有监督的距离度量学习算法研究进展

沈媛媛[1] 严 严[1] 王菡子[1]

摘 要 近年来, 距离度量学习已成为计算机视觉和模式识别等领域最为活跃的研究课题之一. 如何利用训练数据学习得到有效的距离度量来衡量目标之间的相似性是该类研究的关键问题. 针对有监督的距离度量学习问题, 目前已提出了大量的研究算法. 结合近年已发表相关文献对有监督的距离度量学习算法进行了详细的介绍和讨论. 根据样本信息利用方式的不同, 将其划分成基于成对约束和非成对约束的距离度量学习算法, 重点介绍了一些常用的典型算法, 分析了每种算法的原理和优缺点, 最后是未来发展方向和趋势的展望.

关键词 距离度量学习, 马氏距离, 成对约束, 非成对约束

引用格式 沈媛媛, 严严, 王菡子. 有监督的距离度量学习算法研究进展. 自动化学报, 2014, **40**(12): 2673−2686

DOI 10.3724/SP.J.1004.2014.02673

Recent Advances on Supervised Distance Metric Learning Algorithms

SHEN Yuan-Yuan[1] YAN Yan[1] WANG Han-Zi[1]

Abstract Recently, distance metric learning has become one of the most attractive research areas in computer vision and pattern recognition. How to learn an effective distance metric to measure the similarity between subjects is a key problem. A large number of algorithms have been proposed to deal with supervised distance metric learning. This paper reviews and discusses recently developed algorithms for supervised distance metric learning. Based on the partition of pairwise constraints and non-pairwise constraints, some representative algorithms are introduced and their respective pros and cons are analyzed. The prospects for future development and suggestions for further research work are presented in the end.

Key words Distance metric learning, Mahalanobis distance, pairwise constraints, non-pairwise constraints

Citation Shen Yuan-Yuan, Yan Yan, Wang Han-Zi. Recent advances on supervised distance metric learning algorithms. *Acta Automatica Sinica*, 2014, **40**(12): 2673−2686

在计算机视觉和模式识别等领域, 我们通常使用特征向量来表征样本. 在众多衡量特征向量相似性的方法中, 距离度量是最基本的方法. 距离度量学习算法[1−7] 在机器学习、模式识别和计算机视觉等领域有着广泛的应用. 因此, 研究距离度量学习算法具有非常重要的理论意义和应用价值.

距离度量学习 (或相似度学习) 是指利用给定的训练样本集学习得到一个能够有效反映数据样本间距离 (或相似度) 的度量矩阵, 使在基于度量矩阵的新特征空间中, 同类样本的分布更加紧凑, 而不同类样本的分布更加松散.

根据学习方式不同, 距离度量学习算法可分为无监督的距离度量学习算法[8−11] 和有监督的距离度量学习算法[9−16] 两类. 无监督的距离度量学习算法的基本思想是利用降维将原始数据集映射到低维子空间中, 从而获得一个关于原数据集紧致的低维表示. Ye 等[9] 提出了一种自适应距离度量学习 (Adaptive metric learning, AML) 算法, 这种算法结合聚类和距离度量学习的思想, 通过降维使数据具有最大可分性, 并在新的低维子空间中进行有效的距离度量. Chen 等[8] 基于 AML 算法结合核学习技术提出了非线性自适应距离度量学习算法. 有监督的距离度量学习算法的主要思想是利用训练集的样本信息, 通过优化某个目标函数, 得到一个能有效反映样本空间关系的度量矩阵. Wang 等[7] 通过最大化互信息熵来学习距离度量.

根据是否借助相关联的任务协助度量学习, 距离度量学习可分为单任务度量学习[17−20] 和多任务度量学习[21−25] 两类. 单任务度量学习是指针对某一特定任务, 利用训练集中的样本来进行目标模型的学习, 从而求解度量矩阵; 但在实际应用中, 针对

收稿日期 2014-01-20 录用日期 2014-04-10
Manuscript received January 20, 2014; accepted April 10, 2014
国家自然科学基金 (61201359, 61170179), 福建省自然科学基金 (2012J05126), 高等学校博士学科点专项科研基金 (20110121110033) 资助
Supported by National Natural Science Foundation of China (61201359, 61170179), Natural Science Foundation of Fujian Province (2012J05126), and Specialized Research Fund for the Doctoral Program of Higher Education of China (20110121110033)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 厦门大学信息科学与技术学院 厦门 361005
1. School of Information Science and Technology, Xiamen University, Xiamen 361005

某个特定任务的训练样本数目可能不够, 需要联合多个任务之间的信息来进行更为有效的距离度量学习, 即采用多任务度量学习. Zhang 等[21] 和 Li 等[22] 利用迁移学习的思想, 将多个相关任务分为源任务和目标任务来实现迁移度量学习. 其中, Zhang 等[21] 通过计算源任务和目标任务之间的协方差来表征任务间的关系, 进而提出了一种基于距离度量和任务间相关关系联合优化的迁移度量学习 (Transfer metric learning, TML) 算法; 而 Li 等[22] 将迁移度量学习应用于人脸验证. Parameswaran 等[23] 在基于最大边际的最近邻 (Large margin nearest neighbor, LMNN) 算法基础上, 利用多任务学习思想提出了多任务 LMNN 算法来改进训练样本不足情况下的 k 最近邻算法性能. Yang 等[24] 指出任务间保持相对距离的重要性, 并提出了基于几何保持的多任务度量学习算法.

经过多年研究, 距离度量学习算法已取得了长足的进步和发展. 如何通过训练数据学习到有效的距离度量来衡量目标之间的相似性是该类研究的重点. 国内外众多的研究机构, 如美国的密西根州立大学[16, 26]、新加坡的南洋理工大学[27]、澳大利亚的国家信息与通信技术研究机构[28−29]、中国科学院[30−31]、香港中文大学[32] 等都对距离度量学习算法进行了广泛而深入的研究. 鉴于距离度量应用的广泛性以及现有的距离度量学习国内外综述文献较少, 且已有的综述[33] 主要是针对 2006 年以前的距离度量学习算法, 因此有必要对现阶段的度量学习算法进行综述. 本文综述主要针对单任务距离度量学习, 重点讨论有监督的距离度量学习算法. 根据样本信息利用方式的不同, 划分为基于成对约束的距离度量学习算法和基于非成对约束的距离度量学习算法, 本文对这两类距离度量学习算法进行回顾、分析和总结, 期望能够更好地指导未来的研究工作.

1 基于成对约束的距离度量学习算法

根据训练集中两个样本是否属于同一类可以将其划分为等值约束集合 S 和非等值约束集合 D. S 中每个元素表示属于同一类的样本对, D 中每个元素表示属于不同类的样本对. 满足上述条件的先验信息称之为成对约束. 相对于类别信息等其他标签信息, 成对约束更容易被获取[34−36], 所以基于成对约束的距离度量学习被广泛应用.

给定 $\mathbf{R}^d$ 空间的 n 个数据点 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$, 通常采用马氏距离进行距离度量, 即要找到一个度量矩阵 $M \in \mathbf{R}^{d\times d}$ 来衡量样本对 $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 之间的距离:

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} M (\boldsymbol{x}_i - \boldsymbol{x}_j)} \quad (1)$$

为了保证平方根有效性, M 通常为半正定矩阵.

基于成对约束的距离度量学习算法利用成对约束信息来指导度量学习过程, 通过最优化某个目标函数来求解度量矩阵.

按照距离度量学习过程中使用的目标函数模型不同, 基于成对约束的距离度量学习算法分为基于样本对距离和的距离度量学习算法、基于信息论的距离度量学习算法、基于概率论的距离度量学习算法和基于余弦相似度的距离度量学习算法, 并分别进行介绍.

1.1 基于样本对距离和的距离度量学习算法

Xing 等[1] 基于样本对距离和提出了一种经典的距离度量学习算法, 该算法的基本思想是: 在最小化相似对之间马氏距离平方和的同时, 约束非相似对之间的马氏距离和 (令其大于某个阈值). 利用该种方式建立的目标优化函数, 使得在新的度量空间中相似对更加紧凑, 而非相似对更加分离. 其目标函数模型如下:

$$\begin{aligned} &\min_{M\succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in S} d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{s.t.} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in D} d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1 \end{aligned} \quad (2)$$

其中, $M \succeq 0$ 表示 M 为半正定矩阵.

上述问题为典型的凸优化问题. 因此, 上述问题存在全局最优解[37]. Xing 等[1] 将要学习的度量矩阵分成对角矩阵 (一种稀疏矩阵) 和全矩阵两种情况, 并分别进行讨论. 该算法结合了梯度下降算法和迭代映射的思想, 对上述凸优化问题进行求解, 并将其应用于改进 k 均值聚类算法的性能. 该算法实现过程简单可行. 但是由于该算法在整个数据集中构造全部相似对和非相似对, 当数据很多时, 构造出的相似对集合和非相似对集合也会过大, 故难以处理大规模的数据集问题.

Hoi 等[27] 提出了一种半监督的距离度量学习算法. 该算法在基于图论的拉普拉斯正则化框架下, 最小化相似对马氏距离平方和的同时最大化非相似对的马氏距离平方和. 本文将该算法应用于图像检索, 在数据存在噪声时仍可以获得优异的实验效果. Ying 等[38] 提出一种新颖的基于特征值最优化框架的度量学习算法. 该算法将文献 [1] 提出的距离度量学习算法转化为求解相似对距离和不大于某个阈值约束, 同时最大化非相似对之间的最小平方距离的问题, 并进一步简化为最小化对称阵的最大特征值问题进行求解. 相比文献 [1] 提出的度量算法 (每次迭代需要进行矩阵的特征分解), 该算法每次迭代仅需计算矩阵的最大特征向量, 大大加快了算法的执

行速度.

