随机森林分类用于雷达信号预分选新算法研究

刘旭波[1],刘敬蜀[1],刘斌[2],秦令令[2],陈涛[2]

1. 中国人民解放军91977部队,北京 102200
2. 哈尔滨工程大学信息与通信工程学院,哈尔滨 150001

摘要　现代战争中,雷达电子战环境越来越复杂,随着雷达种类的多样化和雷达脉间调制方式的复杂化,对信号分选的识别难度也愈加增大。本研究采用随机森林算法对脉冲描述字特征进行预分选,可自适应的对特征进行选择,并实现分类。随机森林由于可以自动进行特征选择,可对不平衡的数据进行误差平衡等优点,通过多决策树表决方式,可以迅速完成对大量数据快速训练。在脉冲丢失导致的部分特征损失的情况下,仍可以维持识别准确率。通过实验证明了本方法对雷达脉冲描述字特征进行预分选的有效性。

关键词　随机森林;决策树;信号预分选;特征选择

随着现代高科技的发展,空间中的电磁信号愈发复杂多变,密集程度也逐渐增高,其主要表现为空间辐射源的数量多、密度大、信号调制复杂,且分布较广泛[1]。传统的信号分选方法由信号预分选和主分选组成,信号预分选作为信号分选的一部分,主要目的是初步实现信号去交错,降低信号的密度,便于主分选进行处理。传统预分选算法多以载频、脉宽为基础,而随着雷达不断发展,复杂电磁环境下更多使用多功能雷达辐射源替代传统常规辐射源。该类型辐射源具有参数快变、易变,调制方式不断切换的特点[2]。现如今的信号预分选方法已经无法较好地适应复杂环境,很难适应如今复杂的脉间调制方式雷达以及多功能雷达等复杂雷达信号。

针对这样的电子战环境,有学者提出通过改进预分选的算法,如对网格聚类算法或K-Means算法进行改进来提高脉冲信号预分选的准确率[3]。但载频和脉宽分开进行预分选会导致信号由于一维的聚类错误而使其他维的聚类分选发生错误。而将载频和脉宽同时进行聚类的K-Means算法,其复杂度和K簇个数无法确定。

本研究提出的方法将随机森林(random forest)的算法应用于脉冲预分选中,通过输入脉冲描述字数据集。利用随机森林算法可以快速地进行聚类,自适应的完成雷达信号预分选。随机森林是一种机器学习算法,通过利用大量决策树对结果进行预测并投票给出结果,具有较高的准确率,并且对孤立噪声不敏感,训练速度快,能够很好的适应大量高维数据集[4]。在工程化应用中,随机森林简单高效,应用广泛,并能对多个特征进行重要性评分,完成特征选择,在众多领域都取得了较好的成绩。

根据随机森林分类器的分类高准确率作为特征可分判据,对描述字特征重要性进行排序,自适应选择高重要性的特征进行预分选。同时,在训练结束后,随机森林分类器还可以对特征进行重要性评分,得分越高、

收稿日期:2018-12-23;修回日期:2019-06-04

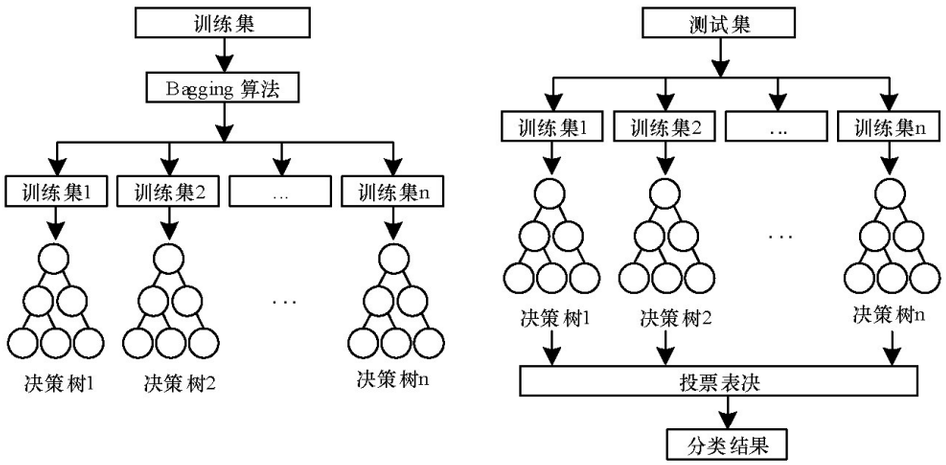
作者简介:刘旭波,工程师,研究方向为信号处理,电子信箱:liujingshux@163.com;陈涛(通信作者),教授,研究方向为电子对抗与信号检测,电子信箱:chentao@hrbeu.edu.cn

