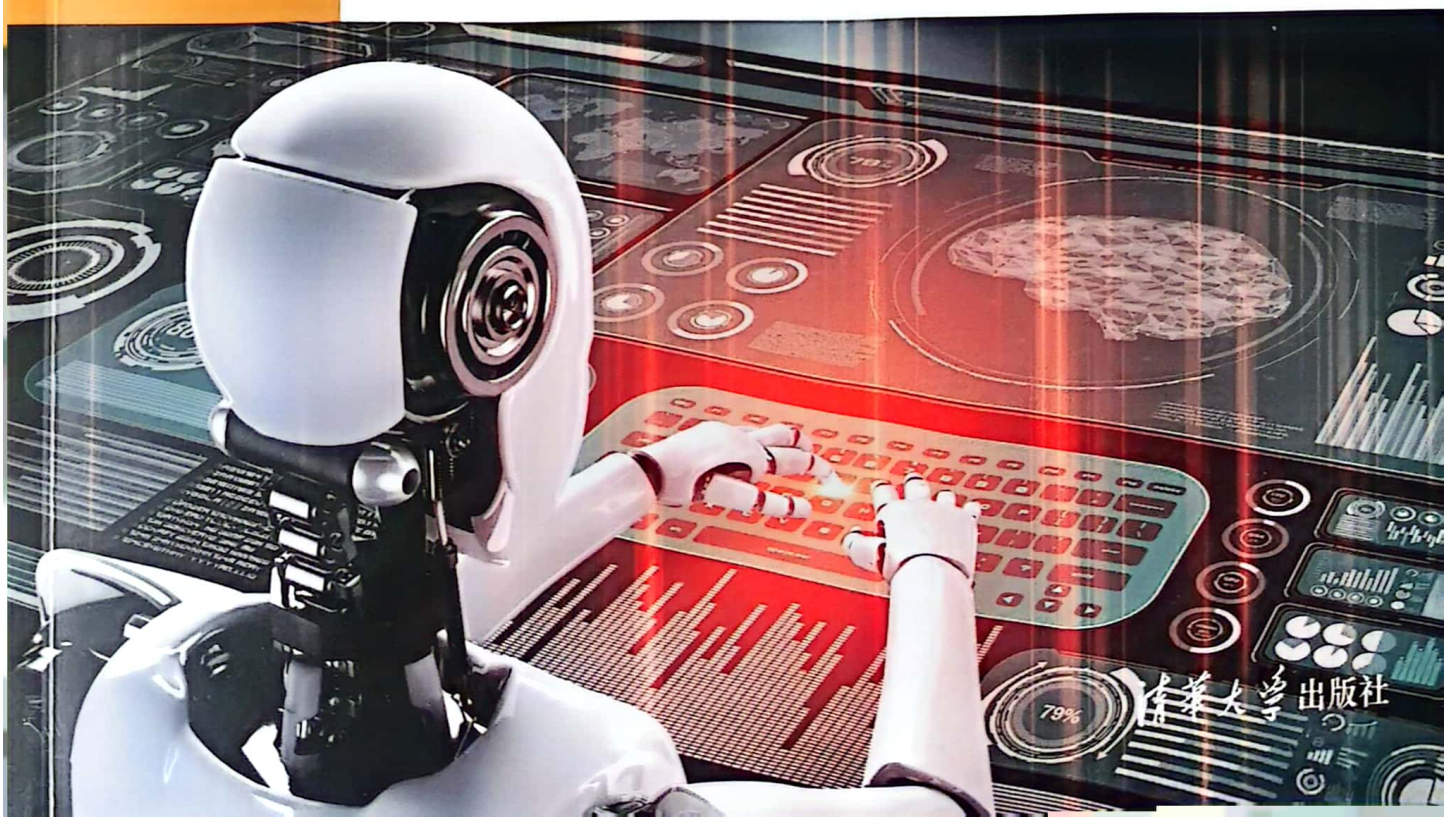
高等学校工业工程类
教指委规划教材

MACHINE LEARNING
INDUSTRIAL BIG-DATA ANALYSIS

机器学习
工业大数据分析

李彦夫 张晨◎编著 宗福季◎主审



清华大学出版社

7.3　5G 通信数据下行传输速率预测

7.3.1　问题背景

第五代移动通信技术（5G）是具有高速率、低时延特点的新一代宽带移动通信技术，对于工业控制、远程医疗、自动驾驶、智慧城市、智能家居、环境监测等对时延和可靠性具有极高要求的垂直行业应用有着极大的推动作用。目前，我国的 5G 通信网正在高速建设当中。对于 5G 通信网络安全性、稳定性和使用速度的预测，是目前工业大数据研究中的一个热点。

针对 5G 的传输特点，5G 基站的建设必须满足“分布式”。据统计，在城市中每几百米、在郊区中每 1000m 左右就需要一个 5G 基站，才能够满足用户 5G 的使用要求。因此，对于城市中的 5G 的数据监控，往往以小区为单位进行。

7.3.2　数据介绍

数据下行传输速率作为影响用户通信上网体验的重要指标，对 5G 网络的效果度量有着重大意义。我们的目标便是对某刚刚完成 5G 基站建设小区的数据下行传输速率进行预测。

