# 脉冲星预测模型及权重最大特征的确定

**摘要：**自上个世纪六十年代人类首次发现脉冲星以来，有关它的研究从未终止过。脉冲星一些不可思议的特征更是为现在物理学特别是高能物理和宇宙学提供了一个天然的平台，为人类研究宇宙的起源、演化和变迁提供了一把天然的钥匙。本文以大量行星发出的辐射信号数据为基础，对信号的八个特征进行了不同程度、不同方面的探索，最后预测了权值最大的特征。在原始数据基础上，本文构建了决策树、逻辑斯蒂回归模型、Adaboost三种模型，得到了在各个模型中每个特征的权值，并检验了本文的预测。最后本文分别计算以及分析了各个模型的精确度和召回率，通过比较得到了最优的模型。

**关键词:** 特征权重；决策树；逻辑斯蒂回归模型；Adaboost模型

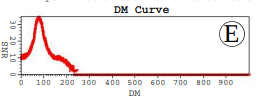
# 数据来源及说明

本文原始数据来源于Kaggle提供的数据。数据是由17898个从宇宙中搜集的射电信号组成，一部分被证实是脉冲星，一部分被证实是不是脉冲星。每一个样本即射电信号都有八个特征，具体如表1:

表1 样本特征的描述

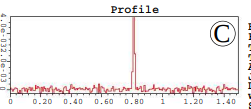
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mean of the integrated profile | Standard deviation of the integrated profile | Excess kurtosis of the integrated profile | 1. Skewness of the integrated profile. | Mean of the DM-SNR curve | Standard deviation of the DM-SNR curve | Skewness of the DM-SNR curve | Skewness of the DM-SNR curve |
| 累积脉冲轮廓的均值 | 累积脉冲轮廓的标准差 | 累积脉冲轮廓的峰度值 | 累积脉冲轮廓的偏度 | 分散度与信噪比曲线的均值 | 分散度与信噪比曲线的标准差 | 分散度与信噪比曲线的峰值 | 分散度与信噪比曲线的倾斜度 |

实际上，在宇宙中存在许多自由电子，这些电子会分散脉冲星射电信号的传播，地球上收到的信号相比于真实信号有一些损失，要解决这一问题，就要找出信号被分散的量即分散度（DM）。在每个假设的分散度情况下，都可以计算出此分散度下信号的信噪比（SNR），从而得到分散度与信噪比的曲线。一般的分散度与信噪比的曲线如下图（1—1）所示:



图（1-1）色散度与信噪比的曲线

从上图可选出信噪比最高的分散度，基于此分散度就能找出候选者曲线。找出的脉冲星候选者信号还只是单个脉冲信号。脉冲星的单个脉冲辐射具有极高的偏振性，脉冲强度和波形彼此不同，且呈现混乱无序变化。所以研究候选者信号的累积轮廓（即将多个单脉冲整合到一起形成新的累积轮廓）更有意义。信号的累积轮廓如下图（1-2）：



图（1-2）脉冲星信号的累积轮廓

从上图可以看出每个累积轮廓像一个正态分布，累积脉冲轮廓的均值就是这个分布的期望值（均值）；累积脉冲轮廓的标准差就是这个分布的标准差；累积脉冲轮廓的峰度值是这个分布于正态分布的陡峭度的比较，其值大于零则此分布比正态分布更陡峭，反之亦然；累积脉冲轮廓的偏度表示此分布与正态分布的偏离度，当偏值大于零则右边的分布大于左边的的分布，反之亦然，但其值等于零时，表明此分布与正态分布的偏度相同。

# 2．总体思路

本文为了实现对脉冲星的预测，建立了三个模型:

1)在原始数据的基础上，计算了各个特征数值的均值和标准差，得到了各个特征的数值分布情况，并假设了在模型中权值最大的特征。

2)在原始数据基础上， 构建了决策树模型和Adaboost模型以及逻辑斯蒂回归模型。得到各个模型中的每个特征的权重以及精确度和召回率。

3)检验1）中的预测并对以上三种模型进行对比与分析，得到最优模型。

# 3．特征数值分布以及重要性预测

各个特征数值分布情况是分析数据的重要前提。首先，本文利用简单的数据统计得到了各个特征数值的描述如图（3-1）:

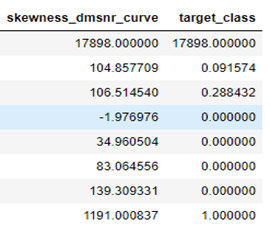
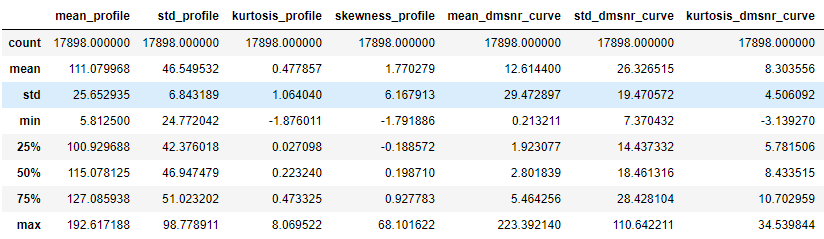
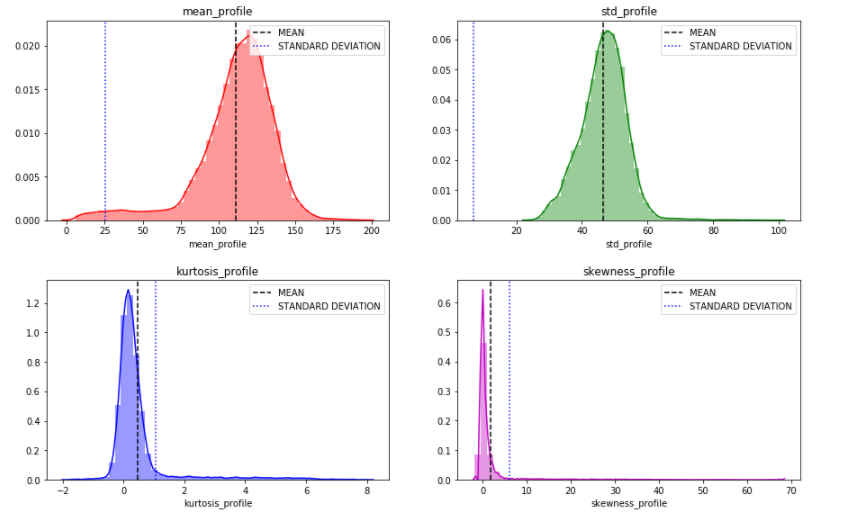
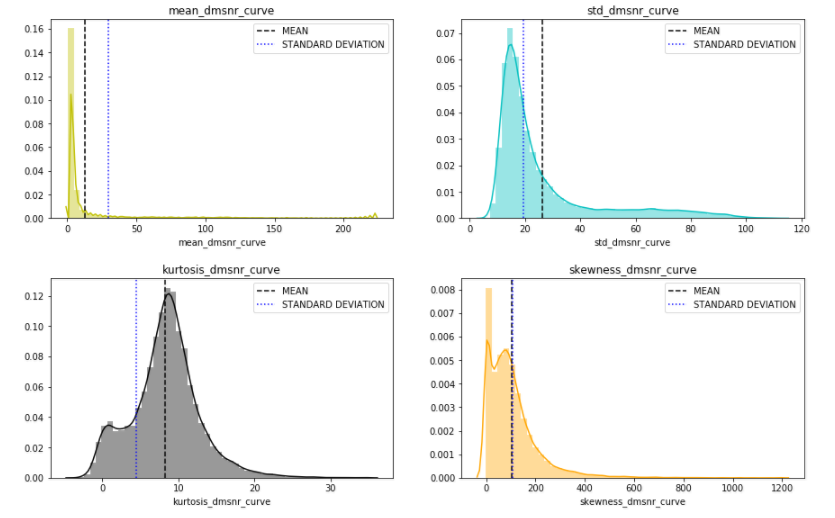


图3-1各个特征的基本信息

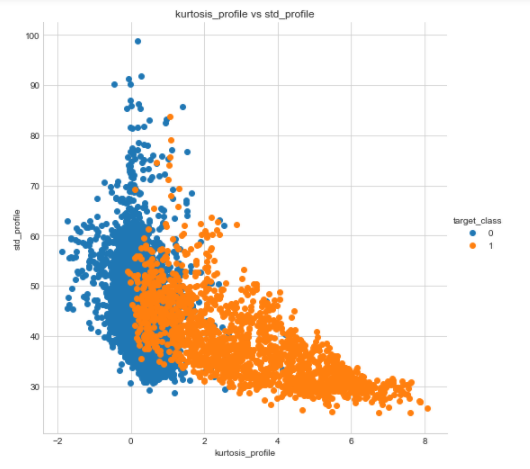
上图展现了各个特征数值的最小值和最大值以及均值，标准差的大小。这些数值只能粗略的描述数据的分布。为得到更直观的分布，本文根据各特征数值分别画了数值分布图，如图（3-2）所示:



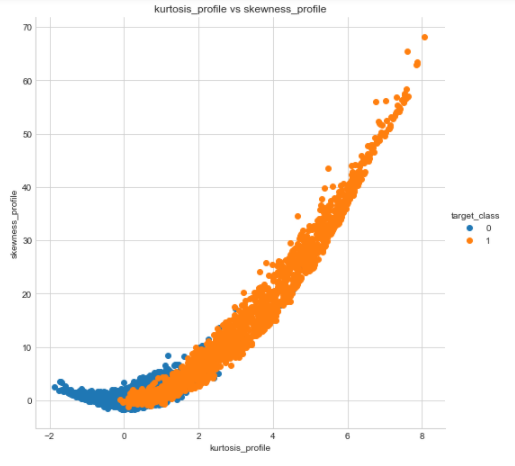


图（3-2）各个特征的数值分布图

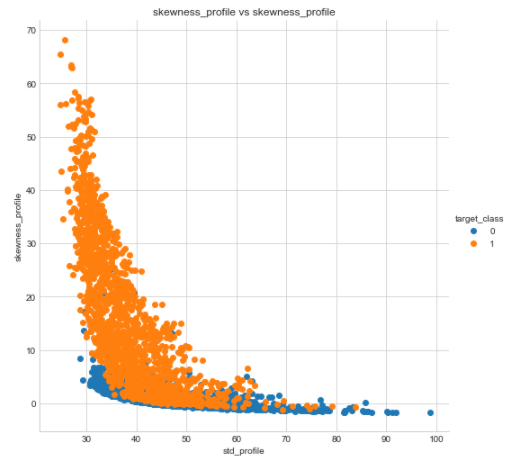
上图中，各个小图的横轴是各个特征的数值，纵轴是此数值在整个数据中的频率，黑虚线是此特征数值的期望值，蓝虚线是特征数值的标准差。可以明显看出只有三个特征（累积脉冲轮廓标准差、峰度值、偏度值）比较符合正态分布，其他的分布较乱。为了更进一步对这三个特征的权值预测，本文对这三个特征进行了对比，如下图所示：



图（3-3）累积轮廓的标准差与峰度值的比较



图（3-4）累积轮廓峰度值和偏度的比较



图（3-5）累积轮廓的标准差和偏度的比较

从上三幅图，可以看出偏度值在其数值的范围内对应的属性值差不多都是1，说明偏度值的变化对属性值的变化影响较小。从图（3-3）标准差和峰度值的比较图中可以知道，当峰度值大于零时，对应的属性值基本上是1，小于零时，对应的属性值基本上是0,而标准差在其变化范围之内，没有一个明显的数值可以作为分界线来区分属性值0和1。

基于以上所有的对比，本文预测累积轮廓的峰度值在所有特征中的权值最大。后文的结果会对此做出验证。

# 4．模型实现

## 4.1逻辑斯蒂回归模型

### 4.1.1逻辑斯回归蒂模型的核心思想

逻辑斯蒂回归模型可用于线性回归以及多类分类问题。本文利用的是二项逻辑斯蒂回归模型（此模型用于二类分类问题）。二项逻辑斯蒂回归模型可由条件概率分布来表示，条件概率的公式如下:



公式（4-1）条件概率的分布

当的计算结果大于时，这个预测样本就属于1类，反之，则属于0类。得到这个条件概率就必须求得的值，二项逻辑斯蒂回归模型中不是具体的数值，而是与样本特征数量所对应的矩阵，如下所示:



公式（4-2）参数的矩阵

公式（4-1）中和的点乘实际上是参数的矩阵和各个特征组成的矩阵的点积。

二项逻辑斯蒂回归模型构建的关键在于求出，可用极大似然法来求出。极大似然法常用与于已知随机变量输出结果来求未知参数的可能取值。对于二项逻辑斯蒂回归模型中为求解参数构建的似然函数如下:



公式（4-3）关于参数的极大似然函数

上述公式中，， 是这个样本的属性值。上述公式是一个乘法，计算量大，一般会用对数处理后来计算，且称处理后的函数为对数似然函数。关于参数的对数似然函数如下：



公式（4-4）关于参数的对数似然函数

将代入上公式后，化简可得如下公式：



公式（4-5）关于参数的对数似然函数的化简式

构建公式（4-5）之后，利用梯度下降法就可以求出关于参数的最优解。

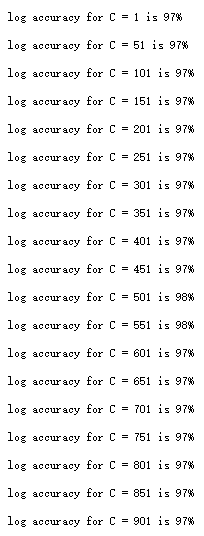
### 4.1.2模型的实现

模型实现采用了scikit-learn包中的LogisticRegression算法。scikit-learn是一个基于Python scikit构建的推荐系统分析模块，提供了大量实现推荐系统的算法，内置了一个功能强大的Cross-validation迭代器，可以非常容易地运行交叉验证程序 。

LogisticRegression函数包含14个参数，都是用来更好求出的最优解。其中最重要的参数是C(正则项)。C用于对函数模型的拟合，缺损值是1，一般取这个值，可得到合适的模型。

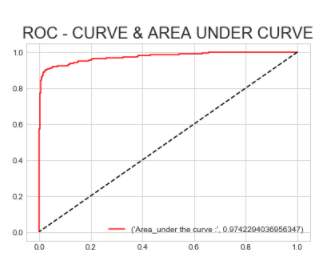
### 4.1.3模型的结果与分析

本文调试逻辑斯蒂模型中的正则项参数C，如下图所示：

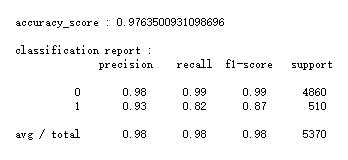


图（4-1）不同参数C对应的模型精确度

从上图中可以知道，参数C在1到1000之间取值时，与之对应的模型精确度精确度变化不是很大。说明当参数C的数值在此区间的时候，模型都没有过拟合或欠拟合。本文最后构建了参数C取值为1的逻辑斯蒂模型，模型的准确率、精确度和召回率如下图所示：



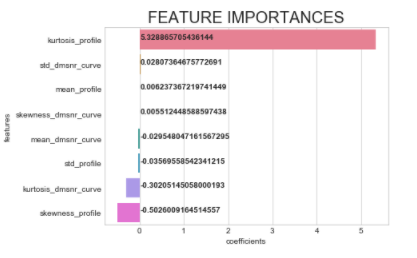
图（4-2）模型准确率的面积表示图



图（4-3）逻辑斯蒂模型的精确度和召回率

从上图可以看出，本文最后构建的模型对0类样本分类的能力是比较强的，对1类样本分类的能力不太强。

得到模型的评价之后，本文利用sklearn.LogisticRegression中自带的特征权重的评估器，得到了在此模型中，各个特征对模型决策的重要性，如下图所示：



图（4-4）逻辑斯蒂模型中各个特征的重要性

从上图中可以得知，第三个特征累积轮廓的峰度值的权值最大，且远远大于其他特征。

## 4.2决策树模型

### 4.2.1决策树模型的核心思想

决策树模型由许多的结点构成。决策树模型在每个结点都会基于某个特征的某个数值为分水岭来分类。新样本按照结点从上到下的顺序，就可以找到对应的属性值。得到最优的决策树模型关键在于每个结点如何找到对应特征以及对应特征数值。解决这个关键一般先使用CART算法计算样本集合的基尼值Gini(D)，其计算公式如下：



公式（4-6）样本集合的基尼值计算公式

从上公式可以看出，实际上，样本集合的基尼值是表示这个样本集合中，一个随机的样本被模型分错的概率。基尼值越小就说明这个样本集合的纯度就越高。计算样本集合的纯度之后，利用计算基于某个特征a划分样本集合D之后的基尼值的计算公式可计算出每个特征在每个数值点的基尼值，其计算公式如下：



