基于 L1 范数凸包数据描述的多观测样本分类算法

胡正平* 王玲丽
(燕山大学信息科学与工程学院 秦皇岛 066004)

摘 要：为建立高维空间样本分布的最佳覆盖为目标来实现覆盖分类，该文提出基于 L1 范数凸包数据描述的多观测样本分类算法。首先对训练集的每个类别以及测试集的多观测样本分别构造凸包模型，这样多观测样本的分类就转化为凸包模型的相似性度量问题。若测试集的凸包模型与训练集无重叠，采用 L1 范数距离测度进行凸包模型之间的相似性度量；若有重叠，采用 L1 范数距离测度进行收缩凸包(reduced convex hulls)之间的相似性度量。然后采用最近邻准则作为多观测样本的分类决策。在 3 个数据库上进行的实验结果，表明该文提出方法对于多观测样本分类具有可行性和有效性。

关键词：模式识别；凸包；L1 范数距离测度；最近邻分类；多观测样本
中图分类号：TP391.41 **文献标识码**： A **文章编号**：1009-5896(2012)01-0194-06
DOI: 10.3724/SP.J.1146.2011.00545

The Classification Algorithm of Multiple Observation Samples Based on L1 Norm Convex Hull Data Description

Hu Zheng-ping Wang Ling-li
(*School of Information Science and Engineering*, *Yanshan University*, *Qinhuangdao* 066004, *China*)

Abstract: In order to construct a high-dimensional data approximate model in the purpose of the best coverage of the distribution of high-dimensional samples, the classification algorithm of multiple observation samples based on L1 norm convex hull data description is proposed. The convex hull for each class in the train set and multiple observation samples in the test set is constructed as the first step. So the classification of multiple observation samples is transformed to the similarity of convex hulls. If the test convex hull and every train hull are not overlapping, L1 norm distance measure is used to solve the similarity of convex hulls. Otherwise, L1 norm distance measure is used to solve the similarity of reduced convex hulls. Then the nearest neighbor classifier is used to solve the classification of multiple observation samples. Experiments on three types of databases show that the proposed method is valid and efficient.

Key words: Pattern identification; Convex hull; L1 norm distance measure; Nearest neighbor classification; Multiple observation samples

1 引言

传统的模式识别解决的分类问题常常仅仅针对测试模式为单观测样本的情况，比如：Eigenface利用特征脸空间对测试模式的单观测样本进行分类；Fisherface通过寻找最有效的分类方向来实现单观测样本的分类。多观测样本可以提供比单观测样本更多的关于测试模式的信息，从而提高分类精度[1]。由此可以预见，多观测样本分类问题将逐渐得到国内外学者的广泛关注。

目前针对多观测样本的分类问题，研究思路主要分为两个大的类别：一类是基于参数模型的方法，例如，文献[2]提出单高斯参数模型，该方法用多变量单高斯参数模型来描述每一个样本集的分布，那么多观测样本的分类问题就转化为样本集之间交叉熵(Kullback-Leibler divergence, KL)的计算。针对非线性分布问题，文献[3]提出利用混合高斯参数模型代替单高斯参数模型对样本集进行密度估计，该方法对非线性流形有更准确的密度估计。基于参数模型的方法缺点是它们需要解决比较复杂的参数估计问题，并且当多观测样本和测试集之间统计特性不明显时，采用这种方法具有一定的局限性。另一类是基于非参数模型的方法，其中具有代表性的是基于子空间的方法，它是根据子空间的相似性度量

2011-06-07 收到，2011-09-19 改回
国家自然科学基金(61071199)，河北省自然科学基金(F2010001297)，中国博士后自然科学基金(20080440124)和第 2 批中国博士后基金(200902356)资助课题
*通信作者：胡正平 hzp@ysu.edu.cn

