智慧物流算法大赛结果分析

南京大学-天才儿童队，程唯，陈立庚

2018 年 7 月 6 日

本文档仅用于第一届“中储智运”杯智慧物流算法大赛的结果分析。队伍来自南京大学，天才儿童队。

1 问题简述

对于题目所述问题，主要目的就是通过历史运价预测未来运价。

对于问题的定位，相比于时序预测问题，其实将其定位成一个回归分析问题更为合适，即从历史数据中拟合一个回归模型，用于计算未来的运价。

2 数据处理

数据集中样本的特征包括日期、品类、货值、运距等，经过分析可以发现每个特征都或多或少会对最后的运价产生影响。但有些特征对结果的影响较小，例如日期，它并不会直接影响到运价，如果直接将日期加入模型进行训练，最终的结果可能并不会太好。其实日期会通过另外的层面间接影响运价，比如季节因素等，但这些因素不好量化，因此决定暂且忽略。

同样的还有经纬度，将经纬度直接加入模型训练也不是一个明智的选择。其实经纬度对运价的影响也是来自于它蕴含的一些信息，比如地形、南北方等，但这些同样不好量化，处理不当可能会影响模型效果，因此也暂且忽略。并且特征中有一个跟地理相关的特征，那就是运距，或许模型可以从运距中学到一些地理因素。

对特征经过筛选之后，最终保留了8个特征用于构造样本，包括一级品类、详细品类、订单类型、交易类型、货值、运距、车型、车长。

接下来，对训练数据集进行扫描发现有37691条样本存在特征缺失的情况，如果对缺失特征填零或经过其他手段处理之后保留这些样本，那必然会对结果产生不好的影响。这些样本占数据集的17.642%，考虑到数据集规模还算大，因

此去掉这些特征缺失的样本也是可以接受的。（**在测试集上做预测时不必去除特征缺失的样本，程序自会处理。**）

扫描过程中还发现数据集中存在一些运价过高或为0的样本，在现实生活中出现这样的运价的几率是非常小的。将运价的合理范围定在大于0且小于1000，最终统计出有240条样本运价异常，占数据集的0.136%，数量较小，可以去除。

去掉所有脏数据之后，最终剩下175714条样本。

3 模型选择

现有主流的数据挖掘算法基本上都可以用于构建回归模型，如线性回归、支持向量回归、深度神经网络、树回归等。考虑到数据集中样本具有复杂的非线性结构，特征之间的关系难以捕捉，而如今火热的深度神经网络在建模高度复杂的非线性关系方面非常有效，且可以非常灵活地学习特征之间的关系，因此决定使用深度神经网络作为此次比赛的回归模型。

在模型构建过程中，本队做了很多尝试，如增大或减小网络深度、调节各种超参、设置多种不同的超参组合等，最终构建出了一个深度为15层的神经网络模型用于运价的预测。

4 结果分析

我们以85比15的比例划分训练集和验证集，最后训练结果如下表所示

Parameter	below 1%	1%-2%	2%-5%	5%-10%	10%-20%	above 20%
number	2466	1429	3474	3081	2327	1649
percent	9.356	5.422	13.180	11.689	8.828	6.256

Table 1: 预测结果高于实际值的数据分布情况

Parameter	below 1%	1%-2%	2%-5%	5%-10%	10%-20%	above 20%
number	2414	1484	2909	2397	1857	871
percent	9.159	5.630	11.036	9.094	7.045	3.305

Table 2: 预测结果低于实际值的数据分布情况

表1和表2分别表示的数据为，预测的结果高于实际的运价和预测的结果低于实际的运价。从数据分布可以看到，还是有共约30%的数据是非常精准的，偏差率在2%以内，大多数的数据分布在偏误2%到10%之间，虽然仍有数据会有

较大的偏差，但是我们可以看到，超过20%误差的数据总量不超过验证集总数的10%，因此，我们的建模结果还是相对不错的。

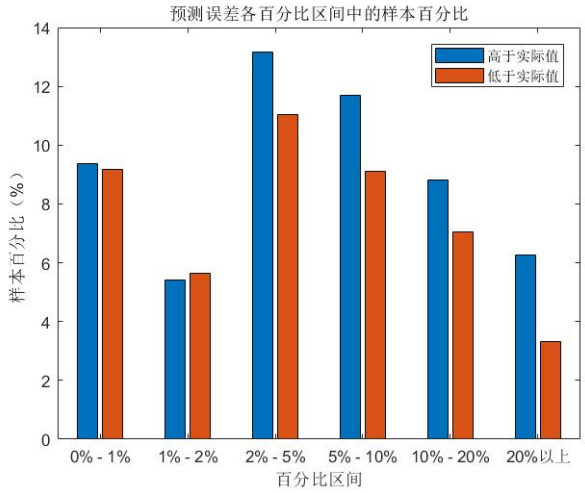

Figure 1: 样本误差情况百分比占比

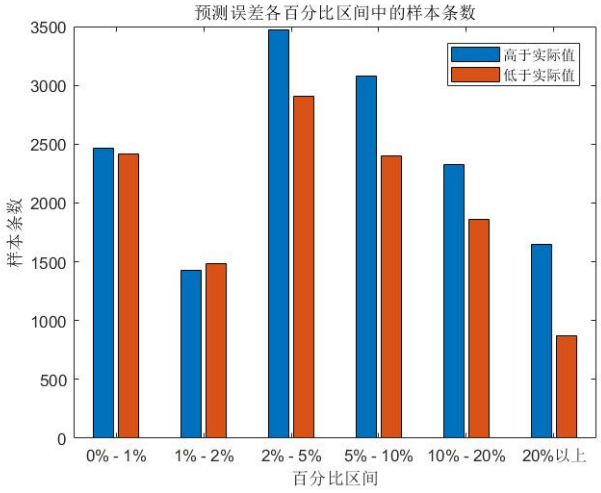

Figure 2: 样本误差情况条数占比

图1和图2两个柱状图分别展示了表1和表2的数据，结果一目了然，大部分样本的预测结果都在较小的误差范围内。

通过对验证集的预测，可以看出模型能够相对准确地预测未来的运价，可以给予决策者一定的参考和借鉴意义。但是在实际情况中，影响价格的因素还有很多，比如一开始被我们忽略的因素，若能找到一个合适的手段将更多的因素加入到模型中来，可能会使模型的预测性能更加有说服力。