DOI：10.13875/j.issn.1674-0637.2019-04-0336-09

基于 t-SNE 算法的 ABPSK 信号个体识别

姚舜禹[1,2,3]，王雪[1,2,3]，邹德财[1,2,3]，李优阳[1,2,3]

（1. 中国科学院 国家授时中心，西安 710600；
2. 中国科学院大学，北京 100049；
3. 中国科学院 精密导航定位与定时技术重点实验室，西安 710600）

摘要：同一通信体系下的 ABPSK（aeronautical binary phase shift keying）信号都有着相同的前导码，传统信号识别方法无法通过相同的前导码部分准确地识别出信号源，且常用信号特征属于高维特征，非常容易引发维度灾难。采用前导码相同的 ABPSK 实际信号采集数据的前导码，使用 t-SNE 算法对实际采集信号的前导码以及双谱进行降维，并且把降维后信号单一特征输入分类器中，不仅有效地利用了信号数据的流形信息，而且显著提升了基于信号单一特征进行信号个体识别的正确率。

关键词：t-SNE 算法流形降维；信号个体识别；维度灾难；信号细微特征

ABPSK signal individual recognition based on t-SNE algorithm

YAO Shun-yu[1,2,3], WANG Xue[1,2,3], ZOU De-cai[1,2,3], LI You-yang[1,2,3]

(1. National Time Service Center, Chinese Academy of Sciences, Xi'an 710600, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Key Laboratory of Precise Navigation and Timing Technology, Chinese Academy of Sciences, Xi'an 710600, China)

Abstract: The ABPSK (aeronautical binary phase shift keying) signals in the same communication systems have the same preamble. The conventional signal recognition methods cannot identify the signal source by the same signal preamble portion accurately. The common signal features are high-dimensional features, and lead to curse of dimensionality easily. Based on the collected ABPSK data with the same preamble, this study used the t-SNE algorithm to reduce the dimension of the signal's preamble and the bispectrum, and the single feature after dimension reduction was also inputted into the classifier, thus it not only effectively used the manifold information of signal data, but also significantly improved the correct rate of signal individual identification by single signal feature.

Key words: t-SNE algorithm of manifold dimension reduction; signal individual recognition; curse of dimensionality; subtle feature

收稿日期：2019-04-19；接受日期：2019-06-05
基金项目：中国科学院青年创新促进会人才资助项目
作者简介：姚舜禹，男，硕士，主要从事空间信号感知研究。

0　引言

信号个体识别是指从采集到将信号源识别出来的过程，识别出信号源个体在电子对抗中有着重要的意义，掌握对方雷达的工作参数等于掌握了对抗的主动权。由于同一通信体系下的 ABPSK（aeronautical binary phase shift keying）信号都有着相同的前导码，属于在相同工作模式下通信辐射源个体识别的问题，在去除掉能量特性之后就只能依靠前导码之间细微特征差异来区分信号源，一般引起差别的原因有信号源内部元器件不稳定性，性能参数具有的非线性。针对常见的细微特征提取方式有 J. Morlet[1]在 1974 年提出的小波变换，黄锷[2]在 1998 年提出的 EMD（empirical mode decomposition）分解等。经过多年的发展，这些提取特征方法衍生出许多变种，这些特征又可以称为指纹特征[3]。然而此类特征在实际应用中一般无法达到理论性能：① 实际应用中许多信号源的机器型号和工作模式完全一样，通过调制参数的差异无法识别出信号源个体；② 样本数量少，通常无法充分地识别出信号源的指纹特征；③ 部分特征对噪声和干扰比较敏感，影响识别个体的能力；④ 高阶谱特征一般处于高维特征空间，周围环境的干扰具有鲁棒性，但是由于处在高维空间，容易引起维度灾难[4]，导致分类识别性能下降；⑤ 一般的线性降维方式如 PCA（principal component analysis）等一般无法找出样本之间的非线性拓扑结构，从而无法很好地寻找到信号源发射的信号样本之间的关系；⑥ 一般的识别都会加入其他特征进行特征融合，这些特征对分类的正确率贡献极大，而单一特征分类的正确率一般比较低。

