文章编号：1001-0920(2009)12-1810-06

一种可最优化计算特征规模的互信息特征提取

谢文彪[1,2]，樊绍胜[1]，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：利用矩阵特征向量分解，提出一种可最优化计算特征规模的互信息特征提取方法. 首先，论述了高斯分布假设下的该互信息判据的类可分特性，并证明了现有典型算法都是本算法的特例；然后，在给出该互信息判据严格的数学意义基础上，提出了基于矩阵特征向量分解计算最优化特征规模算法；最后，通过实际数据验证了该方法的有效性.

关键词：互信息判据；特征提取；特征规模；矩阵特征向量分解
中图分类号：TP274 **文献标识码**：A

Optimization calculation feature scale for mutual information measure feature extraction

XIE Wen-biao[1,2], *FAN Shao-sheng*[1], *FAN Xiao-ping*[2]
(1. School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410004, China; 2. School of Information Science and Engineering, Central South University, Changsha 410083, China. Correspondent: XIE Wen-biao, E-mail: xiewenbiao@tom.com)

Abstract: By using matrix eigenvalue/eigenvector decomposition, an optimization calculation of feature scale is proposed for mutual information measure feature extraction technique. The novel technique based on Gaussian distribution enjoys a good class-separability property with the mutual information. It is proved that the existing algorithm is the specifical example. Then based on the strict mathematic significance of mutual information measure, the method for optimization calculation of the feature scale based on matrix eigenvalue/eigenvector decomposition is given. Finally, real-world data sets demonstrat the effectiveness of the method.

Key words: Mutual information measure; Feature extraction; Feature scale; Matrix eigenvalue/eigenvector decomposition

1 引 言

数据降维是模式识别中十分重要的研究内容，它不但能提高识别速度，更重要的是当训练样本有限或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以降低表达数据的向量维数，已成为模式识别领域中的一个重要课题. 在数据降维领域中，线性判据有着重大影响. 最典型的线性判据算法由 Fisher[1] 提出；Rao[2] 在此基础上将两类判据扩展到多类判据；之后，人们对该方法进行了大量的研究，提出了多个改进方案[3]. 在一定条件下，Fisher 线性判据是一种计算简单的最优判据. 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. 但 Fisher 是在假设数据具有相同方差条件下进行的，无法处理异方差数据. Marco 等[4] 在 Fisher 线性特征提取判据的基础上，提出了权重可变的多类可分性判据；其后又与 Robert 合作[5]，针对异方差问题，提出一种基于 Chernoff 距离类可分性判据，并很自然地扩展到多类判据.

作为高维数据可分性度量的有效工具，互信息被引入特征提取的类可分性判据目标函数中，并产生了特征提取的信息判据分析方法[6]，成功地解决了异方差数据降维问题. 文献[7]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，并通过迭代优化求解特征提取

收稿日期：2009-01-14；**修回日期**：2009-05-01.
基金项目：国家 863 计划项目(2008AA04Z214).
作者简介：谢文彪(1980—)，男，湖南宁乡人，讲师，硕士，从事统计模式识别的研究；樊晓平(1961—)，男，浙江绍兴人，教授，博士生导师，从事机器人、模式识别等研究.

矩阵. 文献[8]假设高维数据特征提取分类具有高斯分布，通过启发式迭代计算类可分性判据函数，克服了罚函数的过度拟合. 最近 Zoran[9] 提出了一种数值优化方法，直接通过数据参数计算特征提取矩阵，在一定程度上克服了互信息判据的计算复杂问题. 互信息作为描述高维数据特征提取分类判据，已成功地解决了诸如脑信号分析[10]、文本分类[11]、图像识别[12]及音频信号识别[13]等模式识别问题.

本文从高维数据降维特性出发，提出一种基于矩阵特征向量分解的互信息特征提取方法，并可最优化计算特征规模.