(典型相关性(canonical correlations)或者主成分角(principal angles))来实现多观测样本的分类，例如：文献[4]提出CMSM(Constraint Mutual Subspace Method)算法，首先把样本集投影到一个约束子空间(只包括对分类最有利的成分)上，然后计算投影之后的样本集之间的相似性。CMSM认为样本分布是线性的，并且约束子空间参数的设置及其维数的选取比较困难。为此文献[5]提出DCC(Discriminant Canonical Correlation)不需要进行维数的选取，首先通过迭代算法得到一个转换矩阵(使类内典型相关性最大同时类间典型相关性最小的矩阵)，然后把样本集投影到转换矩阵上，以此实现多观测样本的分类。为了考虑样本分布的非线性，文献[6]提出KPA(Kernel Principal Angles)，该方法对核函数的选取有一定依赖性，一些研究者认为子空间之间的主成分角重要程度是一样的(权值相同)[4−6]，后来研究发现：不同的主成分角对分类的贡献不同，为此一些研究者提出BoMPA(Boosted Manifold Principal Angles)，首先利用PPCA(Probabilistic Principal Component Analysis)找到局部线性模块，然后得到子空间之间的主成分角，把训练集表示成正负样本特征的形式，利用AdaBoost算法学习到每个弱分类器(主成分角)的权值，采用加权的主成分角进行子空间的相似性度量，从而实现对多观测样本的分类[7]。文献[8]进一步把AdaBoost算法用于多观测样本的分类，提出了Boosted全局和局部主成分角的方法。结合主成分角和局部线性模块的思想，文献[9]提出MMD(Manifold-Manifold Distance)，该方法首先把非线性流形表示成一组局部线性模块的集合，局部线性模块是从一个种子点开始增长，直到打破线性约束的条件，计算线性模块的主成分角以及模块的均值图像之间的欧式距离，MMD就定义为这两个距离的加权和，那么多观测样本的分类就转化为MMD的计算问题，缺点是需要计算欧氏距离和测地线距离，需要设置多个参数，计算量和复杂度较大。文献[10]提出KDT(Kernel Discriminant Transformation)与文献[5]类似，是DCC的非线性形式。近来，一些研究者利用k近邻图来实现多观测样本的分类，提出MASC(MAnifold-based Smoothing under Constrain)算法，该算法在标记传播算法的基础上展开，不足是性能依赖于参数选择，采用L2欧氏距离下的高斯核函数来计算边权值，基于欧氏距离测度难以完全反映数据复杂的空间结构分布特性[11]。以上方法都是直接把样本数据作为特征进行分类，是面向数据的方法。而文献[12]采用稀疏特征(LBP或者Gabor小波)这种基于子空间的识别方法，并提出基于样本集随机森林决策树来进行核函数及其参数的随机学习，以此来进行多观测样本的分类。

近年关于高维空间几何分析的凸包模型研究受到学者的广泛关注，数据的凸包描述可以提供样本分布的直观几何解释，可用凸包模型来估计样本的分布[13−15]。文献[16,17]分别提出了基于SRM自组织多区域覆盖模型和基于高维空间最小生成树自适应覆盖模型来估计样本的分布。基于凸包覆盖的思想，本文提出基于L1范数凸包数据描述的多观测样本分类算法(L1_CHDD)。该方法采用凸包模型来描述训练集的每一类以及测试集的多观测样本，那么多观测样本的分类就转化为凸包模型的相似性度量，以L1范数作为距离测度，进而实现多观测样本的分类。

2　多观测样本分类问题的描述

这里研究的多观测样本的分类问题，其实就是研究以多观测样本作为测试模式的分类问题。多观测样本分类问题的数据集分为两部分，$\boldsymbol{X}=\{\boldsymbol{X}^{(l)},\boldsymbol{X}^{(u)}\}$；$\boldsymbol{X}^{(l)}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l\}\in\Re^d$表示已知类别标记的样本，$d$是样本的维数，$\{y_1,y_2,\cdots,y_l\},y_i\subset\{1,2,\cdots,C\}$是已标记样本的类标；$\boldsymbol{X}^{(u)}=\{\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\}\in\Re^d$表示未标记的多观测样本，这里要求全部样本的类别数$C$是已知的，并且已标记的样本中涵盖了所有的类别。多观测样本分类问题可以看作一个特殊的半监督学习，限制条件是测试集中所有的观测样本属于同一个类别。

3　基于L1范数凸包数据描述的多观测样本分类算法

3.1　基本概念

定义 1(凸包集)　设集合$\boldsymbol{S}\subset\Re^d$，如果对任意$\boldsymbol{x}_1,\boldsymbol{x}_2\in\boldsymbol{S}$和任意$\alpha\in[0,1]$，都有$\alpha\boldsymbol{x}_1+(1-\alpha)\boldsymbol{x}_2\in\boldsymbol{S}$，则称$\boldsymbol{S}$是凸包集。