1.2 基于信息论的距离度量学习算法

基于信息论的距离度量学习算法利用信息论中的相对熵[18] 来学习度量矩阵. 对于度量矩阵 M 的学习, 通常基于如下的假设:

假设 1. 等值约束对之间的距离不大于某个阈值 u (即 $d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\le u$), 而非等值约束对之间的距离不小于某个阈值 l (即 $d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\ge l$) 且 $l\ge u$.

假设 2. 存在先验的度量矩阵 M_0. 对于满足高斯分布的样本集, 使用样本的协方差矩阵来参数化先验矩阵 M_0, 否则使用欧氏距离来参数化先验矩阵 M_0.

基于以上两个假设学习得到的度量矩阵 M, 不仅可以保证训练样本集中的成对约束尽可能满足阈值条件, 同时可以使度量矩阵尽可能接近先验度量矩阵 M_0.

Davis 等[18] 提出基于信息论的距离度量学习 (Information-theoretic metric learning, ITML) 算法. 对于度量矩阵 M, 存在一个协方差矩阵为 M^{-1} 的多元高斯分布 $p(\boldsymbol{x};M)=(1/z)\exp((-1/2)d_M(\boldsymbol{x},\boldsymbol{\mu}))$, 其中 $\boldsymbol{\mu}$ 为均值, z 为归一化因子. 使用相对熵来评价两个多元高斯分布 (基于要学习的度量矩阵 M 和先验矩阵 M_0) 之间的距离, 即:

$$KL(p(\boldsymbol{x};M_0)\|p(\boldsymbol{x};M)) \tag{3}$$

其中, $KL(\cdot)$ 表示相对熵, 又称 Kullbacle-Leibler 散度, 用来表征两个概率分布之间的差异性.

在给定等值约束集合 S 和非等值约束集合 D 的条件下, 距离度量学习问题可以转化为求解以下最优化问题:

$$\begin{aligned}&\min_{M\succeq 0} KL(p(\boldsymbol{x};M_0)\|p(\boldsymbol{x};M))\\ &\text{s.t.}\quad d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\le u,\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\in S\\ &\qquad d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\ge l,\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\in D\end{aligned} \tag{4}$$

上述目标函数的约束条件是为了保证学习到的度量矩阵尽可能满足成对约束的阈值条件; 同时最小化 KL 散度可以使度量矩阵 M 和先验的度量矩阵 M_0 尽可能接近, 进而防止过拟合问题.

通常假定由需要学习的度量矩阵 M 和先验矩阵 M_0 生成的两个高斯分布的均值相同, 则其相对熵可以通过定义在正定矩阵锥上的凸函数 $\phi(X)=\log\det(X)$ 产生的布雷格曼散度进行计算:

$$\begin{aligned}&KL(p(\boldsymbol{x};M_0)\|p(\boldsymbol{x};M))=\frac{1}{2}D_{ld}(M_0^{-1},M^{-1})\\ &D_{ld}(M,M_0)=\operatorname{tr}(MM_0^{-1})-\log\det(MM_0^{-1})-d\end{aligned} \tag{5}$$

其中, $D_{ld}(\cdot)$ 表示布雷格曼散度. $\operatorname{tr}(\cdot)$ 表示矩阵 A 的迹, $\log\det(\cdot)$ 表示矩阵 A 的行列式的对数.

因此, 式 (4) 可转化为

$$\begin{aligned}&\min_{M\succeq 0}\ D_{ld}(M,M_0)\\ &\text{s.t.}\ \operatorname{tr}(M(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}})\le u,\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\in S\\ &\qquad \operatorname{tr}(M(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}})\ge l,\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\in D\end{aligned} \tag{6}$$

为了便于在更广的可行域内求解, ITML 算法引入松弛变量 $\boldsymbol{\xi}$ 并初始化为 $\boldsymbol{\xi}_0$ (其中每个元素的取值方式为等值约束样本对为 u, 非等值约束样本对为 l). 进一步将式 (6) 改写为

$$\begin{aligned}&\min_{M\succeq 0,\boldsymbol{\xi}}\ (D_{ld}(M,M_0)+\gamma D_{ld}(\operatorname{diag}\{\boldsymbol{\xi}\},\operatorname{diag}\{\boldsymbol{\xi}_0\}))\\ &\text{s.t.}\ \operatorname{tr}(M(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}})\le \xi_{i,j},(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S\\ &\qquad \operatorname{tr}(M(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}})\ge \xi_{i,j},(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D\end{aligned} \tag{7}$$

其中, γ 是均衡参数.

式 (7) 可以通过布雷格曼方法进行求解. 该算法采用迭代的方式计算布雷格曼投影, 每次计算在当前解的基础上通过成对约束进行下一步预测, 即:

$$M_{t+1}=M_t+\beta M_t(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}M_t \tag{8}$$

其中, M_t 是第 t 次迭代计算得到的度量矩阵, β 是映射参数, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是约束对.

Davis 等[18] 将 ITML 算法用于 k 最近邻分类和半监督聚类. 实验结果表明了 ITML 算法的优异性能.

但 ITML 算法存在以下问题[39]: 当训练集中的样本为高维数据时, 利用 ITML 算法进行距离度量学习并不可行. 这是因为: 1) ITML 算法的复杂度与数据维度的平方成正比, 从而导致该算法处理高维数据时复杂度过高; 2) ITML 算法学习得到的是一个满秩度量矩阵, 所需学习的参数和数据维度的平方成正比, 因此训练样本集过小时可能产生过拟合问题; 3) 在实际的高维数据集中往往存在某些维度被噪声污染或包含冗余信息的情况, 从而导致该算法无法学习到有效的度量矩阵. 因此, 当训练样本是高维数据时, 通过 ITML 算法学习得到的度量矩阵不能有效地抑制噪声, 还存在求解效率低、容易受到训练数据不足的影响等缺点.

针对高维数据集, 度量矩阵的满秩条件可以被简化. Davis 等[39] 结合稀疏矩阵提出了一种新的基于信息论的算法, 称之为高维低秩距离度量学习 (High-dimensionality low-rank, HDLR) 算法. HDLR 算法的主要思想是通过增加低秩约束来学习度量矩阵, 该算法将式 (6) 转化为

$$\begin{aligned}&\min_{M\succeq 0}\ D_{ld}(M, RR^{\mathrm{T}})\\ &\text{s.t. } d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\le u, \quad (\boldsymbol{x}_i,\boldsymbol{x}_j)\in S\\ &\quad d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)\ge l, \quad (\boldsymbol{x}_i,\boldsymbol{x}_j)\in D\\ &\quad \mathrm{rank}(M)\le k\end{aligned} \tag{9}$$

其中, $M_0 = RR^{\mathrm{T}}$. $\mathrm{rank}(M)$ 表示度量矩阵 M 的秩, k 表示先验矩阵 M_0 的秩.

根据文献 [40] 可知, 求解目标函数过程可以保证 M 的秩不超过基准矩阵 M_0 的秩. 利用上述性质, 只需要保证基准矩阵 M_0 低秩, 就可以使用 ITML 算法来求解式 (9).

Qi 等[41] 结合稀疏技术和信息论提出了另一种稀疏化度量学习算法, 称为稀疏距离度量学习 (Sparse distance metric learning, SDML) 算法. 该算法的主要思想是在 ITML 算法基础上, 通过最小化度量矩阵 M 中非对角元素的 l_1 范数项来实现稀疏. 具体为下列最优化问题:

$$\begin{aligned}&\min_{M\succeq 0}\ (\mathrm{tr}(M_0^{-1}M) - \log\det(M) + \gamma\|M\|_{1,\mathrm{off}} +\\ &\quad \eta l(S,D))\end{aligned} \tag{10}$$

式 (10) 共包含 4 项. 前两项表示度量矩阵 M 与基准矩阵 M_0 的相对熵, 使得 M 尽可能接近基准矩阵 M_0; $\|M\|_{1,\mathrm{off}}$ 是度量矩阵 M 中非对角元素的 l_1 范数项. 通过最小化 l_1 范数项使得 M 尽可能稀疏; 最后一项是损失函数项, 定义 $l(S,D) = (1/2)\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j\in S\cup D)} d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)K_{ij}$, 其中 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S$ 时, $K_{ij}=1$; 否则, $K_{ij}=-1$. γ 和 η 为加权参数. 最小化该损失函数项可以最小化同类样本间距离, 同时最大化不同类样本间距离.

与 ITML 算法相比, SDML 算法一方面通过在目标函数中增加稀疏限制来减少对训练样本数的要求; 另一方面, 该算法提出的优化问题可以通过块坐标下降算法进行求解, 从而大大加快了算法的执行速度.

Cui 等[42] 基于信息论的度量学习算法, 并结合分块思想, 提出了多模块距离度量学习算法. 其主要思想是对数据样本 (如人脸) 进行分块特征提取, 并利用多模块距离度量学习算法有效地整合不同的分块特征, 即针对不同分块学习不同的度量矩阵, 然后使用多个分块的平均度量距离作为样本间的度量距离, 从而有效提高了识别的精度.

1.3 基于概率论的距离度量学习算法

基于概率论的距离度量学习是利用概率论的方法学习度量矩阵. 其基本思想是: 利用训练样本的成对约束信息所建立的概率密度函数来表征目标模型并求解模型参数.