2　异方差数据的类可分性判据

2.1　Fisher 线性判据

Fisher 线性判据因其分析计算上的简单性而广泛地应用于判别特征的提取. 文献[14]在介绍各种变形的基础上给出了统一的定义，即

$$J_F(T)=\mathrm{tr}((T\Sigma_B T^{\mathrm{T}})(T\Sigma_W T^{\mathrm{T}})^{-1}). \tag{1}$$

其中：$\Sigma_B=\sum_{i=1}^{K}p_i(m_i-\overline{m})(m_i-\overline{m})^{\mathrm{T}}$，$\Sigma_W=\sum_{i=1}^{K}p_i\Sigma_i$ 分别为类间散度矩阵和类内散度矩阵. 式中：K 为类别数，m_i 为类 i 的均值，p_i 为类 i 的先验概率，假设整体均值为 $\overline{m}=\sum_{i=1}^{K}p_i m_i$，$\Sigma_i$ 为类 i 的方差. $\overline{R}=TR$，$T:IR^n\to IR^m$ 为 $d\times n$ 满秩特征提取矩阵，它通过对 $\Sigma_W^{-1}\Sigma_B$ 进行特征向量分解计算，避免了复杂的数值优化过程. 但该方法仅计算类间散度 Σ_B，没有考虑各类之间方差差异对类可分性的影响，因而无法准确描述异方差数据的类可分性.

2.2　Chernoff 距离判据

Marco 在 Fisher 线性判据的基础上，提出了一种基于 Chernoff 距离的判别[5]，即

$$\begin{aligned}&J_C(T):=\\&\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}p_i p_j\mathrm{tr}\big((T\Sigma_W T^{\mathrm{T}})^{-1}T\Sigma_W^{1/2}\times\\&\big((\Sigma_W^{-1/2}\Sigma_{ij}\Sigma_W^{-1/2})^{-1/2}\Sigma_W^{-1/2}(m_i-m_j)\times\\&(m_i-m_j)^{\mathrm{T}}\Sigma_W^{-1/2}(\Sigma_W^{-1/2}\Sigma_{ij}\Sigma_W^{-1/2})^{-1/2}\times\\&\frac{1}{\pi_i\pi_j}(\log(\Sigma_W^{-1/2}\Sigma_{ij}\Sigma_W^{-1/2})-\\&\pi_i\log(\Sigma_W^{-1/2}\Sigma_i\Sigma_W^{-1/2})-\\&\pi_j\log(\Sigma_W^{-1/2}\Sigma_j\Sigma_W^{-1/2}))\Sigma_W^{1/2}T^{\mathrm{T}}\big).\end{aligned} \tag{2}$$

其中

$$\begin{aligned}&\pi_i:=p_i/(p_i+p_j),\\&\pi_j:=p_j/(p_i+p_j),\\&\Sigma_{ij}=\pi_i\Sigma_i+\pi_j\Sigma_j.\end{aligned}$$

可见，式(2)中通过分别计算各类间的协方差，并采用概率权重方式计算类间散度，考虑到了方差差异对类可分性的影响，是一种有效的异方差判据. $\overline{R}=TR$，$T:IR^n\to IR^m$ 为 $d\times n$ 满秩特征提取矩阵，其求解类似于 Fisher 线性判据，可通过对一个 $n\times n$ 矩阵特征向量分解求得.