引用格式:刘旭波, 刘敬蜀, 刘斌, 等. 随机森林分类用于雷达信号预分选新算法研究[J]. 科技导报, 2019, 37(13): 93-97; doi: 10.3981/j.issn.1000-7857.2019.13.014

重要性越高、分类越准确、回归误差越小,并且可以依据重要性程度对特征进行取舍,达到降维、优化目的。

1 随机森林

随机森林由 Breiman 在 2001 年提出[5],通过随机的方式建立一个森林,森林里面由决策树组成,每一棵决策树之间没有关联。在得到森林后,当有一个新的样本输入时,森林中的每一棵决策树分别进行预测,判断样本的类别。随机森林算法主要分为3个过程:训练集生成、决策树训练形成随机森林、测试集测试。随机森林分类器的生成与测试如图1所示。



(a) 随机森林生成 (b) 随机森林测试

图1 随机森林分类器产生及测试

Fig. 1 Sketch of random forest classifier generation and test

1.1 训练集生成

随机森林构建过程中,需要对每个决策树提供训练集。对原始训练集进行有放回的抽样方法随机得到不同的训练集,最后的分类结果取决于多棵树的投票结果。

有放回抽样是通过对原始样本进行抽样,并不将其从总体中剔除,这使得训练集之间会出现一些重复。这既保证了不同训练样本之间的相关性,也使得训练集之间具有差异性。

有放回抽样常见的方法有 bagging 和 boosting 两种。随机森林通常使用bagging方法对原始训练集进行随机抽样。该方法保证了训练数据集样本约为三分之二的原始数据集内容。

输入的训练数据集均来自接收机接收到脉冲序列。本文产生特征训练集所选择的特征分别为脉冲描述字特征中的载频、脉宽、到达角、功率等。

1.2 决策树的生成

随机森林分类器由决策树组成,是一个可视树状模型。其包括3种节点:根节点、中间节点、叶子节点(图2)。

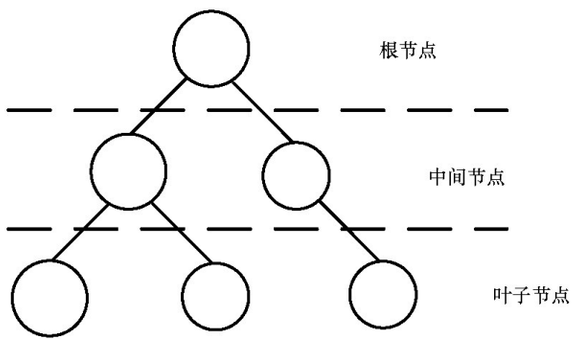


图2 决策树结点

Fig. 2 Node diagram of decision tree

每层节点对应输入数据的某一特征,叶节点对应于输入数据一个类。从根结点出发,依据节点分裂规则,根据输入对象的某一特征进行分裂。最终到达唯一的叶子节点,获得决策树的输出,这个过程就是随机森林中决策树的生成过程[6]。

