

基于流形正则的块增量距离尺度学习算法

王 磊^{1,2}, 刘 铁³, 贾华丁¹

(1. 西南财经大学经济信息工程学院, 四川成都 610074; 2. 西南财经大学金融智能与金融工程重点实验室, 四川成都 610074;
3. IBM 中国研究院, 北京 100193)

摘 要: 在实时应用中, 观测样本通常以数据块的形式依次达到, 传统的批量距离算法难以进行学习. 本文提出一种新颖的利用成对约束关系进行学习的块增量距离尺度算法. 首先给出块增量学习的一般模型, 并通过扩展约束克服其容易“过拟合”的缺陷; 然后引入流形正则项使得学习过程中数据块的局部邻域结构得以保持. 实验结果表明, 本文算法学习的距离尺度在测试精度、计算开销上优于现有的增量距离算法, 并且在存储开销方面显著优于批量距离算法.

关键词: 距离尺度学习; 半监督; 块增量学习; 流形正则

中图分类号: TP391. 4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 05-1131-05

Chunk Incremental Distance Metric Learning Algorithm Based on Manifold Regularization

WANG Lei^{1,2}, LIU Tie³, JIA Hua-ding¹

(1. School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China
2. The Key Laboratory of Financial Intelligence and Financial Engineering, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China
3. IBM China Research Laboratory, Beijing 100193, China)

Abstract: In many real-time applications, observed samples always arrive in the form of chunks stream, traditional batch distance metric algorithms can hardly work well in such scenarios. This paper proposes a novel semi-supervised chunk incremental metric learning algorithm on the basis of the pairwise constraints. One general model is given to learn metric incrementally on the arriving chunks at first with its limitation of over-fitting overcome by utilizing extended constraint sets. Then, a manifold regularization term is used to keep locality adjacency structure of chunks during metric learning. Experimental results indicate superiorities of our algorithm, which obtains better accuracy and lower computation costs than existing incremental metric learning algorithms, and needs much less storage costs than batch ones.

Key words: distance metric learning; semi-supervised; chunk incremental learning; manifold regularization

1 引言

距离尺度学习 (distance metric learning) 是机器学习领域的研究热点之一. 在图像检索、基因分类、文本聚类应用中, 准确度量样本之间的相似度是分类/聚类算法成败与否的关键之一. 传统的欧氏距离 (Euclidean) 将样本的全部特征视为同等重要, 已被证明难以胜任于高维空间中的距离度量^[1,2]. 因此, 自适应的距离尺度学习算法最近几年受到广泛关注^[3].

文献[4,5]利用样本的类标记作为监督信息, 通过计算不同特征在度量相似度过程中的最优权重, 得到性能优异的马氏 (Mahalanobis) 距离尺度. 当监督信息严重

缺乏时, 基于样本之间成对约束关系的半监督距离学习算法成为当前的研究热点. Bar-Hillel 等人提出利用 must-link 约束关系进行马氏距离学习的 RCA 算法^[6]; Xiang 等人通过最大化/最小化 cannot-link/must-link 关系上的距离, 利用迭代方法求解全局最优的马氏距离^[7]. Baghshah 等人在文献[7]的基础上, 进一步考虑了样本的几何拓扑结构对距离尺度的影响^[8]. Yeung 等人利用核函数研究了非线性距离的学习问题^[9]. 然而, 大多数距离尺度算法都是批量算法, 需要在学习之前获取全部离线的观测样本. 在许多实时的分类/聚类应用中, 观测样本通常是以数据块的形式分批依次到达 (见图 1). 并且, 每个数据块中的样本数量、类别数目、约束关系数

量均具有随机性特点。显然,积累所有观测样本将导致巨大的存储开销,因而批量距离算法难以学习实时应用数据。

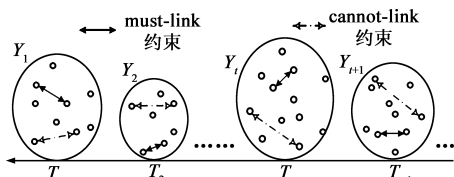


图1 观测样本按块(chunk)分批到达的示意图

增量式距离尺度学习算法近年来引起广泛重视。文献[10,11]在 LogDet 正则意义下,通过最小化实时预测损失推导了距离尺度的增量更新规则。文献[12]通过两次连续投影,给出距离尺度更新的 POLA 算法。然而,这些算法存在明显不足:(1)只能处理观测样本成对到达的情况,且距离尺度学习的效率较低;(2)为度量预测损失,需要估计当前到达样本对之间的目标距离。