鉴于以上一些原因，本文使用双谱变换来提取出信号的细微特征。使用 t 分布随机近邻嵌入算法（t-distributed stochastic neighbor embedding，t-SNE）算法对信号特征进行降维，降低双谱变换结果的维度，缓解因特征维度过高带来的维度灾难问题。由于 t-SNE 一般采用 SGD（stochastic gradient descent）作为优化器，收敛速度慢于 Adam（adaptive moment estimation），收敛函数值一般也大于 Adam，故而引入 Adam 作为 t-SNE 算法的优化器。在 SVM（support vector machine）使用 VC 维（Vapnik-Chervonenkis dimension）较高的核函数进行分类时，使用 t-SNE 实现降维后的特征分类有比较好的效果。

1　信号特征说明及双谱变换介绍

本节介绍 ABPSK 信号和双谱变换。

1.1　ABPSK 信号体制介绍

ABPSK 信号常应用于 INMARSAT 移动通信中，ABPSK 是 DBPSK（differentially conherent BPSK，二相差分键控）的一种特殊形式，是对普通的 BPSK 改进的一种调制方式，借鉴了 QPSK（quadrature phase shift keying）的原理，利用特殊的二进制差分编码和正交调制技术，将 DBPSK 中 180°相位变化转化为 90°相位变化的 ABPSK[5]。对于[−1，1，−1，1，−1，1，1，−1，−1，−1]的码片序列的输出波形如图 1 所示。

前导码一般可以分为同步码和检测码两个部分[6]，本文使用的信号采样率为 6 kHz 的实际采集某个系统的 ABPSK 信号，信号格式为 40 ms 空白保护+150 bit（250 ms）CW+74 bit 0101（123.3 ms）+32 bit UW 独特码+信息码的格式，经过对齐取 I、Q 两路的前 2 460 个点作为 I、Q 两路的前导码特征。

本文中所有信号的前导码部分都是相同的，对齐之后的前导码使用常规的方法（如小波变换、希尔伯特黄变换及分形维数等）都不能很好地识别出信号个体，这就要求找寻一种能够揭示发射机非高斯、非线性的有色噪声的方式。

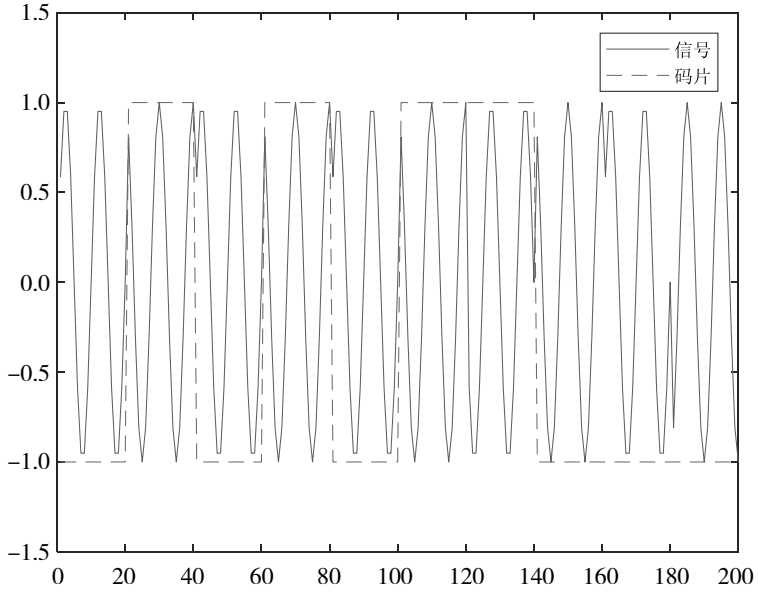

图 1 ABPSK 信号已调波形图

1.2 双谱变换

发射机噪声是雷达无意调制的产物（包括互调频率、谐波频率及一些交叉调制、寄生调制等），由于不同发射装置采用的电路和器件不同造成的发射机噪声不同，这些不规则的非线性、非高斯的有色噪声便是信号源的细微特征[7]，一般的一阶和二阶特征无法揭露这些有色噪声，通常采用高阶累积量的方式来识别这些有色噪声。双谱变换则是最常用的信号细微特征提取算法。

