol{r})}\\ &= \frac{f_{N|S}(\boldsymbol{n}|\boldsymbol{s})\,f_{S|\Omega}(\boldsymbol{s}|\omega_i)\,p_i}{f_{N,S}(\boldsymbol{n},\boldsymbol{s})} = \frac{f_{S|\Omega}(\boldsymbol{s}|\omega_i)\,p_i}{f_S(\boldsymbol{s})}\\ &= P(\omega_i|\boldsymbol{s}),\ \forall i=\{1,2,\cdots,c\},\ \forall \boldsymbol{r} \in IR^n\end{aligned}$$

即 $\varepsilon_R = \varepsilon_S$，则 $\varepsilon_{\overline{R}} = \varepsilon_S$。则通过 $\boldsymbol{T}$ 信号提取的信号空间有 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega}) = \mu(\boldsymbol{S};\boldsymbol{\Omega})$，且有贝叶斯一致优化判别率 $\varepsilon_{\overline{R}} = \varepsilon_S$。

3.3 互信息的特征提取判别

由上所述，给定数据 $\boldsymbol{R} \in IR^n$ 和维数为 m 的特征向

量空间，能够找到一个满秩的变换矩阵 $\boldsymbol{T}\in IR^{m\times n}$，使得互信息 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega})$ 最大，并保证贝叶斯一致优化，即

$$\boldsymbol{T}^{*}=\arg\max_{\boldsymbol{T}\in IR^{m\times n}}\left\{\mu\left(\overline{\boldsymbol{R}};\boldsymbol{\Omega}\right):\overline{\boldsymbol{R}}=\boldsymbol{T}\boldsymbol{R}\right\} \tag{17}$$

其中是 $\boldsymbol{T}$ 一个满秩矩阵。由式(5)得基于互信息的特征提取判别函数

$$\mu\left(\overline{\boldsymbol{R}};\boldsymbol{\Omega}\right)=\frac{1}{2}\left[\lg\left(\left|\boldsymbol{T}\Sigma\boldsymbol{T}^{\mathrm{T}}\right|\right)-\sum_{i=1}^{c}\lg\left(\left|\boldsymbol{T}\Sigma_{i}\boldsymbol{T}^{\mathrm{T}}\right|\right)p_{i}\right] \tag{18}$$

由互信息的线性变换定理和贝叶斯一致优化分析，本文特征提取将原始数据 $\boldsymbol{R}$ 分解成相互独立的高斯数据空间 $\boldsymbol{N}$ 和噪声空间 $\boldsymbol{S}$，通过选择合适的特征提取矩阵 $\boldsymbol{T}$ 能够保证数据空间的互信息不变，并且具有贝叶斯一致优化。

由矩阵链式微分计算[10]，存在对称正定矩阵 $\boldsymbol{\Sigma}\in IR^{n\times n}$ 和满秩矩阵 $\boldsymbol{T}\in IR^{m\times n}$ 有

$$\frac{\partial\lg\left(\left|\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^{\mathrm{T}}\right|\right)}{\partial\boldsymbol{T}}=2\left(\left|\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^{\mathrm{T}}\right|\right)^{-1}\boldsymbol{T}\boldsymbol{\Sigma} \tag{19}$$

于是定义特征提取的互信息梯度 $g(\boldsymbol{T})=\partial\mu/\partial\boldsymbol{T}\in IR^{m\times n}$：

$$g(\boldsymbol{T})=\frac{\partial\mu\left(\overline{\boldsymbol{R}};\boldsymbol{\Omega}\right)}{\partial\boldsymbol{T}}=\left(\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^{\mathrm{T}}\right)^{-1}\cdot\boldsymbol{T}\boldsymbol{\Sigma}-\sum_{i=1}^{c}\left(\boldsymbol{T}\boldsymbol{\Sigma}_{i}\boldsymbol{T}^{\mathrm{T}}\right)^{-1}\boldsymbol{T}\boldsymbol{\Sigma}_{i}p_{i} \tag{20}$$

依据式(18)和式(20)计算互信息 $\mu(\overline{\boldsymbol{R}};\boldsymbol{\Omega})$ 及互信息梯度 $g(\boldsymbol{T})$，利用文献[14]针对向量空间提出一种快速共轭梯度法求解最佳特征提取矩阵。

4 实验及结果分析

本文采用 UCI 通用的高维分类判别数据集 Landsat Satellite 作为实验数据[15]，该数据集共有 6345 个 36 维数据，分为 6 类，其中 4435 作为固定的训练数据，其他为测试数据。为验证算法，在 MATAB7.0 下运用平方(Quadratic)分类器(classify)训练和测试算法，通过选用不同的特征向量规模(m)与文献[3]及典型 Fisher 线性判别的算法作比较，其结果如表 1 所示。

表 1 分类准确率：算法+分类器联合测试 Landsat Satellite 数据集

算法	特征向量规模(m)									
	3	5	9	10	11	12	18	32	34	36
Fisher	86.75	87.30	87.70	87.80	87.65	+87.70	87.20	87.05	**86.95**	**86.85**
文献[3]	87.15	87.45	+87.85	87.40	87.35	87.70	**87.30**	**87.20**	**86.95**	**86.85**
本文	**87.25**	**87.55**	**87.90**	**88.10**	***+88.50**	**88.00**	87.25	87.15	86.90	**86.85**

