安徽农业大学学报, 2015, 42(3): 473-477
Journal of Anhui Agricultural University
[DOI] 10.13610/j.cnki.1672-352x.20150424.005　网络出版时间：2015-4-24　14:33:00
[URL] http://www.cnki.net/kcms/detail/34.1162.S.20150424.1433.005.html

基于机器学习的焦油预测模型研究

郭东锋[1]，刘新民[2]，姚忠达[1]，舒俊生[1]，胡海洲[2*]
(1. 安徽中烟工业有限责任公司技术中心，合肥 230088；2. 中国农业科学院烟草研究所，青岛 266101)

摘　要：为研究卷烟焦油预测模型，以焦油的释放量为研究对象，运用不同的回归方法进行焦油预测研究，以各个模型的标准化均方误差为评判尺度，对各个模型的预测效果进行了比较。结果表明，各模型的预测精度差别较大，整体来看机器学习方法对于焦油的预测精度较高，其中以随机森林算法回归对于焦油的预测精度最高，表现出较高的预测精度和良好的稳定性，其次表现较好的机器学习算法为支持向量机回归方法。因此，在焦油预测应用或研究中可以运用随机森林或其他机器学习方法对焦油进行建模预测。

关键词：机器学习；焦油；回归模型；预测
中图分类号：TS411　　**文献标识码：**A　　**文章编号：**1672–352X (2015)03–0473–05

Study on predicting models for tar yields of cigarettes based on machine learning

GUO Dongfeng[1], LIU Xinmin[2], YAO Zhongda[1], SHU Junsheng[1], HU Haizhou[2]
(1. Technology Center of Anhui Cigarette Industrial Company Co., Ltd., Hefei 230088;
2. Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao 266101)

Abstract: To improve the accuracy of predicting tar yield in cigarettes, several machine learning methods and the ordinary liner regression were used to predict tar yield. The standardized mean square error was set as the criterion to judge the model's predicting accuracy. The results indicated that significant differences among individual regression models were observed. The machine learning methods showed a higher accuracy of predicting tar yield than that of the traditional simple liner regression. Random forest regression performed the best for predicting tar yield in these models and its performance was stable and precise. The second best model should be the support vector machine regression. Thus, machine learning methods could be widely applied in predicting tar yield and other tobacco research areas.

Key words: machine learning; tar yield; regression models; predict

近几年，随着禁烟控烟浪潮的高涨，“降焦减害”日益成为行业的焦点和热点问题，也成为影响卷烟品牌发展的敏感性问题[1-3]。围绕“降焦”科研工作者进行了大量研究，其中包括焦油释放量的预测模型研究[4-5]，有投影寻踪回归分析[6]、线性回归[7-8]、支持向量机[9]等。机器学习是近几年在计算机、概率论、数理统计、算法复杂度等领域迅速发展的交叉学科[10-11]。由于机器学习具有高效、不局限与理想的场景、有别于传统的统计学和概率论、运算速度快及预测精度高等诸多优点[12-13]，在科研领域逐渐成为研究人员的得力研究工具。在焦油预测分析中，运用机器学习开展的研究尚不多见，本研究以卷烟焦油为研究对象，采用集中机器学习方法对焦油进行预测，比较各算法的性能，以期为“降焦减害”工作提供参考。

1　材料与方法

1.1　材料

44 份烤烟烟叶样品，其中，2010 年全国 26 个烤烟产区、6 个品种、17 个不同等级共计 40 份片烟样品，以及巴西、津巴布韦、美国、赞比亚中部叶 4 份。辅材规格：盘纸为华丰 60 CU，嘴棒吸阻 2800 Pa。

收稿日期：2014-11-26
基金项目：安徽中烟工业有限责任公司科技项目“‘黄山’品牌皖南特色烤烟品种筛选与评价”（2013128）和烟叶生产等级结构优化技术研究”（2014125）共同资助。
作者简介：郭东锋，博士，高级农艺师。E-mail：gdf0221@163.com
* **通信作者：**胡海洲，助理研究员。E-mail：huhaizhou@caas.cn

傅里叶变换近红外光谱仪（美国 Thermo Fisher 公司）；RM200A 型吸烟机（德国 Borgwaldt KC 公司）；Aglient 7890A 气相色谱（美国 Agilent 公司）；KBF 系列—P720 恒温恒湿箱（德国 Binder 公司）。