定义 2(凸组合)　设$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\in\Re^d$，如果存在满足$\sum_{i=1}^{n}\alpha_i=1$的非负$\alpha_1,\alpha_2,\cdots,\alpha_n$，使得$\boldsymbol{x}=\sum_{i=1}^{n}\alpha_i\boldsymbol{x}_i$，则称$\boldsymbol{x}$是$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$的一个凸组合。

定义 3(凸包)　设集合$\boldsymbol{S}\subset\Re^d$表示$\Re^d$中$n$个点组成的集合，$\boldsymbol{S}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，则由集合$\boldsymbol{S}$张成的凸包$\mathrm{conv}(\boldsymbol{S})$定义为包含集合$\boldsymbol{S}$的最小凸包集：$\mathrm{conv}(\boldsymbol{S})=\left\{\boldsymbol{x}=\sum_{i=1}^{n}\alpha_i\boldsymbol{x}_i\,\middle|\,\sum_{i=1}^{n}\alpha_i=1,\alpha_i\geq0\right\}$。

3.2　系统组成

本文构造的分类器模型如图1所示，由构造层和判别层两层模型组成。其中，构造层需要对训练集

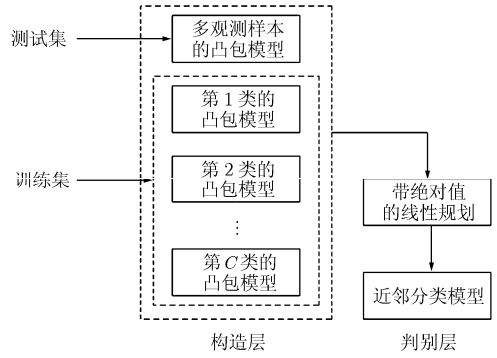

图1 本文算法结构框图

的每一类以及测试集的多观测样本分别建立凸包覆盖模型，凸包模型的相似性度量采用L1范数，然后利用带绝对值的线性规划得到凸包模型之间的距离，判别层采用经典的近邻分类策略。

3.3 基于 L1 范数凸包数据描述的构造层模型

本文构造的高维空间数据覆盖模型依赖与训练集样本或者测试集的多观测样本张成的凸包，即以某类的凸包作为该类数据的覆盖模型，训练集第 c 类的凸包模型 S_c 为

$$\mathrm{conv}(S_c)=\left\{\boldsymbol{x}=\sum_{k=1}^{n_c}\alpha_{ck}\boldsymbol{x}_{ck}\left|\sum_{k=1}^{n_c}\alpha_{ck}=1,\alpha_{ck}\geq 0\right.\right\} \quad (1)$$

其中 n_c 是第 c 类样本的个数，$\boldsymbol{x}_{ck}\in\Re^d$ 表示第 $c\in\{1,2,\cdots,C\}$ 类的第 $k\in\{1,2,\cdots,n_c\}$ 个样本，α_{ck} 是 $\boldsymbol{x}_{ck}$ 的凸组合系数，凸组合系数向量的和是 1，训练集第 c 类的覆盖模型就是第 c 类样本形成的凸包。

测试集多观测样本的凸包模型 S^* 为

$$\mathrm{conv}(S^*)=\left\{\boldsymbol{x}=\sum_{k=1}^{n^*}\alpha_k^*\boldsymbol{x}_k^*\left|\sum_{k=1}^{n^*}\alpha_k^*=1,\alpha_k^*\geq 0\right.\right\} \quad (2)$$

其中 n^* 是多观测样本的个数，$\boldsymbol{x}_k^*$ 是第 k 个多观测样本，α_k^* 是 $\boldsymbol{x}_k^*$ 的凸组合系数，凸组合系数向量的和是 1，测试集的覆盖模型就是多观测样本形成的凸包。

(1)多观测样本的凸包模型与训练集的每一类的凸包模型均没有重叠　给定两个无重叠的凸包模型(即两个凸包模型可分) $\boldsymbol{H}$ 和 $\boldsymbol{H}'$，它们之间的距离定义为 $\boldsymbol{H}$ 中的任意一点和 $\boldsymbol{H}'$ 中的任意一点的 L1 范数距离的最小值，即

$$D(\boldsymbol{H},\boldsymbol{H}')=\min_{\boldsymbol{x}\in\boldsymbol{H},\boldsymbol{y}\in\boldsymbol{H}'}\|\boldsymbol{x}-\boldsymbol{y}\|_1 \quad (3)$$

