文章编号：2096-1472(2020)-11-06-04
DOI:10.19644/j.cnki.issn2096-1472.2020.11.002

多种分类模型在海关智能化风险布控中的应用

金 瑾[1]，王正刚[2]，刘 伟[2]，巫家敏[1]，李 波[1]

(1.成都东软学院，四川 成都 611844；
2.中华人民共和国成都海关，四川 成都 610041)
✉jinjin@nsu.edu.cn; wangzgxs@outlook.com; 45711577@qq.com;
WuJiamin@nsu.edu.cn; li-bo@nsu.edu.cn

摘 要：海关监管部门在风险布控的过程中，需要风险分析人员依据经验对各种商品的风险进行人工分类。本文用Logistic回归模型、决策树模型、随机森林模型等几种的分类模型优化风险布控过程，通过将报关数据输入分类模型得到特定商品的风险评估结果，辅助风险分析人员做出正确判断。通过实验验证这种智能化的方法能够有效克服人工风险布控中的不足，完成智能化风险布控，进一步维护国门口岸安全。

关键词：大数据；机器学习；分类；风险布控
中图分类号：TP183 **文献标识码**：A

Application of Multiple Classification Models in Intelligent Customs Risk Prevention and Control

JIN Jin[1], WANG Zhenggang[2], LIU Wei[2], WU Jiamin[2], LI Bo[2]

(1.*Chengdu Neusoft University*, *Chengdu* 611844, *China*;
2.*Chengdu Customs of the People's Republic of China*, *Chengdu* 610041, *China*)
✉jinjin@nsu.edu.cn; wangzgxs@outlook.com; 45711577@qq.com;
WuJiamin@nsu.edu.cn; li-bo@nsu.edu.cn

Abstract: As part of the process of risk prevention and control, the customs supervision department requires risk analysts to manually classify risks of various commodities based on their experiences of risk management. This paper uses several classification models, such as logistic model, decision tree model, and random forest model, to optimize risk control process. Risk assessment results of specific commodities can be obtained by inputting customs declaration data into the classification models. Thus the results assist risk analysts to make correct judgments. The proposed model is verified through experiments. The results show that it is an intelligent method and can effectively overcome the shortcomings in manual risk control, complete the intelligent risk control, and further maintain the security of national ports.

Keywords: big data; machine learning; classification; risk prevention and control

1 引言(Introduction)

海关通过各种风险布控系统布设处置风险的指令，并由被指定的现场作业环节执行处置和反馈处置结果的风险布控措施[1]。随着海量货物贸易增长，跨境电商新业态发展和国际直邮物品的不断增加，现阶段人工风险判别方式已经很难精准发现高风险货物商品。海关监管部门需要寻找一种智能化的手段，辅助人工风险判别过程，提高口岸查获率，提升人工风险判别的准确度。

风险布控中商品查验问题的本质一个分类问题。而在人工智能领域有很多模型可以处理分类问题。本文旨在探索机

基金项目：2020年度四川省科技厅创新基地（平台）和人才计划《四川对外贸易数据分析与风险防控》（2020JDR0330）.

器学习中分类算法在海关智能化风险分析中的应用，提出用人工智能中成熟的分类模型作用于风险布控流程中的风险判定环节，提高商品布控精度。运用Logistic模型[2]、决策树模型[3]、随机森林模型[4]三种模型对报关数据进行分类，并实验验证三种模型的实验效果，筛选出适合海关报关数据的分类模型。为风险分析人员分析商品数据下达布控指令提供切实有效地辅助决策。

2 分类模型介绍(Introduction to classification models)

在对大量的历史报关数据的分析的过程中，如何选择合适的机器学习算法是建立风险分析模型的重要一环。这一部分我们对三种机器学习经典方法进行分析对比，探索合适的海关报关数据建模的机器学习算法。

2.1 Logistic回归

Logistic回归被应用于多种实际问题的建模，它是分类问题建模的常用方法，同时也被称为其他方法的分类基准。Logistic回归具备以下优点，第一，对自变量的分布没有要求。第二，回归方程易于理解，模型中变量的影响较容易确定，从而方便调整模型。第三，分类的准确性可以被直观地判断。Logistic回归常规的用法是二分类问题，本文采用Logistic回归来解决多分类问题，这里用到了One vs Rest(OvR)和One vs One(OvO)的方法来把二分类问题转换成多分类问题。

