体对于梯度更新影响是很小的, 因此最大相关熵准则拥有良好的鲁棒性.

2.4 算法设计

算法 1. 基于最大相关熵准则的鲁棒度量学习算法

输入: 样本$(x_1,y_1),(x_2,y_2),\cdots,(x_k,y_k)$,学习速率 r, 收敛条件ε, 正则化项系数λ.

输出: L

1) 样本转化为样本对(x_j-x_k,y_t), 当样本对中两个样本是同类时, $y_t=1$; 当两个样本是异类时, $y_t=-1$.

2) 初始化 L 为单位阵 I, 依据公式 (15) 计算初始损失函数值 *loss*.

3) 选择学习速率 r 开始梯度下降, 得到新的目标度量矩阵 *L_new* 和损失函数 *loss1*.

4) 再次梯度下降, 得到新的目标度量矩阵和损失函数 *loss2*.

5) 重复 2)、3) 步骤, 直到满足$|loss1-loss2|/|loss2-loss|<\varepsilon$.

6) 最后得到的 *L_new* 也就是需要的最后输出 L.

3 实验结果与分析

3.1 参数设置

通过公式可以看出, 实验有 4 个主要参数,分别是正则化项系数λ, 控制对距离的敏感程度高斯参数σ, 学习速度 r, 收敛判断条件ε.

λ作为正则项系数, 可以控制模型的复杂程度. σ作为高斯核参数, 可以控制损失函数对距离变化的敏感程度, 学习速度 r 可以调整模型的拟合时间, 好的速度可以使模型快速拟合, 又不至于过拟合或者陷入死循环, 收敛判定条件ε可以给定合适的拟合临界点, 使得拟合的模型更好的收敛, 更好的完成分类任务.

最开始判定的正则项系数λ与高斯参数σ范围是$[10^{-6},10^{3}]$之间, 采用 10 的指数阶变化; 学习速度 r 变化范围在$[10^{-5},1]$之间, 同样采用 10 的指数阶变化; 收敛条件ε取损失函数的当前迭代变化的绝对值除以总的变化的绝对值, 范围在$[10^{-5},10^{-2}]$之间, 也采用 10 的指数阶变化.

由于有 4 个主要参数需要调整, 实验中先调整变化范围小的 r 和ε, 然后将其固定, 再调整变化范围大的λ和σ. 经过在 car evaluation database、teaching assistant evaluation database, balance scale weight & distance database, glass identification evaluation database 这四个数据集上的多次反复试验, 通过取定不同的 r 和ε, 测试变化范围内的λ和σ.发现最佳分类准确率往往在这 4 个核心参数设置在如下变化范围内: 正则项系数λ与高斯参数σ范围是[10^–4,10^2]之间采用 10 的指数阶变化; 学习速度 r 基本可以设定为 10^–4; 收敛条件取损失函数的相对变化, 基本可以设定为 10^–3. 经过实验确定了参数范围后, 将可以极大地降低 metric 学习时间, 同时学习到鲁棒性更好的 metric.

3.2 对比实验简介及对应参数设置

与本文中所提出的基于最大相关熵准则的度量学习算法做对比实验的是 ITML、LMNN 和 RDML 三种算法.

ITML 是 Davis JV 等人在文献[12]中提出的与信息论相关的度量学习算法. 算法的思想是在距离函数约束条件下将两个多元高斯之间的差分相对熵 (KL 散度) 最小化, 从而形成待解问题. 然后将这个问题转化为一个特定的 Bregman 优化问题来表达求解. 实验中需要设置的参数是松弛系数. 实验时设置松弛系数按 10 的指数阶变化, 然后通过验证集选出最佳松弛系数, 最后应用到测试集中. 由于 ITML 每次实验得到的结果并不固定,因此每组做 10 次实验, 然后取平均值.

LMNN 是 Weinberger KQ 等人在文献[15]中提出的经典度量学习算法. 算法的核心思想在于, Metric 是以 k 个最近邻总是属于同一个类, 而不同类的例子是大幅分开为目标来进行训练的. 另外此方法在处理多分类的问题时不需要修改或扩展. 实验中该方法的参数直接通过验证集学习到, 然后应用到测试集中. 由于 LMNN 每次实验得到的结果也不固定, 同样每组做 10 次实验, 然后取平均值.

RDML 是 Jin R 等在文献[18]中提出的度量学习算法. 算法提出在适当的约束下, 正则化距离度量学习可以独立于维度, 使其适合处理高维数据. 在实验中需要设置的参数是正则项系数. 实验时设置正则项系数按 10 的指数阶变化, 然后通过验证集选出最佳正则项系数, 最后应用到测试集中.