Guillaumin 等[43] 基于概率论, 提出了基于逻辑判别距离度量学习 (Logistic discriminant metric learning, LDML) 算法. LDML 算法基于逻辑回归的思想, 使用 S 型函数来表示样本对是否属于等值约束的概率. 样本对属于等值约束的概率定义如下:

$$p_n = p(t_n = 1|\boldsymbol{x}_i,\boldsymbol{x}_j; M, b) = \sigma(b - d_M(\boldsymbol{x}_i,\boldsymbol{x}_j)) \tag{11}$$

其中, $t_n = 1$ 表示样本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S$, $\sigma(z) = (1+\exp(-z))^{-1}$, b 表示阈值. $(1-p_n)$ 表示样本对属于非等值约束的概率. 样本对之间的距离越小, 其属于等值约束的概率越大; 反之, 样本对之间的距离越大, 则属于非等值约束的概率越大.

基于上述概率模型, LDML 算法采用最大似然估计的方法, 建立对数似然函数并用作目标函数, 通过最大化目标函数来求解模型参数. 其目标函数为

$$L = \sum_n t_n \ln p_n + (1-t_n)\ln(1-p_n) \tag{12}$$

上式中的目标函数是平滑的凹函数, 等号右边第 1 项表示属于等值约束的样本对的似然函数, 最大化第 1 项使得属于等值约束的样本对距离更近; 第 2 项表示属于非等值约束的样本对的似然函数, 最大化第 2 项使得属于非等值约束的样本对距离更远. 上式可以使用梯度下降法进行参数求解.

Yang 等[16] 提出了基于贝叶斯框架的主动距离度量学习算法. 该算法的主要思想是在贝叶斯框架下进行距离度量学习, 其基于威沙特先验分布的假设, 对后验概率进行估计并利用训练样本集的前 K 个特征向量的线性组合组成距离度量 M.

Guillaumin 等[43] 和 Yang 等[16] 的距离度量学习算法是通过定义样本对属于等值约束或非等值约束的先验概率建立模型并采用迭代方法求解最优的度量矩阵. 但是针对大规模训练数据上述算法仍然存在复杂度过高的问题.

Kostinger 等[36] 基于高斯分布的假设, 提出了保持简单直接的度量学习 (Keep it simple and straight, KISS) 算法. 该算法通过似然比检验的方法将距离度量问题转化为下式:

$$\delta(\boldsymbol{x}_{ij}) = \log\left(\frac{p(\boldsymbol{x}_{ij}|H_0)}{p(\boldsymbol{x}_{ij}|H_1)}\right) \tag{13}$$

其中, $\boldsymbol{x}_{ij}=\boldsymbol{x}_i-\boldsymbol{x}_j$. H_0 和 H_1 分别定义为样本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D$ 和 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S$ 的假设. 当 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D$ 时, $\delta(\boldsymbol{x}_{ij})$ 取得较大值, 否则 $\delta(\boldsymbol{x}_{ij})$ 取得较小值. 因此, $\delta(\boldsymbol{x}_{ij})$ 可以用来表征样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的距离. 使用均值为 0 的高斯分布对式 (13) 中的概率密度函数建模, 可以进一步化简为

$$\delta(\boldsymbol{x}_{ij})=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\widehat{M}(\boldsymbol{x}_i-\boldsymbol{x}_j) \tag{14}$$

其中, $\widehat{M}=\sum^{-1}_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}-\sum^{-1}_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}$. $\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}$, 表示所有属于等值约束的样本对外积和; $\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}$, 表示所有属于非等值约束的样本对外积和. 对比式 (1) 中的距离公式可以看出, 式 (14) 的中间项正好是对度量矩阵的估计 (即 S 中所有样本对的外积和减去 D 中所有样本对的外积和). 最后将 $\widehat{M}$ 映射到半正定锥面得到度量矩阵 M.

KISS 算法不需要通过复杂的迭代算法计算度量矩阵, 故计算效率更快. 在人脸识别、身份识别等应用的实验结果表明, 对比 ITML、LDML 等算法, KISS 算法在识别的准确率和算法效率上, 都具有更好的效果.

1.4 基于余弦相似度的距离度量学习算法

与传统的基于马氏距离度量的学习算法不同, 基于余弦相似度的距离度量学习算法是将样本向量之间的余弦距离作为样本间的相似度来构建目标函数. 其通过最优化目标函数求解度量矩阵. 假设 L 是变换矩阵, 则余弦相似度的计算公式为

$$CS(\boldsymbol{x}_i,\boldsymbol{x}_j,L)=\frac{(L\boldsymbol{x}_i)^{\mathrm{T}}(L\boldsymbol{x}_j)}{\|L\boldsymbol{x}_i\|\|L\boldsymbol{x}_j\|}=\frac{\boldsymbol{x}_i{}^{\mathrm{T}}L^{\mathrm{T}}L\boldsymbol{x}_j}{\sqrt{\boldsymbol{x}_i{}^{\mathrm{T}}L^{\mathrm{T}}L\boldsymbol{x}_i}\sqrt{\boldsymbol{x}_j{}^{\mathrm{T}}L^{\mathrm{T}}L\boldsymbol{x}_j}} \tag{15}$$

Nguyen 等[44] 提出了基于余弦相似度的距离度量学习 (Cosine similarity metric learning, CSML) 算法, 其主要思想是学习度量矩阵 M, 使得在变换后子空间中利用余弦相似度进行更有效的距离度量. CSML 给出的目标函数如下:

$$f(M)=-\alpha\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}CS(\boldsymbol{x}_i,\boldsymbol{x}_j,M)+\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}CS(\boldsymbol{x}_i,\boldsymbol{x}_j,M)-\beta\|M-M_0\|^2 \tag{16}$$

其中, $CS(\boldsymbol{x}_i,\boldsymbol{x}_j,M)$ 表示样本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 之间的余弦相似度, $M=L^{\mathrm{T}}L$. α 和 β 是平衡因子. $\alpha\geq 0,\beta\geq 0$, α 用来平衡 S 与 D 中元素个数差异较大时对目标函数产生的影响. 通常 $\alpha=|S|/|D|$, 其中 $|S|$ 和 $|D|$ 分别表示等值约束和非等值约束中元素个数. β 用来权衡损失函数项与正则项对目标函数产生的影响. 通常按照层次匹配策略对 β 进行取值.

式 (16) 等号右边第 1 项表示属于 D 中的所有样本对基于余弦相似度进行距离度量得到的相似度和; 第 2 项表示属于 S 中的所有样本对基于余弦相似度进行距离度量得到的相似度和, 即前 2 项构成了损失函数项; 第 3 项是正则项, 通过最小化 M 与先验矩阵 M_0 (其中 M_0 可以是预定义的任意矩阵, 可以定义 M_0 为一稀疏矩阵表示稀疏性先验) 之间的距离来提高目标函数的泛化能力. 该算法使用共轭梯度算法最大化目标函数得到 M, 本文将 CSML 算法应用于人脸验证数据集 LFW[45], 使用局部二值模式 (Local binary pattern, LBP)[46] 特征时, 获得 85.57 % 的识别率.

Cao 等[47] 综合余弦相似度和马氏距离的度量学习算法, 提出了子空间相似度度量学习 (Subspace similary metric learning, Sub-SML) 算法. 该算法的基本思想是使用余弦相似度与马氏距离之差作为新的度量函数, 从而找到一个更具判别性的度量函数. 该算法应用于 LFW 数据集, 使用多种联合特征, 获得了 89.73 % 的识别率.

1.5 小结

本节对基于成对约束的度量学习算法进行讨论. 一些具有代表性的成对约束变量学习算法总结如表 1 所示.

基于样本对距离和的距离度量学习算法简单有效, 能快速找到满足条件的度量矩阵, 但在全局范围构造出的相似对和非相似对集合过大, 故不能用于大规模数据集的度量学习. 基于信息论的距离度量学习算法通过假定数据分布, 在满足约束条件的限制下, 利用信息论的相关理论使得度量矩阵对应的概率分布与基准矩阵对应的概率分布尽可能接近来进行度量矩阵的学习. 由于这类算法在实际应用中是通过迭代的方式进行求解, 因此算法复杂度与训练样本的个数及数据样本的维度都相关. 基于概率论的算法通常假定数据分布是已知的, 并对后验概率进行建模, 从统计的角度表示数据的分布情况, 但也存在学习和计算过程较复杂的缺点. 基于余弦相似度的距离度量学习算法通过余弦距离来表征样本间相似度. 与马氏距离相比, 余弦距离度量往往具有更好的有效性, 但余弦距离度量的目标函数为非凸函数, 得到的解为局部最优.

表 1 典型的基于成对约束的距离度量学习算法的主要思想及目标函数模型

Table 1 The main ideas objective function models of typical metric learning algorithms based on pairwise constraints

算法	年份	主要思想	目标函数模型
基于样本对距离和的距离度量学习算法[1]	2002	在最小化相似对间距离平方和的同时, 约束非相似对间的距离和	样本对距离和
ITML[18]	2007	基于信息论将度量学习转化为最小化两个多元高斯的相对熵	微分相对熵
HDLR[39]	2008	利用低秩技术并使用信息论算法来学习稀疏度量矩阵	微分相对熵
KISS[36]	2012	通过似然比相关理论进行距离度量学习	似然比
LDML[43]	2009	基于后验概率的度量学习最大化对数似然函数	似然函数
CSML[44]	2007	通过余弦距离来表征样本间相似度进而学习度量矩阵	余弦相似度
Sub-SML[47]	2013	利用余弦相似度与马氏距离差作为新的度量函数	余弦相似度、马氏距离

2 基于非成对约束的距离度量学习算法

在距离度量学习中, 除成对约束外, 其他先验信息往往也能在训练过程中得到很好的学习效果. 基于非成对约束的距离度量学习算法大致分为两类: 一类将类别信息 (即训练样本 $\boldsymbol{x}_i$ 和所属类别 y_i) 作为先验知识, 通常被看作基于强监督信息的距离度量学习; 另一类将三元组信息 (即样本间相对关系) 作为先验知识, 其中 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$、$\boldsymbol{x}_k$ 都是数据样本. 三元组 $\{\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k\}$ 表示样本对 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的距离不大于样本对 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_k$ 间的距离, 通常被看作基于弱监督信息的距离度量学习.