2.3　互信息判据

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

$$\begin{aligned}&\mu(R;\Omega)=H(R)-H(R\mid\Omega)=\\&H(R)-\sum_{i=1}^{K}p_i H(R\mid\omega_i).\end{aligned} \tag{3}$$

其中：$H(R)=-\int_R f_R(r)\ln(f_R(r))\mathrm{d}r$，$f_R(r)=\sum_{i=1}^{c}p_i f_{R|\Omega}(r\mid\omega_i)$，$f_{R|\Omega}(r\mid\omega_i)$ 为随机向量的条件分布，ω_i 为类 i 的分布参数. 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[6]. 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[15]. 依据这一假设，式(3)可以改写为

$$\begin{aligned}&\mu(R;\Omega)=H_g(R)-H(R\mid\Omega)=\\&H_g(R)-\sum_{i=1}^{K}p_i H(R\mid\omega_i).\end{aligned} \tag{4}$$

定理 1[16]（高斯分布熵）　具有均值 μ 和方差 Σ 的多元高斯分布数据集 $X=\{X_1,X_2,\cdots,X_n\}$，则

$$\begin{aligned}&H(X)=H(X_1,X_2,\cdots,X_n)=\\&\frac{1}{2}\ln((2\pi\mathrm{e})^n\mid\Sigma\mid).\end{aligned} \tag{5}$$

证明　由多元高斯分布密度函数

$$f(x)=\frac{1}{(\sqrt{2\pi})^n\mid\Sigma\mid^{1/2}}\exp\Big(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\Big),$$

有

$$\begin{aligned}&H(X)=\\&H(X_1,X_2,\cdots,X_n)=\\&-\int f(x)\Big[-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)-\\&\ln((\sqrt{2\pi})^n\mid\Sigma\mid^{1/2})\Big]\mathrm{d}x=\\&\frac{1}{2}\mathrm{E}\Big[\sum_{i,j}(x_i-\mu_i)^{\mathrm{T}}\Sigma_{i,j}^{-1}(x_j-\mu_j)\Big]+\\&\frac{1}{2}\ln((2\pi)^n\mid\Sigma\mid)=\\&\frac{n}{2}+\frac{1}{2}\ln((2\pi)^n\mid\Sigma\mid)=\end{aligned}$$

$$\frac{1}{2}\ln((2\pi e)^n|\Sigma|).$$

因此得证.□

由定理 1,有

$$H_g(R)=\frac{1}{2}\ln((2\pi e)^n|\Sigma|),$$

$$H(R|\omega_i)=\frac{1}{2}\ln((2\pi e)^n|\Sigma_i|).$$

由高斯混合模型定义的性质,有数据总体方差

$$\Sigma=\sum_{i=1}^{K}p_i[\Sigma_i+(m_i-\bar{m})(m_i-\bar{m})^{\mathrm{T}}]. \quad (6)$$

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,于是简化得到的本文的互信息判据为

$$\mu(R;\Omega)=\frac{1}{2}\Big[\ln(|\Sigma|)-\sum_{i=1}^{K}p_i\ln(|\Sigma_i|)\Big]. \quad (7)$$

3 互信息判别的类可分性

类别差异度的类可分性有效的度量[17],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$\Sigma_1=\Sigma_2=\cdots=\Sigma_K$,则互信息计算公式可简化为

$$\mu(R;\Omega)=\frac{1}{2}\ln\frac{|\Sigma_B+\Sigma_W|}{|\Sigma_W|}. \quad (8)$$

其中:$\Sigma_B=\sum_{i=1}^{K}p_i[(m_i-\bar{m})(m_i-\bar{m})^{\mathrm{T}}]$,$\Sigma_W=\sum_{i=1}^{K}p_i\Sigma_i$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.

在考虑均值相似的情况下,两类互信息计算公式可改写为

$$\mu(R;\Omega)=\frac{1}{2}\ln\frac{|\Sigma_W+p_1p_2(m_1-m_2)(m_1-m_2)^{\mathrm{T}}|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}}. \quad (9)$$

其中:$\bar{m}=\sum_{i=1}^{2}p_im_i$,$\Sigma_W=\sum_{i=1}^{2}p_i\Sigma_i$.于是式(9)可进一步修改为

$$\mu(R;\Omega)=\frac{1}{2}\ln\frac{|\Sigma_W|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}}+\frac{1}{2}\ln(1+p_1p_2(m_1-m_2)^{\mathrm{T}}\Sigma_W^{-1}(m_1-m_2)). \quad (10)$$