3.3 UCI 标准数据集上的实验结果

数据预处理: 为了模拟实际噪声影响, 本文在实验中给实验数据集全部加了高斯噪声.具体的加噪方法是: 先选择需要加噪的样本和特征维度, 求取对应特征维度的平均值 *mean* 和方差 *std*, 利用 Matlab 中的随机函数 *normrnd*, 给这些样本的对应维度的特征加上 [0,*std*]之间的噪声. 加噪样本数量比例是 50%, 加噪特征比例是 50%. 在专用的机器学习数据集 UCI 上选取了 4 个数据集 car evaluation database, teaching assistant evaluation database, balance scale weight & distance

database, glass identification evaluation database (下载地址: http://archive.ics.uci.edu/ml/index.php), 在这四个数据集上分别进行实验. 其中 car evaluation database 是汽车评价数据集, 有 6 个特征, 分别是购买价格、维修价格、车门数量、座位数、后备箱大小以及安全性. 汽车种类共有 4 类, 一共 1728 个样本. Teaching assistant evaluation database 是助教评价数据集, 一共 5 个特征, 分别是助教母语、课程指导员、课程、是否夏季课程以及班级大小. 该数据集一共分 3 类, 一共 151 个样本. Balance scale weight & distance database 是天平倾向数据集, 有 4 个特征, 分别是左边重量、左边距中心距离、右边重量、右边距中心距离. 天平倾向种类一共 3 类, 共 625 个样本. Glass identification evaluation database 是玻璃杯评价数据集, 一共 10 个特征. 玻璃杯种类一共分 7 类, 一共 214 个样本.

进行对比实验的是上文中提到的度量学习中领先的三种各具特色的方法 ITML,LMNN,RDML. 将这三种方法与本文提出的算法应用到上述 4 个标准数据集中, 分别学习到合适的度量矩阵, 最后使用简单的 KNN 分类中, 比较实验结果. 实验结果如表 1、2、3、4 所示.

表 1 在 car evaluation database 上的实验结果

	实验一	实验二	实验三	实验四	实验五	平均值
LMNN	0.7284	0.5608	0.7727	0.4290	0.4273	0.5836
RDML	0.7468	0.5757	0.7445	**0.5225**	**0.5191**	0.6217
ITML	0.7156	0.5734	0.7444	0.5183	0.5177	0.6139
Our algorithm	**0.7491**	**0.6069**	**0.8451**	0.5214	0.5145	**0.6474**

表 2 在 teaching assistant evaluation database 上的实验结果

	实验一	实验二	实验三	实验四	实验五	平均值
LMNN	0.4500	0.3789	0.4013	0.4355	0.4342	0.4200
RDML	0.4342	0.3947	0.3421	0.3816	0.4211	0.3947
ITML	0.4395	0.4026	0.3447	0.4105	0.4211	0.4037
Our algorithm	**0.5514**	**0.5327**	**0.4860**	**0.4860**	**0.4474**	**0.5007**

表 3 在 balance scale weight & distance database 上的实验结果

	实验一	实验二	实验三	实验四	实验五	平均值
LMNN	0.7601	0.7204	0.6243	0.6125	0.6294	0.6693
RDML	0.5815	0.6805	0.6709	0.6294	0.6294	0.6383
ITML	0.7482	0.6904	0.5754	0.6157	0.6281	0.6516
Our algorithm	**0.7827**	**0.7252**	**0.6997**	**0.6613**	**0.6326**	**0.7003**

表 4 在 glass identification evaluation database 上的实验结果

	实验一	实验二	实验三	实验四	实验五	平均值
LMNN	0.4358	0.3972	0.4596	0.4945	0.4055	0.4386
RDML	0.4954	0.4587	0.4404	0.5229	0.3945	0.4624
ITML	0.4945	0.4587	0.4606	0.5165	0.3991	0.4659
Our algorithm	**0.5046**	**0.5321**	**0.4862**	**0.5321**	**0.4587**	**0.5027**

如表 1、2、3、4 所示, 在 car evaluation database, teaching assistant evaluation database, balance scale weight & distance database, glass identification evaluation database 这四个数据集上的实验结果证明本文提出的基于最大相关熵准则的度量学习算法相比已有的度量学习算法 LMNN\RDML\ITML 在处理有噪声的数据集时分类准确率更高. 虽然在表 1 的 car evaluation database 上的实验出现偶尔的 RDML 的领先情况, 然而任何方法都不能保证在所有的数据集上都有效, 出现较小的波动是很正常的事情. 可能这个数据集更适合 RDML 或者本次随机加噪的结果刚好对 RDML 的影响较小, 使得 RDML 的分类准确率没有下降很多. 实验中更需要关注的是多次实验的平均结果, 很明显的是在平均结果方面, 本文算法都对另外三种算法保持领先优势.