按照距离度量学习过程中使用的目标函数模型不同, 分为基于信息论的距离度量学习算法、基于最大化分类边际的距离度量学习算法、基于提升学习框架的距离度量学习算法和基于稀疏正则的距离度量学习算法.

2.1 基于信息论的距离度量学习算法

基于信息论的非成对约束距离度量学习算法主要是利用训练样本的类别信息或三元组信息等先验知识, 并通过信息论算法进行距离度量学习的一类算法.

Wang 等[48] 提出了基于信息几何的度量学习 (Information geometry metric learning, IGML) 算法. 主要思想是利用类标签和数据样本集构建两个核矩阵, 其中根据类标签构建的核矩阵是理想矩阵, 根据数据样本集构建的核矩阵是实际矩阵. 该算法使用相对熵度量这两个核矩阵之间的距离, 并将其作为目标函数, 通过最小化目标函数求解度量矩阵. 具体方法如下:

设训练样本集 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$ 且 $\boldsymbol{x}_i \in \mathbf{R}^d$. 类别标签矩阵定义为 $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]$, $\boldsymbol{y}_i = (y_i^1, y_i^2, \cdots, y_i^C) \in \{0,1\}^C$ 是 C 个元素的二元向量, C 为类别数目. 假设每个样本只属于一个类别即 $\boldsymbol{y}_i^{\mathrm{T}}\mathbf{1} = 1$, 其中 $\mathbf{1}$ 是一个全为 1 的向量.

1) 使用类别信息构建第一个核矩阵:

$$K_D = Y^{\mathrm{T}}Y \tag{17}$$

当 $C < n$ 时, K_D 是奇异矩阵. 因此可以通过单位矩阵来平滑 K_D 得到新的核矩阵:

$$\widetilde{K}_D = Y^{\mathrm{T}}Y + \lambda I_n \tag{18}$$

其中, λ 表示平滑参数, I_n 表示 n 维单位阵.

2) 使用样本矩阵和度量矩阵构建第二个核矩阵:

$$K_X = X^{\mathrm{T}}MX \tag{19}$$

其中, M 是要求解的度量矩阵.

3) 利用相对熵将距离度量学习问题转化为下面的最优化问题:

$$\begin{aligned}&\min_{M\succeq 0} d(K_X \| \widetilde{K}_D) = \\ &\quad \min_{M\succeq 0} \left(\mathrm{tr}(\widetilde{K}_D^{-1}X^{\mathrm{T}}MX) - \log|M|\right)\end{aligned} \tag{20}$$

求解式 (20) 可得:

$$M = (X\widetilde{K}_D^{-1}X^{\mathrm{T}})^{-1} \tag{21}$$

IGML 算法具有封闭解. 与其他基于信息论的距离度量学习算法相比, 计算速度更快. 本文将 IGML 算法应用于数据分类和人脸识别, 在 ORL 人脸数据集[49] 上, 取得了 90 % 以上的识别率.

Wang 等[7] 利用先验样本的类别信息构建理想概率分布, 对需要学习的度量矩阵 M 构建实际概率分布. 使用信息熵计算理想概率分布和实际概率分布之间的距离, 提出了一种新的基于信息论的度量矩阵算法. 本文将其应用于单目标和多目标的跟踪上, 实验结果验证了该算法的有效性.

2.2 基于最大化分类边际的距离度量学习算法

基于最大化分类边际的非成对约束距离度量学习算法是在训练样本中使用类别或三元组等非成对约束信息作为先验知识, 利用最大化不同类别边际距离的思想, 来学习度量矩阵的一类距离度量学习算法.

Weinberger 等[4] 利用类别信息作为先验知识, 提出了基于最大边际的最近邻算法. 该算法主要思想是通过最大化不同类别边际距离来改进 k 最近邻的算法性能. 具体算法如下:

$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ 表示 n 个样本组成的训练集, 其中输入 $\boldsymbol{x}_i$ 对应的类别标签为 y_i. 使用标量 $y_{ij} \in \{0,1\}$ 表示类别标签 y_i 和 y_j 是否匹配, 且通常需要学习一个线性转换 $L: \mathbf{R}^d \to \mathbf{R}^d$ 进行如下距离度量:

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} M (\boldsymbol{x}_i - \boldsymbol{x}_j) = \| L(\boldsymbol{x}_i - \boldsymbol{x}_j) \|^2 \tag{22}$$

其中, $M = L^{\mathrm{T}} L$.

对任一输入样本 $\boldsymbol{x}_i$, 其目标邻居定义为符合下列条件的输入样本: 1) 与 $\boldsymbol{x}_i$ 存在着相同类别标签 y_i; 2) 通过式 (22) 计算与 $\boldsymbol{x}_i$ 有着最小距离的输入样本. 对于样本 $\boldsymbol{x}_i$, 可以指定 k 个目标邻居. 在缺乏先验知识时, 通常使用欧氏距离计算 k 最近邻居. 采用标量 $\eta_{ij} \in \{0,1\}$, 当 $\eta_{ij} = 1$ 时, 表示 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的目标邻居; $y_{ij} \in \{0,1\}$, 当 $y_{ij} = 1$ 时, 表示 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 属于同一类. 二元组 y_{ij} 和 η_{ij} 均是固定的 (即在学习过程中不会改变).

根据最大化分类边际思想得到目标函数:

$$\varepsilon(L) = \sum_{ij} \eta_{ij} \| L(\boldsymbol{x}_i - \boldsymbol{x}_j) \|^2 + c \sum_{ijl} \eta_{ij} (1 - y_{il}) \times [1 + \| L(\boldsymbol{x}_i - \boldsymbol{x}_j) \|^2 - \| L(\boldsymbol{x}_i - \boldsymbol{x}_l) \|^2]_+ \tag{23}$$

其中, $[z]_+ = \max(z, 0)$ 且 c 是某个正常数.

式 (23) 中等号右边第 1 项调整所有输入样本与目标邻居间距离, 通过最小化该项使得输入样本与目标邻居间距离尽可能小; 第 2 项调整不同类别间的边际距离, 通过最小化该项使得边际距离最大化.

为了便于在更大的可行域内求解, 引入松弛变量 ξ_{ijl}, 式 (23) 转化为求解下列半正定规划问题:

$$\begin{aligned} \min_{M, \boldsymbol{\xi}} \ & \Big(\sum_{ij} \eta_{ij} d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) + c \sum_{ijl} \eta (1 - y_{il}) \xi_{ijl} \Big) \\ \text{s.t.} \ & \forall (i,j,l), d_M(\boldsymbol{x}_i, \boldsymbol{x}_l) - d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1 - \xi_{ijl} \\ & \forall (i,j,l), \xi_{ijl} \geq 0 \\ & M \succeq 0 \end{aligned} \tag{24}$$

通过标准求解半正定规划算法求解上式得到度量矩阵 M.

Verma 等[50] 基于 LMNN 算法, 提出了一种新的距离度量框架, 将原来针对单类进行分类改进为针对多类进行分类, 即一个样本可以同时属于多个类别, 并将其应用于图像自动标注. Zhang 等[21] 在最大化分类边际目标函数的基础上, 结合多任务学习和迁移学习得到度量矩阵.

2.3 基于提升学习框架的距离度量学习算法

根据半正定矩阵可以分解为多个秩为 1 且迹为 1 的矩阵线性叠加的性质[51], 基于提升学习的距离度量学习算法主要思想是在提升学习框架下, 每次迭代学习一个秩为 1 且迹为 1 的矩阵, 最后将每次学习得到的矩阵线性叠加组成度量矩阵, 即 $M = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$ 且 $U_i = \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$. 取 U_i 所在子空间 $\Omega = \{U | U \succeq 0, \mathrm{tr}(U) = 1, \mathrm{rank}(U) = 1\}$. 样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的度量距离为 $H(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_t \alpha_t h_t(\boldsymbol{x}_i, \boldsymbol{x}_j)$, 其中 h_t 是每次迭代学习的弱假设, α_t 为线性系数.

Bi 等[52] 利用三元组信息作为先验知识, 提出了提升学习算法 (BoostMetric). 该算法的主要思想是在每次迭代过程中最大化三元组之间的相对距离来学习弱假设, 并利用训练误差来学习迭代系数. 具体算法如下:

1) 弱假设 h_t 的学习. 每次迭代过程中弱假设的学习转化为求解下式:

$$\begin{aligned} \max_{\boldsymbol{u}_t} \ & \left| \sum_{(i,j,k) \in \mathcal{T}} D_t(i,j,k) (h_t(\boldsymbol{x}_i, \boldsymbol{x}_k) - h_t(\boldsymbol{x}_i, \boldsymbol{x}_j)) \right| \\ \text{s.t.} \ & h_t(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} U_t (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ & U_t = \boldsymbol{u}_t \boldsymbol{u}_t^{\mathrm{T}} \\ & \| \boldsymbol{u}_t \| = 1 \end{aligned} \tag{25}$$

其中, $\mathcal{T}$ 为三元组集合. $D_t(i,j,k)$ 为样本权值, D_t 赋初值为 $1/m$, m 为三元组个数. 每次迭代更新公式为

$$D_{t+1}(i,j,k) = \frac{[D_t(i,j,k) \exp(\alpha_t (h_t(\boldsymbol{x}_i, \boldsymbol{x}_j) - h_t(\boldsymbol{x}_i, \boldsymbol{x}_k)))]}{Z_t}$$

其中, Z_t 为归一化因子.