公式（4-7）基于特征某个数值的基尼值计算公式

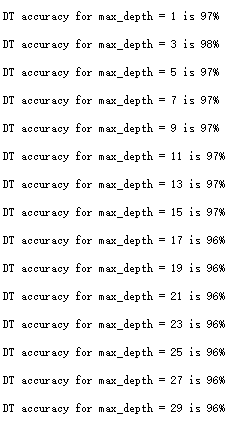
上式中是在a特征中基于某个数值划分的第一个样本集合，是第二个样本集合，是的基尼值；是的基尼值，这两个基尼值可以由公式（4-6）计算得到。选出基尼值最小的特征和数值，从上到下匹配每个结点，可以得到决策树的最优模型。

### 4.2.2模型的实现

模型实现采用了scikit-learn包中的DecisionTreeClassifier算法。DecisionTreeClassifier函数包中包含了10个参数，都是用来优化模型的，其中最重要的是max\_depth参数。调节max\_depth参数的数值使模型挖掘到不同深度，使模型不过拟合或欠拟合，最终得到合适的拟合。

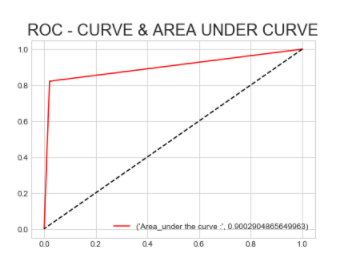
### 4.2.3模型的结果与分析

本文调试了模型中的max\_depth参数，结果如下图：

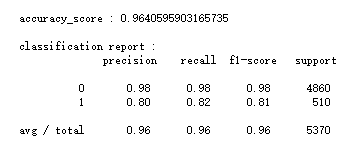


图（4-5）决策树模型参数max\_depth不同数值对应的准确率

从上图中可以看出在参数max\_depth取1~30之间的值对应精确度相差不大，且正确率都比较高。所以，本文最后构建的决策树模型中参数max\_depth的取值为缺损值，即参数max\_depth等于0。得到最终模型后，本文使用最终模型对测试数据集进行预测，并得到了模型的分数（正确率）、精确度以及召回率，具体数值如下图：



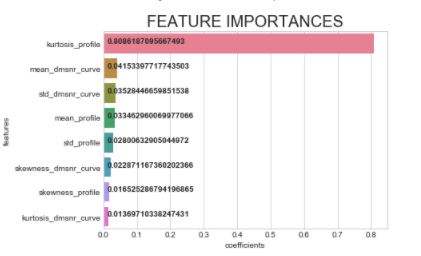
图（4-6）决策树模型的准确率



图（4-7）决策树模型的精确度和召回率

从上图中可以看出本文构建的决策树模型对0类的分类的能力是比较高的，而对1类样本分类的能力不如一类。

最后，本文使用sklearn.tree函数包中的选取特征重要性的模块，得到了此模型中各个特征的权值，如图所示：



图（4-8）决策树模型中各个特征的重要性

从上图中很明显看出累积轮廓的峰度值在此模型中的权值最大，即在预测新样本属性的时候，考虑此特征的数值的重要性最大。得到的这个结果和前文预测的一样。

## 4.3Adaboost模型

### 4.2.1Adaboost模型的核心思想

Adaboost是一个组合多个弱分类器(基本分类器)，构成一个强的分类器。得到最优的Adaboost模型的关键是在增加弱分类器的每一轮计算过程中，如何改变样本的概率密度和每个弱分类器的分类。在第一轮计算中，每个样本的权值都是一样的。利用第一个弱分类器分类之后，就知道那些被分错的样本，则这些养的权值会被加重，计算公式如下：



公式（4-8）分类误差率的计算



公式（4-9）第m个弱分类器前面的系数



公式（4-10）第m个弱分类器中每个样本的权重

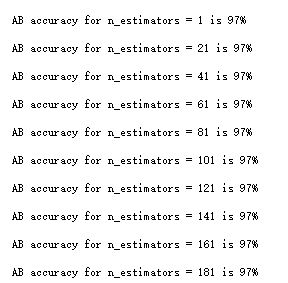
是第m轮计算中被误分类样本的权重之和，是基于每个分类器的分类误差率计算得到的权重。是m+1轮计算中各个样本的权重。以此计算，当没有出现错分样本（即分类误差率为零）的时候，得到的模型就是最优的Adaboost模型，其函数式如下所示：