由二阶泰勒级数$\|x\|\approx 0$时,有

$$\log(1+\alpha x^{\mathrm{T}}Qx)\approx\alpha x^{\mathrm{T}}Qx.$$

于是式(10)最终修改为

$$\mu(R:\Omega)=\frac{1}{2}\log\frac{|\Sigma_W|}{\prod_{i=1}^{2}|\Sigma_i|^{p_i}}+\frac{p_1p_2}{2}(m_1-m_2)^{\mathrm{T}}\Sigma_W^{-1}(m_1-m_2). \quad (11)$$

这就是文献[5]所提出的Chernoff距离类可分性判据.由式(8)和(11)可知,Fisher线性判别和基于Chernoff距离的异方差判别都是本文式(7)的特例,这表明互信息具有良好的统计物理意义,是一种通用的类可分性判据.

4 互信息判据特征提取的理论基础

互信息的引入建立了高维数据和分类信息之间内在的关系,大大提高了数据分类的精确度.但在特征提取时,极大似然、罚函数的方法及文献[9]的数值优化计算需要大量的迭代运算;同时互信息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,得到计算简单且物理意义更加明确的信息判据特征提取以进行数据降维.

4.1 互信息的线性变换定理

定理 2[16](线性变换熵) 存在系数a,有

$$H(aX)=H(X)+\ln(|a|).$$

证明 设$Y=aX$,则$f_Y(y)=\frac{1}{\alpha}f_x(\frac{y}{a})$,且有

$$H(aX)=-\int f_Y(y)\log f_Y(y)\mathrm{d}y=-\int\frac{1}{a}f_x(\frac{y}{a})\ln\Big[\frac{1}{a}f_x(\frac{y}{a})\Big]\mathrm{d}y=-\int f_x(x)\ln f_x(x)\mathrm{d}x+\ln(|a|)=H(X)+\ln(|a|).$$

因此得证.□

推论 1[16](矩阵变换熵) 由定理 2 易知,当A为非奇异线性变换矩阵时,有

$$H(AX)=H(X)+\ln(|A|). \quad (12)$$

定理 3(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

$$\mu(\bar{R};\Omega)=\mu(R;\Omega). \quad (13)$$

其中:$\bar{R}=TR$,$T:IR^n\to IR^n$是一个非奇异变换矩阵.

证明 由推论1可知,对于任意非奇异矩阵T,有$H(TR)=H(R)+\ln(|T|)$.又根据定义,有

$$\mu(\bar{R};\Omega)=H_g(R)+\ln(|T|)-[H(R|\Omega)+\ln(|T|)]=\mu(R;\Omega).$$

于是得证.□

定理 4(正交不变性) 对于任意满秩变换矩阵$T:IR^n\to IR^m(m<n)$,存在一个正交变换矩阵$E\in$

$IR^{m\times n}$，使得 $\bar{R}\triangleq TR$，$\tilde{R}=ER$，有

$$\mu(\bar{R};\Omega)=\mu(\tilde{R};\Omega). \tag{14}$$

证明　根据奇异值分解理论，存在正交矩阵 $U\in IR^{m\times m}$ 和 $V\in IR^{n\times n}$，使得 $U^{\mathrm{T}}TV=[\Lambda\mid 0_{m\times d}]$. 其中：$\Lambda\in IR^{m\times m}$ 是由变换矩阵 T 的特征值构造的对角矩阵，$d=n-m$，$\mathrm{rank}(T)=m$. 令 $\tilde{R}=\Lambda^{-1}U^{\mathrm{T}}\bar{R}=\Lambda^{-1}U^{\mathrm{T}}TR$，由定理 3 有 $\mu(\bar{R};\Omega)=\mu(\tilde{R};\Omega)$. 又令 $E=\Lambda^{-1}U^{\mathrm{T}}T\in IR^{m\times n}$，则 $EV=\Lambda^{-1}U^{\mathrm{T}}TV=[I_m\mid 0_{m\times d}]$，有 $EV(EV)^{\mathrm{T}}=EVV^{\mathrm{T}}E^{\mathrm{T}}=EE^{\mathrm{T}}=I_m$，$E$ 是一个正交矩阵. □