式 (25) 存在着封闭解, 最优解 $\boldsymbol{u}_t$ 为矩阵 $\sum_{(i,j,k) \in \mathcal{T}} D_t(i,j,k) ((\boldsymbol{x}_i - \boldsymbol{x}_k)(\boldsymbol{x}_i - \boldsymbol{x}_k)^{\mathrm{T}} - (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}})$ 中最大绝对值的特征值所对应的特征向量.

2) 线性系数 α_t 的学习. 利用弱假设训练误差不大于 1/2. 类似自适应提升学习算法, 取 $\alpha_t = \ln(\epsilon_-/\epsilon_+)/2$ 或 $\alpha_t = \ln((1+r)/(1-r))/2$. 其中 ϵ_- 和 ϵ_+ 分别表示三元组中度量正确的样本比例和三元组中度量错误的样本比例. $r = \sum_{(i,j,k)\in\mathcal{T}} D_t(i,j,k)(h_t(\boldsymbol{x}_i,\boldsymbol{x}_k) - h_t(\boldsymbol{x}_i,\boldsymbol{x}_j))$ 且 h_t 被正规化到 $[0,1]$.

在式 (1) 和式 (2) 的迭代学习过程中, 度量 H 可以表示为 $H(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\sum_t \alpha_t U_t)(\boldsymbol{x}-\boldsymbol{y})$.

Shen 等[51] 利用三元组信息作为先验信息, 提出了度量提升学习算法 (MetricBoost). 该算法的主要思想是在每次迭代过程中, 利用三元组之间相对距离的指数损失作为损失函数并引入正则项来学习弱假设. 当数据维度增加时, 使用距离提升学习算法大大降低了传统算法的训练时间.

2.4 基于稀疏正则的距离度量学习算法

在目标识别、图像检索等计算机视觉应用中, 需要处理的数据通常都是高维的. 一方面, 针对数据的高维度以及包含噪声等不利因素, 度量学习算法若考虑学习一个计算复杂度较高且无法有效抑制噪声的全矩阵, 则通常需要增加数据的稀疏性或低秩性条件来提升学习的鲁棒性; 另一方面, 目前大部分存在的稀疏度量学习算法[41, 53] 并没有直接将稀疏作为目标函数. 为了更容易求解其往往增加了约束条件 (如增加对角约束[53]), 但是该做法并不能保证获得稀疏问题的最优解.

针对上述问题, Huang 等[30] 基于相对距离约束, 提出了一种普适的稀疏距离度量学习 (Generalized sparse metric learning, GSML) 框架. 该框架主要思想是结合稀疏条件来学习一个度量矩阵, 并通过约束样本间的相对距离, 保证映射到转换空间后样本之间的距离关系保持不变. 具体算法如下:

该算法度量学习的目标是学习一个满足特定三元组约束 $\mathcal{T} = (i,j,k)|f(\boldsymbol{x}_i,\boldsymbol{x}_j) \le f(\boldsymbol{x}_i,\boldsymbol{x}_k)$ (即使得样本 $\boldsymbol{x}_j$ 比样本 $\boldsymbol{x}_k$ 更接近于样本 $\boldsymbol{x}_i$) 的半正定度量矩阵, 其中 $f(\cdot)$ 表示距离度量函数. 因此结合稀疏正则项, 距离度量学习算法可表示为

$$\begin{aligned}
&\min_{M,\boldsymbol{\xi}} \ \Big(\sum_t \xi_t + \gamma \left\|M\right\|_{(2,1)}\Big)\\
&\text{s.t. } \forall (i,j,k)\in\mathcal{T}, \|\widehat{\boldsymbol{x}_i} - \widehat{\boldsymbol{x}_j}\|_2^2 \le \|\widehat{\boldsymbol{x}_i} - \widehat{\boldsymbol{x}_k}\|_2^2 + \xi_t\\
&\quad \forall t, \xi_t \ge 0\\
&\quad M \succeq 0
\end{aligned} \tag{26}$$

其中, M 是度量矩阵; $\|M\|_{(2,1)}$ 是 M 的混合范数, $\|M\|_{(2,1)} = \sum_l \|M_l\| = \sum_l (\sum_k M_{lk}^2)^{\frac{1}{2}}$; $\|M_l\|$ 表示 M 的第 l 行行向量; 令 $M = A^{\mathrm{T}}A$. $\widehat{\boldsymbol{x}_i} = A^{\mathrm{T}}\boldsymbol{x}_i$, 表示原数据 $\boldsymbol{x}_i$ 在新的转换空间的映射; γ 表示加权参数.

上式目标函数共两项. 等号右边第 1 项联合约束式 (1) 保证了在新的转换空间中原数据依然能够保持距离关系, 即数据样本之间的相对距离关系保持不变, 相似对之间的距离比非相似对之间的距离更近. 第 2 项是稀疏正则项. 对于高维数据, 若转换矩阵 A 第 l 行行向量 $\|A_l\| = 0$, 则数据在转换空间下第 l 维为 0, (即 $\widehat{\boldsymbol{x}_i}^{l} = A_l\boldsymbol{x}_i = 0$), 使用 l_1 正则化稀疏正则项表示为 $\sum_l \|A_l\|$ (即 $\|A\|_{2,1}$). 由 $M = A^{\mathrm{T}}A$ 可知 $\|A\| \equiv 0 \Leftrightarrow \|M\| \equiv 0$, 故进一步使用 $\|M\|_{2,1}$ 表示稀疏正则项, 保证学习到的度量矩阵满足一定的稀疏性质.

通过引入预定义矩阵 L, 进一步将式 (26) 转化为

$$\begin{aligned}
&\min_{M,\boldsymbol{\xi}} \ \Big(\sum_t \xi_t + \gamma \mathrm{tr}(LM)\Big)\\
&\text{s.t. } \forall (i,j,k)\in\mathcal{T}, (\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}M(\boldsymbol{x}_i-\boldsymbol{x}_j) \le\\
&\quad (\boldsymbol{x}_i-\boldsymbol{x}_k)^{\mathrm{T}}M(\boldsymbol{x}_i-\boldsymbol{x}_k) + \xi_t\\
&\quad \forall t, \xi_t \ge 0\\
&\quad M \succeq 0
\end{aligned} \tag{27}$$

根据上述框架中 L 取值不同, 该框架可以包含以下若干算法: 1) 如果取预定义矩阵为单位阵 (即 $L = I$), 则上述框架可以进一步转化为稀疏距离度量学习算法[53]; 2) 如果取预定义矩阵为相似样本对的协方差矩阵 (即 $L = 1/|S|\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}$, 其中 $|S|$ 表示相似对的数目), 则上述框架可以转化为基于最大边际的最近邻算法[4]; 3) 如果取预定义矩阵为度量矩阵 (即 $L = M$), 则上述框架可以表示支持向量机; 4) 对于部分基于成对约束的非稀疏度量学习算法[1, 34], 通过对松弛变量改写再加上稀疏正则项, 则可以将原来的非稀疏度量学习算法转化为 GSML 框架下的稀疏度量矩阵学习算法.

Huang 等[54] 提出了一种新颖的基于稀疏距离度量核回归算法, 并使用度量矩阵的混合范数作为正则项来建立稀疏回归模型, 获得了优异的效果. Bah 等[55] 通过相对距离约束使得数据样本间的距离关系映射到转换空间后保持不变. 该算法结合了 l_1 正则项和 l_0 范数约束来学习一个稀疏度量矩阵.

2.5 小结

本节对基于非成对约束的距离度量学习算法进行了讨论. 一些具有代表性的非成对约束距离度量学习算法总结如表 2 所示. 非成对约束通常利用类别标签或三元组信息作为先验知识学习度量矩阵. 基于类别标签的学习通常可以获得较好的实验效果, 但在很多情况下, 需要对数据进行人工标定, 标定过

表 2 典型的基于非成对约束的距离度量学习算法的主要思想及目标函数模型

Table 2 The main ideas and objective function model of typical metric learning algorithms based on non-pairwise constraints

算法	年份	主要思想	目标函数模型
IGML[48]	2009	利用样本类标签信息构建两个核矩阵, 利用相对熵度量核矩阵之间的距离	微分相对熵
LMNN[4]	2006	基于最大化类间距离最小化类内距离思想改进 k 邻居算法	最大化分类边际
TML[21]	2010	结合多任务学习和迁移学习的度量矩阵	最大化分类边际
MetricBoost[48]	2011	基于提升学习框架, 最大化三元组间相对距离学习一个弱假设度量矩阵	提升学习框架
BoostMetric[49]	2012	基于提升学习框架, 利用三元组间相对距离的指数损失学习一个弱假设度量矩阵	提升学习框架
GSML[30]	2011	结合三元组相对距离约束及稀疏正则学习稀疏度量学习矩阵	稀疏框架

程相对耗时. 而三元组信息相对来说容易获取, 在实际学习过程中也取得良好的实验效果, 因此, 对三元组信息的利用方式进行进一步探索, 具有重要意义. 实际上三元组信息可以看成是有类别标签的一个特例.