公式（4-11）最优的Adaboost模型分类器的函数式

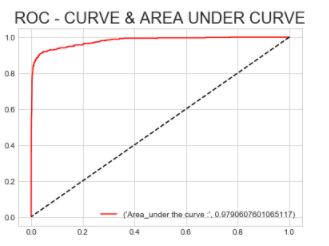
### 4.2.2模型的结果与分析

n\_estimators是Adaboost模型中优化模型中重要的参数，其的数值是计算的轮数和弱分类器的数量。本文对此参数进行了取值为1到200之间不同的调试，调试结果如下图所示：

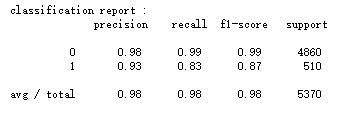


图（4-9）Adaboost模型参数n\_estimators不同数值对应的准确率

从上图中可以看出，当n\_estimators取值为1和取值为181的时候，准确率变化不大。说明当只用一个弱分类器（即决策树）的时候，Adaboost模型对样本的分类能力已经比较强，所以本文最后构建的模型是n\_estimators取缺损值（即取值为50）的Adaboost模型。得到最后的模型之后，本文计算了模型的准确率、进精确度以及召回率，结果如下图所示：



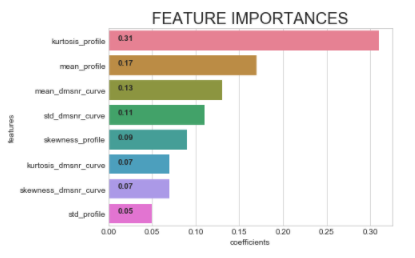
图（4-10）Adaboost模型的准确率



图（4-11）Adaboost模型的精确度和召回率

从上图中可以知道，本文构建的Adaboost模型擅长对0类样本的分类，对1类的分类能力良好，但并是特别擅长。

最后，本文利用sklearn.Adaboost函数包中选取特征重要性的模块，得到了在此模型中各个特征的权重，如下图所示：



图（4-12）Adaboost模型中各个特征的权值

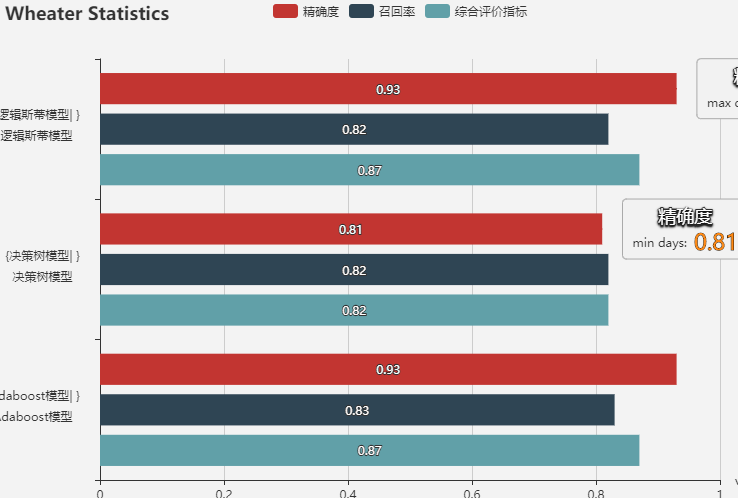
从上图中很明显看出累积轮廓的峰度值在此模型中的权值最大，即在预测新样本属性的时候，考虑此特征的数值的重要性最大。得到的这个结果和逻辑斯蒂回归模型、决策树模型的结果一样。

## 4.4权值最重特征的选择

从各个模型的特征权重可以看出，三个模型中权值最大的特征都是累积轮廓的峰度，而且每个模型中累积轮廓的峰值的权重都比其他的特征的权重大很多。这说明对新样本分类的时候，可以凭借此特征的取值进行样本进行粗略的预测。这一结果也证明了前文的预测是正确的。

# 5．模型的比较与分析

从上可以知道，本文构建的三个模型对数据的总体精确度、召回率以及综合评价指标的分数相差不大，无法区分各个分类器对此数据集的实用性，为解决这问题，本文对各个模型对1类分类的能力进行了比较，如下图所示：



图（4-13）各个模型对1类分类的能力比较

从上图中可以看出，Adaboost模型对1类的分类比其它两个模型略胜一筹。

# 6.结语

本文基于原始数据构建了决策树和Adaboost以及逻辑斯蒂回归三个模型，每个模型对总数据集的精确度和召回率都比较高且数值相近，但各个模型对1类的分类的能力的不同，从这个方面比较得到了最优模型Adaboost。这个模型可以更好预测脉冲星。