本文对半监督增量距离尺度算法进行深入研究。首先提出针对数据块进行增量距离学习的一般模型,并在尺度渐进更新的假设下,通过扩展约束集避免了块增量学习过程中的“过拟合”问题;然后,引入流形正则项,使得数据块的局部邻域结构在学习过程中得以保持。最后,给出了基于流形正则的块增量距离尺度学习算法,并分析了其空间和时间复杂度。

2 半监督距离尺度学习

已知两个观测样本 $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$, 它们之间的马氏距离可以按照如下公式计算:

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j)} \quad (1)$$

其中, $\mathbf{A} \in \mathbb{R}^{n \times n}$ 是待学习的马氏距离尺度, 满足半正定性质。利用矩阵分解技术^[13], 将 $\mathbf{A}$ 分解成 $\mathbf{A} = \mathbf{P} \cdot \mathbf{P}^T$, 即 $\mathbf{P} \in \mathbb{R}^{n \times d}$ 是与 $\mathbf{A}$ 相关的距离投影矩阵。因此, 马氏距离可重写为:

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\|\mathbf{P}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2} \quad (2)$$

给定规模为 N 的样本集 $C = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^n$, 且令 S 和 D 分别是它们的 must-link 和 cannot-link 成对约束集。半监督距离尺度学习的基本原则是: 在距离尺度 $\mathbf{A}$ 下, 满足 must-link 约束关系的样本更靠近, 而满足 cannot-link 约束的样本更加远离。设 $l(\mathbf{P}, S, D)$ 依据此原则定义的损失函数, 则最优尺度 $\mathbf{A}$ 或者 $\mathbf{P}$ 可通过下面的二次优化问题求解:

$$\begin{aligned} \mathbf{P}^* &= \arg \max_{\mathbf{P} \in \mathbb{R}^{n \times d}} l(\mathbf{P}, S, D) \\ &= \arg \max_{\mathbf{P} \in \mathbb{R}^{n \times d}} \left(\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} \|\mathbf{P}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \eta_s \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} \|\mathbf{P}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 \right) \\ &= \arg \max_{\mathbf{P} \in \mathbb{R}^{n \times d}} (\text{tr}(\mathbf{P}^T \mathbf{D} \mathbf{P}) - \eta_s \text{tr}(\mathbf{P}^T \mathbf{S} \mathbf{P})) \end{aligned} \quad (3)$$

式(3)中, $\text{tr}(\cdot)$ 是矩阵的迹算子; 参数 $\eta_s > 0$ 用于权衡约束集 S 和 D 对于损失的贡献; 等式约束 $\mathbf{P}^T \mathbf{P} = \mathbf{I}$ 用于避免出现退化解。此外, S 和 D 是关于约束集的协方差矩阵, 定义如下:

$$\mathbf{S} = \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (4)$$

$$\mathbf{D} = \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (5)$$

3 基于流形正则的块增量距离尺度学习

3.1 块增量距离尺度学习的一般模型

若所有样本按图 1 所示的数据块依次到达, 即 $Y_1, Y_2, \dots, Y_t, \dots$ 。在每个时间点 T_t , 数据块由三元组表示, 块的大小为 N_t 。

令 $\mathbf{P}_{t-1}$ 表示时间点 T_{t-1} 已学习的距离投影矩阵, 则块增量距离尺度学习目标是: 利用 Y_t 中的监督信息 (即成对约束集) 和已知投影矩阵 $\mathbf{P}_{t-1}$, 更新时间点 T_t 的新投影矩阵 $\mathbf{P}_t$, 以增量方式实时学习距离尺度。