表 1、2、3、4 已经证明了本文提出的算法在处理受到高斯噪声的影响的标准数据集分类问题时, 可以有效地提高分类准确率. 接下来, 详细地拓展实验, 在 glass identification evaluation database 上进行 2 组进阶实验, 观察比较随着噪声的变化, 不同的度量学习算法对比本文提出的算法, 分类准确率的变化趋势.

进阶试验 1, 固定加噪样本数量比例为 50%, 在不同加噪特征比例上进行试验, 不同比例均进行 3 次实验, 然后取平均值, 实验结果如图 2 所示.

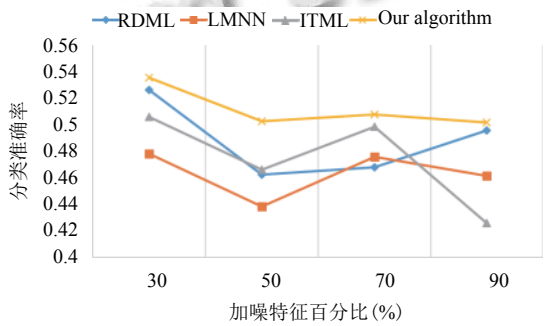

图 2 分类准确率随加噪特征百分比变化图

从图 2 中可以看出, 随着加噪特征百分比的增多, 本文的算法通常都领先其他算法, 而且随着加噪特征百分比的增多, 识别准确率的变化不像其他算法那样有较大的波动, 识别准确率依然保持稳定, 说明本文算法的鲁棒性良好.

进阶试验 2, 固定加噪特征比例为 50%, 在不同加噪样本数量比例上进行实验, 同样的不同比例均进行

3 次实验, 然后取平均值.实验结果如图 3 所示.

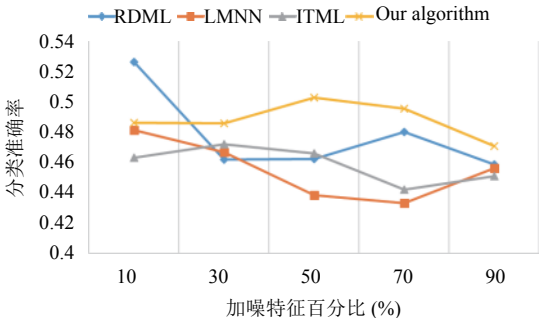

图 3 分类准确率随着加噪样本数量百分比变化图

从图 3 中可以看出, 随着加噪样本数量百分比的增多, 本文的算法的分类准确率逐渐开始领先其他算法, 并且随着夹杂噪声的样本数量的增加, 分类准确率并不会产生较大的波动, 变化较为平稳. 这说明本文的算法对噪声的鲁棒性是更好的.

3.4 人脸数据集 YALEB 实验

在上述 UCI 的 4 个标准数据集上验证算法有效性之后, 接下来将在标准人脸数据库上验证. 实验选定 YALEB 数据集 (下载地址: http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html 或者 http://download.csdn.net/download/fantacy08/8077733). 数据集由耶鲁大学提供, 该数据集经常用于人脸检测和人脸识别. 数据集一共包含 38 个人, 每个人有不同光照条件下的 64 张人脸图片, 每张图片经过裁剪后是 32×32 的图片. 下载的人脸图片需要进行一定的预处理, 本文的处理方法分为两步. 其一是加噪部分, 仍然和上述 4 个小数据集的加噪方法一样, 加噪样本数量百分比和加噪特征百分比都选择 50%; 其二是降维部分, 本文选择 PCA 降维处理, 提取前 100 个特征用于接下来的分类实验.

表 5 在 YaleB 上的实验结果

	实验一	实验二	实验三	实验四	平均值
LMNN	0.7833	0.7752	0.7963	0.7970	0.7880
RDML	0.8048	0.7883	0.7974	0.8156	0.8015
ITML	0.8214	**0.8247**	0.8189	0.8278	0.8232
Our algorithm	**0.8255**	0.8246	**0.8205**	**0.8296**	**0.8251**

从表 5 中也可以看出, 本文提出的算法相比已有的 LMNN, RDML, ITML, 在处理这样的加噪人脸时, 分类准确率同样得到了提高. 并且实验结果比较稳定, 基本不会因受到噪声的干扰而产生较大的波动. 说明了本文算法的鲁棒性经过实验验证是成功的. 虽然在表 5 中的实验二出现偶尔的 ITML 领先的情况, 实验中要考虑到 ITML 每次实验的结果不固定, 通常是 10 次实验的平均值, 因此可能出现选中的 10 次实验的分类准确率都较高, 进而引起平均准确率的提升. 另外单次的实验无法反映一般性, 多次实验的平均值是更为重要的判断依据. 从表 5 中可以明显看出 4 次实验的平均值中, 本文的算法保持领先优势.

接下来, 与上述 glass 数据集一样, 进行拓展试验, 观察人脸分类准确率随着加噪特征百分比的变化趋势. 实验结果如图 4 所示.