定理 5（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

$$\mu(\hat{R};\Omega)\leqslant\mu(R;\Omega). \tag{15}$$

其中：$\hat{R}=TR$，$T:IR^n\to IR^m(m<n)$ 是一个满秩矩阵.

证明　假设 $T^{\perp}\in IR^{(n-m)\times n}$ 为矩阵 T 的行生成零空间矩阵，定义变换矩阵 $\tilde{T}$ 和随机向量 $\tilde{R}$ 分别为

$$\tilde{T}=\begin{bmatrix}T\\T^{\perp}\end{bmatrix}\in IR^{n\times m},\ \tilde{R}=\tilde{T}R=\begin{bmatrix}TR\\T^{\perp}R\end{bmatrix}=\begin{bmatrix}\hat{R}\\\hat{N}\end{bmatrix}.$$

显然 $\tilde{T}$ 是满秩矩阵，有 $\mu(\tilde{R};\Omega)=\mu(R;\Omega)$，且

$$\mu(\tilde{R};\Omega)=\mu(\hat{R},\hat{N};\Omega)=$$
$$H_g(\hat{R},\hat{N})-H(\hat{R},\hat{N}\mid\Omega).$$

根据熵的链式法则和高斯条件熵，有

$$H_g(\hat{N}\mid\hat{R})\geqslant H(\hat{N}\mid\hat{R})\geqslant$$
$$H(\hat{N}\mid\hat{R},\Omega),$$

则

$$\mu(\hat{R};\Omega)=H_g(\hat{R})+H_g(\hat{N}\mid\hat{R})-$$
$$[H(\hat{R}\mid\Omega)+H(\hat{N}\mid\hat{R},\Omega)]=$$
$$\mu(\hat{R};\Omega)+H_g(\hat{N}\mid\hat{R})-H(\bar{N}\mid\hat{R},\Omega)\geqslant$$
$$\mu(\hat{R};\Omega).$$ □

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，能更好地体现数据处理上的物理意义.

4.2　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. 贝叶斯分类误差的计算公式为

$$P_R(\varepsilon)=1-\sum_{i=1}^{K}\int_{R_i}p_if_{R|\Omega}(r\mid\omega_i)\mathrm{d}r. \tag{16}$$

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

$$i^*=\arg\max_{1\leqslant i\leqslant c}P(\omega_i\mid r_0). \tag{17}$$

其中：r_0 为无标记的观测数据，$P(\omega_i\mid r_0)$ 为分类后验概率. 从而定义 $\varepsilon_R=\min P_R(\varepsilon)$ 贝叶斯误差，且有可逆线性变换不变性[9].

为分析互信息的贝叶斯优化，假设高维数据降维后具有高斯分布特性，将数据分解成独立高斯信号空间和噪声空间，即 $S\mid\Omega\in IR^m$ 和 $N\mid\Omega\in IR^d$ 且 $f_{N|S,\Omega}(n\mid s,\omega_i)=f_{N|S}(n\mid s)$，$\forall i=\{1,2,\cdots,c\}$，$\forall s\in IR^m$，$\forall n\in IR^d$. 则有

$$R\mid\omega_i=M\begin{bmatrix}S\\N\end{bmatrix}\begin{matrix}\vdots\\\vdots\end{matrix}\begin{matrix}\omega_i\\\omega_i\end{matrix},\ \forall i=\{1,2,\cdots,c\}, \tag{18}$$