假设第 c 类的样本 $\boldsymbol{X}_c=(\boldsymbol{x}_{c1},\boldsymbol{x}_{c2},\cdots,\boldsymbol{x}_{cn_c})$，第 c 类的凸组合系数向量 $\boldsymbol{\alpha}_c=[\alpha_{c1},\alpha_{c2},\cdots,\alpha_{cn_c}]^{\mathrm{T}}$，多观测样本 $\boldsymbol{X}^{(u)}=\boldsymbol{X}^*=(\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,\cdots,\boldsymbol{x}_{n^*}^*)$，多观测样本的凸组合系数向量 $\boldsymbol{\alpha}^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_{n^*}^*]^{\mathrm{T}}$，则 $\{S_c\}=\boldsymbol{X}_c\boldsymbol{\alpha}_c$，$\{S^*\}=\boldsymbol{X}^*\boldsymbol{\alpha}^*$，那么这两个凸包模型之间的 L1 范数距离是

$$D(S_c,S^*)=\arg\min_{\boldsymbol{\alpha}_c,\boldsymbol{\alpha}^*}\|\boldsymbol{X}_c\boldsymbol{\alpha}_c-\boldsymbol{X}^*\boldsymbol{\alpha}^*\|_1 \quad (4)$$

其中 $\sum_{k=1}^{n_c}\alpha_{ck}=1$，$\sum_{k=1}^{n^*}\alpha_k^*=1$，$\alpha_{ck}\geq 0$，$\alpha_k^*\geq 0$。

令 $\boldsymbol{X}=(\boldsymbol{X}_c\ -\boldsymbol{X}^*)$，$\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\alpha}_c\\ \boldsymbol{\alpha}^*\end{pmatrix}$，$\boldsymbol{e}_c=[1,1,\cdots,1]_{1\times n_c}^{\mathrm{T}}$，$\boldsymbol{e}^*=[1,1,\cdots,1]_{1\times n^*}^{\mathrm{T}}$，$\boldsymbol{\theta}_c=[0,0,\cdots,0]_{1\times n_c}^{\mathrm{T}}$，$\boldsymbol{\theta}^*=[0,0,\cdots,0]_{1\times n^*}^{\mathrm{T}}$，$\boldsymbol{\theta}=\begin{pmatrix}\boldsymbol{\theta}_c\\ \boldsymbol{\theta}^*\end{pmatrix}$，那么式(4)可转化为

$$\begin{aligned}D(S_c,S^*)&=\arg\min_{\boldsymbol{\beta}}\|\boldsymbol{X}\boldsymbol{\beta}\|_1=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{d}\left|\sum_{j=1}^{n_c+n^*}\boldsymbol{x}_{ij}\boldsymbol{\beta}_j\right|\\&=\arg\min_{\boldsymbol{\beta}}\underbrace{(1,\ 1,\ \cdots,\ 1)}_{d}\mathrm{abs}(\boldsymbol{X}\boldsymbol{\beta}),\end{aligned}$$

$$\text{s.t.}\ \ \boldsymbol{e}_c^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_{n_c}\end{pmatrix}=1,\ \boldsymbol{e}^{*\mathrm{T}}\begin{pmatrix}\boldsymbol{\beta}_{n_c+1}\\ \boldsymbol{\beta}_{n_c+2}\\ \vdots\\ \boldsymbol{\beta}_{n_c+n^*}\end{pmatrix}=1,\ \boldsymbol{\beta}\geq\boldsymbol{\theta} \quad (5)$$

式(5)是一个带绝对值的线性规划问题，d 是样本的维数。若 $\boldsymbol{\beta}^\circ=[\boldsymbol{\alpha}_1^\circ,\boldsymbol{\alpha}_2^\circ,\cdots,\boldsymbol{\alpha}_{n_c}^\circ,\boldsymbol{\alpha}_{n_c+1}^\circ,\cdots,\boldsymbol{\alpha}_{n_c+n^*}^\circ]^{\mathrm{T}}$ 是式(5)的解，那么这两个可分凸包模型之间的距离为

$$D(S_c,S^*)=\underbrace{(1,\ 1,\ \cdots,\ 1)}_{d}\mathrm{abs}(\boldsymbol{X}\boldsymbol{\beta}^\circ)$$

