

文章编号: 2095-4980(2016)03-0336-05

基于标签重构的弹载毫米波距离像识别算法

王剑桥, 李跃华, 陈建飞

(南京理工大学 电子工程与光电技术学院, 江苏 南京 210094)

摘要: 毫米波引信通过发射宽带信号获得目标的精细结构信息, 然而在非合作目标的探测与识别过程中, 由于缺乏目标的类别信息, 大量的目标样本无法得到充分的利用。针对这一问题, 将基于拉普拉斯得分(LS)的监督特征选择算法推广到半监督情况, 得到基于标签重构的拉普拉斯得分算法(LRLS), 并应用到非合作目标的识别中。LRLS的理论框架与LS相同, 并利用标签重构技术获得半监督情况下的图拉普拉斯矩阵。为了更好地描述高维目标样本的相似性, 在标签重构的过程中使用测地距离代替欧氏距离。实验结果表明, 相对于传统的特征选择算法, LRLS能够得到更好的识别效果。

关键词: 标签重构; 特征选择; 拉普拉斯得分; 半监督学习

中图分类号: TN957.52

文献标识码: A

doi: 10.11805/TKYDA201603.0336

Label reconstruction based missile-borne MMW HRRP recognition algorithm

WANG Jianqiao, LI Yuehua, CHEN Jianfei

(School of Electronic Engineering and Optoelectronic Technology, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, China)

Abstract: Millimeter wave fuze obtains the detailed structure information by transmitting a wideband signal. However, in the process of non-cooperative detection and recognition, a large number of target samples could not be fully utilized due to the lack of target class information. To this end, Laplacian Score(LS) is extended to Label Reconstruction based Laplacian Score(LRLS) and applied to the case of semi-supervised learning. Under the framework of LS, LRLS utilizes the label reconstruction technique to calculate the Laplacian matrix. In order to better describe the similarity between the high-dimensional samples, the Euclidean distance is replaced with the geodesic distance. The experimental results show that LRLS performs better than the traditional methods.

Key words: label reconstruction; feature selection; Laplacian Score; semi-supervised learning

毫米波引信通过发射大带宽信号来实现高距离分辨。随着发射信号带宽的增大, 引信能够达到分米级甚至厘米级的距离分辨能力。此时目标占据十几到几十个距离单元, 即目标的有效距离像为十几到几十维的向量。而为了得到目标完整的距离像, 通常会用一个有冗余的距离窗来截取, 使得用于目标识别的距离像不仅包含了目标信息, 而且包含了一段无用的数据, 使得距离像的维数变得更高。而距离像又受目标姿态、雷达参数和环境的影响, 使得距离像呈现出高度的非线性特性。因此, 距离像识别本质上是高维非线性数据的分类问题。

距离像的每一个维度可以看作是目标的一个特征, 维数越高, 描述目标的特征也就越多。然而, Hughes^[1]指出, 当样本数量有限时, 随着特征数的增加, 分类精确度先增加后减小。也就是说, 存在一个最优的特征数, 使得分类精确度可以达到最大。若特征数过多, 反而会导致分类精确度下降。

特征选择从原始的特征集合中选择有利于分类的特征子集。通过特征选择, 无关和冗余的特征被删除, 数据的维数得到降低, 同时描述数据的模型会变得更加精确^[2]。一个好的特征子集能使新的样本满足类间距离大而类内方差小, 即不同类间的特征远离, 而同类内的特征致密聚集。拉普拉斯得分(LS)^[3]是近些年提出的一种非线性特征选择算法, 它根据特征的局部保持能力来选择特征。LS 是一种监督或非监督的特征选择算法。然而, 引信

收稿日期: 2014-11-13; 修回日期: 2015-09-21

基金项目: 国家部委基金资助项目(51305050102)

通常要对很多非合作目标进行识别,目标样本的类别信息难以全部获得。本文介绍一种基于标签重构的拉普拉斯得分算法(LRLS),将LS推广到半监督应用场合,使之适用于非合作目标的探测与识别。