对于均值为零的连续信号 $x(t)$，三阶相关函数 $C_{3x}(\tau_1,\tau_2)$ 定义如下：

$$C_{3x}(\tau_1,\tau_2)=\int_{-\infty}^{+\infty} x^*(t)x(t+\tau_1)x(t+\tau_2)\mathrm{d}t\ , \tag{1}$$

式（1）中，τ_1 和 τ_2 为自相关操作滑动窗口的时间间隔。

连续信号 $x(t)$ 的双谱表示为

$$B(f_1,f_2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} C_{3x}(\tau_1,\tau_2)\exp(-\mathrm{j}(f_1\tau_1+f_2\tau_2))\mathrm{d}\tau_1\mathrm{d}\tau_2\ 。 \tag{2}$$

对于一个离散时间能量有限的确定信号，将双谱定义为

$$B(f_1,f_2)=X(f_1)X(f_2)X^*(f_1+f_2)\ , \tag{3}$$

式（3）中，$X(f)$ 为信号序列 $x(t)$ 的离散傅里叶变换，$B(f_1,f_2)$ 为双谱变换的结果。

1.3 核函数

根据统计机器学习的观点，存在一个非线性变换，使得在低维空间中不可分的样本可以通过某种非线性变换映射到另一个空间，样本在这个空间中是线性可分的[8]。

假设在原空间中有一组样本 $x_1,x_2,\cdots,x_n,x_i$，通过一个非线性映射投影到一个新的空间形成的点 $\varphi(x_i)$，这个空间是一个希尔伯特空间，两个样本在这个空间的内积形成的函数称为核函数[9]，生成的空间称为再生核希尔伯特空间，表示为

$$K(x_i,x_j)=<\varphi(x_i),\varphi(x_j)>\ 。 \tag{4}$$

根据 Mercer 定理，任何半正定的函数都可以作为核函数[10]，常用的核函数有：

① 线性核。线性核是最简单的核函数，核函数的数学公式如下：

$$K(x,y)=x^{\mathrm{T}}y\ 。\tag{5}$$

② 多项式核。多项式核是一种非标准核函数，它非常适合于正交归一化后的数据，其具体形式如下：

$$K(x,y)=(ax^{\mathrm{T}}y+c)^{d}\ 。\tag{6}$$

③ RBF 核函数。RBF 核函数的性能对参数十分敏感，以至于有一大把的文献专门对这种核函数展开研究，其数学形式如下：

$$K(x,y)=\exp(-\frac{\|x-y\|_2^2}{2\sigma^2})\ 。\tag{7}$$

核函数有以下特点：① 核函数的引入缓解了维度灾难，大大减小了计算量，可处理高维输入。② 不需要知道非线性映射 $\varphi(x_i)$ 函数形式和具体参数。③ 核函数的参数和形式的变化会隐式地改变从样本空间到再生核希尔伯特空间的映射，进而对再生核希尔伯特空间的性质产生影响，最终改变各种核函数方法的性能。

核函数方法可以和不同的算法进行结合，形成多种基于核函数技术的方法，且这两部分的设计并不冲突，并可以为不同的应用选择各种不同的核函数和算法。

2　t-SNE 算法

为了解决维度灾难问题，提高分类器识别正确率，在此引入 t-SNE 算法[11]对信号特征进行降维。

2.1　模型介绍

t-SNE 算法是 L. V. D. Maaten[12]在 2008 年提出的，其前身是 G. Hinton[13]在 2002 年提出的 SNE 算法，因为 SNE 算法损失函数为 KL 散度并且衡量映射近邻的方式是高斯分布，会造成映射概率的非对称性和拥堵问题，所以引入 t 分布来解决拥堵问题。

假设数据集 X，它共有 N 个数据点。每一个数据点 x_i 的维度为 D，我们希望降低为 d 维。在一般可视化的条件下，d 的取值为 2，即在平面上表示出所有数据。t-SNE 通过原始数据之间的欧氏距离转化为概率来表征相似性：

$$p_{j|i}=\frac{\exp(-\left\|x_i-x_j\right\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\left\|x_i-x_k\right\|^2/2\sigma_i^2)}\ 。\tag{8}$$