不失一般性, 我们假设距离尺度随着数据块的到达呈渐进变化的趋势, 任意两个相邻数据块上学习的距离尺度彼此相似, 即关于 $\forall \mathbf{x}_i, \mathbf{x}_j \in C_t$ 和任意小的正实数 δ , 满足

$$|d_{A_t}(\mathbf{x}_i, \mathbf{x}_j) - d_{A_{t-1}}(\mathbf{x}_i, \mathbf{x}_j)| \leq \delta \quad (6)$$

设 S_t 和 D_t 是在约束集 S_t 和 D_t 上计算的协方差矩阵, 块增量尺度学习的一般模型如下:

$$\mathbf{P}_t^* = \arg \max_{\mathbf{P} \in \mathbb{R}^{n \times d}} (l(\mathbf{P}, S_t, D_t) - \eta_r l(\mathbf{P}, \mathbf{P}_{t-1})) \quad (7)$$

其中, $l(\mathbf{P}, S_t, D_t)$ 是距离尺度在数据块 Y_t 上的损失; 而 $r(\mathbf{P}, \mathbf{P}_{t-1})$ 是用于保证尺度渐进更新的正则项, 正则参数为 $\eta_r > 0$ 。可见, 该优化模型在增量更新距离尺度时, 不仅考虑数据块内的成对约束关系, 而且考虑距离尺度渐进更新的性质。文献[10]和[11]采用了类似的模型更新距离尺度。

然而, 由于数据块中样本集和约束集的规模均具有随机性, 当 N_t 或者 $|S_t|, |D_t|$ 取较小值时, 对高维样本直接用公式(7)计算 $\mathbf{P}_t^*$ 容易产生“过拟合”现象, 即学习的距离尺度过分拟合 Y_t 中的约束集, 从而导致学习性能的降低。

克服“过拟合”问题的有效途径之一是增大约束集的规模。考虑到实时应用中, 相邻数据块一般具有相同或者相似 (考虑概念漂移的情况) 的分布函数, 可以利用相邻数据块中的约束集扩展当前块的约束集。

令 $S_{t-1}, D_{t-1}, S_{t-2}, D_{t-2}, \dots$ 表示时间点 $T_{t-1}, T_{t-2}, \dots$ 的 must-link 和 cannot-link 约束集, 我们可以将当前块的约束集扩展为: $\hat{S}_t = \sum_{i=0}^v S_{t-i}$ 和 $\hat{D}_t = \sum_{i=0}^v D_{t-i}$, 其中, v 表示邻近窗口的大小。并且, 扩展后的协方差矩

阵 $\hat{S}_t$ 和 $\hat{D}_t$, 定义如下:

$$\hat{S}_t = S_t + \sum_{i=1}^v \rho^i S_{t-i} \quad (8)$$

$$\hat{D}_t = D_t + \sum_{i=1}^v \rho^i D_{t-i} \quad (9)$$

这里, $\rho \leq 1$ 是衰减因子, 目的是限制较近时间点的数据块才能对扩展协方差矩阵产生明显影响. 它的取值范围通常为 $\rho \in [0.1, 0.5]$.

3.2 基于流形正则的块增量距离尺度学习算法

为给出模型(7)中正则函数的具体形式, 我们采用和文献[8]类似的方法: 在更新距离尺度时, 不仅考虑样本间的成对约束关系, 而且保持样本集的几何流形结构. 但与前者不同的是, 我们采用局部邻域结构来描述样本集的内蕴流形^[1].

在时间 T_t , 设 A_{t-1} 是当前已知的距离尺度, 利用其计算新到达数据块 Y_t 中所有样本间的距离, 并生成关于 Y_t 的邻接图. 其中, 图上任意边的权重 $W_{t-1}(i, j)$ 由如下的热核函数计算:

$$W_{t-1}(i, j) = \begin{cases} \exp\left(-\frac{\|P_{t-1}^T(x_i - x_j)\|^2}{\sigma}\right), & \text{if } x_j \in \text{knn}(x_i) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

这里, $\text{knn}(\cdot)$ 表示 x_i 的 k 最近邻集合; σ 是与样本相关的常量. 显然, 邻接矩阵 $W_{t-1} \in \mathbb{R}^{N_t \times N_t}$ 体现了样本集 C_t 的局部邻域结构, 其基本元素为 $W_{t-1}(i, j)$.

