采用概率密度比值估计的距离度量学习

吕清秀 李弼程 高毫林
(解放军信息工程大学信息系统工程学院，郑州，450002)

摘 要：现有的距离度量学习算法都是假设训练数据和测试数据服从相同的分布，但是该假设在实际中不一定成立。当训练数据和测试数据的分布不同时，利用训练数据学习得到的度量函数可能难以适用于测试数据。针对上述问题，本文在 NCA(Neighbourhood Components Analysis)度量学习方法的基础上，通过引入概率密度比值对目标函数加权，提出了一种采用概率密度比值估计的距离度量学习方法(Distance metric learning with Probability Density Ratio Estimation，DML-PDR)。在 UCI 数据集和 Corel 图像库上的 KNN 分类实验表明，新方法克服了传统度量学习方法的不一致问题，提高了分类的准确率。

关键词：距离度量学习；半正定规划；概率密度比值估计；图像分类
中图分类号：TP391 **文献标识码**：A **文章编号**：1003-0530(2013)05-0607-08

Distance metric learning with Probability Density Ratio Estimation

LV Qing-xiu LI Bi-cheng GAO Hao-lin
(Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou, China, 450002)

Abstract: Previous distance metric learning algorithms assume that the training data and test data have the same distribution, but the assumption may be not always true in practice. When the training data and test data have different distribution, the distance metric learned from the training data may be not fit for test data. In order to resolve above-mentioned problem, this paper propose a novel distance metric learning with probability density ratio estimation based on NCA(Neighbourhood Components Analysis), which weight the objective function by applying the probability density ratio. The KNN classification on UCI data sets and Corel images demonstrate that the new method resolve the inconsistent of traditional distance metric learning.

Key words: distance metric learning; semi-definite programming; probability density ratio estimation; images classification

1 引言

大量的机器学习方法，如 k 近邻、径向基函数网络、支持向量机等[1]分类方法以及 k-means 等聚类[2]方法，其性能好坏主要由相似性度量方式决定。尽管欧氏距离简单，但是由于没有考虑特征不同维度的区分性，也没有考虑到不同维度之间的相关性。与欧式距离相比，学习式的距离度量是通过训练数据学习一个能够反映样本空间特性的距离函数，包括有监督的距离度量学习和无监督的距离度量学习及半监督的距离度量学习。监督的距离度量学习主要是利用标注样本学习到一个反映样本语义关系的度量函数，使得语义上相近的样本之间距离较近，反之则较远；无监督的距离度量学习是学习一种潜在的低维流形，使得观测数据的几何关系(如距离)被保留下来，多指一些降维方法。而半监督的距离度量学习则是在利用标注数据训练的同时，也利用了大量未标注数据的分布信息。

近些年，有监督的距离度量学习技术研究已

收稿日期：2012-11-19；修回日期：2013-03-05
基金项目：全军军事学研究生课题资助项目(YJS1062)

经取得了很大的进展。2003 年,Xing[3] 提出了一种运用半正定规划(Semi-definite Programming, SDP)进行距离度量学习的方法,通过最小化类内数据之间的距离,同时使得类间数据分离。2006 年,Goldberger 等人[4] 提出了一种近邻成分分析方法(Neighbourhood Components Analysis, NCA),它通过最小化分类错误概率学习一种距离度量。2007 年, Davis 等人[5] 又提出了一种基于信息论的距离度量学习方法(Information-Theoretic Metric Learning,ITML),它利用了最小化多高斯之间的相关信息熵,并提出将距离度量学习转化为 Bregman 最优化问题。2009 年,Weinberger 等人[6] 将 NCA 进行了扩展,提出了一种大间隔最近邻(LargeMargin Nearest Neighbor,LMNN),同年,刘博等[7] 将特征分解运用到距离度量学习中用于降低度量学习的算法时间复杂度。2010 年,吴磊等[8] 提出了概率相关成分分析方法。该方法利用概率信息作为边信息进行距离学习。2012 年,Ying 等[9] 提出了一种特征值最优化的距离度量学习方法,用于减少算法时间复杂度。

上述有监督度量学习只利用了有限的标注数据,且常会遇到训练数据不足的问题,而实际中却有大量未标注的数据存在。半监督学习算法通过对未标注数据加以利用,以获得更准确的模型。目前已有很多半监督学习算法,Yeung 等[10] 提出了一种基于核的半监督距离度量学习方法,但没有利用拓扑结构。Baghshah 等[11] 通过保留类似于 LLE[12] 的局部关系学习距离度量,然而该方法没有考虑到聚类信息(Topo-preserved)。Steven[13] 等提出了一种 laplacian 正则化度量学习(LRML),将样本点的近邻看作相似点,联合已有标注数据学习距离度量。但 LRML 没有考虑到分布的概率密度信息。