基于信息论的算法利用生成式模型, 通常是假定数据分布已知, 在实际应用中通过迭代方式进行求解, 但存在学习和计算过程比较复杂的缺点. 基于最大间隔是判别式模型, 其基本思想是寻找不同类别之间的最优分类面来区分各类数据. 在实际应用中能清晰地分辨出多类或某一类与其他类别的差异, 更适用于多类别的识别问题. 基于提升学习的距离度量学习算法则是基于提升学习框架将多次迭代学习的弱假设模型进行线性叠加组成最终的距离度量. 该类算法将复杂的半正定矩阵求解问题转化为多次求解秩为 1 的简单矩阵求解问题, 可以大大降低算法的复杂度. 普适的稀疏距离度量学习框架不仅包含了目前流行的多种距离度量学习算法[4, 53], 且可以有效地扩展一些经典的非稀疏距离度量学习算法[1, 34] 至稀疏的版本. 该框架对于稀疏距离度量学习的进一步研究具有重要指导意义.

3 距离度量学习的应用及其作用

距离度量学习的应用主要包括聚类[1, 56−57]、分类[58−60] 以及计算机视觉[1, 42−44, 61−69] 等领域. 距离度量学习在这些领域有如下作用.

1) 距离度量学习在保持数据几何结构的同时, 利用降维的方式将原始数据集映射到低维子空间中. 在子空间中同类样本之间的距离更近, 不同类样本之间的距离更远, 这对于改进聚类算法的性能具有重要的作用. Xing 等[1] 通过构建样本的相似对和非相似对进行距离度量的学习, 利用学习到的度量矩阵计算样本间距离, 从而提高 k 均值聚类算法的性能. Xiang 等[3] 基于样本的散布矩阵提出了一种有效的距离度量学习算法, 并将其用于改进 k 均值聚类算法的性能.

2) 在各种数据分类过程中, 有效地度量样本之间的距离, 进而判断样本的属性信息往往是关键步骤之一. Weinberger 等[4] 提出了基于最大边际的最近邻算法, 利用学习到的距离度量提高 k 最近邻分类器的精度. Wang 等[58] 对每个数据点学习局部度量矩阵, 使用目标邻居的局部度量加权线性和来表示数据点的局部度量, 从而改进最近邻分类器性能.

3) 距离度量对于改进各种实际计算机视觉应用问题也具有显著效果. 这类问题的本质即计算两幅图像的相似度. 图像经过特征提取后获得的数据通常在一个高维特征空间中, 但是由于数据各个维度的重要性不同, 且高维数据往往存在着噪声等干扰因素, 使得通过欧氏距离、余弦距离等简单几何距离获得的相似度并不准确 (这些距离通常将数据的各个维度等同对待). 而根据距离度量学习算法得到的度量矩阵则利用数据各维度重要性不同, 对特征空间的数据进行加权, 从而获得更好的相似度度量. Guillaumin 等[43] 基于尺度不变特征变换特征 (Scale-invariant feature transform, SIFT) 的特性, 利用逻辑判别距离度量学习算法得到有效的距离度量, 在新的度量空间中通过人脸距离比对进行人脸验证. Jiang 等[70] 将邻近成分分析应用到目标跟踪中, 利用学习得到的距离度量在视频中的每一帧计算目标模板与候选目标之间的相似度. Chang 等[61] 将提出的度量学习算法用于基于内容的图像检索, 并使用学习到的距离度量计算测试图像与待查询图像之间的相似度, 从而选出与测试图像相似度最高的前若干幅图像作为检索结果. Tran 等[67] 将 LMNN 算法应用于行为识别, 通过度量学习算法改进行为分类过程. Lebanon[69] 利用度量学习计算文档之间的相似度, 其文档分类结果明显优于传统的基于词频–逆向文件频率的余弦相似度的分类结果. 表 3 列出了一些代表性的距离度量学习的应用场景及其作用.

表 3 距离度量学习的应用场景及作用

Table 3 Application scenarios and the corresponding functionalities of distance metric learning

应用场景	作用
聚类	改进 k 均值 等聚类算法的性能[1, 56−57]
分类	提高分类器 (如 k 最近邻分类器 等) 的精度[58−60, 71]
人脸验证	在新的度量空间中通过人脸距离比对进行人脸验证[42−44, 72−73]
图像检索与分类	利用学习到的度量矩阵计算测试图像与待查询图像之间的相似度[61−63]
目标跟踪	利用学习到的度量矩阵计算目标模板与候选目标之间的相似度[64−66]
行为识别	利用度量学习改进行为分类性能[67−68]
文本分类	利用度量学习计算文档之间的相似度[39, 69]

4 现有算法及性能评估

本节列举了典型距离度量学习算法在人脸识别和 k 最近邻分类器分类 2 个任务上的性能评估结果. 针对人脸识别任务, 我们给出了在常用的 LFW 人脸数据集上的结果比较; 针对 k 最近邻分类器分类任务, 给出了在常用的 UCI[74] 机器学习数据集上的性能评估.

4.1 实验数据集

LFW 人脸数据集是进行人脸验证的公共数据集, 其中包括来自 5 749 个人的 13 233 幅人脸图片, 且有 1 680 个人包含 2 张或者多张人脸图片. 每幅图片大小为 250 像素 ×250 像素. UCI 数据集是机器学习领域常用的标准测试数据集, UCI 数据集包含了数据的属性和类别. 通常文献中使用某种分类算法 (如 k 最近邻算法) 对数据进行分类, 将实验结果与数据结果进行对比计算分类正确 (或错误) 率. 我们在 UCI 数据集上选择了 Iris、Wine、Segmentation、Ionosphere、Optdigits、Usps 共 6 种常用的数据集.

4.2 实验结果及分析

表 4 和表 5 是一些典型的距离度量学习算法在 LFW 和 UCI 数据集上的结果, 其中在 LFW 上给出了不同算法的识别率 (平均准确率和方差), 在 UCI 上给出了不同算法的分类错误率, 其中 k 表示最近邻的数目.

实验比较的距离度量学习算法包括 LDML[43]、ITML[18]、基于特征值最优化框架的度量学习算法 (DML-eig)[38]、KISS[36]、CSML[44]、基于成对约束的多度量学习 (Pairwise-constrained multiple metric learning, PMML) 算法[42]、Sub-SML[47]、LMNN[4]、IGML[48]、基于信息几何的核度量学习算法[48]、在线的正则化距离度量学习算法[26]、BoostMetric[51]、基于特征值优化的最大边际的最近邻算法[38].

实验结果表明, 距离度量学习算法在人脸验证和分类中具有良好的实验性能. 对于 LFW 人脸验证集, 使用 PMML 算法和 Sub-SML 算法都达到了 89 % 以上的准确率. 对于 UCI 数据集, 基于提升学习思想的提升度量学习算法及 DML-eig 算法都取得了较低的分类错误率. 需要指出的是, UCI 数据集的实验结果也表明了距离度量学习算法的性能有一定的不可预见性, 如 IGML 算法在 Iris 数据集上的分类错误率低于 3 % , 而在 Ionosphere 数据集上的分类错误率却高达 16.6 %. LMNN 算法在 Optdigits 数据集上的错误率仅为 1.6 %, 而在 Iris 数据集上的错误率为 4.5 %, 高于其他算法. 该现象说明在实际应用中, 为取得最佳性能应该根据数据集的属性, 选择合适的距离度量学习算法.

表 4 各种算法在 LFW 数据集上的识别率 (%)

Table 4 The recognition rates obtained by different algorithms on LFW dataset (%)

算法	特征	准确率和方差
LDML[43]	SIFT	77.50±0.0050
LDML[43]	LBP	80.65±0.0047
ITML[18]	SIFT	78.12±0.0230
ITML[18]	LBP	79.98±0.0039
DML-eig[38]	SIFT	81.27±0.0230
DML-eig[38]	LBP	82.28±0.0041
KISS[36]	SIFT	83.08±0.0056
KISS[36]	LBP	83.37±0.0054
CSML[44]	LBP	85.57±0.0052
PMML[42]	SFRD	89.35±0.0050
Sub-SML[47]	SIFT	85.55±0.0061
Sub-SML[47]	LBP	89.73±0.0053

5 结束语

本文详细介绍了目前有监督的距离度量学习算法的研究进展, 对各种典型算法进行了介绍和讨论, 分析了各种算法的优缺点.