根据邻域保持投影方法的思想^[1], 在 Y_t 上更新距离尺度时应该能保持样本集原有的邻域结构; 对于破坏邻域结构的投影矩阵 P 应该给予较大的惩罚. 因此, 我们采用如下的正则函数:

$$r(P, P_{t-1}) = \frac{1}{2} \sum_{i,j=1}^{N_t} \|P^T(x_i - x_j)\|^2 W_{t-1}(i, j) \quad (11)$$

分析可知, 公式(11)和距离尺度渐进变化的基本假设是一致的. 因为在尺度 A_{t-1} 下相互邻近的样本对, 若在更新后的尺度 A_t 下相互远离, 必将产生很大的正则罚值.

通过定义对角矩阵 $B_{t-1} \in \mathbb{R}^{N_t \times N_t}$ 和 Laplacian 矩阵 $L_{t-1} = B_{t-1} - W_{t-1}$, 我们对式(11)进行简化. 其中,

$$\begin{aligned} B_{t-1}(i, i) &= \sum_{j=1}^{N_t} W_{t-1}(i, j), \\ r(P, P_{t-1}) &= \sum_{i=1}^{N_t} x_i^T P P^T x_i B_{t-1}(i, i) \\ &\quad - \sum_{i,j=1}^{N_t} x_i^T P P^T x_j W_{t-1}(i, j) = \text{tr}(X^T P P^T X L_{t-1}) \end{aligned} \quad (12)$$

这里, $X \in \mathbb{R}^{n \times N_t}$ 是数据块 Y_t 中样本的数据矩阵.

将式(8)、(9)和(12)代入式(7), 可以得到扩展约

束集和考虑几何流形结构后的块增量距离尺度学习模型.

$$\begin{aligned} P_t^* &= \arg \max_{P \in \mathcal{I}} (\text{tr}(P^T \hat{D}_t P) - \eta_s \text{tr}(P^T \hat{S}_t P) \\ &\quad - \eta_r \text{tr}(P^T X L_{t-1} X^T P)) = \arg \max_{P \in \mathcal{I}} (\text{tr}(P^T M P)) \end{aligned} \quad (13)$$

式中, 矩阵 M 定义为 $M = (\hat{D}_t - \eta_s \hat{S}_t - \eta_r X L_{t-1} X^T)$.

注意到, M 是实对称矩阵, 根据矩阵优化理论^[13], 式(13)中目标函数最优值是 M 的前 d 个最大的非负特征值之和, 而最优解 P_t^* 由它们对应的 d 个特征向量构成, 即

$$P_t^* = [v_1, v_2, \dots, v_d] \in \mathbb{R}^{n \times d} \quad (14)$$

其中, $v_1, v_2, \dots, v_d, \dots, v_n$ 是与矩阵 M 的 n 个特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq \dots \geq \lambda_n$ 相应的特征向量.

综上, 我们得到一种块增量式距离学习算法, 用于处理观测样本按块依次到达的情况, 称为“基于流形正则的块增量距离尺度学习算法”, 或 MR-CIML. 其主要步骤如算法 1 所示.

算法 1 MR-CIML 算法

输入: 观测数据块 $Y_1, Y_2, \dots, Y_t, \dots$; 初始距离尺度 $A_0 = I$, 初始约束集 $S_0 = D_0 = \emptyset$; 参数, η_s, η_r ;
输出: 马氏距离尺度 $A_1, A_2, \dots, A_t, \dots$

For $t = 1, 2, 3, \dots$

- 按式(4)和(5)计算协方差矩阵 S_t 和 D_t ;
- 按式(8)和(9)计算扩展协方差矩阵 $\hat{S}_t$ 和 $\hat{D}_t$;
- 利用尺度 A_{t-1} 在数据块 Y_t 上构建邻接图, 按式(10)计算邻接矩阵 W_{t-1} , 并计算 Laplacian 矩阵 L_{t-1} ;
- 利用特征值分解技术对模型(13)进行求解, 并由式(14)得到其最优解 P_t^* ;
- 更新距离尺度为 $A_t = P_t^* (P_t^*)^T$.