在很多有监督的学习中,标注的训练数据和测试数据分布不同的情况经常存在。例如,训练数据不足,或者训练数据是在有偏抽样下获得的,或由于环境变化而相应的测试数据的分布也发生了变化,在这些情况下,运用训练数据得到的度量矩阵不一定适用于测试数据。如图 1 所示,对于训练数据来说,第一维 x 轴具有更多的区分信息,然而对于测试数据来说,第二维 y 轴却包含更多的区分信息。因此,使得训练数据中的相似约束对靠近的度量函数在数据分布改变后不一定能继续使得这些相似约束对仍然靠近。这就是距离度量学习的不一致问题。

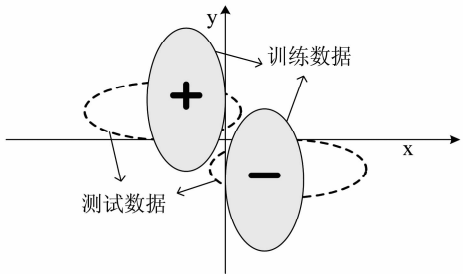

图 1 对于该二类数据的分类来说,在训练数据中,x 轴具有更多的区分信息;而在测试数据中 y 轴具有更多的区分信息。从训练数据中学习得到的度量函数在用于测试数据的时候有偏差

Fig. 1 For the two classes data classification. On the training dataset, the x-axis is crucial to distinguish the two classes. However, on the test set the y-axis is more important instead. The metric learned from the training set will have bias when generalized to test set

数据挖掘和机器学习等领域的研究中经常遇到训练数据和测试数据分布不同的问题。例如在迁移学习[14] (transfer learning)、多任务学习[15] (multi-task learning)、离群点检测(outlier detection)[16]、条件密度估计以及概率分类等[17] 实际问题中,通常都需要估计两个分布不同的样本空间的概率密度的比值[18]。但是鲜有对距离度量学习中训练数据和测试数据分布不同问题的研究。事实上,距离的概念和分布非常密切。欧式距离和高斯分布相关,而曼哈顿距离(Manhattan)和 Laplace 分布相关。因此,分布的改变将影响度量学习的泛化能力,使得先前的学习结果不能适用于分布改变后的数据。在协变量转移(covariate shift,训练数据和测试数据分布 $p(x)$ 不同而输的条件概率 $p(y|x)$ 保持不变)情况下,传统的度量学习方法就不能得到一致的结果,即使训练数据量很大也不能得到最优解。常用的补偿由协变量转移引起的偏差的方法是根据样本重要性对损失函数进行重加权,权重

是测试数据和训练数据的概率密度比值。

本文主要研究协变量转移情况下距离度量学习的不一致性问题,提出了一种采用概率密度比值估计的距离度量学习方法(Distance metric learning with Probability Density Ratio Estimation, DML-PDR),新方法利用直接概率密度比值估计方法计算出测试数据和训练数据的分布的概率密度比值,然后将该比值引入近邻成分分析中对样本进行加权。该方法解决了距离度量学习的不一致问题。最后我们在 UCI 数据集和 Corel 图像库上进行了 KNN 分类实验,将本文方法和几种经典的度量学习方法进行了比较,验证了其有效性。

2　距离度量学习及不一致性问题

2.1　距离度量学习

监督框架下的距离度量学习可以通过含有标注信息的训练数据学习一个能够反映样本语义关系的距离函数,使得相似的样本之间的距离减小,非相似的样本之间的距离增大。本文方法主要基于 NCA 距离度量学习。NCA 以概率的方式定义点的软邻域(soft neighborhood),然后通过最大化训练样本的留一法分类错误率学习距离度量矩阵。

首先令度量矩阵 $\boldsymbol{Q}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$,从而保证度量矩阵的半正定条件。$\boldsymbol{A}$ 相当于变换矩阵,通过变换矩阵 $\boldsymbol{A}$ 对样本进行变换,在变换后的度量空间里相似的样本之间的欧式距离减小,非相似样本之间的欧式距离增大。设由向量 $\boldsymbol{x}_i \in R^m$ 构成的训练集合为 $C=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$。$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的距离为:

$$\begin{aligned} d(\boldsymbol{x}_i,\boldsymbol{x}_j) &= (\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{x}_i-\boldsymbol{x}_j) \\ &= (\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j) \end{aligned} \tag{1}$$