其中 $M\in IR^{n\times n}$ 为非奇异矩阵. 令 $\bar{R}\triangleq TR$，$T\in IR^{m\times n}$ 是使 $\mu(\bar{R};\Omega)$ 值最大的满秩矩阵，由定理 3 有 $\mu(\bar{R};\Omega)=\mu(R;\Omega)$，$\varepsilon_R=\varepsilon_{\bar{R}}$. 令 $T^{\perp}\in IR^{(n-m)\times n}$ 为矩阵 T 的零行生成矩阵，则有

$$\tilde{T}=\begin{bmatrix}T\\T^{\perp}\end{bmatrix}=M^{-1}\in IR^{n\times n},$$

$$\tilde{R}=\tilde{T}R=\begin{bmatrix}TR\\T^{\perp}R\end{bmatrix}=\begin{bmatrix}S\\N\end{bmatrix}.$$

由定理 1 可知 $\mu(\tilde{R};\Omega)=\mu(R;\Omega)$，且 $\varepsilon_{\tilde{R}}=\varepsilon_R$. 依据链式熵法则，有

$$\mu(R;\Omega)=\mu(S,N;\Omega)=$$
$$\mu(S;\Omega)+H_g(N\mid S)-H(N\mid S,\Omega).$$

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

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

即 $\varepsilon_R=\varepsilon_S$，则 $\varepsilon_{\bar{R}}=\varepsilon_S$.

结论 1　通过 T 提取的信号空间有 $\mu(\bar{R};\Omega)=\mu(S;\Omega)$，且有贝叶斯一致优化判别率 $\varepsilon_{\bar{R}}=\varepsilon_S$.

4.3　矩阵特征向量分解的互信息特征提取

由上节所述，给定数据 $R\in IR^n$，能够找到一个满秩的变换矩阵 $T\in IR^{m\times n}$，使得互信息 $\mu(\bar{R};\Omega)$ 最大，并保证贝叶斯一致优化，即

$$T^* = \arg \max_{T \in IR^{m\times n}} \{\mu(\bar{R};\Omega): \bar{R} = TR\}, \quad (19)$$

其中 T 是一个满秩矩阵. 由式(7)可得基于互信息的特征提取判别函数为

$$\mu(\bar{R};\Omega) = \frac{1}{2}\Big[\ln(|\ T\Sigma T^{\mathrm{T}}\ |) - \sum_{i=1}^{c} p_i \ln(|\ T\Sigma_i T^{\mathrm{T}}\ |)\Big]. \quad (20)$$

由互信息的线性变换定理和贝叶斯一致优化分析,通过选择满足式(19)的特征提取矩阵 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. 本文借鉴文献[5]的方法,通过矩阵特征向量分解选择总体方差和各类方差前 m 个特征值,以保证互信息判据函数值最大. 其基本原理如下:

给定正定对称矩阵 $\Sigma \in R^{n\times n}$,由矩阵特征向量分解原理知,存在特征值矩阵 $\Lambda = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 和特征向量矩阵 V,其中 V 为单位正交矩阵且 $|\ V\ | = |\ V^{\mathrm{T}}\ | = 1$. 则由矩阵对数算术运算,有

$$\ln|\ \Sigma\ | = \ln(|\ V\Lambda V^{\mathrm{T}}\ |) = |\ V\ ||\ \mathrm{diag}(\ln(\lambda_1),\ln(\lambda_2),\cdots,\ln(\lambda_n))\ ||\ V^{\mathrm{T}}\ | = \ln\Big(\prod_{i=1}^{n}\lambda_i\Big). \quad (21)$$

因此,在给定特征规模 m 下,由互信息判据原则容易选定式(20)的 m 个最大特征值满足式(19),并获得最大互信息. 与现有的其他互信息判据不同,由式(21)可同时以大于1特征值的个数为特征规模,一步完成最优化规模计算,且满足

$$T^* = \arg \max_{m\in\{i:0<i<n\}} \{\arg \max_{T\in IR^{m\times n}} \{\mu(\bar{R};\Omega): \bar{R} = TR\}\}. \quad (22)$$

5 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[18]. 该数据集共有6345个36维数据,分为6类,其中4435个作为固定的训练数据,其他为测试数据. 为验证算法,采用欧氏距离分类器训练和测试算法,通过选用不同的特征向量规模(m)与文献[5]的算法作比较. 其结果如表1和图1所示.