(2)多观测样本的凸包模型与训练集的某一类或者多类凸包模型有重叠　给定两个有重叠的凸包模型(即两个凸包模型不可分) $\boldsymbol{H}$ 和 $\boldsymbol{H}'$，如果采用式(3)，它们之间的距离是 0。本文通过改变凸包模型组合系数的上下界构造收缩凸包，来进行有重叠凸包模型之间的相似性度量。对于训练集的第 c 类和测试集，本文构造的收缩凸包有如下形式：

$$\begin{aligned}&R\_\mathrm{conv}(S_c,\mu)\\&=\left\{\boldsymbol{x}=\sum_{k=1}^{n_c}\alpha_{ck}\boldsymbol{x}_{ck}\left|\sum_{k=1}^{n_c}\alpha_{ck}=1,0\leq\alpha_{ck}\leq\mu\right.\right\}\end{aligned}$$

$$R\_\mathrm{conv}(S^*,\mu)=\left\{\boldsymbol{x}=\sum_{k=1}^{n^*}\alpha_k^*\boldsymbol{x}_k^*\left|\sum_{k=1}^{n^*}\alpha_k^*=1,0\leq\alpha_k^*\leq\mu\right.\right\}$$

其中 α_{ck}，$\boldsymbol{x}_{ck}$，n_c，α_k^*，$\boldsymbol{x}_k^*$，n^* 的参数意义同上，$\mu<1$ 是人工设置的使两个凸包模型可分的参数。μ 的选取方法：按照步长减少两个凸包模型的上界，如果凸包模型之间的距离仍然是 0，按照步长继续减少凸包模型的上界，直到凸包模型之间的距离不是 0。一般地，$\mu=1/n_c$，$\mu=1/n^*$ 分别表示训练集第 c 类的中心和测试集的中心。如果 μ 的取值接近 $\max(1/n_c,1/n^*)$ 时，两个凸包模型之间的距离还是 0，则认为这两个凸包模型属于包含关系，那么，多观测样本就属于当前训练集凸包模型的类别。

对于有重叠的凸包模型通过构造收缩凸包，实现了凸包模型的可分性，然后采用如情况(1)所示的带绝对值的线性规划，得到收缩凸包之间的相似性度量(距离)。

3.4 最近邻分类判别层模型

多观测样本属于同一个类别，它们的凸包模型与训练集每个类别的凸包模型之间的距离，已经通过带绝对值的线性规划得到，用 $D(S^*,S_c)$ 表示多观测样本凸包模型与训练集第 c 类凸包模型之间的距离。那么，多观测样本所属类别的判别规定是：

如果 $D(S^*,S_k)=\arg\min_c D(S^*,S_c),\ c=1,2,\cdots,C$，则判决 $\{\boldsymbol{X}^{(u)}\}\in k$。

3.5 基于L1范数凸包模型数据描述的多观测样本分类算法(L1_CHDD)和1NN算法、SVM算法的比较与分析

L1_CHDD与1NN(1-Nearest Neighbor)分类算法相比，都是通过比较距离，然后以最近邻原则作为分类决策。但是，1NN分类算法比较的是点与点的距离(测试点和训练点)，而L1_CHDD算法比较的是模型与模型的距离(多观测样本凸包模型与训练集的凸包模型)。

L1_CHDD算法与SVM算法相比，都是通过样本的凸包模型来构造分类器。但是，SVM算法的支持向量是边界上的特征向量，是实在存在的样本点的线性组合，而L1_CHDD算法通过带绝对值的线性规划得到的使凸包模型距离最小的点对不一定是已知的样本点，有可能是虚拟样本点。