为了降低距离样本较远点对样本分类结果的影响,增强距离样本较近点对分类结果的影响,NCA 采用局部距离 $\exp(-\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2)$ 来学习到适合 KNN 分类的最优解 $\boldsymbol{A}$。对于样本 $\boldsymbol{x}_i$,除该样本外的任何样本 $\boldsymbol{x}_j$ 成为 $\boldsymbol{x}_i$ 近邻的概率可表示为:

$$p_{ij}=\frac{\exp(-\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2)}{\sum_{k\neq i}\exp(-\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_k\|^2)} \qquad p_{ii}=0 \tag{2}$$

设 c_j 为样本 $\boldsymbol{x}_j$ 的类别,将所有和 $\boldsymbol{x}_i$ 为一类的样本记为 $C_i=\{j \mid c_i=c_j\}$,则 $\boldsymbol{x}_i$ 被正确分类的概率为

$$p_i=\sum_{j\in C_i}p_{ij} \tag{3}$$

则全部被正确分类的样本个数的期望为

$$f(\boldsymbol{A})=\sum_{i=1}^{n}p_i=\sum_{i=1}^{n}\sum_{j\in C_i}p_{ij} \tag{4}$$

这样,满足 $f(\boldsymbol{A})$ 取最大值的度量矩阵 $\boldsymbol{A}$ 就是学习的目标度量矩阵。变换矩阵 $\boldsymbol{A}$ 可以采用基于梯度下降的最优化方法得到。由于 $f(\boldsymbol{A})$ 可微,对 $\boldsymbol{A}$ 的导数表示如下:

$$\frac{\partial f}{\partial \boldsymbol{A}}=2\boldsymbol{A}\sum_{i=1}^{n}\left(p_i\sum_{k\neq i}p_{ik}(\boldsymbol{x}_i-\boldsymbol{x}_k)(\boldsymbol{x}_i-\boldsymbol{x}_k)^{\mathrm{T}}-\sum_{j\in C_i}p_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\right) \tag{5}$$

最大化目标函数 $f(\boldsymbol{A})$ 等同于让每个样本 $\boldsymbol{x}_i$ 被正确分类的概率趋近于 1,即对如下目标函数取最大值。

$$g(\boldsymbol{A})=\sum_{i=1}^{n}\log(p_i)=\sum_{i=1}^{n}\log\Big(\sum_{j\in C_i}p_{ij}\Big) \tag{6}$$

通过最大化目标函数 $g(\boldsymbol{A})$ 可以得到更为准确的度量矩阵 $\boldsymbol{A}$,同时我们注意到 $g(\boldsymbol{A})$ 的梯度相比 $f(\boldsymbol{A})$ 更为简单。

$$\frac{\partial g}{\partial \boldsymbol{A}}=2\boldsymbol{A}\sum_{i=1}^{n}\left(\sum_{k\neq i}p_{ik}(\boldsymbol{x}_i-\boldsymbol{x}_k)(\boldsymbol{x}_i-\boldsymbol{x}_k)^{\mathrm{T}}-\frac{\sum_{j\in C_i}p_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}}{\sum_{j\in C_i}p_{ij}}\right) \tag{7}$$

2.2　度量学习中的不一致性问题

上述距离度量学习假设训练数据的分布和测试数据的分布是一致的,从而保证了学习得到的度量函数的泛化能力。然而当训练数据和测试数据的分布不一致时,在训练数据中通过最优化目标函数求得的参数不一定能使得目标函数在测试数据中仍然最优。实际中训练数据由于需要人工标注,因此训练数据的规模都比较小,而测试数据规模比较大,因此训练数据和测试数据分布不同的情况在度量学习中经常存在。通

常运用概率密度比值解决这种最优化问题中的不一致性问题。

设训练数据 $\boldsymbol{x}_{train}$ 的概率密度为 $p_{tr}(\boldsymbol{x})$，测试数据 $\boldsymbol{x}_{test}$ 的概率密度为 $p_{te}(\boldsymbol{x})$，设含待求解参数 $\boldsymbol{\theta}$ 的目标函数为 $g(\boldsymbol{x},\boldsymbol{\theta})$，则目标函数在测试数据中的期望值为 $\int g(\boldsymbol{x},\boldsymbol{\theta})p_{te}(\boldsymbol{x})d\boldsymbol{x}$，学习的目的是让得到的目标函数在测试数据中最优。而具体求解过程中是利用训练数据学习参数，目标函数的期望值是 $\int g(\boldsymbol{x},\boldsymbol{\theta})p_{tr}(\boldsymbol{x})d\boldsymbol{x}$。为了解决这种问题，引入权重 $w(\boldsymbol{x})=\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}$ 对训练数据的分布进行修正，尽可能消除参数估计的偏差，解决不一致问题，如下式所示：

$$\int g(\boldsymbol{x},\boldsymbol{\theta})p_{te}(\boldsymbol{x})d\boldsymbol{x}=\int g(\boldsymbol{x},\boldsymbol{\theta})w(\boldsymbol{x})p_{tr}(\boldsymbol{x})d\boldsymbol{x} \tag{8}$$

