## 3.3 分类算法 Classification algorithm

首先对收集到的文件进行分类，将提取出来的特征作为训练数据集，在这个时候，随机森林（random forests）在分类上表现了很好的优势，有效且误报率极低，并易于使用,可以很快的对数据进行分类。随机森林分类方法给出的结果是基于很多棵分类树判断结果的集合展现，每一个决策树都是在训练数据中随机选择生成的，因此，随机森林总的来说是一个集成分类器，他使用 bagged training data, 通过随机选择的特征子集,，并使用该节点的训练数据，确定每个节点处的最佳分割来创建树中的每个节点。此外，每棵树都是基于一个独立的特征子集，最后，在分类过程中每一个树的投票来确定最终结果。

Fristy,we calassify for colleged files,use extracted feature as training data sets ,At this time,Random Forests show to advantagutive

First, we classify the collected files by extracted features which are used as the training dataset. Through the process, Randon Forests Algorithm we used performs well in classification, with high effectiveness and low false positive rate. It is easy to use and can classify data rapidly.

Firstly, We classify the collected files and use the extracted features as training data sets. At this time, Random Forests has good advantages in classification, Effective and low false positive rate, easy to use, you can quickly classify the data. The result of the random forest classification method is based on the collective representation of the judgment results of many classification trees. Each decision tree is generated randomly in the training data. Therefore, the random forest is an integrated classifier in general. Using bagged training data, each node in the tree is created by randomly selecting a subset of features, and using the node's training data to determine the best segmentation at each node. In addition, each tree is based on a separate subset of features. Finally, each tree is voted in the classification process to determine the final result.

AI引擎的重要 组成部分之一是算法，我们选取了几个实用性较好的算法来比较包括KNN邻近算法，NNET神经网络，RF随机森林和SVM支持向量机，经过多次的训练与分类实验，发现随机森林准确率高，误报率低，低延时鲁棒性良好，等和可解析性等优势，于是将其作为我们选定作为默认算法。并且经过我们的一百多次的验证，在特征发生改变的时候，随机森林准确率依然趋于一个稳定的值。

One of the important components of the AI engine is the algorithm. We have selected several algorithms with good usability to compare the KNN neighboring algorithm, the NNET neural network, the RF random forest and the SVM support vector machine, after many trainings and classification experiments. We found random forests with high accuracy, low false alarm rates, good robustness with low latency, etc., and their advantages such as resolvability. We chose this as our default algorithm. And after more than one hundred validations of ours, the random forest accuracy rate still tends to a stable value when the characteristics change.

Table 3 算法准确率对比

|  |  |  |  |
| --- | --- | --- | --- |
| SVM | NNET | KNN | RF |
| 75.23% | 82.41% | 97.12% | 99.64% |

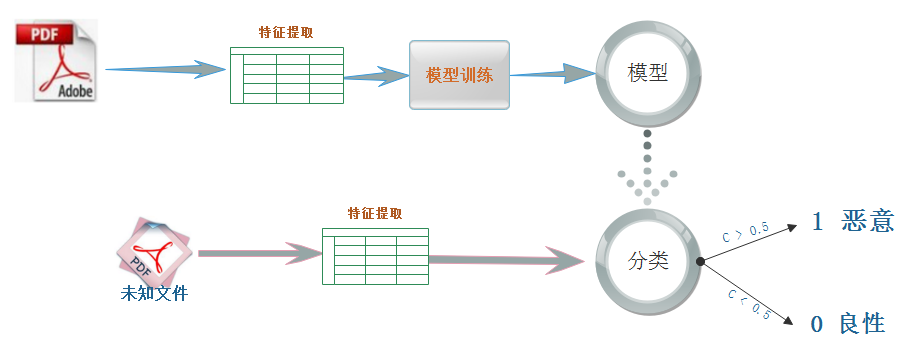
## 3.4 模型构建

所提出的基于机器学习的恶意PDF文档检测的方法包括以下两个步骤，如图1所示：

1. 提取文件特征。此为基本的预处理步骤，对PDF文件的结构、内容和元数据进行解析，并做相应的向量计算，提取为一个二维的特征集，使得这些特征可以进入到基于机器学习的模型中进行训练分类。

2. 学习和分类。我们会随机选取数据的80%进行训练，通过训练之后保存训练模型，然后使用20%的文件进行预测分类，从中计算出模型的准确率，误报率等信息。

图1 机器学习的基本框架



在我们的实验当中，一共对模型进行了4次更新，其中最开始的模型（model1）是使用peepdf为解析工具，然后通过一些特征计算与特征量化，使其可以用于机器学习训练与预测，当我们的特征数提取到133个的时候，这些特征包括有基于结构的（count\_font、size、count\_startxref），内容信息的（title\_oth、subject\_lc）和metadata(producer\_oth、producer\_len)的一些静态属性,并通过多次实验核对比较选取性能和精度较好的算法(随机森林RF)，但这样存在问题，可是经过我们的研究发现，由于在一开始使用peepdf进行解析的时候，只有一半的文件可以被解析到，所以我们重新选取了解析器mimicus[2]，这个工具可以解决之前因为结构缺陷或混淆而不能正常解析的问题，我们使用了mimicus 对之前的文件进行解析，对所有的数据（20万）均能正常解析，并做特征提取。

在模型2的训练中，我们先使用平衡的数据集进行训练与预测，其中包含2万恶意样本与2万良性样本，在总的数据集中进行随机抽选一共4万个样本用于model2 的训练与预测。并且从Model2 开始我们就使用mimicus 对文件进行特征提取，一共提取特征135个。我们的主要算法还是使用准确率较高的随机森林。通过算法调优后对数据进行训练与预测，使Model2的检测率提高到99.99%，误报率降低为0.012%，并经过多次随机选择数据进行训练预测后，准确率依然可以维持在99.99%。

经过Model2 的算法调优和数据筛选后，我们把训练数据增加到十万级别，来对模型进行训练。我们的训练是在本地进行的，在4核4G的CPU上，训练10万个样本仅需要56s, 当训练样本达到十万级别的时候，Model2准确率持续稳定在99.81%，误报率为0.086%。