如果以数据点 x_i 为中心的高斯分布所占的概率密度 $p_{j|i}$ 为标准选择近邻，那么 $p_{j|i}$ 就代表 x_i 将选择 x_j 作为它的近邻。对于相近的数据点，条件概率 $p_{j|i}$ 是相对较高的，然而对于分离的数据点，几乎是无穷小量（高斯分布的方差 σ_i^2 由预先设置的参数困惑度决定）。

因为 KL 散度是非对称度量，所以 p_{ij} 的表达式如下：

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}\ 。\tag{9}$$

在低维空间下，我们使用更加偏重长尾的 t 分布的方式来将距离转换为概率分布，使得高维度下中低等的距离在映射后能够有一个较大的距离。q_{ij} 为 y_i 和 y_j 在低维数据点映射的相似概率，使用 t 分布的 q_{ij} 如下：

$$q_{ij}=\frac{(1+\left\|y_i-y_j\right\|^2)^{-1}}{\sum_{k\neq l}(1+\left\|y_i-y_j\right\|^2)^{-1}}\ 。\tag{10}$$

其损失函数如下：

$$C = KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} 。 \tag{11}$$

对损失函数求梯度的结果如下：

$$\frac{\partial C}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \left\| y_i - y_j \right\|^2)^{-1} 。 \tag{12}$$

2.2 Adam 优化器

t-SNE 一般采用的优化器为 SGD，由于随机梯度下降收敛速度较慢，收敛过后的损失函数值一般也比较大，本文采用 Adam 算法作为优化器。Adam 优化算法基本上就是将 Momentum 和 RMSprop 结合在一起，Adam 算法如下[14]。

$$\begin{cases} v_{\mathrm{d}y_i} = \beta_1 v_{\mathrm{d}y_i} + (1-\beta_1)\dfrac{\partial C}{\partial y_i} \\ v_{\mathrm{d}y_i}^{\mathrm{corr}} = \dfrac{v_{\mathrm{d}y_i}}{1-(\beta_1)^t} \\ s_{\mathrm{d}y_i} = \beta_2 s_{\mathrm{d}y_i} + (1-\beta_2)(\dfrac{\partial C}{\partial y_i})^2 \\ s_{\mathrm{d}y_i}^{\mathrm{corr}} = \dfrac{s_{\mathrm{d}y_i}}{1-(\beta_1)^t} \\ y_i = y_i - \alpha \dfrac{v_{\mathrm{d}y_i}^{\mathrm{corr}}}{\sqrt{s_{\mathrm{d}y_i}^{\mathrm{corr}}} + \varepsilon} \end{cases} 。 \tag{13}$$

式（13）中，Adam 参数说明如下：y_i 为第 i 个样本在低维空间的映射坐标向量，为最终求解的目标；C 为损失函数；$v_{\mathrm{d}y_i}$ 为带动量的梯度方向，即速度方向。$v_{\mathrm{d}y_i}^{\mathrm{corr}}$ 为经过修正的 $v_{\mathrm{d}y_i}$；$s_{\mathrm{d}y_i}$ 为微分平方的加权平均数；$s_{\mathrm{d}y_i}^{\mathrm{corr}}$ 为经过修正的 $s_{\mathrm{d}y_i}$。

Adam 超参数说明如下：β_1 和 β_2 是控制指数加权平均的超参数；α 是学习率；ε 是一个非常小的正数，作用是为了避免分母为 0。

2.3 模型优化

对损失函数公式（7）的优化采用 Adam 方式进行优化，算法详细过程如下：① 设置参数困惑度以及迭代次数 T；② 设置 Adam 优化算法的超参数；③ 计算在给定困惑度条件下的条件概率 $p_{j|i}$；④ 计算 p_{ij}，为了计算方便，当 $i = j$ 时，p_{ij} 取 0.000 000 01；⑤ 用方差很小的正态分布初始化所有的 y_i；⑥ 重复迭代计算低维度下的 q_{ij} 并使用 Adam 算法更新所有的 y_i；⑦ 重复迭代超过最大迭代次数 T 后结束。