3 采用概率密度比值估计的距离度量学习

由 2.2 所述可知运用概率密度比值对目标函数加权可以解决不一致问题。可以认为每个训练样本点的概率密度比值相当于该 1 点代表的测试样本个数乘以一常数项，因为概率密度比值与密度比值之比为常数，而常数项对于最优化问题无关。基于以上分析，本部分首先将概率密度比值引入了 NCA 中对其进行了改进，然后讨论了概率密度比值函数的估计方法。

3.1 本文方法

设 $w(\boldsymbol{x})$ 为测试样本和训练样本的概率密度比值函数，对于一个确定的样本 $\boldsymbol{x}_i$，则除该样本外的任何样本 $\boldsymbol{x}_j$ 成为 $\boldsymbol{x}_i$ 近邻的概率关于测试样本分布的期望值可表示为：

$$p'_{ij}=\frac{w(\boldsymbol{x}_j)\exp(-\left\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\right\|^2)}{\sum_{k\neq i}w(\boldsymbol{x}_k)\exp(-\left\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_k\right\|^2)} \tag{9}$$

设 c_j 为样本 $\boldsymbol{x}_j$ 的类别，将所有和 $\boldsymbol{x}_i$ 为一类的样本记为 $C_i=\{j\mid c_i=c_j\}$，$\boldsymbol{x}_i$ 被正确分类的概率为

$$p'_i=\sum_{j\in C_i}p'_{ij} \tag{10}$$

同理，式(4)中的 $f(\boldsymbol{A})$ 变为下式，相当于测试样本中全部被正确分类的样本个数的期望值除以一常数

$$f'(\boldsymbol{A})=\sum_{i=1}^{n}w(\boldsymbol{x}_i)p'_i \tag{11}$$

相应的式(6)中的 $g(\boldsymbol{A})$ 变为：

$$g'(\boldsymbol{A})=\sum_{i=1}^{n}w(\boldsymbol{x}_i)\log(\sum_{j\in C_i}p'_{ij}) \tag{12}$$

同时得到 $g'(\boldsymbol{A})$ 的梯度为

$$\frac{\partial g'}{\partial \boldsymbol{A}}=2\boldsymbol{A}\sum_{i=1}^{n}w(\boldsymbol{x}_i)\left(\sum_{k\neq i}p'_{ik}(\boldsymbol{x}_i-\boldsymbol{x}_k)(\boldsymbol{x}_i-\boldsymbol{x}_k)^{\mathrm{T}}-\frac{\sum_{j\in C_i}p'_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}}{\sum_{j\in C_i}p'_{ij}}\right) \tag{13}$$

变换矩阵 $\boldsymbol{A}$ 可以通过对 $g'(\boldsymbol{A})$ 采用基于梯度下降的最优化方法得到。

3.2 概率密度比值估计

权重 $w(\boldsymbol{x})$ 是解决距离度量学习一致性问题的关键，本文采用了 LSIF[12]（least-squares importance fitting）直接估计概率密度比值，相比分别估计训练数据概率密度 $\hat{p}_{tr}(\boldsymbol{x})$ 和测试数据的概率密度 $\hat{p}_{te}(\boldsymbol{x})$ 再求 $\hat{w}(\boldsymbol{x})$ 的方法可以避免更多误差的引入。LSIF 主要是将概率密度比值估计问题转化为一个二次函数的最小值问题，具体思想如下：

首先对 $w(\boldsymbol{x})$ 进行建模：

$$\hat{w}(\boldsymbol{x})=\sum_{l=1}^{b}\alpha_l\varphi_l(\boldsymbol{x}) \tag{14}$$

$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_b)^{\mathrm{T}}$ 是从样本中学习到的参数，$\{\varphi_l(\boldsymbol{x})\}_{l=1}^{b}$ 是基函数，且：

$$\varphi_l(\boldsymbol{x})>0,\ \forall \boldsymbol{x}\in\mathbb{R}^d,l=1,2,\ldots,b \tag{15}$$