所用的数据集有十余个属性，为一系列通信性能参数，包括连接中断率、同频切换成功率、下行平均 MCS、拥塞率、小区下行流量、小区上行流量、下行 PDCP 丢包率等。使用这些参数来预测用户下行平均数据传输速率。这是一个典型的回归预测问题。

表 7.2 展示了若干个所用数据集的情况。

7.3.3　数据预处理

数据预处理分为以下几个步骤。

（1）数据清洗。这个数据集中存在大量全部为 0 的无效数据，以及部分属性缺失的数据，我们需要除去这些无效数据行。最终，得到 12 242 个有效数据条。

（2）特征选择。本数据集最大的特点，是数据的属性数目非常多。因此，选择出一些合适的特征，不仅有利于训练效果的提升，也有利于训练效率的提升。

常用的特征选择手段有 L1 正则项选择、相关性计算、单变量模型构建等。我们使用 Lasso 的 L1 正则项选择法进行特征选择，该方法利用 L1 正则项可以天然地得到稀疏的特性，进行变量选择。最终，我们选出了 8 个属性来进行后续的建模。

（3）标准化。从上文的数据中可知，每个属性的取值范围差异都非常大，这将会对结果产生影响。我们对每个属性进行 Min-Max 标准化，将值映射到 $[0,1]$ 范围内，以便于后续训练。

（4）**训练集分割**。打乱数据集条目之间的顺序。我们把 60%的数据作为训练集，20%的数据作为验证集，20%的数据作为测试集。

表 7.2　部分数据展示

下行平均 MCS	上行 PDCP 丢包率/%	PDCCH 占用率/%	乒乓切换比例/%
5825.676	4.503	16.489	0.762
4491.394	5.761	12.801	0.559
2183.053	5.409	10.01	0.406
22 218.455	3.172	4.354	0.048
5288.64	2.941	10.827	0.511

上行 PRB 占用率/%	CS 呼叫建立成功率/%	RRC 拥塞率/%	RRC 建立成功率/%
5.085	99.892	0	100
18.024	97.886	0	99.959
21.576	99.977	0	99.977
1.679	100	0	100
11.584	99.894	0	99.894

小区下行流量/GB	小区上行流量/GB	上行平均干扰	下行 PDCP 丢包率/%	下行平均传输速率/（Mb/s）
1140.082	109.919	−119.25	0.001	40 841.4
3064.717	283.559	−109.75	0	4573.217
6268.42	403.275	−119	0	18 667.616
249.278	15.468	−119	0	42 410.181
1794.825	177.366	−110.5	0	13 433.083

7.3.4　模型构建

我们利用传统的广义线性模型来解决这个问题，利用传统的线性回归进行实验。将选择出的 8 个特征加上截距项 1 组合成自变量 $\boldsymbol{X}$，将预测目标（数据下行平均传输速率）作为因变量 y。由于数据规模过大，传统的求解析解的方式会带来过高的复杂度。因此，需要使用优化算法来逼近结果。本案例中，我们使用梯度下降法和随机梯度下降法来处理，具体流程如 3.1.4 节的算法 8 和算法 9 所示。

我们随后利用套索回归（lasso regression）以及岭回归（ridge regression）进行实验，通过 K 折交叉验证的方式，选择出最佳的 λ 系数值。关键算法的 Python 代码如下所示。

```
from sklearn.linear_model import Lasso
alpha = 0.1
lasso = Lasso(alpha=alpha)
y_pred_lasso = lasso.fit(X_train, y_train).predict(X_test)
r2_score_lasso = r2_score(y_test, y_pred_lasso)
print(lasso)
```

```
print("r^2 on test data : %f" % r2_score_lasso)
```

最终训练结果如表 7.3 所示。

表 7.3　回归模型训练结果

模型名称	线性回归	岭回归	Lasso 回归
Accuracy	0.709	0.752	0.748

我们再利用多项式回归进行实验，高次多项式会造成自变量数的指数增长，仅二阶多项式模型就有 $C_8^2 = 28$ 个交叉项，更不用说高次模型。这么大的自变量规模会极速加剧模型的过拟合，不利于训练模型。训练结果如表 7.4 所示。

表 7.4　多项式回归模型训练结果

模型名称	线性回归	岭回归	立方回归
Accuracy	0.709	0.752	0.748

7.3.5　结果分析

结果表明，二阶多项式回归模型取得了最好的效果。这表明，变量之间的内在关系不是简单的线性关系，因此，单纯的线性模型很难取得比较好的预测结果。

7.3.6　总结

本项目利用线性回归模型、多项式回归模型对 5G 通信系统中的下行流量进行了预测，取得了较好的结果。线性回归模型的简洁性，使得这个案例的拟合过程非常快，并且结果的解释性很好。

然而，这个结果的拟合效果依然不是特别理想。线性回归模型虽然有着简洁、解释性强的优点，但同时也存在着模型复杂度不够、欠拟合严重的问题。在实际项目中，稳定性、解释性和效果之间往往是一种 Trade-off 的关系，对于一些特别追求解释性的场景，往往就很难应用效果很好的神经网络模型。直到今天，在工程领域还有许多问题是使用传统的机器学习来解决的。虽然深度网络发展迅速，但传统机器学习模型直到今天仍有用武之地。深度学习论文中很多创新性的思路也是从传统机器学习模型中得到的。

第 7 章　能源、电信系统相关案例.rar