R 语言机器学习之旅：实战与分析

李婷

（东北大学 数学与统计学院）

摘　要：本文运用 R 语言进行相关建模分析。首先，分别对 fifa.csv、sales.csv 和 loan.csv 数据集单独进行预处理，并采用随机森林模型进行训练和评估，结果显示模型在 loan 数据集上表现最佳，具有较高的分类准确率、Kappa 系数和 AUC 值。其次，针对 weather.csv 数据集中存在的问题进行修正，经数据清理，共保留了 582 条数据，且 Weather 列仅留下了 5 个类别，在此基础上使用 SVM、贝叶斯分类和 Bagging 集成学习进行训练，同时通过网格搜索对 SVM 的参数进行优化。实验结果显示，SVM 在分类准确率、宏召回率、宏精确率和 F1 分数等指标上均胜出，在测试集上的分类准确率达到 81%。此后，借助方差分析，判断出风速对 Weather 列所示的天气具有显著影响。最后，本文将 R 语言与 neo4j 结合，构建了一个简单的知识图谱。

关键词：随机森林；SVM；方差分析；知识图谱

中图分类号：请查阅中图分类号　**文献标志码：**A

Journey into Machine Learning with R: Practical Experience and Analysis

LI Ting

(School of Mathematics & Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei 066004, China. E-mail: 1424743982@qq.com)

Abstract： This article conducts comprehensive modeling analysis using the R programming language. Initially, individual preprocessing was performed on the fifa.csv, sales.csv, and loan.csv datasets, followed by training and evaluation using a random forest model. The results revealed the model's exceptional performance on the loan dataset, exhibiting high accuracy, Kappa coefficient, and AUC values. Subsequently, issues within the weather.csv dataset were addressed through meticulous data cleansing, retaining 582 records. The Weather column was refined to only encompass five categories. Building upon this, training utilized SVM, Bayesian classification, and Bagging ensemble learning techniques, with SVM parameters optimized through grid search. Experimental outcomes showcased SVM's superiority across metrics such as accuracy, macro recall, macro precision, and F1 score, achieving an 81% accuracy on the test set. Additionally, employing variance analysis, the study identified significant influences of wind speed on the weather patterns indicated in the Weather column. Finally, the article integrated R language with neo4j, constructing a simplistic yet insightful knowledge graph.

Key words： random forest; SVM; analysis of variance; knowledge graph

“机器学习”一词是 Arthur Samuel 在 1959 年正式提出[1]，定义机器学习为“能够使计算机系统具备学习能力，而无需明确地进行编程”的领域[2]。在随后的几十年中，机器学习不断演进，涌现出各种突出的算法，包括 Vapnik 提出的支持向量机[3]、Leo Breiman 提出的随机森林[4]和 Bagging 集成学习[5]等。时至今日，机器学习这座大厦仍不断完善中，吸引着国内外无数科学家和学者为之着迷，共同推动机器学习在当下的发展。

本文选用 R 语言作为编程语言，在给定的数据集下实现多种机器学习算法，并分析数据之间的关联性。

1.1　fifa、sales、loan 数据预处理

1.1　fifa 数据集预处理

fifa 数据集中包含了来自 Argentina 等 32 个国家的足球队员信息，完整记录了 736 位足球运动员出生日期、身高、体重等 12 条特征，未存在缺失数据（如图 1 所示）。

图 1 所展示的字段中，team、position、birth_date、shirt_name、club、league 和 name 这八列为字符型，number、height、weight、age 和 caps

收稿日期：2023-11-01

基金项目：R 语言结课论文。

作者简介：李婷(2003−),女,湖南省娄底人, 东北大学学生,本科生.