在对概率密度比值函数进行建模后，我们通过让下面的平方差最小来确定参数 $\{\alpha_l\}_{l=1}^{b}$：

$$\begin{aligned}J_0(\boldsymbol{\alpha})&=\frac{1}{2}\int(\hat{w}(\boldsymbol{x})-w(\boldsymbol{x}))^2p_{tr}(\boldsymbol{x})d\boldsymbol{x}\\&=\frac{1}{2}\int\hat{w}(\boldsymbol{x})^2p_{tr}(\boldsymbol{x})d\boldsymbol{x}-\int\hat{w}(\boldsymbol{x})w(\boldsymbol{x})p_{tr}(\boldsymbol{x})d\boldsymbol{x}\\&\quad+\frac{1}{2}\int w(\boldsymbol{x})^2p_{tr}(\boldsymbol{x})d\boldsymbol{x}\\&=\frac{1}{2}\int\hat{w}(\boldsymbol{x})^2p_{tr}(\boldsymbol{x})d\boldsymbol{x}-\int\hat{w}(\boldsymbol{x})p_{te}(\boldsymbol{x})d\boldsymbol{x}\\&\quad+\frac{1}{2}\int w(\boldsymbol{x})^2p_{tr}(\boldsymbol{x})d\boldsymbol{x}\end{aligned} \tag{16}$$

上式中第二项的训练样本概率密度 $p_{tr}(\boldsymbol{x})$ 和 $w(\boldsymbol{x})$ 的分子抵消后剩下 $p_{te}(\boldsymbol{x})$。第三项是常数，因此求最小

值时忽略。上式可进一步转化为凸二次规划问题,由文献[18]所述算法可以得到唯一的全局最优解。

本文的基函数 $\varphi_l(\boldsymbol{x})$ 采用的是高斯核函数 $\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_l\|^2}{2\sigma^2})$,核函数的中心 $\boldsymbol{x}_l$ 是来自训练样本和测试样本的一些点,本文采用的是随机选择 $b=\min(100,n_{te})$ 个测试样本分别作为 b 个核函数的中心,基函数参数的选择对估计的效果影响很大,本文采用交叉验证法(cross-validation)求得最优的核宽度 σ 和调整参数。

4　实验分析

本实验是在一台内存为 1G,主频为 2.4GHz 的 PC 机上实现的,实验环境为 matlab2009。分别采用了 UCI Machine Learning Repository[19] 中选取的若干组数据集和 Corel 图像数据库的部分数据(如图 2 所示)进行了 3NN 分类实验。所对比的方法包括有监督的距离度量学习方法 ITML[5]、NCA[4] 和半监督的距离度量学习方法 Topo-preserved[12]、LRML[13]。通过分类错误率比较各度量方法的性能。分类错误率的计算公式如下:

错误率=(错误分类图像数/测试图像总数)×100%　　(17)

4.1　UCI 数据集上的实验

UCI 数据集是常用于机器学习及数据挖掘的标准数据库,很多度量学习方面的算法都用其进行算法性能评价[4],[5],[7],[10]。表 1 列出了所采用的 UCI 各数据集名称、样本总数、特征维数、样本类别数、训练样本数及训练样本所占百分比。分类结果如图 2 所示。

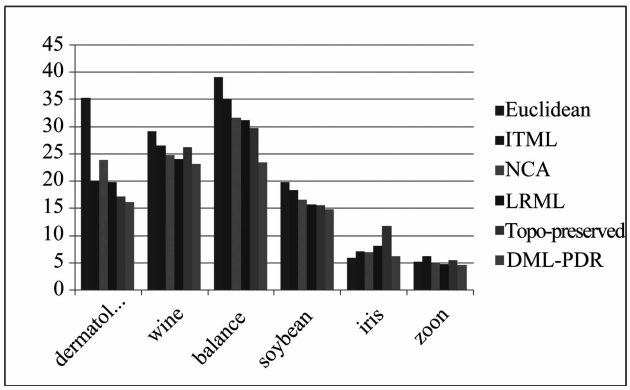

图 2　UCI 数据集上的分类错误率比较

Fig. 2　Comparison of classification error rate on UCI data sets

表 1　UCI 数据集的相关信息

Tab. 1　Description of the UCI data sets

数据集名称	样本总数	特征维数	样本类别数	训练样本数	训练样本百分比
dermatology	358	34	6	30	8.38
wine	178	13	3	15	8.43
balance	625	4	3	15	2.4
soybean	562	35	15	75	13.35
iris	150	4	3	15	10
zoo	101	16	7	35	34.65