1.3 节点分裂算法

生成决策树离不开节点分裂算法,包括ID3算法、C4.5算法、以及CART算法等,判别规则分别有信息最

大增益,信息增益率,以及*Gini*指数等,最常见的是CART算法。

随机森林的弱分类器使用的是CART树,CART决策树又称分类回归树。每一个非叶节点只能引伸出两个分支,因此也被称为二叉决策树。

CART算法分裂规则采用的是*Gini*指标最小原则,基尼系数的选择标准就是每个子节点达到最高的纯度,即落在子节点中的所有投票都属于同一个分类,此时基尼系数最小,纯度最高,不确定度最小。计算集合D的不纯度,获得*Gini*指标,如式(1)所示:

$$Gini(D)=1-\sum_{i=1}^{m}P_i^2 \tag{1}$$

式(1)中P_i表示类别i在集合D中的概率。*Gini*(D)表示样本集的*Gini*系数。

计算分裂划分后的*Gini*系数,假设集合D被划分成两个子集D_1、D_2,获得此次分裂的*Gini*系数为:

$$Gini_{split}(D)=\frac{n_1}{n}Gini(D_1)+\frac{n_2}{n}Gini(D_2) \tag{2}$$

式(2)中n_1为满足集合D的样本个数,n_2为不满足集合D的样本个数,n为集合D的总个数。

通过利用式(2)计算得出根据不同特征作分裂的*Gini*系数,选择*Gini*指标最小的特征作为本次分裂的最佳选择。*Gini*系数反应的是数据集D中随机选取两个样本,为不同类别的概率,因此越小的*Gini*(D),表明数据具有相同类别的纯度越高。

通过*Gini*指标分别计算每个变量的各种切分或组合情况,找出该变量的最佳切分或组合点,根据比较各个变量的切分或组合点,最终找出最佳变量和该变量的切分或组合点。

由于CART算法总是将当前样本集分割为两个子样本集,使得生成的决策树的每个非叶结点都只有两个分支。因此,在这里选用的是CART算法生成决策树对特征进行分类[7]。

2　基于随机森林的信号预分选算法

本文随机森林的决策树是利用CART分类算法生成的,节点分裂时的分裂规则是*Gini*系数最小原则。生成过程如下:

(1) 读取原始输入训练数据,采用bagging有放回地随机抽取K个新的自助样本集。作为训练集;

(2) 通过每次抽样得到的样本集生成决策树;

(3) 随机选择d个特征,计算每个特征的*Gini*系数评分;

(4) 选在最小*Gini*指标的划分作为分裂特征,重复上述特征,直到达到预先设定的停止准则。每棵决策树不进行剪枝;

(5) 重复上述步骤,直至形成ntree决策树;

(6) 多棵决策树构成森林,然后就可以对未知类别的样本进行分类,最后的输出结果由森林中各决策树的多数投票决定。

基于随机森林的信号预分选步骤如下。

(1) 将接收机接收到的脉冲序列作为原始数据输入;

(2) 利用脉冲描述字特征生成N棵决策树组成随机森林;

(3) 将需要分类的未知样本输入到已经构建好的随机森林中,根据随机森林中各决策树分类器投票结果的简单多数投票法来获得最终的分类结果。

具体流程如图3所示。

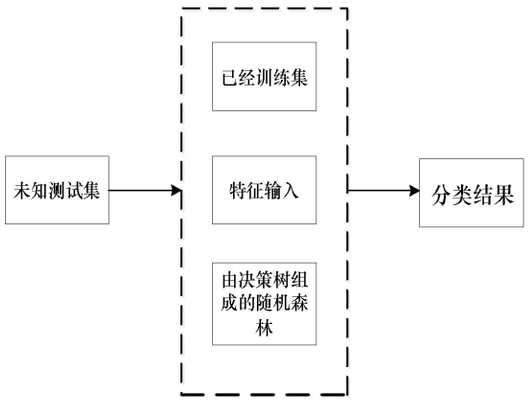


图3　随机森林训练分类

Fig. 3　Random forest training classification

3　仿真实验

3.1　随机森林分类器训练

复杂环境下,辐射源脉冲由于脉冲密度大,被动接收机等原因,将造成脉冲丢失,在对脉冲重要性进行度量时,采用不同丢失率下脉冲序列对其进行模拟,用以仿真真实电磁环境下的辐射源信号。

分类器使用的训练集是根据上文中介绍的特征所组成的4维向量,提取于4种不同的脉间调制类型信号,包括:常规、捷变频、脉组捷变、脉宽捷变雷达提取特征的脉冲序列。训练集参数如表1所示随机进行选取,常规标签为[1 0 0 0]。