上述为本文所用 Adam 算优化器进行优化的详细步骤，经过上述步骤所得到的 y_i 为第 i 个样本在低维空间的映射坐标向量，所有的 y_i 构成了低维空间中所有的样本集合。

3 实验结果及分析

本文采用前导码相同的 ABPSK 实际信号采集数据，同一个个体发出的信号没有固定频率，采样带宽 1 kHz，采样时长约 24 h，采样率为 6 kHz，经过人工标注标签，共 10 类 235 个样本，为同一调制方式的不同个体，信噪比在 5~20 dB 之间。

3.1　数据的预处理

本文采用 ABPSK 基带信号进行识别，数据及特征提取的预处理步骤如下：① 通过所给出的标签对信号在时域上进行提取；② 信号的时间对齐；③ 分离 I、Q 两路信号；④ 截取前导码的长度为 2 640 个点；⑤ 信号的时间对齐；⑥ 对前导码进行能量归一化；⑦ 对前导码进行双谱变换。

经过数据预处理，得到 ABPSK 信号前导码特征以及双谱变换后的双谱特征。

3.2　信号前导码及双谱幅度谱 t-SNE 降维结果

分别将已经去除掉能量特性的 I、Q 两路的前导码及双谱幅度图降至 2 维，降维之后的结果如图 2 至图 4 所示，其中的数字为信号的类别编号。

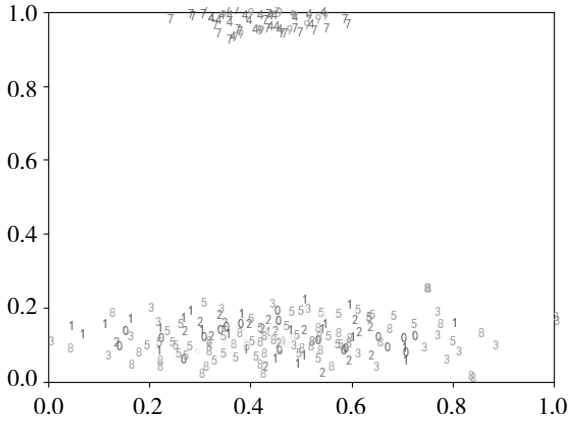

图 2　I 路前导码降至二维的结果

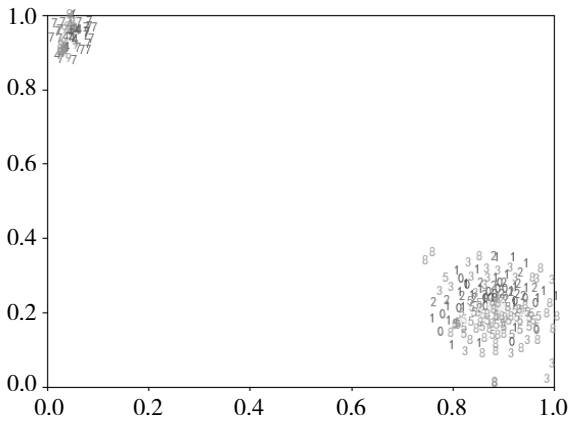

图 3　Q 路前导码降至二维的结果

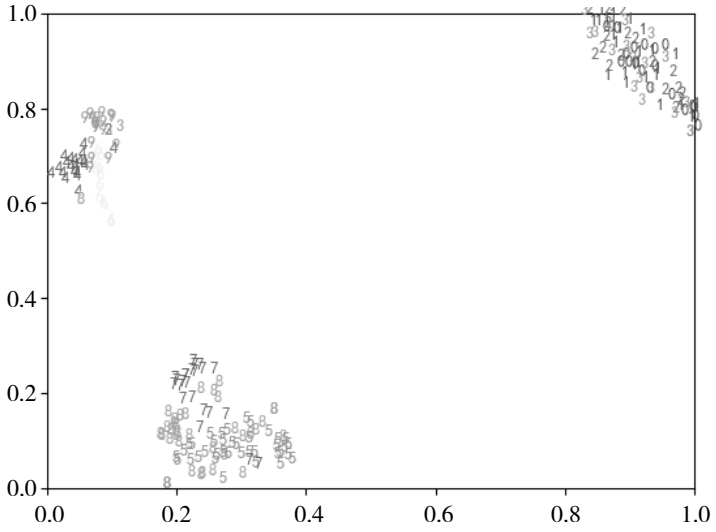

图 4　双谱幅度图降至二维结果

在降至 2 维后，能够明显看出信号的聚类情况，样本在 2 维空间之间的远近反映了样本在原始空间之间的距离的远近，显示了 t-SNE 在信号数据可视化方面性能强大。从图 2 至图 4 中可以看出前导码降维过后明显分成两组，且组内样本不易区分，经过双谱变换和 t-SNE 降维后，样本分成 3 组，且组内样本区比前导码更易区分。