4 实验仿真

为验证本文提出算法的有效性，分别在Binary手写体数据库，ETH-80物体识别数据库以及ORL人脸识别数据库上进行了实验，并将提出的算法与单观测样本分类算法1NN, SVM以及多观测样本分类算法L2_CHDD，文献[9]的MMD，文献[11]的MASC进行对比。L1_CHDD和L2_CHDD算法中步长 $\mathrm{step}=0.01$，SVM和MASC采用高斯核函数，核宽带分别为 $\sigma=0.1$，$\sigma=0.5$，MMD中局部线性模块的阈值参数 $\mathrm{th}=1.1$，PCA子空间的维数是保持能量的96%，两种距离的加权比值 $\lambda=0.5$，MASC和MMD中近邻个数 $k=8$ (对手写数字和物体识别实验)和 $k=4$ (对人脸识别实验)。

4.1 Binary手写体识别实验

本组实验的数据来源于Binary手写体数据库。Binary手写体数据库包括0–9和A–Z共36组数据，每组数据包括39个二值样本，每一个样本都归一化到 20×16 大小。

该数据库的36个类别均作为已知类，训练样本由每组数据均随机抽取的10个样本组成，而对于某一类别的多观测样本在剩余的29个样本中按照一定的步长随机抽取。该实验比较了所有的算法对不同数目多观测样本的识别率，以此来评估所有算法对不同数目多观测样本的鲁棒性，即对测试模式的多观测样本按照步长 $R=5$ 进行抽取 $m=[10:R:20]$。对不同数目的多观测样本 $\boldsymbol{X}^{(u)}$，每个类别均做10次随机实验，实验结果中的每一点都是36个类别的360次随机实验的均值。实验结果如表1所示。

分析表 1 的实验结果，可以得到如下结论：多观测样本分类算法的识别率要高于单观测样本分类算法 1NN, SVM 的识别率，因为多观测样本可以提供更多的关于某一个类别的信息，把多观测样本看作一个整体来进行分类，获得的识别率要更高；基于凸包的分类算法优于 MMD 和 MASC, MMD 首先找到局部线性模块，然后根据定义的流形到流形的距离来实现多观测样本的分类，该方法对多个参数有较大的依赖性；MASC 算法根据 k 近邻图，在标记传播算法的基础上，把寻找最优类标矩阵的计算转化为离散目标函数优化问题，来实现多观测样本的分类，但是 MASC 中近邻个数的选取是非自适应的，并且权值的计算采用 L2 欧氏距离下的高斯核函数，依赖与参数并且基于欧氏距离测度难以完全反映数据复杂的空间结构分布特性；L1_CHDD 算法通过构造凸包覆盖模型来描述样本的分布，采用 L1 范数距离测度，把多观测样本的分类转化为带绝对值的线性规划问题，L2_CHDD 算法采用 L2 范数作为距离测度，转化为凸二次规划问题来实现多观测样本的分类；这两个方法不依赖于参数，都

表1 在Binary手写体数据库的识别率百分比及标准差(括号内的数值)

多观测样本的个数	1NN	SVM	MMD	MASC	L2_CHDD	L1_CHDD
10	60.9167(1.7619)	61.4167(1.3824)	91.1111(2.1911)	92.5000(2.6352)	95.2778(1.8749)	95.8333(1.4640)
15	61.1296(1.2267)	61.1481(1.0756)	91.9444(2.0496)	94.7222(2.4322)	96.1111(1.9422)	96.3889(1.8749)
20	60.9444(1.5820)	61.1944(1.4015)	93.6111(1.3418)	95.0000(1.7568)	97.2227(0.2418)	96.6667(1.1712)

是通过转化为优化问题来实现分类的，只是优化的方法不同而已。由实验结果知 L1_CHDD 的识别率在多观测样本个数为 20 的时候略低于 L2_CHDD 算法。

4.2 ETH-80物体识别实验

本组实验的数据来源于ETH-80数据库。该数据库包括8个种类如图2所示，每个种类又包括10个不同的物体类，每个物体类由41个样本组成。图3展示了其中一个物体类horse的41个样本。该实验采用的是已裁剪过的数据，大小是 128×128 。为了实验的方便，实验中将图像归一化为 32×32 大小。

该数据库的8个种类均作为已知类，每个种类的训练样本由其10个不同的物体类均随机抽取的10幅图像组成，对于某一种类中的某一物体类的多观测样本由剩余的31幅图像按照一定的步长随机抽取，即对测试模式的多观测样本按照步长 $R=10$ 进行抽取 $m=[10:R:30]$ 。对不同数目的多观测样本，每个类别均做10次随机实验，图中的每个值都是800次实验的均值。实验结果如表2所示。