由图 2 分类结果可以看出,除在 iris 数据集上外,所有度量学习方法均优于基本的欧式距离,半监督的度量学习方法略优于监督的度量学习方法。但都不及本文采用概率密度比值估计的距离度量学习方法(DML-PDR)。由此可以看出本文方法在训练样本较少时,通过概率密度比值对目标函数加权,得到了更为可靠的距离度量函数。为了进一步分析训练样本和测试样本分布不一致情况下本文算法的度量学习性能,我们在 Corel 图像集上人为设置不同的情况的训练样本进行度量学习,通过分类错误率评价度量算法性能。

4.2　Corel 数据集上的试验

本文试验采用的 Corel 图像集包括印第安部落、海滩、古建筑、公交车、恐龙、花、大象、马、美食、雪山 10 类共 1000 幅图像(图 3 所示)。Corel 图像数据库拥有以下特点: 1. 图像光照强度、拍摄角度等有很大不同; 2. 对于目标物体类图像,物体的出现位置,背景条件各不相同; 3. 对于自然景观类图像,光照条件景观样式等等也有很大差异。因此该数据库图像在检索、分类过程中难度很大, 经常使用于图像检索与分类的实验。

对于 Corel 图像集中的图像,本实验提取了一种压缩域特征,通过对图像的 DCT 系数处理得到每个 DCT 块内像素的均值 u 和标准方差 σ。按其统计参数 $u-\sigma$ 将 YCbCr 三个分量的亮度空间 Y 分成 28 个子空间,将两个色度 CbCr 空间分别分成 15 个子空间。然后对 YCbCr 各个分量中的 DCT 块落到相应子空间的个数分别进行统计并相应的进行归一化。最后得到一个 28 维的直方图和两个 15 维直方图连接起来的 58 维直方图。因为归一化后的直方图特征比较小,因此本文给每维特征都乘了常数 10,来增大 NCA 中局部距离的作用。

图 3 实验中 Corel 数据库的 10 类图像实例

Fig. 3 Representative sample image of 10 classes in Corel database for the experiment

为了进一步分析本文方法对度量学习中不一致问题的有效性,我们对 Corel1000 图像集提取特征后,分别从印第安部落、海滩、古建筑、公交车、恐龙、花、大象、马、美食、雪山 10 类的图像中分别按如下六种方案进行了随机抽样以作为训练数据,试验分别进行 10 次,取平均值。方案 1:18, 5, 2, 2, 3, 10, 2, 3, 4, 11 ;方案 2,每类 6 幅;方案 3:30, 5, 2, 15, 3, 30, 16, 25, 14, 40;方案 4,每类 18 幅;方案 5:10, 27, 45, 15, 53, 5, 55, 30, 20, 40;方案 6,每类 30 幅。方案 1 和方案 2,方案 3 和方案 4,及方案 5 和方案 6 的区别都是抽样总数相同,但一个是有偏抽样,一个是均匀抽样。方案 3、4 比方案 1、2 的抽样总数多,方案 5、6 比方案 3、4 的抽样总数多。用于 3NN 分类判别的已知类别的每类图像数为抽样总数除以 10,剩余图像为待分类图像。

Corel 图像集上的分类错误率如图 4 所示

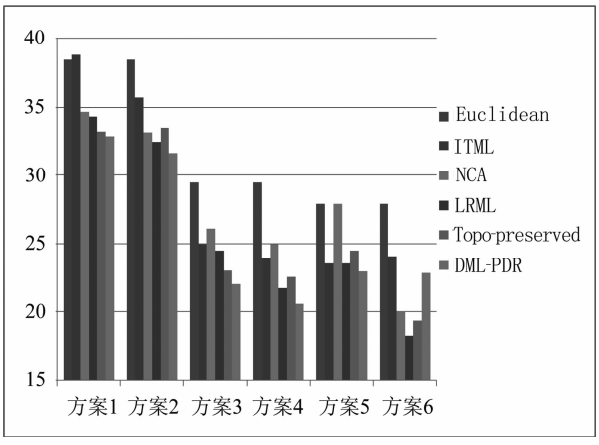

图 4 Corel 图像集上的分类错误率比较

Fig. 4 Comparison of classification error rate of Corel data sets

由图 4 的实验结果可以看出,在抽样总数相同下,均匀抽样比有偏抽样的分类错误率略低,有的度量学习方法在有偏抽样情况下分类效果还没有欧式距离的分类效果好(如在方案 1 中,ITML 没有欧式距离的分类错误率低)。这说明在测试数据和训练数据分布不同时,度量学习存在不一致性。同时在 Corel 图像集上 DML-PDR 度量性能也优于半监督的度量学习方法 LRML、Topo-preserved 和监督的度量学习方法 ITML、NCA。