1.2 方法

1.2.1 烟叶化学成分检测 按照标准 YC/T 159-2002《烟草及烟草制品 水溶性糖的测定 连续流动法》中规定的方法测定烟叶中总糖和还原糖，按照标准 YC/T 160-2002《烟草及烟草制品 总植物碱的测定 连续流动法》，YC/T 161-2002《烟草及烟草制品 总氮的测定 连续流动法》，YC/T 173-2003《烟草及烟草制品 钾的测定 火焰光度法》和 YC/T 162-2002《烟草及烟草制品 氯的测定 连续流动法》中规定的方法分别测定烟叶中的烟碱、总氮、总钾和总氯。

1.2.2 卷烟主流烟气成分检测 采用同一规格辅材（“三纸一棒”）按照成品卷烟标准，将每种烟叶样品卷制成试验卷烟。样品的制备按 GB/T 606.1-2004《卷烟 第 1 部分：抽样》进行。按 GB/T 6447-2004《烟草及烟草制品 调节和测试的大气环境》调节卷烟水分。按 GB/T 9069-2004《卷烟 用常规分析用吸烟机测定总粒相物和焦油》测定焦油量。

1.3 算法简介

1.3.1 普通最小二乘回归（ordinary linear square）线性回归是利用数理统计中的回归分析，来确定 2 种或 2 种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛，可参考统计学相关书籍。鉴于此，此处不做详细介绍。

1.3.2 回归树（regression tree） 决策树（decision tree）是数据挖掘中的重要算法，用于回归分析则成为分类回归树（regression tree）。它描述给定预测向量值 X 后，变量 Y（Y 连续型为回归，Y 离散型为分类）条件分布的一个灵活的方法，对噪声数据具有很好的鲁棒性。具体可参考文献[12-13]。

1.3.3 Boosting 回归（boosting） Boosting 是一种提高任意给定学习算法准确度的方法。它的思想起源于 Valiant 提出的 PAC（Probably approximately correct）学习模型。Boosting 方法每个单个的分类器的识别率不一定很高，但他们联合后的结果有很高的识别率，具体可参考文献[14-18]。

1.3.4 自助抽样（Bagging） Bagging 是 bootstrap aggregating（自助抽样合集）的缩写，它首次介绍是在 Breiman 第 1 批用于多分类器集成算法，具体可参考文献[13-14,19-21]。

1.3.5 随机森林回归(Random forest regression,RF) 随机森林又叫 Random trees，是一种由多棵决策树组合而成的联合预测模型，天然可以作为快速且有效的多类分类模型，可参考文献[11,14,17]。

1.3.6 支持向量机回归(Support vector machine regression) 支持向量机回归是 Cortes 和 Vapnik 等于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中，可广泛地应用于统计分类以及回归分析。参考文献[10,14,22,24-26]。

1.3.7 模型预测精度评价方法 为了比较各模型的预测精度，本例定义预测值的标准化均方误差为评价各模型精度的指标，此处标准化均方误差（NMSE）的定义为：

$$NMSE=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-Y'\right)^2}$$

式中：Y_i 为实测值，$\overline{Y}$ 为实测值均值，Y'为从训练集得到模型对一个数据集的预测值。

1.4 数据处理方法

数据处理在 excel 2003 中实现、统计及作图均在 R3.0 软件中完成，对于数据集采用 5 折交叉验证（5-fold cross validation）。模型建立过程中通过随机建立的 5 个训练集建立相应方法的训练模型，对训练集合测试集分别得到 5 个标准化均方误差（NMSE），求 5 个 NMSE 均值。

2 结果与分析

2.1 样品统计描述及相关分析

烟叶常规化学成分及卷烟主流烟气成分的统计描述结果（表 1）表明：焦油释放量分布均为平顶峰，左偏态峰。烟叶化学成分样体间存在较为广泛的变异，其中总氯变异最大，达到了 62.40%，且总氯含量数据分布右偏态尖顶峰，离散程度较大；总氮变异最小，变异系数达到 12.37%，烟碱、总糖、还原糖、总钾含量变异在 20%~30%；其余指标均为平顶峰。说明烟叶样本间常规化学成分的差异较为明显。由相关分析（图 1）可知，焦油的释放量与烟叶中烟碱、总氮和总钾含量存在显著的相关关系。