这五列为数值型。可见，fifa 各列数据的类型较为杂乱，为了便于后续的建模分析，需要事先进行数据预处理。

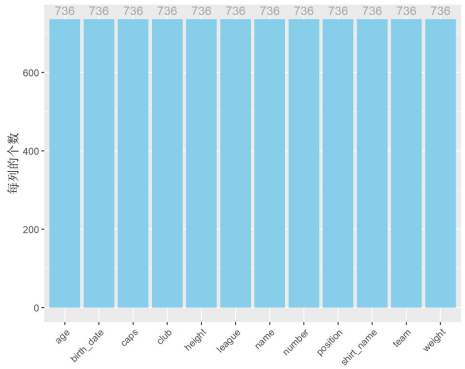

图 1 fifa 数据集各列的取值个数

Fig.1 Number of values in each column of fifa data set

- 选取 position 列（各队员在足球比赛中的位置标识）作为标签列，该列共有四种类型：GK 为守门员，DF 为后卫，MF 为中场，FW 为前锋。这些标识代表了球员在足球比赛中的不同位置和角色；
- shirt_name、name 和 club 列类似于球员的个人标识信息，具有特定性，对后续的分析帮助不大，故本文剔除这三列；此外，由于 age 列更能准确地表达球员的年龄信息，故剔除“birth_date”列；
- 剩余的字符列中，将 team 和 league 这两列的数据转换为整数，以替代原本的字符数据。

1.2 sales 数据集预处理

sales 数据集中提供了不同地区和国家的产品销售信息。经统计，原数据中共有 5000 条记录，11 个特征字段。各列的缺失值情况如图 2 所示，其中 Item_Type 存在 11 个缺失值，Order_ID 存在 3 个缺失值。

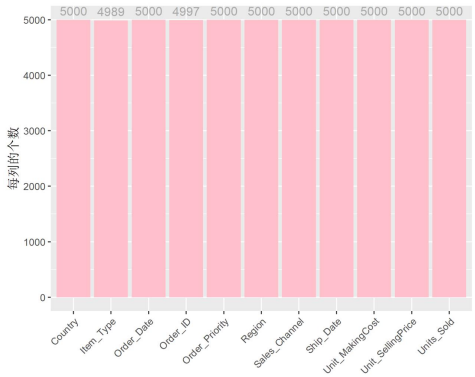

图 2 sales 数据集各列的取值个数

Fig.2 The number of values in each column of the sales dataset

图 2 所展示了字段中，Region、Country、Item_Type 、Sales_Channel 、Order_Priority 、Order_Date 和 Ship_Date 这七列为字符型，其余为数值型。可见，sales 的数据类型也不一致，需要进行数据处理。

- 选定 Order_Priority 作为标签列，共包含 C、H、L、M 这四个字母代码；
- 剔除 Item_Type 和 Order_ID 列缺失值所在的行；
- 剔除 Order_ID 列；
- 以 1970 年 1 月 1 日为标准，分别计算出每条记录中 Order_Date 和 Ship_Date 的日期距离天数；
- 将剩余的字符列通过编码转换成数值型。

对 sales 数据集进行如上处理后，共有 4986 条信息，9 个特征字段和 1 个标签列。

1.3 loan 数据集预处理

loan 数据集记录了与借贷有关的信息，原数据集共包含 614 条借款记录，以及对应的 13 个特征字段。经统计，部分特征所记录的信息存在缺失，具体情况如图 3 所示。

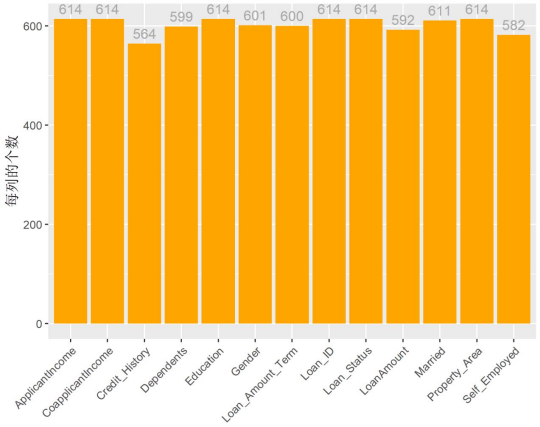

图 3 loan 数据集各列的取值个数

Fig.3 Number of values in each column of loan dataset

loan 原数据中，Loan_ID、Gender、Married、Eduaction、Self_Employed、Property_Area 和 Loan_Status 这七列为字符型，Dependents 列包含部分字符，其余均为数值型，可见也存在数据类型不一致的现象，需要对相关数据进行处理。

- 选定 Loan_Status 列作为标签列，包含 N 和 Y 两种贷款状态；
- 剔除所有缺失值所在的行；
- 剔除 Loan_ID 列；
- 将剩余的字符型通过编码转换成数值型。
- 单独处理 Dependents 列，将每单个字符“3+”转换成“3”。

2 随机森林的评估结果

随机森林（Random Forest, RF）是一种强大的集成分类算法，其核心思想在于通过构建多个决策树来有效降低单个决策树的过拟合风险。每个决策树都在不同的样本和特征子集上进行训练，引入这种随机性可以降低算法的方差，从而提高模型的泛化能力。