表 1 分类准确率:算法+分类器联合测试结果

算法	特征向量规模 m					
	5	10	11	15	20	36
Fisher	87.30	87.80	87.70	*87.80	87.60	86.85
文献[5]	86.95	86.70	86.50	*87.20	87.10	86.85
文献[9]	87.55	88.10	*88.20	87.30	87.15	86.85
本文	87.55	88.10	*88.20	87.30	87.15	86.85

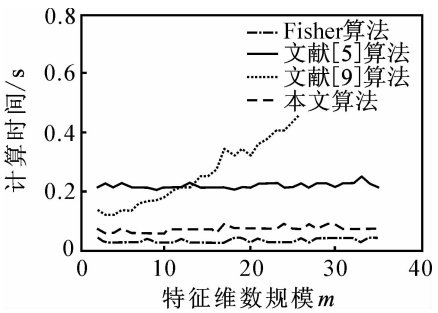

图 1 计算特征提取矩阵时间代价

从表1可以看出,本文算法和文献[9]的算法在降维情况下($m = 5,10,11$)能够得到较高的分类精度,表明互信息判据在数据降维后服从高斯混合分布的假设,说明式(7)考虑了均值差异对类可分性的影响,能够更精确地描述异方差高维数据的分类信息. 从图1可以看出,文献[5],Fisher 算法以及本文算法采用矩阵特征向量分解的方法计算特征提取矩阵具有并行计算的特性,不受特征规模(m)的影响,并且在考虑异方差的情况下,本文算法优于文献[5]的算法;本文算法本质上和文献[9]采用数值优化算法一样,但后者的算法计算代价随特征向量规模的增大而增大,并且始终高于本文算法;本文算法能一步完成特征规模的选择($m = 11$).

6 结 论

本文在互信息判据的基础上,提出了一种基于高斯分布假设的特征提取判据算法. 算法充分考虑了高维数据降维后的高斯分布特点,运用互信息描述数据与高斯分类参数信息关系,得到了简明的判据函数,同时还考虑了均值差异和异方差的情况,是一种通用的特征提取信息判据. 本文进一步证明了互信息的线性变换定理,并推证了具有贝叶斯一致优化的特征提取框架. 最后通过实验验证了本文算法在精确度上优于文献[5]的算法,在时间计算上优于文献[9]的算法,并能完成一步计算优化特征规模.

参考文献(References)

[1] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(1): 179-188.

[2] Rao C R. The utilization of multiple measurements in problems of biological classification[J]. J of the Royal Statistical Society, Series B, 1948, 10(2): 159-203.

[3] 范玉刚, 李平, 宋执环. 基于非线性映射的 Fisher 判别分析[J]. 控制与决策, 2007, 22(4): 384-388.
(Fan Y G, Li P, Song Z H. Fisher discriminant analysis based on nonlinear mapping[J]. Control and Decision, 2007, 22(4): 384-388.)

[4] Loog M, Duin R P W, Haeb-Umbach R. Multiclass linear dimension reduction by weighted pairwise fisher criteria[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(7): 762-766.

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

[6] Kenneth E Hild II, Deniz Erdogmus, Kari Torkkola. Feature extraction using information-theoretic learning [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1385-1392.

[7] Padmanabhan M, Dharanipragada S. Maximizing information content in feature extraction[J]. IEEE Trans on Speech Audio Processing, 2005, 13(4): 512-519.