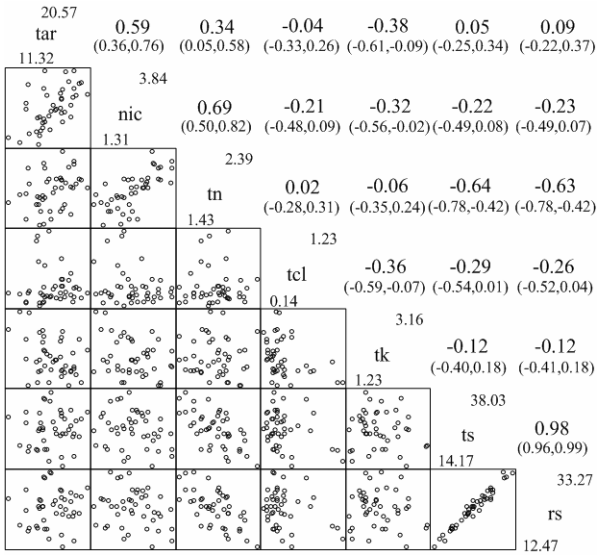
2.2 模型构建

首先构建线性回归模型，由图 1 可知自变量间存在多重共线性，因此回归方法采用逐步回归，最终化学成分中只有烟碱与焦油关系显著，线性方程模型诊断见表 2 及表 3。

表 1　烟叶常规化学与卷烟主流烟气成分统计描述

Table 1　Descriptive statistics of the main chemical compositions in flue-cured tobacco leaves and main stream smoke

指标 Index	变幅 Amplitude of variation	均值 Mean	方差 Variance	偏度 Skewness	峰度 Kurtosis	标偏 *SD*	正态性 Normality (sig)	变异/% Variation
焦油量/mg·cig^{-1} Tar yield	11.32～20.57	16.49	4.41	–0.22	–0.39	2.10	0.85	12.73
烟碱/% Nicotine	1.31～3.84	2.58	0.45	0.06	–0.75	0.67	0.62	26.02
总糖/% Total sugar	14.17～38.03	26.04	32.49	–0.04	–0.52	5.70	0.84	21.89
还原糖/% Reducing sugar	12.47～33.27	23.18	26.02	–0.04	–0.42	5.10	0.74	22.01
总氯/% Total Cl	0.14～1.23	0.41	0.07	1.90	3.30	0.26	0.06	62.40
总钾/% Total K	1.23～3.16	1.97	0.24	0.61	–0.21	0.49	0.12	24.69
总氮/% Total N	1.43～2.39	1.88	0.05	0.00	–0.51	0.23	0.74	12.37



左下角为相关散点矩阵，右上角为相关系数(pearson)及置信区间

The left bottom of the figure means related scatter dot matrix, and upper-right corner displays correlation coefficient (pearson) and confidence interval

图 1　焦油量与烟叶常规化学成分相关矩阵图

Figure 1　Correlation analysis scatter plot between tar and chemical components

表 2　焦油与常规化学回归模型验证

Table 2　Model checking for regression models of tar

R^2	调整 R^2 Adjusted R^2	估计标准误差 Estimated *SE*	*Sig.*	Durbin-Watson
0.3511	0.3356	1.7113	<0.001	1.939

表 3　焦油回归模型回归系数检验

Table 3　ANOVA for regression models of tar

指标 Index	系数 Coefficient	标准误差 *SE*	*t*	*Sig.*
常量 Major element	11.7037	0.7992	10.016	<0.001
烟碱 Nicotine	1.8567	0.1969	13.900	<0.001

其次，以焦油为研究目标变量，以烟叶常规化学成分为自变量，运用回归树、基于模型的 boosting 回归、自助抽样（bagging）回归、随机森林回归和支持向量机回归算法对焦油释放量进行了预测建模，集中机器学习方法均进行 5 折交叉验证，构建训练集和测试集。模型预测精度评价采用 NMSE 指标进行对比分析。

各模型训练及预测结果见表 4 及图 2，从训练集预测结果来看随机森林回归算法预测精度最高，其焦油预测结果的标准化均方误差仅为 0.1436，其次为基于模型的 boosting 方法回归，其焦油预测结果的标准化均方误差为 0.2894，其他模型针对训练集的回归预测精度依次为普通最小二乘回归（*NMSE*=0.4881）＞自助抽样回归（*NMSE*=0.4927）＞支持向量机回归（*NMSE*=0.5613）＞回归树（*NMSE*=0.5747）；从测试集训练的精度来看，仍然是随机森林回归方法的测试集预测精度最高，其测试集标准化均方误差仅为 0.7534，其次为支持向量机回归的测试集预测精度达到了 0.7574，而简单线性回归和回归树两种算法对测试集的预测精度超过了 1，由于标准化均方误差为预测值和均值的标准化均方误差的比值，因此，如果其超过 1 说明预测精度很低，模型已经不适合用来进行预测。其他模型的测试集预测精度依次为自助抽样回归＞基于 boosting 模型的回归方法。