1 拉普拉斯得分

假设有 n 个样本 $X=[x_1, x_2, \dots, x_n]$, x_i 的第 r 个特征用 f_{ri} 表示。为了刻画局部几何结构,LS 构造了一个近邻图 G 。在非监督场合,如果 2 个样本 x_i 和 x_j 是近邻,则样本 x_i 和 x_j 在图 G 中有边连接。在监督场合,标签信息是已知的,则在图 G 中属于同一类的样本有边连接。然后,计算权重矩阵 S 。如果样本 x_i 和 x_j 是近邻,则有 $S_{ij} = \exp(-\|x_i - x_j\|^2 / t)$; 否则,有 $S_{ij} = 0$ 。对于第 r 个特征,定义:

$$f_r = [f_{r1}, f_{r2}, \dots, f_{rn}]^T, \quad D = \text{diag}(S1), \quad L = D - S \quad (1)$$

式中: $1 = [1, \dots, 1]^T$; L 为图拉普拉斯矩阵。令

$$\tilde{f}_r = f_r - \frac{f_r^T D 1}{1^T D 1} 1 \quad (2)$$

则第 r 个特征的拉普拉斯得分可以通过式(3)得到:

$$L_r = \frac{\tilde{f}_r^T L \tilde{f}_r}{\tilde{f}_r^T D \tilde{f}_r} \quad (3)$$

2 基于标签重构的拉普拉斯得分

从LS的推导过程可以看出,特征的拉普拉斯得分与矩阵 D 和 L 有关,而矩阵 D 和 L 又可通过权重矩阵 S 得到。换句话说,权重矩阵 S 唯一决定了拉普拉斯得分。在监督和非监督的情况下,权重矩阵都可以容易得到。可是,在实际中,获得训练集的全部标签有时是非常困难的,最常见的情况是只有少量的标签已知的样本和大量标签未知的样本。因此,不能使用监督的方法。但是如果使用非监督的方法,已知的标签将会被丢弃。

将样本矩阵 X 重新写为 $X = [X^L, X^U]$, 其中 X^L 表示前 l 个标签已知的样本, X^U 表示剩下的 $n-l$ 个标签未知的样本。首先使用标签重构技术^[4-5]来估计 X^U 的标签。标签重构的基本假设是如果 2 个样本是相似的,那么它们的标签也应该是相似的。在模式识别领域中,2 个样本的相似性通常用它们之间的距离表示。受稀疏重现(Sparse Represent)^[6-7]技术的启发,利用所有的标签已知样本来重构标签未知的样本。对于标签未知样本 x_i^U , 使用重构权重 $w_{Ui,j}$ 来重构 x_i^U , 则 x_i^U 可表示为:

$$x_i^U = w_{Ui,1} x_1^L + \dots + w_{Ui,l} x_l^L + \varepsilon_{Ui} \quad (4)$$

为了得到最优权重,引入约束条件:

$$\sum_{j=1}^l w_{Ui,j} = 1 \quad (5)$$

则重构误差可以表示为:

$$\varepsilon_{Ui} = x_i^U - (w_{Ui,1} x_1^L + \dots + w_{Ui,l} x_l^L) = \sum_{j=1}^l w_{Ui,j} (x_i^U - x_j^L) \quad (6)$$

通过最小化式(6)可以得到最优权重。从式(6)可以看出,最优权重取决于标签未知样本与每一个标签已知样本之间的相似度。如果用欧氏距离来度量相似度,则重构误差可以写为:

$$\varepsilon_{Ui} = \sum_{j=1}^l w_{Ui,j} d_E(x_i^U, x_j^L) \quad (7)$$

式中 $d_E(x_i^U, x_j^L)$ 表示样本 x_i^U 与样本 x_j^L 之间的欧氏距离。在线性特征空间中,欧氏距离能够很好地描述样本之间的相似性。然而,现实中的很多高维观测数据呈现出非线性特性。测地距离考虑了样本的结构信息,因此在非线性空间中,测地距离能够更好地反映出 2 个样本之间的相似性。这里,使用测地距离代替欧氏距离来度量非线性特征空间中的样本相似性。测地距离的计算过程如下:

对于每个 x_i , 寻找与其欧氏距离最小的 k 个点为近邻,构造邻接图 G 。当 x_j 是 x_i 的近邻时,图 G 中有边 $x_i x_j$ 。并设测地距离 $d_G(x_i, x_j) = d_E(x_i, x_j)$, 否则设 $d_G(x_i, x_j) = \infty$ 。对于 $l=1, 2, \dots, N$, 最终的测地距离可以表示为:

$$d_G(x_i, x_j) = \min\{d_G(x_i, x_j), d_G(x_i, x_l) + d_G(x_l, x_j)\} \quad (8)$$

令样本 x_i^U 与样本 x_j^L 之间的测地距离为 $d_G(x_i^U, x_j^L)$, 则重构误差可以表示为:

$$\varepsilon_{Ui} = \sum_{j=1}^l w_{Ui,j} d_G(x_i^U, x_j^L) \quad (9)$$

在流形上, 标签重构的假设可以描述为: 如果 2 个样本间的测地距离越小, 这 2 个样本属于同一个类的可能性就越大。假设样本矩阵 $\mathbf{X}$ 中共有 c 个类, 则标签未知样本 x_{Ui} 属于第 p 个类的概率为:

$$L_{Ui,p} = \sum_{j=1}^l w_{Ui,j} \delta(La(x_j^L) - p) \quad (10)$$

式中: $La(x_j^L)$ 表示样本 x_j^L 的标签; $\delta(\cdot)$ 表示单位脉冲函数。定义标签矩阵 $\mathbf{L} \in \mathbf{R}^{n \times c}$, 其中元素 L_{ij} 表示样本 x_i 属于第 j 个类的概率。对于标签已知的样本 x_i^L , 如果 x_i 属于第 p 类, 则有:

$$L_{ij} = \begin{cases} 1, & j = p \\ 0, & j \neq p \end{cases} \quad (11)$$

对于标签未知的样本 x_i^U , 标签矩阵中对应的行可以通过式(10)计算得到。此时, 连通图 G 中的边矩阵 $\mathbf{G}_E$ 可以表示为:

$$\mathbf{G}_E = \mathbf{L}\mathbf{L}^T \quad (12)$$

在监督的情况下, 通过式(12)得到的 $\mathbf{G}_E$ 监督拉普拉斯得分得到的 $\mathbf{G}_E$ 相同; 在半监督的情况下, $\mathbf{G}_{Eij}$ 指出了样本 x_i 和样本 x_j 之间存在边的概率。此时, 样本 x_i 和样本 x_j 之间的权重 S_{ij} 可以表示为:

$$S_{ij} = G_{Eij} \exp\left(-\frac{d_G(x_i, x_j)}{t}\right) \quad (13)$$

式中 t 为距离权重系数, 用于调整距离对权重矩阵 $\mathbf{S}$ 的影响。得到权重矩阵 $\mathbf{S}$ 后, 可以得到矩阵 $\mathbf{D}$ 和图拉普拉斯矩阵 $\mathbf{L}$ 。代入式(3)即可得到对应的特征得分。

3 实验结果与分析

实验数据采用 3 种不同地面目标的仿真距离像。3 种地面目标分别为 T32 重型坦克、KV-1 重型坦克、wz551 轮式装甲车的简化模型, 其散射点分布简化模型见图 1, 图中黑色圆点表示强散射点。

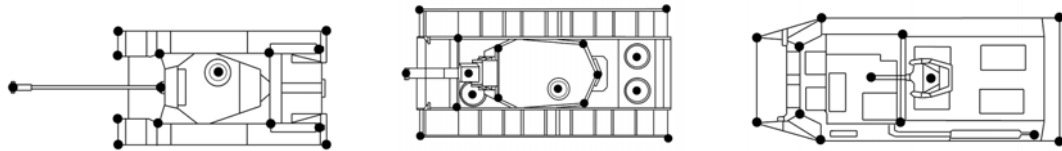


Fig.1 Simplified scattering centre models of the three kinds of simulation ground targets