[8] Jose Miguel Leiva-Murillo, Antonio Artés-Rodríguez. Maximization of mutual information for supervised linear feature extraction[J]. IEEE Trans on Neural Networks, 2007, 18(5): 1433-1440.

[9] Zoran Nenadic. Information discriminant analysis: Feature extraction with an information-theoretic objective[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(8): 1394-1407.

[10] Moritz Grosse-Wentrup, Martin Buss. Multiclass common spatial patterns and information theoretic feature extraction[J]. IEEE Trans on Biomedical Engineering, 2008, 55(8): 1991-2000.

[11] 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 Y, Li J T, Wang B, et al. A study on constraints for feature selection in text categorization[J]. J of Computer Research and Development, 2008, 45(4): 596-602.)

[12] 赵峙江，赵春晖，张志宏. 一种新的 PCNN 模型参数估算方法[J]. 电子学报，2007，35(5)：996-1000.
(Zhao Z J, Zhao C H, Zhang Z H. A new method of PCNN's parameter's optimization[J]. Acta Electronica Sinica, 2007, 35(5): 996-1000.)

[13] 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 G, Chen X M. An audio feature extraction method taking class information into account[J]. J of Computer Research and Development, 2006, 43(11): 1959-1964.)

[14] Andrew R Webb. Statistical pattern recognition[M]. 2nd ed. West Sussex: John Wiley and Sons Ltd, 2002: 123-165.

[15] Rasmussen C E, Williams C K I. Gaussian processes for machine learning[M]. Cambridge: The MIT Press, 2006: 171-188.

[16] Cover T M, Thomas J A. Elements of information theory[M]. 2nd ed. New York: Wiley InterScience, 2006: 243-260.

[17] 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 H L, Kong F S, Li Y X. An analysis of diversity measures in clustering ensembles[J]. Chinese J of Computers, 2008, 30(8): 1315-1324.)

[18] Arthur Asuncion, David Newman. UCI Repository of Machine Learning Databases [EB/OL]. [2008]. http://archive.ics.uci.edu/ml/.

(上接第 1809 页)

[5] 谢乃明，刘思峰. 一种新的弱化缓冲算子[J]. 中国管理科学，2003，11(增)：46-48.
(Xie N M, Liu S F. A new applicative weakening buffer operator[J]. Chinese J of Management Science, 2003, 11(S): 46-48.)

[6] 党耀国，刘思峰，刘斌，等. 关于弱化缓冲算子的研究[J]. 中国管理科学，2004，12(2)：108-111.
(Dang Y G, Liu S F, Liu B, et al. Study on the weakening buffer operators[J]. Chinese J of Management Science, 2004, 12(2): 332-336.)

[7] 党耀国，刘斌，关叶青. 关于强化缓冲算子的研究[J]. 控制与决策，2005，20(12)：1332-1336.
(Dang Y G, Liu B, Guan Y Q. Study on the strengthening buffer operator[J]. Control and Decision, 2005, 20(12): 1332-1336.)

[8] 党耀国，刘思峰，米传民. 强化缓冲算子性质的研究[J]. 控制与决策，2007，22(7)：730-734.
(Dang Y G, Liu S F, Mi C M. Study on characteristics of the strengthening buffer operators[J]. Control and Decision, 2007, 22(7): 730-734.)

[9] 关叶青，刘思峰. 基于不动点的强化缓冲序列算子及其应用[J]. 控制与决策，2007，22(10)：1189-1192.
(Guan Y Q, Liu S F. Sequence of strengthening buffer operator and its application based on fixed point[J]. Control and Decision, 2007, 22(10): 1189-1192.)

[10] 崔杰，党耀国. 一类新的弱化缓冲算子的构造及其应用[J]. 控制与决策，2008，23(7)：741-750.
(Cui J, Dang Y G. A kind of new weakening buffer operators and their applications[J]. Control and Decision, 2008, 23(7): 741-750.)