从表 1 实验结果可以看出：文献[3]和本文算法在分类准确度上优于 Fisher 算法，这表明考虑到了异方差对分类的影响，能提高分类精度；本文算法在低维情况下能够得到较高的分类精度，并且在 m=11 时获得最高分类准确率，表明互信息判据能够考虑到均值对类可分性的影响能更精确描述高维数据的分类信息，在降维下符合高斯分布；在选取合适的降维特征提取，文献[3]的算法(m=9)和本文算法(m=11)都能得到最佳的分类精度。从图 1 试验结果可以看出，在小特征规模下本文算法时间代价小于文献[3]的算法，说明作为特征提取判据本文的互信息梯度优化在时间上有一定优势。

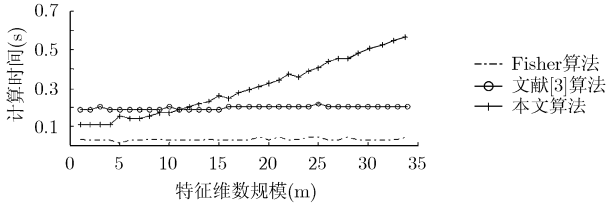

图 1 计算特征提取矩阵时间代价

5 结论

本文在研究 Fisher 线性分类判别和基于互信息的信息判别的基础上提出了一种基于互信息梯度的特征提取矩阵求解算法，克服了混合模型的极大似然拟合和罚函数复杂迭代计算。本文提取的互信息判别同时考虑到方差和均值对类可分析的影响，并证明 Fisher 线性判别和文献[3]异方差判别是本文提出的互信息判别函数的特例。最后通过实验验证了算法，下一步将研究特征空间大小(m)对分类性能的影响，以期获取最佳的分类性能。

参 考 文 献

[1] 王文胜, 陈伏兵, 杨静宇. 一种基于奇异值分解的特征抽取方法[J]. 电子与信息学报, 2005, 27(2): 294–297.
Wang Wen-sheng, Chen Fu-bing, and Yang Jing-yu. A method of feature extraction based on SVD [J]. *Journal of Electronics & Information Technology*, 2005, 27(2): 294–297.

[2] Loog M, Duin R P W, and Haeb-Umbach R. Multiclass linear dimension reduction by weighted pairwise Fisher criteria [J]. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 2001, 23(7): 762–766.

[3] Loog M and Duin R P W. Linear dimensionality reduction via a heteroscedastic extension of LDA: The chernoff criterion [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(6): 732–739.

[4] Hild II K E, Erdogmus D, and Torkkola K. Feature extraction using information-theoretic learning [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(9): 1385–1392.

[5] Padmanabhan M and Dharanipragada S. Maximizing information content in feature extraction [J]. *IEEE Transactions on Speech Audio Processing*, 2005, 13(4): 512–519.

[6] Leiva-Murillo J M and Artés-Rodríguez A. Maximization of mutual information for supervised linear feature extraction [J]. *IEEE Transactions on Neural Networks*, 2007, 18(5): 1433–1440.

[7] Grosse-Wentrup M and Buss M. Multiclass common spatial patterns and information theoretic feature extraction [J]. *IEEE Transactions on Biomedical Engineering*, 2008, 55(8): 1991–2000.

[8] 徐燕, 李锦涛, 王斌等. 文本分类中特征选择的约束研究[J]. 计算机研究与发展, 2008, 45(4): 596–602.
Xu Yan, Li Jin-tao, and Wang Bin. A study on constraints for feature selection in text categorization [J]. *Journal of Computer Research and Development*, 2008, 45(4): 596–602.

[9] 陈刚, 陈莘萌. 一种考虑类别信息的音频特征提取方法[J].计算机研究与发展, 2006, 43(11): 1959–1964.
Chen Gang and Chen Xin-meng. An audio feature extraction method taking class information into account [J]. *Journal of Computer Research and Development*, 2006, 43(11): 1959–1964.

[10] Fukunaga K. Introduction to Statistical Pattern Recognition [M]. New York: Academic Press, 1990: Appendix A (Derivatives of Matrices)

[11] Rasmussen C E and Williams C K I. Gaussian Processes for Machine Learning [M]. Massachusetts. The MIT Press, 2006: 171–188.

[12] Cover T M and Thomas J A. Elements of Information Theory (Second Edition) [M]. New Jersey. Wiley InterScience, 2006: 243–260.

[13] 罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报, 2008, 30(8): 1315–1324.
Luo Hui-lan, Kong Fan-sheng, and Li Yi-xiao. An analysis of diversity measures in clustering ensembles [J]. *Chinese Journal of Computers*, 2008, 30(8): 1315–1324.

[14] Frandsen P E, Jonasson K, Nielsen H B, and Tingleff O. Unconstrained Optimization (Third Edition) [M]. Denmark. IMM, 2004, Chapter 4.

[15] Asuncion A and Newman D. UCI repository of machine learning databases [R]. http://archive.ics.uci.edu/ml/, 2008.

谢文彪：　男，1980 年生，讲师，研究方向为数据特征提取、分布拟合及其优化.

樊绍胜：　男，1966 年生，教授，研究方向为智能控制.

费洪晓：　男，1968 年生，副教授，研究方向为智能信息处理.