3.3　能量对分类结果的影响

对 235 个样本打乱顺序进行随机划分，165 个训练样本，70 个测试样本，分别用带有未能量归一化

的特征和能量归一化过后的特征，正确率为 10 次随机选择测试样本正确率的平均值，使用 SVM 进行分类，试验使用了 3 种 Mercer 核函数进行试验，特征包括 I、Q 两路前导码及双谱幅度图，测试结果示于表 1。

表 1 有无能量特性对于分类正确率的影响

信号特征及 SVM 使用的核函数	10 次平均正确率/%								
	I 路前导码			Q 路前导码			双谱幅度图		
	线性核	多项式核	RBF 核	线性核	多项式核	RBF 核	线性核	多项式核	RBF 核
能量归一化使用原始数据分类	19.71	19.14	16.00	20.71	20.71	17.86	60.00	43.00	14.57
未经过能量归一化原始数据分类	53.71	42.71	14.57	57.86	40.00	18.29	67.29	44.14	17.57

从表 1 可以看出，能量对分类的正确率影响特别大，能量归一化过后分类难度明显变大，传统识别方法都会把信号的参数（如中心频率、码频率等）与未能量归一化的特征进行特征融合，分类正确率一般都很高。

3.4 维度对分类结果的性能影响

依旧随机分配 165 个训练样本和 70 个测试样本，使用 t-SNE 降维方式进行降维，降到不同的维度试图寻找出最佳的维度，正确率是 10 次正确率的平均值，示于表 2。

表 2 能量归一化特征降维后正确率提升对比

信号特征及 SVM 使用的核函数	10 次平均正确率/%								
	I 路前导码			Q 路前导码			双谱幅度图		
	线性核	多项式核	RBF 核	线性核	多项式核	RBF 核	线性核	多项式核	RBF 核
能量归一化原始特征数据分类	19.71	19.14	16.00	20.71	20.71	17.86	60.00	43.00	14.57
能量归一化特征降维后分类	31.14	35.57	37.86	34.43	44.86	45.14	65.71	64.43	65.43
最佳降维维度	200 维	3 维	5 维	200 维	3 维	4 维	200 维	5 维	6 维
正确率提升百分比	11.34	16.43	21.86	13.72	24.15	27.28	5.71	21.34	50.86

从表 2 中可以看出，降维在 SVM 核函数 VC 维较高的情况下显著提升了分类的正确率，缓解了因维度过高带来的过拟合问题。在 SVM 使用线性核分类时，I、Q 两路前导码和双谱幅度图降维后分类正确率分别比原始信号提高了 11.34%、13.72%和 5.71%，3 种特征在 200 维时分类平均正确率最高。在 SVM 使用多项式核进行分类时，双谱幅度图降维在 5 维时正确率最高，比原始信号提高了 21.34%，Q 两路前导码都降维在 3 维时正确率最高，比原始信号分别提高了 16.43%和 24.15%。在 SVM 使用 RBF 核进行分类时，双谱幅度图降维在 6 维时分类正确率最高，比原始数据正确率提高了 50.86%，I、Q 两路前导码降维最佳维度分别为 5 维和 4 维，正确率比原始信号分别提高了 21.86% 和 27.28%。