由图 2－图 4 及表 4 可以看出，从标准偏差来看机器学习的预测结果较线性模型离散性普遍有所降低，同时也可以看出除了回归树模型较为特殊外，机器学习模型预测值与实测焦油值有较好的拟合，由相关分析（预测值与真实值）可以看出：虽然各个模型预测值与实测值之间相关都达到了极显著水平，但是相关关系强弱顺序为随机森林回归（0.89）＞boosting 回归（0.74）＞bagging 回归（0.69）＞支持向量机回归（0.63）＞普通最小二乘回归（0.59）

表 4 机器学习方法性能训练结果比较

Table 4 Comparison of the results with different machine learning methods

机器学习方法 Machine learner method	训练集的标准化均方误差 *NMSE*	测试集的标准化均方误差 *NMSE*	均值 Mean	标准偏差 *SD*	变异系数/% *CV*
简单线性回归 Simple liner regression	0.4881	1.1989	16.49	2.10	7.55
回归树 Regression tree	0.5747	1.1721	16.49	1.24	6.27
基于模型的 Boosting *m*-boosting	0.2894	0.9205	16.62	1.04	8.02
自助抽样 Bagging	0.4927	0.8231	16.61	1.33	6.46
随机森林回归 Random forest regression	0.1436	0.7534	16.51	1.07	7.58
支持向量机回归 Support vector machine regression	0.5613	0.7574	16.65	1.26	6.46

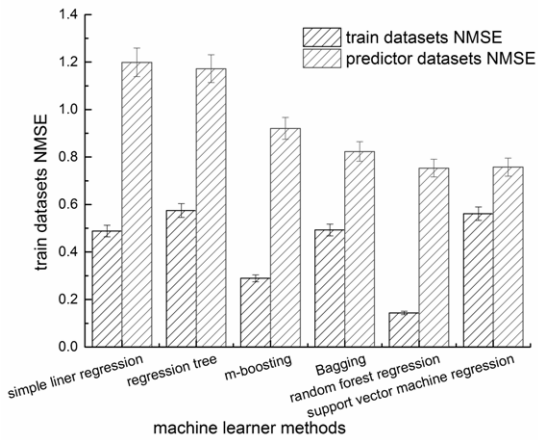


图 2 6 种模型训练集及预测集预测精度比较

Figure 2 Accuracy comparison based on six machine learning models

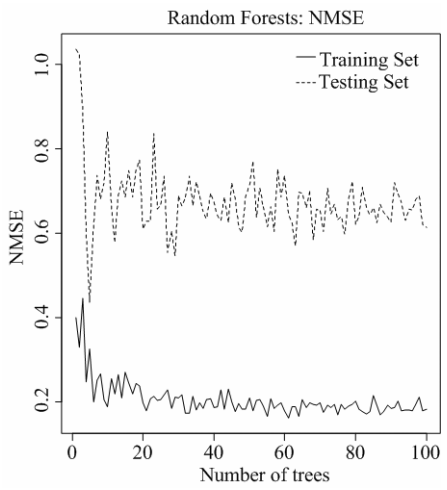


图 3 随机森林预测性能稳定性分析

Figure 3 Stability analysis of random forest model

＞分类回归（0.52）。由于回归树模型源自决策树的剪枝思想，因此在其对焦油连续型变量的预测结果不符合我们对焦油预测的预期。

3 小结与讨论

通过对焦油用机器学习算法进行预测分析，综合各种算法的预测精度以及对于测试集的稳定性来看，随机森林回归算法对于焦油的预测有较高的预测精度，且表现出良好的稳定性能，其次是支持向量机回归对于焦油的预测精度较好，回归树算法对焦油的预测出现了过拟合现象（图 4）。因此，在烟草分析中可以适当引入机器学习算法，尤其是随机森林算法，可以有效提高对于连续型响应变量的预测效能。