End For

我们对算法 1 的空间复杂度进行分析. 容易计算, 在时间点 T_t , 存储 Y_t, A_{t-1}, S_{t-1} 和 D_{t-1} 需要的存储空间为 $\max\{n \cdot N_t, n^2, N_t^2\}$; 计算协方差矩阵 $S_t, D_t, \hat{S}_t$ 和 $\hat{D}_t$ 需要的空间为 $O(n^2)$; 存储邻接图、矩阵 W_{t-1} 以及 L_{t-1} 需要的空间为 $O(N_t^2)$; 求解模型(13)需要的空间为 $\max\{O(n^2), O(n \cdot N_t)\}$. 因此, 若数据块的最大规模为 $N_{\max}$, 算法 1 的存储空间复杂度为 $\max\{O(n^2), O(N_{\max}^2)\}$.

接着, 我们对其时间复杂度进行分析. 计算矩阵 $S_t, D_t, \hat{S}_t$ 和 $\hat{D}_t$ 所需的时间开销为 $O(n^2 \cdot \max\{|S|, |D|\})$; 构造邻接图以及计算矩阵 W_{t-1} 和 L_{t-1} 需要的时间约为 $O(n^2 \cdot N_t^2)$; 求解模型(13)的时间开销约为 $O(n^3)$. 因此, 算法 1 每次更新距离尺度的时间复杂度为 $\max\{O(n^3), O(n^2 \cdot N_{\max}^2)\}$.

4 实验与分析

为检验本文提出的 MR-CIML 算法的性能,我们将它用在模式分类任务中度量样本之间的距离,并比较它与 Xiang 的批量算法^[7],Davis 的增量算法^[10],以及 POLA 增量算法^[12]在测试误差、存储和计算开销等性能指标上的优劣.实验在四个高维图像数据集上进行(如表 1 所示).其中,Extended Yale-B 包含 28 个不同人脸在 9 种姿态、64 种光照条件下的图像^[14];CMU PIE 包含 68 个不同人脸在 13 种姿态、43 种光照条件、4 种表情下的图像^[15];ORHD 和 PRHD 是两种典型的 UCI 机器学习数据集*,由不同书写者提供的手写体阿拉伯数字(0~9)图像构成.此外,我们将 Extended Yale-B 和 CMU PIE 数据集的每幅原始图像缩减成 32×32 像素的 256 色灰度图像.实验环境为一台 3GHz CPU,4GB 内存的主机,所有算法均采用 C++ 程序实现.

实验中,将每个数据集按 2:1 的比例随机分割成训练集和测试集,并利用训练样本的类标记模拟出成对约束集 S 和 D ,规模约为训练集规模 10%.上述距离尺度算法在训练集上学习距离尺度 A 后,支持向量机使用这些距离尺度分别在训练集上训练分类器,然后统计它们在测试集上取得的测试误差.其中,支持向量机采用如下的高斯核函数:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{(\mathbf{x}_i - \mathbf{x}_j)^T A (\mathbf{x}_i - \mathbf{x}_j)}{2\sigma^2}\right) \quad (15)$$

数据集	样本总数	特征数	类别数
Extended Yale-B	16,128	1,024	28
CMU PIE	41,368	1,024	68
UCI optical recognition of handwritten digits (ORHD)	5,620	64	10
UCI pen-based recognition of handwritten digits (PRHD)	10,992	16	10

MR-CIML 算法的参数和 η_r 在实验中参照文献[8]分别设置为 2.0 和 0.2.为模拟依次到达的数据块(假设规模固定为 $\bar{N}$),采用无放回随机抽样的方式从训练集中生成数据块,依次输入到算法(ORHD 和 PRHD 设置 $\bar{N} = 100$,其它为 $\bar{N} = 500$).成对约束集在每个数据块内按 10% 的比例随机生成.为简便起见,邻近窗口和衰减因子统一设置为 $\nu = 3$ 和 $\rho = 0.1$.此外,支持向量机分类器的模型参数对所有距离尺度均设置为相同的值,并由 5-fold 交叉验证方法选取.