此部分，采用随机森林算法，对经过前文处理的三个数据集（即 fifa、sales 和 loan）进行分析，借助了准确率、Kappa 系数和 AUC 这三个指标，以评估随机森林在不同数据集上的分类性能，并记录运行时间（Time，单位：秒），具体结果见表 1。此外，随机森林在三个数据集上的交叉验证通过，表明未存在过拟合现象，证明表 1 中得到的结果有效。

表 1 随机森林在 fifa、sales 和 loan 数据集上的评估结果

Tab.1 Evaluation results of random forest on fifa, sales and loan data sets

	准确率	Kappa 系数	AUC	Time
fifa	0.5586	0.3772	0.5929	0.3617
sales	0.2462	-0.0078	0.4940	5.0928
loan	0.7895	0.4385	0.6938	0.2428

根据表 1，综合分析来看，随机森林在 loan 数据集上表现最为出色，以 0.7895 的分类准确率明显优于 fifa 数据集的 0.5586 和 sales 数据集的 0.2462。值得一提的是，在运行时间方面，loan 数据集的处理速度远快于其他两个数据集，显示出更低的时间复杂度。然而，评估指标的差异也突显出随机森林分类在 sales 数据集上的不足。具体而言，随机森林在 sales 数据集上的表现较差，不仅准确率和 AUC 均低于 0.5，而且 Kappa 系数接近 0，这表明模型未能很好地捕捉到 sales 数据集中标签列和特征列之间的关联。

3 weather 数据预处理

weather 数据集主要记录了不同的天气状况，共有 8784 条天气信息，每条信息对应着 8 个特征字段，且不存在缺失值。为简化流程，将日期时间（Data/Time）视为无关因素并予以剔除，将 Weather 列作为分类的标签列，用于后续分析。

3.1 异常值处理

利用箱线图对数值特征列进行可视化，如图 4 所示，除了 Temp_C 和 Dew Point Temp_C 列以外，其余列或多或少都存在一定数量的异常值，为避免异常值对模型稳健性的影响，剔除包含异常值的所有行，剩下的数据集中共有 5082 条记录。

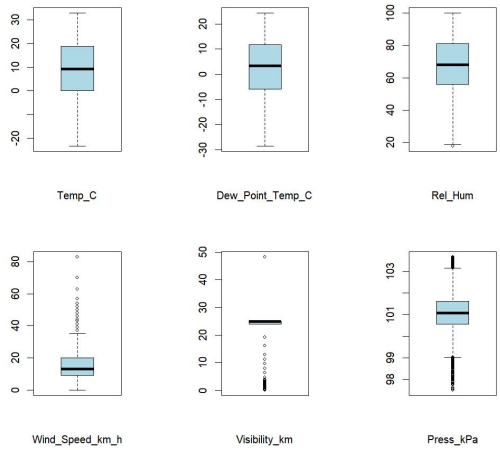

图 4 weather 数值列的可视化箱线图

Fig.4 Visual box diagram of weather numerical column

3.2 数据集划分

异常值处理后，标签列 Weather 下各类别的数量如图 5 所示。

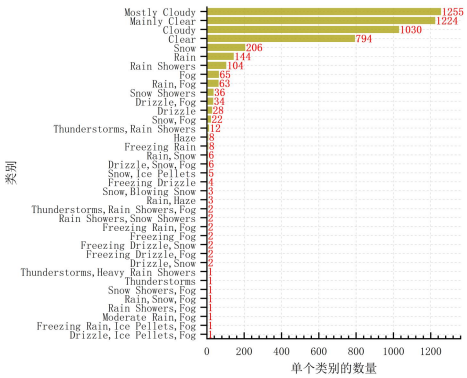

图 5 异常值处理后 weather 列各类别的数量

Fig.5 Number of categories in weather column after outlier processing.