表1 训练集参数

Table 1 Training set parameters

输入参数(时间精度:12.5 ns)					
调制类型:	载频/MHz	脉宽/μs	功率/dBm	到达角/°	脉冲重复时间/μs
常规、捷变频、脉宽捷变、脉组捷变	700~1400	10~20	−80~−60	30~60	150~300

表2为调制类型的具体参数表,其它参数在表1中随机进行选取。

表2 输入特殊参数

Table 2 Complex parameters

类型	跳变个数	跳变间隔	误差/%
常规	0	0	1
捷变频	32	15 MHz	1
脉宽捷变	8	3 μs	1
脉组变频	6	10 MHz	1

对于随机森林,参数的调整不会对其有很大的波动,相比于神经网络,采用默认的参数也可以达到较好的效果。随机森林分类器训练时通过GridSearchCV网格搜索选择最优的训练参数,如表3所示。其中max_features设置的值越高,随机森林的速度越慢,同时也会影响随机森林的多样性,降低预测准确率。n_estimators的设置,一般而言,数值越大准确性越高,但会牺牲分类的速度。min_samples_leaf该值越小,训练树划分的越精细,同样对噪声数据越敏感。通过参数设置生成10棵决策树。

表3 随机森林参数设置

Table 3 Random forest parameter settings

参数名	参数值	说明
n_estimators	100	选取合适个数平衡性能和误差率
cirterion	Gini	CART算法
max_features	Log2	最多同时考虑3组特征
max_depth	5	树的最大深度
min_samples_leaf	25	叶子节点最小样本数
min_samples_split	10	特征划分节点所需最小样本

通过训练得到的特征贡献度评分如表4所示,f_1、f_2、f_3、f_4分别为到达角、载频、幅度、脉宽特征。

由表4评分可以看出,特征f_1、f_2对随机森林重要性的贡献度最高,对结果分类的影响力较强,而特征f_3和f_4重要性较低。

表4 特征重要性参数

Table 4 Feature importance parameters

特征	f_1	f_2	f_3	f_4
分数	0.310329	0.276197	0.220027	0.106945

3.2 复杂脉间调制类型预分选

将随机森林预分选新方法与传统使用载频和脉宽的方法进行分选结果对比,并将分选结果进行对比,传统预分选算法采用改进的K-means算法。输入信号如图4所示。

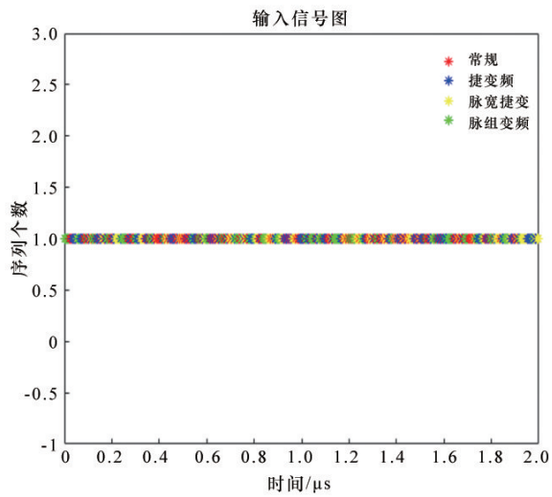


图4 输入脉冲序列

Fig. 4 Input pulse sequence

图4中所示为以上形成的脉冲序列,横坐标为时间,纵坐标为序列个数。分别使用传统预分选算法和随机森林新方法就该脉冲序列进行预分选分类。传统分选聚类结果如图5所示。

图5中所示,横坐标为时间,纵坐标为分选出类别编号,可以看出传统方法分选出7个类别,其中脉宽捷变和脉组变频的信号发生了严重错误。脉宽捷变分类成5个类别,并与脉组变频混叠。用随机森林的方法对该脉冲序列进行训练分类,结果如图6所示。

图6可以看到,在使用随机森林算法自适应选择特征进行预分选,信号很干净的被分成4个类别,分别为类别1常规雷达,类别2捷变频雷达,类别3脉宽捷变雷达,类别4脉组变频雷达。验证了该算法有效性。