我们将上述实验重复 10 次,表 2 统计了上述四种距离尺度算法学习的距离以及传统欧氏距离用于图像分类任务时取得的平均测试误差.图 2 和图 3 详细描述了 CMU PIE 和 ORHD 数据集上测试误差随着已到达样本数量的变化规律.

表 2 比较五种距离尺度在图像分类任务中的测试误差(×100%)

距离尺度	Extended Yale-B	CMU PIE	ORHD	PRHD
Xiang's ^[7]	4.22 ± 0.63	2.51 ± 0.28	1.29 ± 0.15	1.02 ± 0.07
Davis's ^[10]	7.45 ± 0.99	4.20 ± 0.44	1.76 ± 0.12	1.38 ± 0.11
POLA ^[12]	10.07 ± 1.36	4.41 ± 0.60	1.90 ± 0.16	1.38 ± 0.11
MR-CIML	4.05 ± 0.48	1.95 ± 0.24	1.24 ± 0.09	1.05 ± 0.08
Euclidean	14.18 ± 1.70	7.19 ± 0.69	3.43 ± 0.22	1.97 ± 0.16

从表 2 的实验结果可以看出,距离尺度算法学习的距离均显著优于传统的欧氏距离,说明了距离尺度学习的重要性.MR-CIML 算法获得的测试精度在所有情况下显著优于另外两种增量算法,且略优于 Xiang 的批量算法.这表明本文算法通过块增量的方式能够学习获得最优的距离尺度,两点原因可以给予解释:(1)采取的扩展约束集的方法能够有效避免“过拟合”的出现;(2)利用流形正则项进一步提升了算法的性能(和文献[8]的结论一致).图 2 和图 3 的结果进一步揭示了 MR-CIML 算法优于另外两种增量算法的原因:前者学习的测试误差能够很快接近批量算法的水平,而后者则下降较为缓慢.

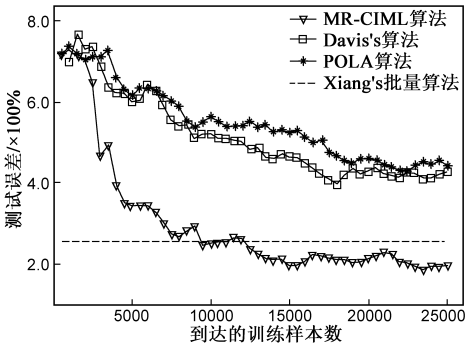


图2 测试误差随着到达样本数的变化(CMU PIE数据集)

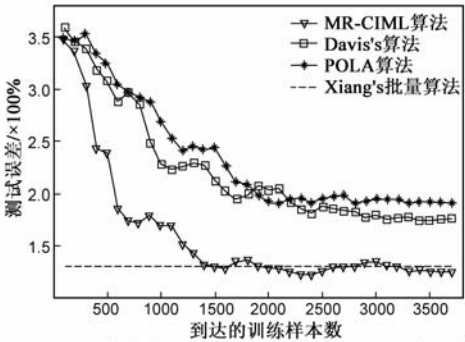


图3 测试误差随着到达样本数的变化(ORHD数据集)

此外,表 3 和表 4 分别统计了实验过程中求解上述距离尺度所需要的存储开销和计算开销.

* UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]

表 3 比较四种距离尺度算法所需的存储开销 (M Byte)

距离尺度算法	Extended Yale-B	CMU PIE	ORHD	PRHD
Xiang's	245.74	516.13	1.57	1.05
Davis's	1.72	1.98	0.08	0.01
POLA	2.46	2.29	0.14	0.10
MR-CIML	4.40	4.71	0.21	0.07

表 4 比较四种距离尺度算法所需的计算开销 (秒)

距离尺度算法	Extended Yale-B	CMU PIE	ORHD	PRHD
Xiang's	65	142	2.4	1.9
Davis's	424	776	28.6	25.2
POLA	501	1,105	32.3	44.1
MR-CIML	132	283	3.4	4.0

可见,由于仅需要存储单个数据块的样本,MR-CIML 算法需要的存储开销远远低于 Xiang 的批量算法.在计算开销方面,本文算法显著优于另外两种距离尺度算法,出现该现象的原因是:后者在每一对成对约束样本到达时都需要更新距离尺度,且每次更新的时间复杂度至少为 $O(n^3)$,造成总计算开销较高;而本文算法仅在数据块达到时才更新距离.