OvR的原理是分别取一种样本作为一类，余下的看作另一类，这样N个样本就形成了N个二分类问题。使用逻辑回归算法对N个数据集训练，得到N个模型的结果。将待预测的样本传入N个模型中，取概率最高的模型对应的分类结果为最终的预测结果。

OvO的原理是，从N个样本中每次挑选出两种类型，一共有C_N^2种二分类情况，使用逻辑回归对这么多种二分类进行预测，种类最多的样本类型被认为是样本的最终类型。

2.2 决策树模型

决策树模型是机器学习最具影响力的分类和预测方法之一。它具有清晰的可解释性，而且性能良好。决策树算法从出现开始经过不断改进，出现了ID3、C4.5等算法，现在较为成熟的是C5.0算法[5]。C5.0引入了Boosting技术。决策树具有以下性质：

(1)每个内部节点都是一个分割属性。

(2)每个分支节点都由上层节点具体分类原则决定。

(3)每个叶结点都是一个类别。

选择合适的分割属性是不同决策树算法的关键问题，也因分割属性的不同演化出了不同的决策树算法。鉴于本文研究的目的是对海关风险布控等级进行分类，因此选择较成熟的C5.0算法。

2.3 随机森林模型

决策树模型虽然性能良好，但仍然存在一些缺点，比如容易发生过拟合，陷入局部优解。在决策树的基础上产生了随机森林算法。随机森林算法在生成单模决策树的基础上，采用Bootstap抽样方法训练多个样本。随机森林的基础分类器为CART。因为由多棵树来共同做出决策，所以增加了算法泛化能力和抗噪能力。

随机森林的具体步骤：

(1)采用bootstrap抽样方法生成N个训练样本集。

(2)随机抽取部分特征作为待选特征。

(3)生成N棵决策树，形成随机森林。

(4)通过所有的树对分类目标进行投票，投票最多的分类即为最后的分类。

3 模型实证分析(Empirical analysis of models)

3.1 数据处理

本文数据集为海关自有数据集。在所有的字段中，重点关注的字段为原产地、重量、价格、商品描述、企业、HS编码。在数据预处理阶段，对缺失值进行填充，删除不必要的字段。数据的最终标签为风险等级划分，风险等级分为10级，因此建模时需使用机器学习中的多分类模型。

样本被划分为不同的数据集，分别用来训练模型，测试模型，验证模型。本文采用随机抽取的方式，把数据集按照8:2来划分。在样本划分的时候需首先可视化出所训练样本的分布图，查看各个属性的取值分布和总样本集中各属性的取值分布，以查看数据集是否平衡。

对日期数据进行处理，运用正则(\d+)-(\d+)-(\d+)取出年月日，用取出来的年月日代替原来的列。对数据进行转换问题，将一样的数据转换成数值代替，在做的过程中发现有很多为空的数据是不做处理，直接将为空的数据定义为一类数据从而用一个特定的值将其代换，因为数据为空也是一种数据特征。

对于缺失值，处理的方法主要是运用了python中的字典。现将每列特征值中数据存入一个字典的key，并且赋不一样的value值，这样每一个特征列中相同特征值列就变成了一个键值对。然后在依次遍历一遍每个特征列的特征值，并将字典中每一个以特征值作为key的value赋值给对应的特征值。从而达到数据的转换。(例如：一个特征列中的值为[“苹果”，“香蕉”，“苹果”，“鸭梨”，“火龙果”，“香

蕉”，“苹果”]，通过特征列生产字典{“苹果”:0}，{“香蕉”:1}，{“鸭梨”:2}，{“火龙果”:3}，特征列中有四类特征值，就生产了四个键值对。然后再将特征列的每一个特征值作为key，找到对应的value，将value赋值给原来的特征值。通过转换和新的特征列变成了([0,1,0,2,3,1,0])。