显然，Weather 列的不同类别之间存在着明显的数量不平衡问题。当某些类别的样本数量较少时，模型往往无法充分学习它们的特征，从而在分类任务中容易出现错误判断；相反，当某个类别的样本过多时，模型可能会过度拟合这些样本，而忽视其他类别，导致欠拟合。鉴于这一问题，本文采取了一种双端平衡策略，观察到类别数量在 50 到 500 之间的差异较小，因此选择保留这一范围内的类别。

最终，筛选出了 582 条数据，留下的各特征列之间的皮尔逊相关系数热力图如图 6 所示，易知，Dew.Point.Temp_C 列和 Temp_C 列之间存在

着非常强的正相关性。为了减小多重共线性的影响，剔除 Dew.Point.Temp_C 列。

表 2 Weather 列的剩余类别

Tab.2 Remaining categories of Weather column

类别	Fog	Rain	Rain Showers	Rain,Fog	Snow
数量	65	144	104	63	206

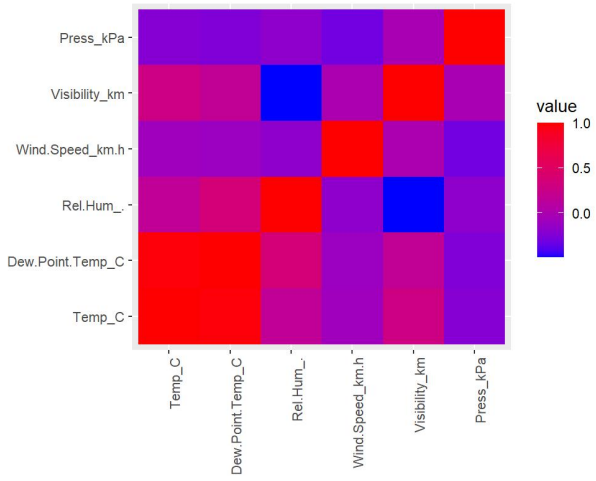

图 6 筛选后特征列之间的相关系数图

Fig.6 Correlation coefficient diagram between filtered feature columns

3.3 标准化

本文采用 Z-Score 方法标准化各特征列，消除量纲的影响。

4 三类模型的评估结果

此部分，本文采用三种不同的算法进行分类测试，分别为贝叶斯分类、网格优化的 SVM 和 Bagging，并将分类准确率（Accuracy）、宏平均召回率（Recall）、宏平均精确率（Precision）、宏平均分数（F1 score）作为评价指标，且各模型在数据集上均通过了交叉验证，确保未出现过拟合现象。具体评估结果如图 7 所示。

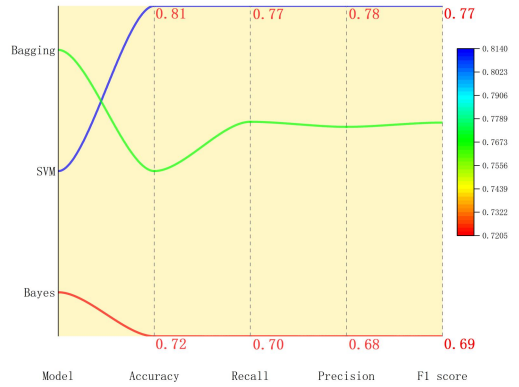

图 7 不同模型的性能评估结果

Fig.7 Performance evaluation results of different models

根据图 7 的结果，观察到 SVM 模型在数据集中表现出色，各项评估指标均明显优于 Bayes 和 Bagging 模型。

表 3 各模型的运行时间（单位：秒）

Tab.3 Running time of each model (unit: seconds)

	Bayes	SVM	Bagging
训练时长	0.0058	0.0505	0.2302

此外，从表 3 中的训练时长可以看出，各模型的时间复杂度相差不大。因此，综合评估来看，对于 weather 数据集，模型的性能排名依次为 SVM > Bagging > Bayes。

5 方差分析

方差分析（ANOVA）是一种用于比较多个组之间是否存在显著差异的统计方法，被广泛应用于确定不同因素对某个连续性因变量产生显著影响的程度。本文使用方差分析来判别各特征对天气（Weather）的显著性大小，具体结果如表 3 所示。

表 4 weather 数据集的方差分析结果

Tab.4 Results of variance analysis of weather data set

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp_C	1	366.7	366.7	292.62	< 2e-16 ***
Rel Hum_%	1	43.4	43.4	34.62	6.79e-09 ***
Wind Speed_km/h	1	69.7	69.7	55.60	3.28e-13 ***
Visibility_km	1	15.0	15.0	12.01	0.00057 ***
Press_kPa	1	4.9	4.9	3.87	0.04962 *
Residuals	576	721.9	1.3		