5 结论

针对批量距离尺度算法难于进行实时学习的问题,本文提出一种新颖的块增量半监督距离尺度算法.它利用随机达到的数据块中的成对约束关系进行学习,并通过约束集扩展解决了由于成对约束数量不足而引起的“过拟合”现象.此外,该算法采用邻域保持投影技术,在学习过程中引入正则项以保持数据块的流形结构.实验结果表明,本文算法学习的距离尺度在测试精度、计算开销方面明显优于其它增量算法,且需要的存储开销显著低于批量算法.

参考文献

[1] He Xiaofei, Niyogi P. Locality preserving projections [A]. Advances in Neural Information Processing Systems [C]. Cambridge: MIT Press, 2003. 153 – 160.

[2] 刘中华, 周静波, 等. 距离保持投影非线性降维级数的可视化与分类 [J]. 电子学报, 2009, 37(8): 1820 – 1825.

Liu Zhonghua, Zhou Jingbo, et al. Non-linear dimensionality reduction techniques of distance preserving projection for visualization and classification [J]. Acta Electronica Sinica, 2009, 37(8): 1820 – 1825. (in Chinese)

[3] Liu Yang, Rong Jin. Distance Metric Learning: A Comprehensive Survey [R]. Department of Computer Science and Engineering, Michigan State University, 2006.

[4] Goldberger J, Hinton G, Salakhutdinov R. Neighborhood components analysis [A]. Advances in Neural information Processing System [C]. Cambridge: MIT Press, 2004. 513 – 520.

[5] Weinberger K, Blitzer J, Saul L. Distance metric learning for large margin nearest neighbor classification [A]. Advances in Neural information Processing System [C]. Cambridge: MIT Press, 2006. 1473 – 1480.

[6] Bar-Hillel A, Hertz T, Shental N, et al. Learning a Mahalanobis metric from equivalence constraints [J]. Journal of Machine Learning Research, 2005, 6(6): 937 – 965.

[7] Xiang Shiming, Nie Feiping, Zhang Changshui. Learning a Mahalanobis distance metric for data clustering and classification [J]. Pattern Recognition, 2008, 41(12): 3600 – 3612.

[8] Baghshah M S, Shouraki S B. Semi-supervised Metric Learning using Pairwise Constraints [A]. Proc of IJCAI'09 [C]. San Francisco: Morgan Kaufmann Publishers, 2009. 1217 – 1222.

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

[10] Davis J V, Kulis B, et al. Information theoretic metric learning [A]. Proc of ICML'07 [C]. Corvalis: Omni Press, 2007. 209 – 216.

[11] Jain P, Kulis B, et al. Online metric learning and fast similarity search [A]. Advances in Neural Information Processing Systems [C]. Cambridge: MIT Press, Cambridge, 2008. 761 – 768.

[12] Shai S S, Yoram S, Andrew Y N. Online and batch learning of pseudo-metrics [A]. Proc of ICML'04 [C]. Banff: ACM Press, 2004. 743 – 750.

[13] Golub G H, Loan C F. Matrix Computations (3rd) [M]. Baltimore: the Johns Hopkins University Press, 1996. 308 – 469.

[14] Georgiades A, Belhumeur P, et al. From few to many: illumination cone models for face recognition under variable lighting and pose [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643 – 660.

[15] Wang Wei, Shan Shiguang, Gao Wen, et al. An improved active shape model for face alignment [A]. In Proc. of ICMI'02 [C]. Pittsburgh: IEEE Computer Society, 2002. 523 – 528.

作者简介



王磊男, 1978 年生于河南信阳. 2007 年毕业于电子科技大学, 获工学博士学位. 现为西南财经大学经济信息工程学院讲师, 主要从事模式识别、人工智能、数据挖掘方面的研究工作.

E-mail: wanglei_t@swufe.edu.cn

刘铁男, 1979 年生于湖北武汉. 2007 年毕业于西安交通大学, 获工学博士学位. 现为 IBM 中国研究院分析与优化部门高级研究员, 主要从事数据挖掘、金融计算方面的研究工作.