其次我们注意到训练样本较少情况下的距离度量学习的分类错误率比训练样本数较多情况下的分类错误率要高,这说明了较少的训练样本难以反映测试样本的分布。DML-PDR 在训练样本数比较少的时候相比其他度量学习方法更好的降低了分类错误率。当然我们也看到在均匀抽样的个别情况中,DML-PDR 的分类错误率也有比其他度量学习方法高的时候(如在方案 6 中,DML-PDR 就比 NCA 及 LRML 和 Topo-preserved 的分类错误率高),这是因为此时抽样本来就是均匀的,训练样本和测试样本的分布很接近,概率密度比值估计反而引入了误差。

此外,我们还观察分析了不同抽样情况下概率密度比值函数 $\hat{w}(\boldsymbol{x})$ 的估值(如图 5、6、7 所示)。

图 5、6、7 为不同抽样总数下均匀抽样和有偏抽样的概率密度比值估计的对比图(注意:图中 x 轴的不同整数用于标记不同的样本,且同一图中,对于两种不同的抽样方案,同一整数未必代表同样的样本)。由图可以看出有偏抽样中测试样本与训练

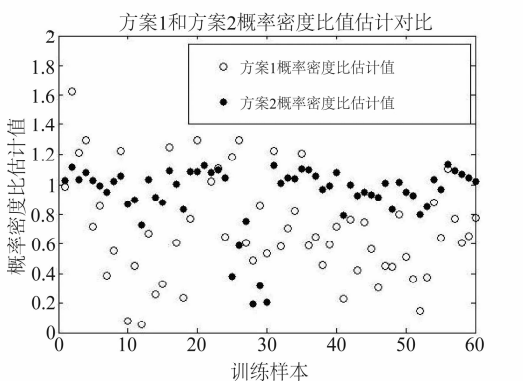

图 5　训练样本总数为 60 时有偏抽样与均匀抽样概率密度比值估计比较

Fig. 5　Comparison of probability density ratio estimation value between biased sampling and uniform sampling when the number of training samples is 60

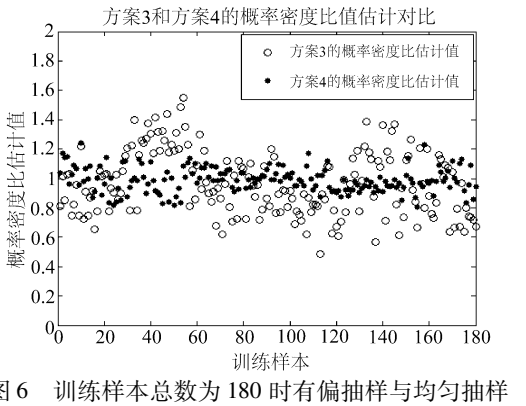

图 6　训练样本总数为 180 时有偏抽样与均匀抽样概率密度比估计值比较

Fig. 6　Comparison of probability density ratio estimation value between biased sampling and uniform sampling when the number of training samples is 180

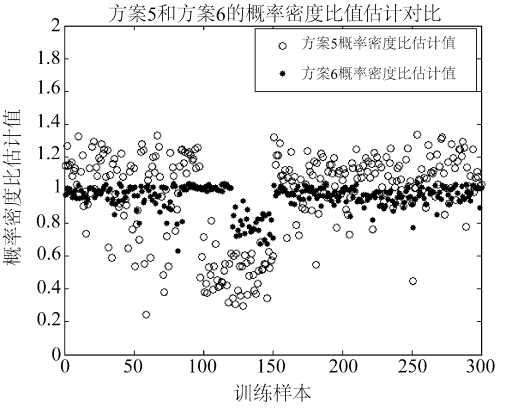

图 7　训练样本总数为 300 时有偏抽样与均匀抽样概率密度比估计值比较

Fig. 7　Comparison of probability density ratio estimation value between biased sampling and uniform sampling when the number of training samples is 300

样本的概率密度比值比较分散,与 1 比较远,说明它们的分布很不相同。均匀抽样中的测试样本和训练样本的概率密度比较集中,且接近与 1,说明他们的分布更加接近。同时应该注意,由于图像底层特征的一些局限,如噪声、特征的随机性、语义鸿沟等,所以特征的分布也不能精确反应语义上的图像分布。其次,概率密度比值估计本身也会因核函数的选择不够恰当存在一定的偏差。