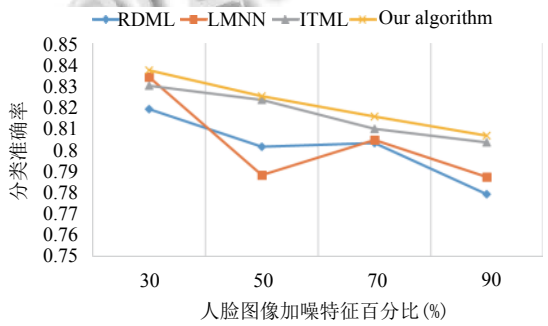

图 4 人脸分类准确率随人脸加噪特征百分比变化图

从图 4 可以看出, 本文提出的算法在处理加噪人脸图像时, 分类准确率往往保持领先, 同时对比另外三种度量学习方法, 分类准确率变化很稳定, 不会大幅度变化, 鲁棒性良好.这说明本文算法不仅在处理 UCI 上的一些小型数据集有很好的效果, 对于大型的人脸数据集同样有良好的效果.

4 总结

本文引入信息论中的最大相关熵准则, 提出了基于最大相关熵准则的鲁棒度量学习算法. 经过实验验证, 该方法在处理噪声环境中的分类问题时, 有优秀的表现, 是一个可行的算法. 之后, 我们计划将 MCC 引入深度学习, 也可以将 MCC 与深度学习, 度量学习三者结合起来, 期望得到效果更好的鲁棒度量学习方法或者深度学习框架.

参考文献

1 周志华. 机器学习. 北京: 清华大学出版社, 2016.
2 李航. 统计学习方法. 北京: 清华大学出版社, 2012.
3 谢剑斌, 兴军亮, 张立宁, 等. 视觉机器学习 20 讲. 北京: 清华大学出版社, 2015.

4 Yang L. Distance metric learning: A comprehensive survey [Thesis]. Michigan: Michigan State University, 2006.
5 He R, Zheng WS, Hu BG. Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1561–1576. [doi: 10.1109/TPAMI.2010.220]
6 Zhao W, Chellappa R, Phillips PJ, *et al*. Face recognition: A literature survey. ACM Computing Surveys, 2003, 35(4): 399–458. [doi: 10.1145/954339]
7 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.
8 战扬, 金英, 杨丰. 基于监督的距离度量学习方法研究. 信息技术, 2011, (12): 21–23. [doi: 10.3969/j.issn.1009-2552.2011.12.006]
9 Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2003, 4: 119–155.
10 Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500): 2319–2323. [doi: 10.1126/science.290.5500.2319]
11 Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, British Columbia, Canada. 2001. 585–591.
12 Davis JV, Kulis B, Jain P, *et al*. Information-theoretic metric learning. Proceedings of the 24th International Conference on Machine Learning. Corvalis, OR, USA. 2007. 209–216.
13 Globerson A, Roweis ST. Metric learning by collapsing classes. In: Weiss Y, Sch lkopf B, Platt J, eds. Advances in Neural Information Processing Systems. Cambridge, MA, USA. MIT Press, 2006. 451–458.
14 Goldberger J, Roweis S, Hinton G, *et al*. Neighbourhood components analysis. In: Saul LK, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems. Cambridge, MA, USA. MIT Press, 2005. 513–520.
15 Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 2009, 10: 207–244.
16 Tsang IW, Cheung PM, Kwok JT. Kernel relevant component analysis for distance metric learning. Proceedings of 2005 IEEE International Joint Conference on Neural Networks. Montreal, QB, Canada. 2005. 954–959.
17 Kim TK, Kittler J. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 318–327. [doi: 10.1109/TPAMI.2005.58]
18 Jin R, Wang SJ, Zhou Y. Regularized distance metric learning: Theory and algorithm. In: Bengio Y, Schuurmans D, Lafferty JD, et al, eds. Advances in Neural Information Processing Systems. Bethesda, MD, USA. MIT Press, 2009.
19 Erdogmus D, Principe JC. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. IEEE Transactions on Signal Processing, 2002, 50(7): 1780–1786. [doi: 10.1109/TSP.2002.1011217]
20 Liu WF, Pokharel PP, Principe JC. Correntropy: Properties and applications in non-Gaussian signal processing. IEEE Transactions on Signal Processing, 2007, 55(11): 5286–5298. [doi: 10.1109/TSP.2007.896065]
21 Shalev-Shwartz S, Singer Y, Ng AY. Online and batch learning of pseudo-metrics. Proceedings of the Twenty-first International Conference on Machine Learning. Banff, Alberta, Canada. 2004. 94.
22 Bousquet O, Elisseeff A. Stability and generalization. Journal of Machine Learning Research, 2002, 2: 499–526.