图2 ETH-80数据库

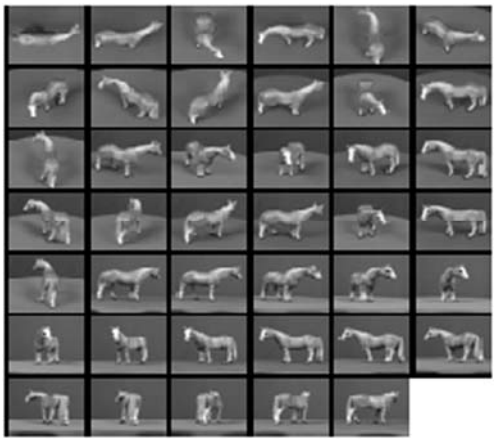

图3 其中一个horse模型的41个样本

实验结果表明，每个算法的识别率会随着多观测样本数目的增加而提高，增加多观测样本的数目会提供更多的信息，因此算法识别率会提高；MMD采用局部线性模块来表示非线性流形，从而实现多观测样本的分类，依赖于太多的参数；MASC采用L2欧式距离测度寻找样本的近邻来构造图，但是欧氏距离下的近邻往往并非同类样本；L1_CHDD的识别率优于L2_CHDD，采用L1范数距离测度的带绝对值的线性规划的优化结果优于凸二次规划，这说明基于L1范数距离测度的多观测样本分类算法的有效性。

4.3 ORL人脸识别实验

本组实验数据来源于ORL(Olivetti Research Lab)的标准人脸图像库。ORL标准人脸库包括400幅灰度图像，共40人，每人10幅，分辨率为 92×112 ，含有光照、表情、姿态等的变化。为了实验的方便，实验中将图像双三次差值为 32×32 大小，并按列展成1维样本向量。

(1)该数据库的40个类别均作为已知类，训练集由每个人均随机抽取的5幅图像组成，而对于某一类别的多观测样本由剩余5幅图像组成。该实验对每个人的多观测样本均做10次随机实验，所以表中的每个值都是400次实验的均值。实验结果如表3所示。

分析表3的实验结果，可以得到如下的结论：L1_CHDD 算法和 L2_CHDD 算法优于 MMD, MASC算法，L1_CHDD和L2_CHDD用凸包模型来覆盖样本的分布，采用L1范数和L2范数距离测度进行凸包之间的相似性度量，进行多观测样本的分类，获得了较好的识别率；L1_CHDD的识别率略高于L2_CHDD，说明采用L1范数距离测度来处理多观测样本分类问题的可行性。

(2)为了验证算法对噪声的鲁棒性，本文在原始

表2 在ETH-80数据库的识别率百分比及标准差(括号内的数值)

多观测样本的个数	MMD	MASC	L2_CHDD	L1_CHDD
10	83.2500(8.2937)	82.3750(8.6597)	96.3750(4.8868)	96.8750(3.8661)
20	89.2500(7.5734)	87.5000(8.3648)	97.1250(3.7154)	97.2500(3.3349)
30	91.6250(4.8679)	90.3750(6.3003)	97.2500(3.2943)	98.1250(2.9921)

表3 在ORL人脸数据库的识别率百分比和标准差

算法	MMD	MASC	L2_CHDD	L1_CHDD
识别率(标准差)	96.0000(3.1623)	97.2500(1.8447)	99.5000(1.0541)	99.7500(0.3536)

训练集(原始训练集由每个人均随机抽取的4幅图像组成，而对于某一类别的多观测样本在剩余图像中随机抽取5幅组成)中加入噪声：对于第c个类别的噪声，从其余类别的剩余样本中随机抽取一个加入第c个类别中。该实验对每个人的多观测样本均做10次随机实验，表中的每个值都是400次实验的均值。实验结果如表4所示。

表4 有噪声ORL人脸数据库的识别率百分比和标准差

算法	L2_CHDD	L1_CHDD
识别率(标准差)	98.5000(1.2910)	99.2500(1.2076)