对于模型的评价，针对具体的数据，把有风险的商品判断为低风险的商品称为A类错误，把无风险的商品判断为高风险的商品称为B类错误，一般情况下，我们希望A类错误尽可能低。因此在模型的定量和定性分析中，我们会重点对比两类错误。

3.2 Logistic回归模型

(1)Logistic回归正则化

在Python scikit-learn包中，逻辑回归用了正则化的方法。正则化方法可选L1或L2。本文选用L1正则化作为逻辑回归的正则化方法。

(2)多分类方法分析

传统的Logistic回归解决的是二分类问题，而通过改进使其可以解决的问题由二分类扩展为多分类。在scikit-learn包中默认有OvR和OvO两种。根据本文的具体问题，海关数据分析对精度要求稍高，而对时间要求为次要考虑因素，因此通过分析两种算法的优劣，最终选取OvO作为分类方法。

3.3 决策树C5.0模型

初步模型的建立采用Python的scikit-learn包，参数的选择整体使用默认参数，但是因为树的深度对算法的性能影响较大，因此对于树的深度设计了相关实验，实验结果如图1所示，从实验结果中可知当树的深度为16时，性能最佳，因此决策树的深度为16。

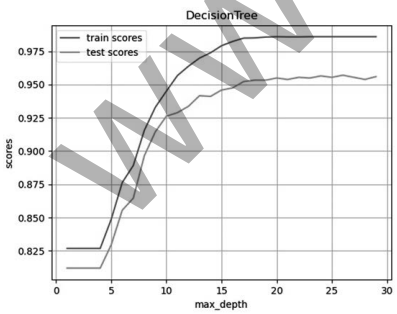

图1 决策树参数选择

Fig.1 Decision tree parameter selection

3.4 随机森林模型

(1)生成随机森林

样本按照8:2分为训练集及测试集，通过Python scikit-learn包中的RandomForestClassifier建立初始随机森林模型，默认的评价指标为基尼系数，深度为40，树的个数为300棵，从结果来看，训练集正确率为88%，测试集正确率为88%。而模型还可以通过调参来进一步提高准确率。

(2)参数的调整

随机森林模型中参数的设置也是至关重要的。第一个重要的参数为决策树的数量。理论上，树的数量越多，参与投票的分类器数量越多，从而具有更高的分类准确率，但是另一方面，树的数量越多运行的速度会越慢。本文通过多组实验分析结果得出当数量为300时分类性能最佳。而树的深度也是一个重要的参数，通过实验对比得出图2，由图可知，当随机森林中树的深度为15时，性能达到最佳。

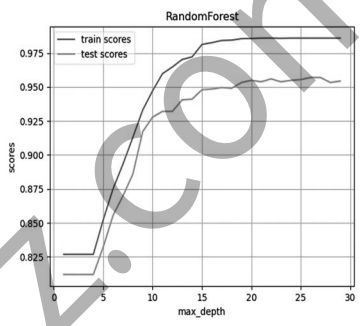

图2 随机森林参数选择

Fig.2 Random forest parameter selection

4 分类模型评估(Performance evaluation of classification models)

4.1 评估指标

由于这是一个“分类”问题，所有分类器将使用以下评价指标来评价。

混淆矩阵如表1所示，又被称为错误矩阵，被用来评价分类器的性能。在多分类任务中，混淆矩阵同样适用。

表1 混淆矩阵

Tab.1 Confusion matrix

真实值 \ 预测值	+	−
+	TP	FN
−	FP	TN

在混淆矩阵中，TP表示实际为正样本，结果为正样本；FN表示实际为正样本，预测结果为负样本；FP表示实际为负，预测结果为负；TN表示实际为负，预测为负。对于多分类问题，通过分析混淆矩阵可以查看模型预测错误的地方。