表 5 不同算法在 UCI 数据集上的分类错误率 (%)

Table 5 The classification error rates obtained by different algorithms on UCI datasets (%)

算法	Iris	Wine	Segmentation	Ionosphere	Optdigits	Usps
LMNN ($k=4$)[4]	4.5±2.1	4.1±1.8	14.7±1.9	15.0±1.9	1.6±0.3	—
ITML ($k=4$)[18]	4.3±2.7	7.7±3.0	16.6±5.0	11.1±2.6	2.1±0.3	—
IGML ($k=4$)[48]	2.7±1.7	5.0±1.6	12.9±3.4	16.6±1.8	3.2±0.3	—
基于信息几何的核度量学习算法 ($k=4$)[48]	3.9±2.8	6.1±1.9	12.4±3.5	14.2±1.6	1.4±0.2	—
在线正则化距离度量学习算法 ($k=3$)[26]	3.2±1.3	1.8±1.1	12.9±2.2	—	2.9±0.4	—
提升度量学习算法 ($k=3$)[51]	3.56±2.52	2.31±2.18	4.21±0.48	—	1.38±0.33	3.34
DML-eig ($k=3$)[38]	3.11±1.15	1.35±1.30	2.97±0.55	—	1.45±0.22	3.66
基于特征值优化的最大边际的最近邻算法 ($k=3$)[38]	4.00±2.30	2.88±1.87	3.61±0.83	—	1.43±0.42	3.13

不同的距离度量学习算法适用于不同的任务和应用场景, 很难找到一种普遍适用于所有问题的距离度量学习算法. 在实际应用中, 应该根据不同的应用场合来选择不同的算法. 目前距离度量学习研究已经取得了一定的成绩, 但仍有些内容值得去进一步探索和研究:

1) 多特征融合距离度量. 目前大部分度量学习算法主要针对某种特定特征进行度量学习[75−76], 或分别计算多种特征的度量距离后通过特定分类器进行简单的融合[38, 44]. 在实际应用中, 选择提取不同的特征可以表征样本数据不同的属性. 因此, 在以后的研究中, 可以考虑对于样本数据提取多种不同的特征, 在一个更为有效的多特征融合的框架中进行距离度量的学习.

2) 在线距离度量学习. 在线距离度量学习可以很好地解决训练样本集不断增加所带来的大规模数据学习问题, 是解决这类问题的一种很好的思路. 目前在线度量学习大部分主要应用在目标跟踪[64, 77]中. 但好的在线距离学习策略对于图像检索及目标识别等问题也应该同样有效. 因此, 在以后的研究中, 可以考虑在更多领域中扩展现有的距离度量学习算法的在线应用.

3) 多种距离度量技术的综合应用. 距离度量学习过程中产生的不同问题常常使用不同的技术来解决. 例如, 半监督技术[78−79] 可以用于解决训练样本有限的问题; 在线技术[80−83] 可以用于解决训练样本不断增加而带来的大规模学习问题; 稀疏技术[31, 53, 84] 和低秩技术[85] 可以用于解决特征冗余问题. 因此, 在以后的研究中, 可以考虑将多种技术综合起来应用到距离度量学习中, 针对不同的实际应用实现更有效的距离度量学习性能.

4) 鲁棒距离度量学习. 目前大部分的距离度量学习算法都假设获得的训练集是准确无误差的. 而实际获取的训练集往往是包含大量噪声以及多结构的数据, 甚至可能包括了一些标定错误的训练样本. 利用现有的距离度量学习算法学习得到的度量矩阵效果往往较差. 因此, 在今后的研究中, 可以考虑将鲁棒统计学相关技术[86−88] 引入距离度量学习算法中, 从包含离群数据中正确估计模型的参数, 克服离群数据和多结构的影响, 进而学习一个更加鲁棒有效的距离度量.

References

1 Xing E P, Ng A Y, Jordan M I, Russell S. Distance metric learning with application to clustering with side-information. In: Proceedings of the 2003 Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2003. 521−528

2 Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighbourhood components analysis. In: Proceedings of the 2004 Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2004. 513−520

3 Xiang S M, Nie F P, Zhang C S. Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 2008, **41**(12): 3600−3612

4 Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 2009, **10**: 207−244

5 Mensink T, Verbeek J, Perronnin F, Csurka G. Metric learning for large scale image classification: generalizing to new classes at near-zero cost. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: IEEE, 2012. 488−501

6 Feng Z, Jin R, Jain A. Large-scale image annotation by efficient and robust kernel metric learning. In: Proceedings of the 2013 International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 1609−1616

7 Wang X Y, Hua G, Han T X. Discriminative tracking by metric learning. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer, 2010. 200−214

8 Chen J H, Zhao Z, Ye J P, Liu H. Nonlinear adaptive distance metric learning for clustering. In: Proceedings of the 2007 International Conference on Knowledge Discovery and Data Mining. California, USA: ACM, 2007. 123−132

9 Ye J P, Zhao Z, Liu H. Adaptive distance metric learning for clustering. In: Proceeding of the 2007 Computer Society Conference on Computer Vision and Pattern Recognition. Minnesota, USA: IEEE, 2007. 1−7

10 Cinbis R G, Verbeek J, Schmid C. Unsupervised metric learning for face identification in TV video. In: Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 1559−1566

11 Wang B, Jiang J Y, Wang W, Zhou Z H, Tu Z W. Unsupervised metric fusion by cross diffusion. In: Proceedings of the 2012 Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2997−3004

12 Mignon A, Jurie F. CMML: a new metric learning approach for cross modal matching. In: Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, Korea: Springer, 2012. 14−27

13 Cao B, Ni X C, Sun J T, Wang G, Yang Q. Distance metric learning under covariate shift. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain: AAAI, 2011. 1204−1210

14 Guillaumin G, Verbeek J, Schmid C. Multiple instance metric learning from automatically labeled bags of faces. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer, 2010. 634−647

15 Baghshah M S, Shouraki S B. Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data. *Pattern Recognition*, 2010, **43**(8): 2282−2292

16 Yang L, Jin R, Sukthankar R. Bayesian active distance metric learning. In: Proceedings of the 23th Conference on Uncertainty in Artificial Intelligence. Vancouver, Canada: AUAI Press, 2007. 442−449

17 Cevikalp H. Distance metric learning by quadratic programming based on equivalence constraints. In: Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010. 3352−3355

18 Davis J V, Kulis B, Jain P, Sra S, Dhillon I S. Information-theoretic metric learning. In: Proceedings of the 24th International Conference. Oregon, USA: ACM, 2007. 209−216

19 Wang J, Do H, Woznica A, Kalousis A. Metric learning with multiple kernels. In: Proceedings of the 2001 Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2011. 1170−1178

20 Baghshah M S, Shouraki S B. Semi-supervised metric learning using pairwise constraints. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. California, USA: IJCAI, 2009. 1217−1222

21 Zhang Y, Yeung D Y. Transfer metric learning by learning task relationships. In: Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining. Washington, USA: ACM, 2010. 1199−1208

22 Li W, Zhao R, Wang X G. Human reidentification with transferred metric learning. In: Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, Korea: Springer, 2012. 31−44

23 Parameswaran S B, Weinberger K Q. Large margin multi-task metric learning. In: Proceedings of the 2010 Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2010. 1867−1875

24 Yang P P, Huang K Z, Liu C L. A multi-task framework for metric learning with common subspace. *Neural Computing and Applications*, 2013, **22**(7−8): 1337−1347

25 Yang P P, Huang K Z, Liu C. Geometry preserving multi-task metric learning. *Machine Learning*, 2013, **92**(1): 133−175

26 Jin R, Wang S J, Zhou Y. Regularized distance metric learning: theory and algorithm. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2009. 862−870

27 Hoi S C H, Liu W, Chang S F. Semi-supervised distance metric learning for collaborative image retrieval. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA: IEEE, 2008. 1−7

28 Shen C H, Kim J, Wang L. Scalable large-margin Mahalanobis distance metric learning. *IEEE Transactions on Neural Networks*, 2010, **21**(9): 1524−1530

29 Shen C H, Kim J, Wang L. A scalable dual approach to semidefinite metric learning. In: Proceedings of the 24th Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2011. 2601−2608

30 Huang K Z, Ying Y M, Campbell C. GSML: a unified framework for sparse metric learning. In: Proceedings of the 9th International Conference on Data Mining. Florida, USA: IEEE, 2009. 189−198

31 Huang K Z, Ying Y M, Campbell C. Generalized sparse metric learning with relative comparisons. *Knowledge and Information Systems*, 2011, **28**(1): 25−45

32 Liu W, Hoi S C H, Liu J Z. Output regularized metric learning with side information. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 358−371

33 Yang L, Jin R. Distance Metric Learning: A Comprehensive Survey, Technical Report, Michigan State University, USA. 2006, 1−51

34 Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning*, 2005, **6**: 937−965

35 Mignon A, Jurie F. PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proceedings of the 2012 International Conference on Computer Vision and Pattern Recognition. Providence RI: IEEE, 2012. 2666−2672

36 Kostinger M, Hirzer M, Wohlhart P, Roth P M, Bischof H. Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2012. 2288−2295

37 Boyd S P, Vandenberghe L. *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

38 Ying Y M, Li P. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 2013, **13**(1): 1−26

39 Davis J V, Dhillon I S. Structured metric learning for high dimensional problems. In: Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: ACM, 2008. 195−203

40 Kulis B, Sustik M A, Dhilon I S. Learning low-rank kernel matrices. In: Proceedings of the 23rd International Conference on Machine Learning. USA: ACM, 2006. 505−512

41 Qi G J, Tang J H, Zha Z J, Chua T S, Zhang H J. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In: Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009. 841−848

42 Cui Z, Li W, Xu D, Shan S G, Chen X L. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In: Proceedings of the 2013 Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 3554−3561

43 Guillaumin M, Verbeek J, Schmid C. Is that you? Metric learning approaches for face identification. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 498−505

44 Nguyen H V, Bai L. Cosine similarity metric learning for face verification. In: Proceedings of the 10th Asian Conference on Computer Vision. Queenstown, New Zealand: Springer, 2010. 709−720

45 Huang G B, Mattar M, Berg T, Erik L M. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report, University of Massachusetts, Amherst, USA. 2007, 1−11

46 Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(7): 971−987

47 Cao Q, Ying Y M, Li P. Similarity metric learning for face recognition. In: Proceedings of the 2013 International Conference on Computer Vision. Sydney: IEEE, 2013. 2408−2415