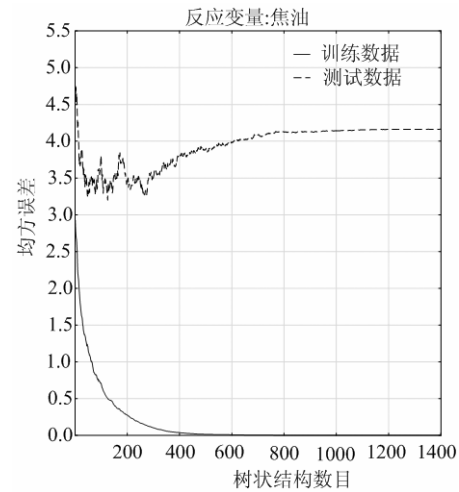


图 4 回归树模型预测性能稳定性分析

Figure 4 Stability analysis of regression tree model

虽然机器学习有诸多优点，但是如果参数设置、组合设置等不当，则会存在模型的过拟合问题。由于本研究仅限于此 5 种机器学习方法，且对其工作原理研究还不够，因此在烟草科研工作中的拓展应用尚有待继续深入研究。

参考文献：

[1] 汪银生, 路国行, 王晓婷. 论“反吸烟运动”在烟草科技进步中的地位和作用[J]. 中国烟草科学, 1999(4): 44- 46.

[2] 雷樟泉, 杨进, 储国海, 等. 我国卷烟降焦历程回顾、现状与展望[J]. 烟草科技, 2003(5): 29-31.

[3] 朱尊权, 穆怀静, 方淑杰. 我国卷烟焦油的现状和问题[J]. 烟草科技, 1987(6): 18-19.

[4] 王允白. 烤烟原料总粒相物与烟叶内在化学成分关系及预测模型研究[J]. 中国烟草学报, 1998(2): 1-5.

[5] 梁德成, 王德吉, 邱道尹, 等. 卷烟焦油预测研究[J].

东南大学学报: 自然科学版, 2009(S1): 195-198.
[6] 殷发强, 李丹, 宋旭艳, 等. 投影追踪回归(PPR)法建立卷烟焦油预测数学模型[J]. 烟草科技, 2009(6): 15-18.
[7] 张志刚, 王二彬, 苏东赢. 卷烟常规化学成分与焦油的线性回归分析[J]. 烟草科技, 2003(11): 32-33.
[8] 刘华. 卷烟材料与焦油量关系的回归设计与分析[J]. 烟草科技, 2008(5): 9-11.
[9] 王德吉, 李广才, 栗卫军. 基于信息几何的卷烟焦油SVM(支持相量机)预测[J]. 中国烟草学报, 2009(4): 22- 25.
[10] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.
[11] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.
[12] Gammerman A. Machine Learning: Progress and Prospects[M]. Royal Holloway: University of London, 1996.
[13] Kubat M. Introduction to machine learning. Lecture Notes in Computer Science, 1992, 617: 104-138.
[14] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[15] Rätsch G, Onoda T, Müller K R. Soft margins for adaboost[J]. Machine Learning, 2001, 43(3): 287-320.
[16] Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization[J]. Machine Learning, 2000, 40(2):139-157.
[17] Chernick M R.Bootstrap Methods:A Guide for Practitioners and Researchers[M]. 2nd ed. New York: Wiley, 2007.
[18] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]//Proceedings of the thirteenth international conference on machine learning. 1996.
[19] Breiman L, Friedman J, Olshen R, et al. Classification and regression trees (Wadsworth Statistics/Probability)[M]. CRC Press, 1984.
[20] Mooney C Z, Duval R D. Bootstrapping: A nonparametric approach to statistical inference[M]. SAGE Publications Inc, 1993.
[21] Good P. Permutation, Parametric and Bootstrap Tests of Hypotheses[M]. Springer-Verlag New York Inc, 2005.
[22] Wang J Q, Tao Q, Wang J. Kernel projection algorithm for large-scale SVM problems[J]. Journal of Computer Science and Technology, 2002, 17(5): 556-564.
[23] Joachims T. Transductive inference for text classification using support vector machines[C]//Proc 16th international conference on machine learning. 1999: 200-209.
[24] Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002, 46(1/3): 131-159.
[25] Sebald D J, Bucklew J A. Support vector machine techniques for nonlinear equalization[C]//IEEE transactions on signal processing. 2000.
[26] Valiant L G. A theory of the learnable [J]. Communications of the ACM, 1984, 27(11): 1134-1142.