方差分析（ANOVA）表中不同列的解释如下。

- Df 表示自由度，通常用于衡量模型中可变参数的数量。在方差分析中，有两种类型的自由度，分子自由度（组间）和分母自由度（组内）。分子自由度表示因素之间的差异，而分母自由度表示误差或随机性。在表 4 中，weather 各特征列均有 1 个分子自由度，而“Residuals”表

示误差项，有 576 个分母自由度。

- Sum Sq 表示各组或因素的平方和，展示了观察值与平均值之间的差异的总和。在表 4 中，“Sum Sq”列显示了每个特征的平方和。
- Mean Sq 表示平均方差，是平方和与相应的自由度的比率，即各组的平均方差。在表 4 中，“Mean Sq”是通过将“Sum Sq”除以相应的自由度计算得出的。
- F value 表示 F 统计量，用于比较各组之间差异的统计值，是分子均方与分母均方的比值，通常用于判断各组之间是否存在显著性差异。在表 4 中，“F value”显示了方差分析的 F 统计值，用于衡量各特征对天气的影响程度。
- Pr(>F)表示 p 值，用于衡量 F 统计值的显著性水平，用于确定观察到的差异是否具有统计显著性。在表 4 中，“Pr(>F)”表示了每个特征对天气影响的 p 值。较小的 p 值意味着差异更显著，本文在此处用于判断特征的影响是否显著。

综合上述分析，可知特征的显著性由大到小依次为 Wind Speed_km/h > Temp_C > Rel Hum_% > Visibility_km > Press_kPa，此结果意味着风速的影响对天气的影响最为显著，而能见度和压力的影响相对较小。建议人们在考虑户外活动、旅行或天气变化时，密切关注当天的风速。根据本文分析的结果，尽管能见度和压力的影响相对较小，但是本文未考虑实际中的各种复杂因素，其他因素对天气的影响仍不可忽视，实际中，最好的选择是各有侧重地综合考虑各方面。

6 R 语言的扩展

通过理论课程的学习，可知 R 语言适合统计分析和进行数据的可视化。此外，笔者了解到，知识图谱是一种用于表示和组织知识的数据结构。尽管它们属于不同的领域，但可以通过特定的应用和工具将它们结合起来，以实现更强大的数据分析和知识管理。限于篇幅，本文不去做大量扩展，仅侧重于向读者展示知识图谱的美妙之处（实际的用途不限于本文所述）。笔者以林语堂编写的《苏东坡传》为背景，创建大文学家苏轼的家庭关系网络。

首先，本文使用 R 语言提取了苏轼的关系网络数据，并数据导入 neo4j 中，以绘制知识图谱，如图 8 所示。

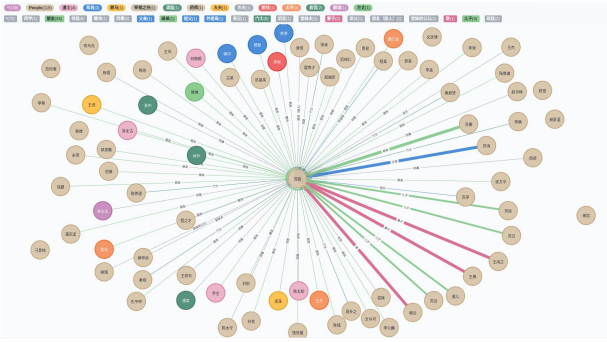

图 8 知识图谱的构造

Fig.8 Construction of knowledge map

7 结论

1)不是所有的模型都能在同一个数据集上展现出良好的性能，也并非所有的数据集都适用于同一个模型；

2)R 语言本身是一门统计学语言，若与其他领域结合可挖掘出更大的潜在价值。

参考文献

[1] 徐浩然,许波,徐可文.机器学习在股票预测中的应用综述[J].计算机工程与应用,2020,56(12):19-24.
(Xu Haoran, Xu Bo, Xu Kewen. Overview of the application of machine learning in stock forecasting [J]. Computer Engineering and Application, 2020,56(12):19-24.)

[2] 羡一鸣. 基于机器学习的中长期和短期电力负荷预测方法研究[D].河北工业大学,2022.
(Xian Yiming. research on medium and long-term and short-term power load forecasting method based on machine learning [D]. Hebei university of technology, 2022.)

[3] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

[4] Breiman L.Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[5] Belsley, D., Kuh, E., & Welsch, R. (1980). “Regression Diagnostics”, John Wiley and Sons.