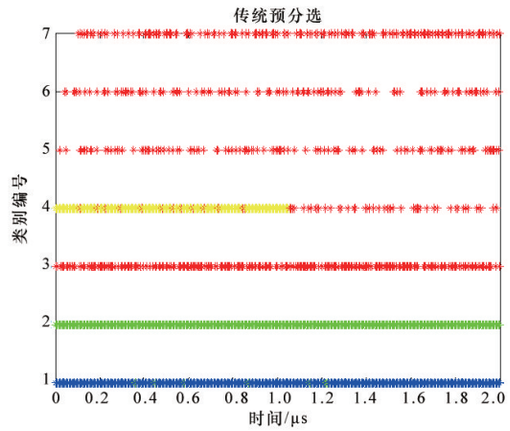


图5　传统载频、脉宽分类

Fig. 5　Traditional carrier frequency, pulse width clustering

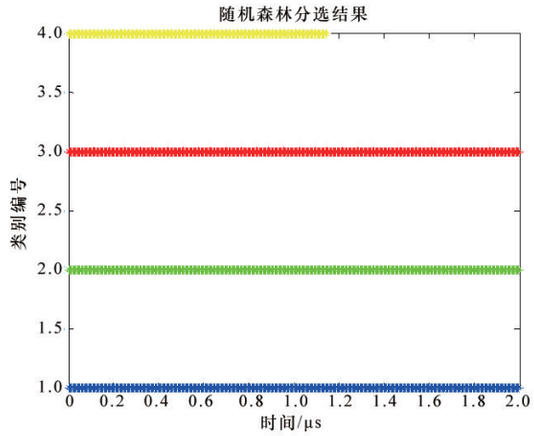


图6　随机森林分类

Fig. 6　Random forest classification

4　结论

提出了基于随机森林算法进行信号预分选的方法。通过与传统预分选使用的载频、脉宽特征进行比较实验,证实在相同输入信号下,使用随机森林算法对脉冲描述字特征进行预分选,可准确完成信号的预分类。对于多部复杂调制雷达信号,可以达到较好的分选识别效果。通过实验仿真,验证了算法的可行性。

参考文献(References)

[1] 刘海军, 樊昀, 李悦, 等. 多功能雷达建模中的雷达字提取技术研究[J] 国防科技大学学报, 2010, 32(2): 91-96.

[2] 周一鹏, 王星, 田园荣, 等. 基于极值序列特征集的雷达PRI调制模式识别算法[J]. 现代雷达, 2016, 38(5): 37-41.

[3] 何佩佩, 唐霜天, 匡华星. 一种基于层次划分聚类的雷达信号分选算法[J]. 现代防御技术, 2016, 44(4): 51-55.

[4] 胡国兵, 刘渝, 邓振淼, 等. 复杂体制脉冲重复间隔调试方式识别[J]. 数据采集与处理, 2010, 25(6): 722-726.

[5] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报 (工学版), 2014, 44(1): 137-141.

[6] Chen S, Jiang Q, Pan J F. A novel method for the modulation type recognition of radar PRI[J]. Aerospace Electronic Warfare, 2012, 28(1): 31-34.

[7] 刘歌, 张国毅, 于岩. 基于随机森林的雷达信号脉内调制识别[J]. 电信科学, 2016, 32(5): 69-76.

A novel alogrithm of signal pre-sorting based on random forest

LIU Xubo[1], LIU Jingshu[1], LIU Bin[2], Qin Lingling[2], CHEN Tao[2]

1. Unit 91977 of People's Liberation Army of China, Beijing 102200, China

2. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China

Abstract　In modern warfare, the radar electronic warfare environment is more and more complicated. With the variety of radar types and the complexity of radar inter-pulse modulation, the difficulty in identifying signal sorting is increasing. This paper proposes a random forest algorithm to sort the characteristics of pulse descriptors, which can daptively select features and achieve classification. Random forest can make the error balance of unbalanced data. Meanwhile through the multi-decision tree voting method, it can quickly complete the rapid training of large amounts of data. In the case of a pulse loss, the recognition accuracy can still be maintained. Experimental results show that the proposed method is effective in presorting radar pulse descriptors.

Keywords　random forest; decision tree; radar signal sorting; feature selection

(编辑　徐丽娇)