另外，我们将在数据集上使用“交叉验证”(在我们的实验中使用10倍交叉验证)方法，并取平均精度。这将使我们对分类器的准确性有一个整体的认识。

使用“管道”的方法结合所有预处理和主要分类器的计算。ML“管道”将所有处理阶段包装在一个单元中，并充当

“分类器”本身。通过这种方式，所有阶段都变得可重用，并可以形成其他“管道”。

跟踪每一种方法在构建和测试中的总时间。

对于以上内容，主要使用来自Python的scikit-learn包来实现。

4.2 模型性能比较分析

各个模型在最终的数据集上的准确率评价如表1所示。基中A类错误率和B类错误率为将原来的10分类问题简化为二分类问题得到的结果。这个结果可以辅助决策。所有的结果为模型运行20次得到的均值数据。

表2 三类模型参数调整后错误率与准确率的对比分析

Tab.2 Contrast of the three models

模型	A类错误率	B类错误率	总正确率	运行时间
Logistic	1.8732	5.3647	0.8326	1.376517
C5.0	0.5314	3.2456	0.9479	0.217938
随机森林	0.6219	4.2198	0.9581	2.579914

从所有分类器中可以清楚地看出，就精确度而言，随机森林性能最佳。但是如果我们把“花费的时间”和“准确度”放在一起，那么C5.0是一种较好的选择。但我们也看到了如何使用简单的线性分类器，如“逻辑回归”与适当的特征工程，以提供更好的准确性。其他分类器不需要那么多的特性工程工作。

它依赖于可用的数据需求、用例和数据工程环境来选择一个完美的分类器。结合具体的问题可以选择合适的分类模型。如果对实时性要求较高决策树可以考虑，如果对精度要求高随机森林是一个较好的模型。

进一步地，对于随机森林模型，为了分析哪些类别容易误判，并找出原因，实验得出了随机森林的混淆矩阵。如图3所示，从混淆矩阵我们可以看出，大多数出现错误就是被误判成了0，比较明显的为0和13，0容易被误判成13。通过查看原始数据标签含义可知，0和13本身具有很高的相似性，需要人为进行区分。因此排除了这个因素，随机森林可以达到更高的准确率。

```
[[1518    0    0    2    0    0   10    0    0    0    0]
 [   0    7    0    0    0    1    0    0    0    0    0]
 [   0    0   10    0    0    0    0    0    0    0    0]
 [   4    0    1   10    0    0    0    0    0    0    0]
 [   0    0    0    0    6    0    0    0    0    0    0]
 [   6    1    0    0    0  243   18    0    0    1    0]
 [  13    0    0    0    0   21    8    0    0    0    0]
 [   0    0    0    0    0    0    0    2    0    0    0]
 [   1    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    1]]
```

图3 随机森林的混淆矩阵

Fig.3 Confusion matrix

5 结论(Conclusion)

本文基于几种常见的多分类模型，对海关风险数据进行定性定量分析。其主要贡献如下：

其一，筛选出对海关风险数据影响比较大的几种特征。通过本文的特征筛选对于海关数据分析具有指导性影响。

其二，论文用Logistic回归，决策树C5.0、随机森林等机器学习模型对海关数据进行分析，发现了各种模型的优劣。

其三，论文虽然以建模分析为主，但是包含了数据挖掘的全部流程。对于其他的政务数据分析建模具有指导性作用。

实验验证了这种利用智能化的方法克服人工风险布控中的不足的有效性，海关系统可以通过智能化风险布控，进一步维护国门口岸安全。

参考文献(References)

[1] 黄锦霞,张璇,郝振宇.深圳海关:“大数据”打造立体风险防控[J].中国海关,2018(1):26-28.

[2] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression[M]. John Wiley & Sons, 2013.

[3] Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data[J]. Remote sensing of environment, 1997,61(3):399-409.

[4] Pandya R, Pandya J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning[J]. International Journal of Computer Applications, 2015,117(16):18-21.

[5] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002,2(3):18-22.

作者简介：

金　瑾(1988-)，女，硕士，讲师.研究领域：人工智能，大数据.

王正刚(1982-)，男，硕士，工程师.研究领域：人工智能，信息系统.

刘　伟(1969-)，女，本科，工程师.研究领域：人工智能，信息系统.

巫家敏(1976-)，男，博士，教授.研究领域：人工智能，大数据.

李　波(1981-)，男，博士，副教授.研究领域：人工智能，大数据.