平均正确率随着维度变化的趋势如图 5 至图 7 所示。

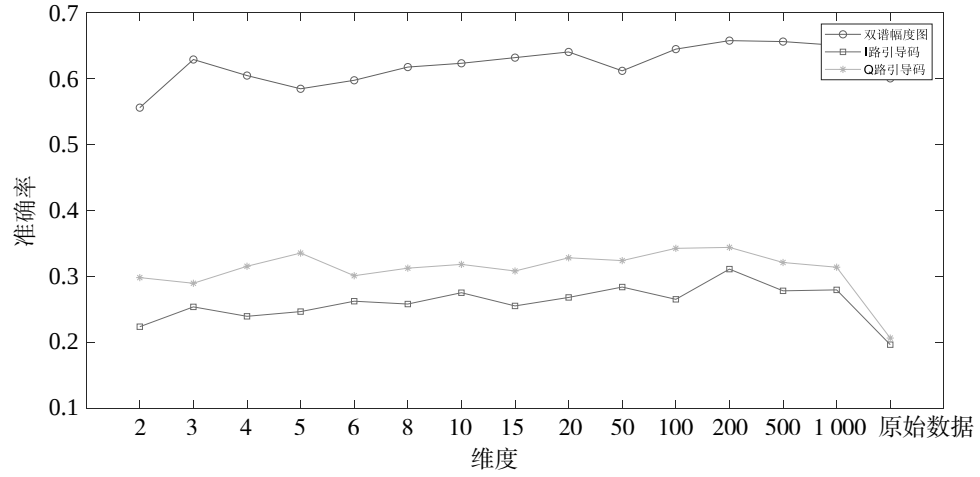

图 5　3 种特征在各个维度下线性核分类正确率的平均值

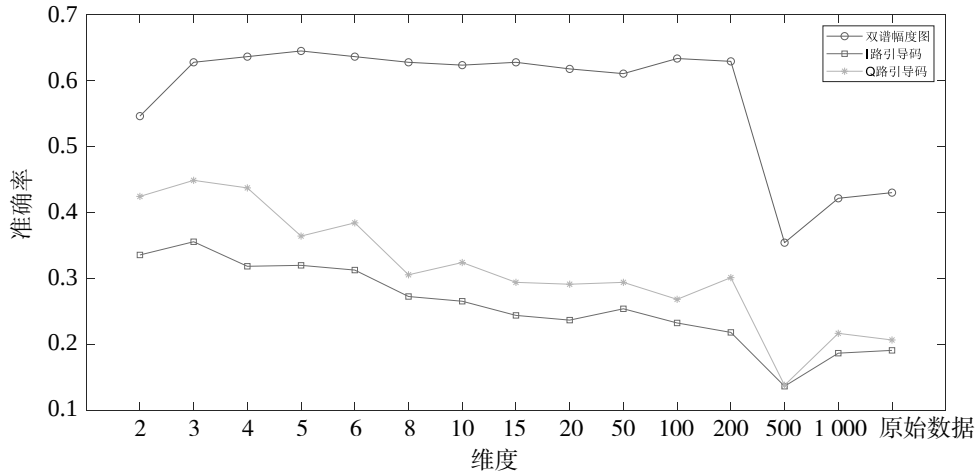

图 6　3 种特征在各个维度下多项式核分类正确率的平均值

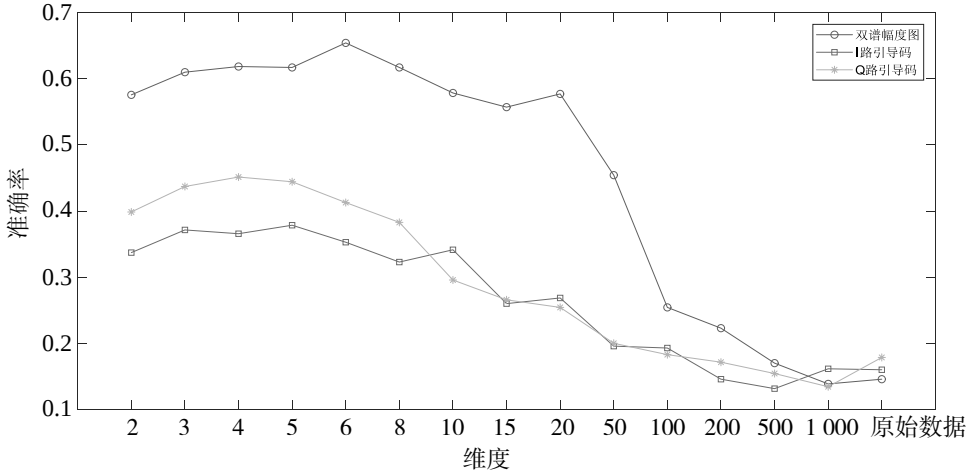

图 7　3 种特征在各个维度下 RBF 核分类正确率的平均值

从图 5 至图 7 可以看出，双谱幅度图明显比原始信号的前导码分类效果更好。从图 5 可以看出在 SVM 使用线性核函数分类时维度对分类的正确率趋势较为缓和，在前 200 维随着维度上升正确率缓慢提

升，所有特征在 200 维时正确率最高。从图 6 可以看出在使用多项式核函数分类时双谱幅度图在降维小于 200 维时正确率没有明显的变化，在维度大于 200 维时有明显的正确率下降，在 5 维时正确率最高。I、Q 两路前导码降维随着维度的上升正确率逐渐的下降，都在 3 维时正确率最高。从图 7 可以看出使用 RBF 核函数分类时在双谱幅度图及原始信号前导码降维小于 6 维时正确率较平稳，超过 6 维时有明显的下降趋势，双谱幅度图在 6 维时分类正确率最高，I、Q 两路前导码最佳维度分别为 5 维和 4 维。

4 结语

本文使用 t-SNE 降维方式对 3 种特征进行降维，研究了 3 种特征随着维度的变化正确率的变化趋势，并研究了 SVM 使用常用核函数在维度变化时对正确率的影响。试验结果表明对于 VC 维比较高的 RBF 核函数和多项式核函数，在特征维度较低的情况下不容易过拟合，降维过后不会出现支持向量过多的情况，也解决了维度灾难问题。本文使用的双谱变换就是信号领域中常用人工提取特征方式，优势是可以在样本量较小的情况下训练出分类模型，不需要用使用大量的计算资源（如大量的 GPU）花费几个月时间训练出复杂模型的参数。

但使用 t-SNE 降维仍有部分不足：① 采用距离进行测算可能无法分辨某些信号；② 计算复杂度为 $o(n^2)$，若不用并行化速度会非常慢。希望在后续的研究中能够优化细微特征提取方法来提升识别的正确率。

参考文献：

[1] DAUBECHIES I. The wavelet transform, time-frequency localization and signal analysis[J]. IEEE Transactions on Information Theory, 2015, 36(5): 961–1005.

[2] CHEN Y K, FOMEL S. EMD-seislet transform [J]. 2017, 83(1): 27–32.