5　总结

本文提出的采用概率密度比值估计的距离度量学习方法较好的解决了训练数据和测试数据分布不同情况下距离度量学习的不一致性问题,提高了度量学习的泛化能力,在 UCI 数据集和 Corel 图像库上进行的分类实验验证了本文方法的有效性。由于概率密度比值估计和度量学习都是一个最优化问题,算法复杂度比较高,因此如何减少算法复杂度有待于进一步的研究。

参考文献

[1]　高恒振, 万建伟. 基于聚类核函数的最小二乘支持向量机高光谱图像半监督分类[J]. 信号处理, 2011, 27(2): 276-280.
GAO Heng-zhen, WAN Jian-wei. Semi-supervised classification of hyperspectral image based on clustering kernel and LS-SVM[J]. Signal Processing, 2011, 27(2): 276-280. (in Chinese)

[2]　叶敏超, 钱沄涛. 基于聚类的图像稀疏去噪[J]. 信号处理, 2011,27(10): 1593-1598.
YE Min-chao, QIAN Yun-tao. Clustering based sparse model for image denoising[J]. Signal Processing, 2011, 27(10): 1593-1598. (in Chinese)

[3]　E Xing, A Ng, et al. Distance Metric Learning, with Application to Clustering with Side-information[C]//Proceedings of Neural Information Processing Systems. MA USA, 2003:505-512.

[4]　Jacob Goldberger, Sam Roweis, et al. Neighbourhood components analysis[C]//In Advances in Neural Information Processing Systems. Washington: MIT Press, 2005:103-110.

[5]　J Davis, B Kulis, et al. Information-theoretic metric learning[C]//In Proceedings of the International Conference on Machine Learning . Florida, USA:ACM, 2007.

209-216.

[6] Weinberger, L K Saul. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009,10(12):207-244.

[7] 刘博. 距离度量学习理论与应用研究[D]. 合肥:中国科学技术大学, 2010.
Bo Liu. The study of theory and application of distance metric learning[D]. He Fei: University of Science and Technology of China, 2010. (in Chinese)

[8] 吴磊. 视觉语义分析:从底层特征表达到语义距离学习[D]. 合肥:中国科学技术大学, 2010.
Lei Wu. Visual language analysis: from low level feature [D]. He Fei: University of Science and Technology of China, 2010. (in Chinese)

[9] Y Ying, P Li. Distance metric learning with eigenvalue optimization[J]. Journal of Machine Learning Research, 2012, 13(1): 1-26.

[10] D Yeung, H Chang. A kernel approach for semi-supervised metric learning[J]. IEEE Transactions on Neural Net-works, 2007, 18(1):141-149.

[11] M Baghshah, S Shouraki. Semi-supervised metric learning using pairwise constraints[C]// Proceedings of the 21st international jont conference on Artifical intelligence, 2009: 1217-1222.

[12] D Roweis, L Saul. Nonlinear dimensionality reduction by local linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[13] Steven Hoi,W Liu. Semi-supervised distance metric learning for collaborative image retrieval and clustering[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2010, 6(3):153-159.

[14] Masashi S,Makoto Y. Direct Density-ratio Estimation with Dimensionality Reduction via Least-squares Hetero-distributional Subspace Search[J]. Neural Net works, 2011, 24(2):183-198.

[15] Bickel S, Bogojeska J, etal. Multi-task learning for HIV therapy screening[C]//Proceedings of 25th Annual International Conference on Machine Learning. New York, 2008:56-63.

[16] Sugiyama M. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting[J]. IEICE Transactions on Information and Systems, 2010,93(10): 2690-2701.

[17] Masashi S,Taiji S. Direct Importance Estimation for Covariate Shift Adaptation[J]. Annals of the Institute of Statistical Mathematics, 2008, 60(4): 699-746.

[18] Takafumi K,Shohei H. A Least-squares Approach to Direct Importance Estimation[J]. Journal of Machine Learning Research, 2009, 10(12):1391-1445.

[19] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

作者简介

吕清秀 男,1985年11月出生于甘肃环县。毕业院校为解放军信息工程大学,获得的学位为学士学位、现为解放军信息工程大学硕士研究生学员。主要研究方向为智能信息处理、图像检索。
E-mail:lv_q_x@163.com

李弼程 男,1970年7月出生于湖南衡南。毕业院校为解放军国防科技大学,获得的学位为博士学位,现担任博导、教授。主要研究方向为智能信息处理。
E-mail:bichen@163.com

高毫林 男,1979年出生于河南新郑。毕业院校为解放军信息工程大学,获得的学位为硕士学位,现为解放军信息工程大学博士研究生学员。主要研究方向为智能信息处理。E-mail: 53431023@qq.com