分析表4的实验结果，可以得到如下的结论：L1_CHDD算法与L2_CHDD算法相比，对噪声有更好的鲁棒性，L2范数距离测度比L1容易受到噪声的影响。

5 结束语

针对某一特定模式的多观测样本的分类问题，利用凸包模型来逼近高维空间样本的分布，把多观测样本的分类问题转化为凸包模型之间的相似性度量，提出基于L1范数凸包数据描述的多观测样本分类算法。该算法首先根据凸包的知识得到训练集的每一类以及测试集的多观测样本的凸包模型，然后判断凸包模型之间的距离是否为0，分别采用带绝对值的线性规划对凸包模型或者收缩凸包模型进行相似性度量，从而实现多观测样本的分类。在物体识别和人脸识别的实验上，证明了以L1范数作为距离测度的合理性和有效性，并且L1比L2有更好的噪声鲁棒性。

参 考 文 献

[1] Kim T K, Kittler J, and Cipolla R. On-line learning of mutually orthogonal subspaces for face recognition by image sets [J]. *IEEE Transactions on Image Processing*, 2010, 19(4): 1067–1074.

[2] Shakhnarovich G, Fisher J W, and Darrel T. Face recognition from long-term observations[C]. European Conference on Computer Vision(ECCV), San Diego, CA USA, 2002, 3: 851–868.

[3] Arandjelovic O, Shakhnarovich G, Fisher J, *et al*.. Face recognition with image sets using manifold density divergence [C]. IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), San Diego, CA USA, 2005, 1: 581–588.

[4] Fukui K and Yamaguchi O. Face recognition using multi-viewpoint patterns for robot vision [C]. 11th International Symposium on Robotics Research, Siena, Italy, 2003, (15): 192–201.

[5] Kim T K, Kittler J, and Cipolla R. Discriminative learning and recognition of image set classes using canonical correlations[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 1005–1018.

[6] Wolf L and Shashua A. Learning over sets using kernel principal angles[J]. *Machine Learning Research*, 2003, 4(10): 913–931.

[7] Kim T K, Arandjelovic O, and Cipolla R. Boosted manifold principal angles for image set-based recognition[J]. *Pattern Recognition*, 2007, 40(9): 2475–2484.

[8] Li X, Fukui K, and Zheng N N. Image-set based face recognition using boosted global and local principal angles[C]. 9th Asian Conference on Computer Vision(ACCV), Xi'an, 2010: 323–332.

[9] Wang R P, Shan S G, Chen X L, *et al*.. Manifold-manifold distance with application to face recognition based on image set[C]. IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), Anchorage, Alaska, USA, 2008: 1–8.

[10] Chu W S, Chen J C, and Lien J J. Kernel discriminant transformation for image set-based face recognition[J]. *Pattern Recognition*, 2011, 44(8): 1567–1580.

[11] Kokiopoulou E, Pirillos S, and Frossard P. Graph-based classification of multiple observation sets[J]. *Pattern Recognition*, 2010, 43(12): 3988–3997.

[12] Bonde U D, Kim T K, and Ramakrishnan K R. Randomised manifold forests for principal angle-based face recognition [C]. 10th Asian Conference on Computer Vision (ACCV), Queenstown, New Zealand, 2011: 228–242.

[13] Takigawa I, Kudo M, and Nakamura A. Convex sets as prototypes for classifying patterns[J]. *Engineering Applications of Artificial Intelligence*, 2009, 22(1): 101–108.

[14] Liu Z B, Liu J G, and Pan C. A novel geometric approach to binary classification based on scaled convex hulls[J]. *IEEE Transactions on Neural Networks*, 2009, 20(7): 1215–1220.

[15] Zhou X F, Jiang W H, and Tian Y J. Kernel subclass convex hull sample selection method for SVM on face recognition [J]. *Neuro-computing*, 2010, 73(10-12): 2234–2246.

[16] 胡正平, 贾千文. 基于 SRM 自组织多区域覆盖的可拒绝近邻分类算法研究[J]. 电子与信息学报, 2009, 31(2): 293–296.

[17] 胡正平, 许成谦, 贾千文. 基于高维空间最小生成树自适应覆盖模型的可拒绝分类算法[J]. 电子与信息学报, 2010, 32(12): 2895–2900.

胡正平： 男，1970 年生，在站博士后，教授，硕士生导师、目前研究方向为模式识别.

王玲丽： 女，1986 年生，硕士生，研究方向为多观测样本分类模型.