图 1 3 种地面目标散射点分布的简化模型

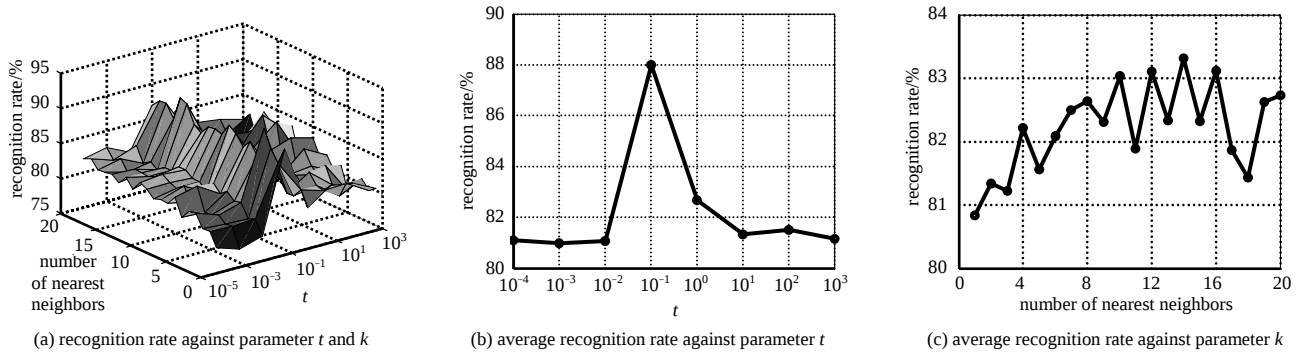


Fig.2 Recognition rate of LRLS against parameter t and k

图 2 不同参数选择下的 LRLS 性能比较

仿真参数为: 载波 35.07 GHz, 带宽 400 MHz, 俯角 30° , 方位角从 0° 到 179.5° , 每隔 0.5° 取一幅距离像, 每幅距离像包含 256 个特征。杂波采用相同参数的毫米波探测器实测的水泥地面杂波。分别从每类目标的 360 幅距离像中随机选择出 180 个距离像作为训练样本, 并从其中随机选择 60 个样本作为标签已知样本, 剩下的 120 个样本作为标签未知样本。未被选中的 180 个样本作为测试样本。分别从 256 个特征中选出 10, 20, ..., 250 个特征进行分类, 分类器选用最近邻分类器。每个实验重复 50 次, 取平均识别率作为评估算法性能的标准。对比算法

为方差得分(Variance Score, VS)^[8]、LS 和约束拉普拉斯得分(Constrained Laplacian Score, CLS)^[9]。为了利用类别信息, VS 和 LS 使用监督版本,即将标签已知样本作为训练集而忽略标签未知的样本。由于 CLS 使用的是对约束,首先将标签信息转化为对约束信息,再执行 CLS 算法。

首先进行 LRLS 的参数选择实验。LRLS 有 2 个输入参数:近邻数 k 和参数 t 。图 2(a)显示了不同地杂波环境下的平均识别率与近邻数和参数 t 的关系。近邻数的取值范围为 1,2,...,20, 参数 t 的取值范围为 $10^{-4}, 10^{-3}, \dots, 10^3$ 。图 2(b)和图 2(c)分别对近邻数和参数 t 进行了统计平均,得到平均识别率与参数 t 和近邻数的关系。从图 2(b)中可以看出,参数 t 的变化对 LRLS 算法的性能影响较大,当 t 取 0.1 时, LRLS 的平均识别率最高。图 2(c)显示了平均识别率与近邻数的关系,从中可以看出,近邻数对算法性能的影响不如参数 t 那么大。在下面的实验中, LRLS 的参数选择为: $k=14, t=0.1$ 。

下面进行不同算法的性能对比实验。VS,LS,CLS 与 LRLS 在不同特征数下的识别率见图 3。表 1 详细地比较了 4 种算法的平均识别率、标准差、最大识别率和最优特征数等 4 个性能指标。