48 Wang S J, Jin R. An information geometry approach for distance metric learning. In: Proceedings of the 2009 International Conference on Artificial Intelligence and Statistics. Florida, USA: AISTATS, 2009. 591−598

49 Samaria F S, Harter A C. Parameterisation of a stochastic model for human face identification. In: Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision. Sarasota, USA: IEEE, 1994. 138−142

50 Verma Y, Jawahar C V. Image annotation using metric learning in semantic neighbourhoods. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 836−849

51 Shen C H, Kim J, Wang L, Hengel A. Positive semidefinite metric learning using boosting-like algorithms. *Journal of Machine Learning Research*, 2012, **13**: 1007−1036

52 Bi J B, Wu D J, Lu L, Liu M Z, Tao Y M, Wolf M. AdaBoost on low-rank PSD matrices for metric learning. In: Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011. 2617−2624

53 Rosales R, Fung G. Learning sparse metrics via linear programming. In: Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006. 367−373

54 Huang R Q, Sun S L. Kernel regression with sparse metric learning. *Journal of Intelligent and Fuzzy Systems*, 2013, **24**(4): 775−787

55 Bah B, Becker S, Cevher V, Gozcu B. Metric learning with rank and sparsity constraints. In: Proceedings of the 2014 International Conference on Acoustics, Speech, and Signal Processing. Florence, Italy: IEEE, 2014, 21−25

56 Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the 21th International Conference on Maching Learning. New York: ACM, 2004. 81−88

57 Zou Peng-Cheng, Wang Jian-Dong, Yang Guo-Qing. Distance metric learning based on side information autogeneration for time series. *Journal of Software*, 2013, **24**(11): 2642−2655
(邹朋成, 王建东, 杨国庆. 辅助信息自动生成的时间序列距离度量学习. 软件学报, 2013, **24**(11): 2642−2655)

58 Wang J, Woznica A, Kalousisi A. Parametric local metric learning for nearest neighbor classification. In: Proceedings of the 2012 Annual Conference on Neural Information Processing Systems. Nevada, USA: MIT Press, 2012. 1610−1618

59 Liu Song-Hua, Zhang Jun-Ying, Xu Jin, Jia Hong-En. Kernel-kNN: a new kNN algorithm based on informational energy metric. *Acta Automatica Sinica*, 2010, **36**(12): 1681−1688
(刘松华, 张军英, 许进, 贾宏恩. Kernel-kNN: 基于信息能度量的核 k–最近邻算法. 自动化学报, 2010, **36**(12): 1681−1688)

60 Gao Jun, Wang Shi-Tong, Wang Xiao-Ming. Contextual-distance metric based Laplacian maximum margin criterion. *Acta Automatica Sinica*, 2010, **36**(12): 1661−1673
(皋军, 王士同, 王晓明. 基于语境距离度量的拉普拉斯最大间距判别准则. 自动化学报, 2010, **36**(12): 1661−1673)

61 Chang H, Yeung D Y. Locally smooth metric learning with application to image retrieval. In: Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−7

62 Yang L, Jin R, Mummert L, Sukthankar R, Goode A, Zheng B, Hoi S C H, Satyanarayanan M. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(1): 30−44

63 Zhao K, Liu W, Liu J Z. Optimal semi-supervised metric learning for image retrieval. In: Proceedings of the 20th Multimedia Conference. New York: ACM, 2012. 893−896

64 Cong Y, Yuan J S, Tang Y D. Object tracking via online metric learning. In: Proceedings of the 19th International Conference on Image Processing. Orlando, USA: IEEE, 2012. 417−420

65 Jiang N, Liu W Y, Wu Y. Order determination and sparsity-regularized metric learning adaptive visual tracking. In: Proceedings of the 2012 International Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 1956−1963

66 Yao Zhi-Jun, Liu Jun-Tao, Lai Zhong-Yuan, Liu Wen-Yu. An improved Jensen-Shannon divergence based spatiogram. *Acta Automatica Sinica*, 2011, **37**(12): 1464−1473
(姚志均, 刘俊涛, 赖重远, 刘文予. 一种改进的 JSD 距离的空间直方图相似度度量及目标跟踪. 自动化学报, 2011, **37**(12): 1464−1473)

67 Tran D, Sorokin A. Human activity recognition with metric learning. In: Proceedings of the 2008 European Conference on Computer Vision. Marseille, France: Springer, 2008. 548−561

68 Kliper-Gross O, Hassner T, Wolf L. One shot similarity metric learning for action recognition. In: Proceedings of the 2011 Similarity-Based Pattern Recognition. Berlin, Heidelberg: Springer, 2011. 31−45

69 Lebanon G. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(4): 497−508

70 Jiang N, Liu W Y, Wu Y. Adaptive and discriminative metric differential tracking. In: Proceedings of the 2011 International Conference on Computer Vision and Pattern Recognition. CO, USA: IEEE, 2011. 1161−1168

71 Zhang Y N, Zhang H C, Nasrabadi N M, Huang T S. Multi-metric learning for multi-sensor fusion based classification. *Information Fusion*, 2013, **14**(4): 431−440

72 Yan Yan, Zhang Yu-Jin. State-of-the-art on video-based face recognition. *Chinese Journal of Computers*, 2009, **32**(5): 878−886)
(严严, 章毓晋. 基于视频的人脸识别研究进展. 计算机学报, 2009, **32**(5): 878−886)

73 Gao Quan-Xue, Gao Fei-Fei, Hao Xiu-Juan, Cheng Jie. Image Euclidean distance-based two-dimensional local diversity preserving projection. *Acta Automatica Sinica*, 2013, **39**(7): 1062−1070
(高全学, 高菲菲, 郝秀娟, 程洁. 基于图像欧氏距离的二维局部多样性保持投影. 自动化学报, 2013, **39**(7): 1062−1070)

74 Frank A, Asuncion A. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml. April 1, 2014

75 Liu M Z, Vemuri B C. A robust and efficient doubly regularized metric learning approach. In: Proceedings of the 12th European Conference on Computer Vision, Florence, Italy: Springer, 2012. 646−659

76 Ebert S, Fritz M, Schiele B. Active metric learning for object recognition. In: Proceedings of the 2012 Pattern Recognition. Graz, Austria: Springer, 2012. 327−336

77 Tsagkatakis G, Savakis A E. Online distance metric learning for object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2011, **21**(12): 1810−1821

78 Yu J, Wang M, Tao D C. Semi-supervised multiview distance metric learning for cartoon synthesis. *IEEE Transactions on Image Processing*, 2012, **21**(11): 4636−4648

79 Niu G, Dai B, Yamada M. Information-theoretic semi-supervised metric learning via entropy regularization. In: Proceedings of the 29th International Conference on Machine Learning. Edinburgh, UK: ACM, 2012. 89−96

80 Chechik G, Sharma V, Shalit U, Bengio S. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning*, 2010, **11**: 1109−1135

81 Adrián P S, Francesc J F, Miguel A H. Passive-aggressive online distance metric learning and extensions. *Progress in Artificial Intelligence*, 2013, **2**(1): 85−96

82 Cong Y, Liu J, Yuan J S, Luo J B. Self-supervised online metric learning with low rank constraint for scene categorization. *IEEE Transactions on Image Processing*, 2013, **22**(8): 3179−3191

83 Jain P, Kulis B, Dhillon I S, Grauman K. Online metric learning and fast similarity search. In: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2009. 761−768

84 Ying Y M, Huang K Z, Compbell C. Sparse metric learning via smooth optimization. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2009. 2214−2222

85 Li Z C, Liu J, Yu J, Tang J H, Lu H Q. Low rank metric learning for social image retrieval. In: Proceedings of the 20th ACM International Conference on Multimedia. Japan: ACM, 2012. 853−856

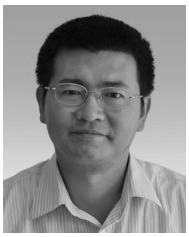
86 Zha Z J, Mei T, Wang M, Wang Z F, Hua X S. Robust distance metric learning with auxiliary knowledge. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. San Francisco: AISTATS, 2009. 1327−1332

87 Huang K H, Jin R, Xu Z L, Liu C L. Robust metric learning by smooth optimization. In: Proceedings of the 26th Uncertainty in Artificial Intelligence. California, USA: AUAI Press, 2010. 244−251

88 Lim D, McFee B, Lanckriet G R G. Robust structural metric learning. In: Proceedings of the 2013 International Conference on Machine Learning. Atlanta, USA: ACM, 2013. 615−623

沈媛媛 厦门大学信息科学与技术学院硕士研究生. 2010 年获安徽大学学士学位. 主要研究方向为距离度量学习和模式识别.
E-mail: shenyuanyuan1989@gmail.com
(**SHEN Yuan-Yuan** Master student at the School of Information Science and Technology, Xiamen University. She received her bachelor degree from Anhui University in 2010. Her research interest covers distance metric learning and pattern recognition.)

严 严 厦门大学信息科学与技术学院副教授. 主要研究方向为计算机视觉和模式识别. 本文通信作者.
E-mail: yanyan@xmu.edu.cn
(**YAN Yan** Associate professor at the School of Information Science and Technology, Xiamen University. His research interest covers computer vision and pattern recognition. Corresponding author of this paper.)



王菡子 厦门大学信息科学与技术学院教授. 主要研究方向为鲁棒统计, 计算机视觉和图像与视频处理.
E-mail: hanzi.wang@xmu.edu.cn
(**WANG Han-Zi** Professor at the School of Information Science and Technology, Xiamen University. His research interest covers robust statistics, computer vision, and image and video processing.)