[3] 黄健航, 雷迎科. 基于深度学习的通信辐射源指纹特征提取算法[J]. 信号处理, 2018, 34(1): 31–38.

[4] KEOGH E, MUEEN A. Curse of dimensionality[M]. Boston: Springer, 2017: 314–315.

[5] 胡铁乔. ABPSK 原理及实现[J]. 中国民航大学学报, 1999, 17(1): 1–7.

[6] 方晗, 缪蔚, 洪志良. 基于前导码的 WLAN 802.11b 频偏估计算法[J]. 电路与系统学报, 2007, 12(3): 143–146.

[7] 陈昌孝, 何明浩, 朱元清, 等. 基于双谱分析的雷达辐射源个体特征提取[J]. 系统工程与电子技术, 2008, 23(6): 1046–1049.

[8] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[9] 张莉, 周伟达, 焦李成. 尺度核函数支撑矢量机[J]. 电子学报, 2002, 30(4): 527–529.

[10] M G. Mercer kernel-based clustering in feature space[J]. IEEE Transactions on Neural Networks, 2002, 13(3): 780–784.

[11] ZHAN W W, WANG B, XUE J, et al. Adaptive weighted t-SNE algorithm and application in dimensionality reduction of human brain network state observation matrix [J]. Application Research of Computers, 2018, 35(7): 2055–2058.

[12] MAATEN L V D, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579–2605.

[13] HINTON G, ROWEIS S. Stochastic neighbor embedding[J]. Advances in Neural Information Processing Systems, 2002, 41(4): 833–840.

[14] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. 3rd International Conference for Learning Representations, San Diego: [s.n.], 2014.