从图 3 中可以看出, CLS 和 LRLS 的识别率在大多数情况下比 VS 和 LS 要高得多,而 VS 和 LS 的性能相似。随着特征数增加, LRLS 最先达到最大识别率,其次是 LS,CLS 和 VS。LRLS 和 CLS 的识别率均随着特征数的增加先增大而后减小,而 VS 和 LS 的识别率在特征数较少时没有明显变化,当特征数超过 130 时识别率才开始上升。下面对不同算法的性能进行定量分析。从表 1 中可以看出, LRLS 的平均识别率比 VS,LS 和 CLS 分别高出 12.10%,11.26%和 3.32%,而最大识别率则分别高出 8.95%,8.70%和 3.04%。这说明在该实验条件下, LRLS 比 VS,LS 和 CLS 能取得更好的性能。同时, LRLS 的最优特征数要比其他 3 种算法小得多,这也减少了后续分类算法的计算量。然而, LRLS 在训练阶段需要计算测地距离,计算量比 VS,LS 和 CLS 要大得多。因此, LRLS 并不适合实时分类的应用场合。

4 结论

本文利用标签重构技术将拉普拉斯得分推广到半监督应用场合,以适应弹载非合作目标探测与识别的需求。LRLS 利用 LS 的理论框架,同时利用标签重构技术获得半监督情况下的图拉普拉斯矩阵。为了更好地描述非线性特征空间中的样本,使用测地距离代替欧氏距离来度量样本之间的相似性。仿真实验表明,相比于 VS,LS 和 CLS, LRLS 在平均识别率和最大识别率上均有较大的优势,同时具有更低的最优特征数。这意味着在接下来的分类过程中需要更少的计算量。然而, LRLS 在训练过程中需要计算测地距离,计算量相比于其他 3 种算法要大得多。如何更快地获得测地距离将是进一步研究的重点。

参考文献:

- [1] HUGHES G. On the mean accuracy of statistical pattern recognizers[J]. IEEE Transactions on Information Theory, 1968, 14(1):55-63.
- [2] LIU H, MOTODA H. Feature Selection for Knowledge Discovery and Data Mining[M]. Boston: Kluwer Academic Publishers, 1998.
- [3] HE X, CAI D, NIYOGI P. Laplacian score for feature selection[C]// Advances in Neural Information Processing Systems 18. [S.l.]: MIT Press, 2005:507-514.
- [4] WRIGHT J, YANG A Y, GANESH A, et al. Robust face recognition via sparse representation[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2009, 31(2):210-227.

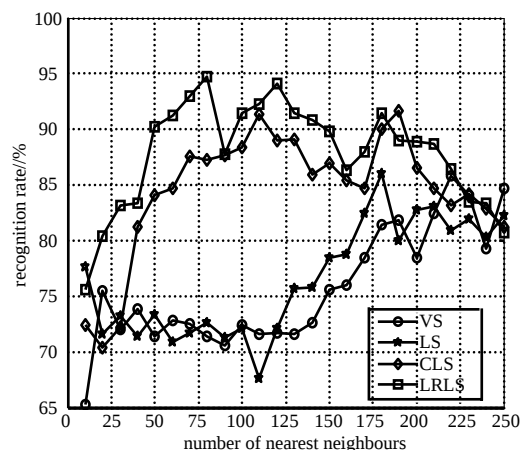


Fig.3 Recognition rate of VS, LS, CLS, LRLS against the number of nearest neighbors

图3 不同特征数下 VS,LS,CLS 与 LRLS 的识别率比较

表1 VS,LS,CLS 与 LRLS 的性能比较
Table1 Performance comparison of VS,LS,CLS and LRLS

	VS	LS	CLS	LRLS
average recognition rate/%	75.73	76.57	84.51	87.83
standard deviation	5.30	5.05	5.58	4.78
best recognition rate/%	85.79	86.04	91.70	94.74
optimal number of features	220	180	190	80