面向物流场景的数据分析报告

|  |  |
| --- | --- |
| 团队名称： | 黄金矿工 |
| 团队成员： | 罗浩、罗如瑜、丁博 |
| 指导老师： | 聂高峰 |
| 所在学校： | 北京邮电大学 |
| 参赛组别： | 本科A组 |

摘 要

为了破解交通物流行业中存在的物流资源浪费问题，帮助网络货运平台降本增效，本项目通过数据分析、机器学习与数学建模的方法，对客户物流数据集和车辆基础数据集从运量走势、车货匹配、线路分布三个方面进行分析。在运量走势方面，通过整理客户历史运单数据得到两地之间的运量时间序列，使用Holt-Winters时间序列平滑法对客户未来的运输需求进行较好的预测。在车货匹配方面，使用数学建模与机器学习相结合的方法分析客户历史数据，对客户未来的运货需求提供车辆推荐。在线路分布方面，对客户过往运单的线路数据进行统计分析；模型使用统计与数学建模相结合的方法，通过对路线热门度、流行度和长度三项指标加权评分的方法推荐至多三条得分最高的路线。本项目通过上述三个方面的预测或推荐，为网络货运平台的资源调配和线路优化提供参考方案，并给出一个参考方案的示例。

关键词：运量预测、车货匹配、线路优化、机器学习、数学建模

目录

[摘 要 I](#_Toc56758803)

[第1章 物流领域问题场景 1](#_Toc56758804)

[第2章 分析目的及思路 3](#_Toc56758805)

[2.1 运量走势 3](#_Toc56758806)

[2.1.1 分析目的 3](#_Toc56758807)

[2.1.2 分析思路 3](#_Toc56758808)

[2.2 车货匹配 4](#_Toc56758809)

[2.2.1 分析目的 4](#_Toc56758810)

[2.2.2 分析思路 4](#_Toc56758811)

[2.3 线路分布 4](#_Toc56758812)

[2.3.1 分析目的 4](#_Toc56758813)

[2.3.2 分析思路 4](#_Toc56758814)

[第3章 数据分析与建模 6](#_Toc56758815)

[3.1 运量走势 6](#_Toc56758816)

[3.1.1 整理两地之间运量随月份变化的时间序列 6](#_Toc56758817)

[3.1.2 数据分析 8](#_Toc56758818)

[3.2 车货匹配 10](#_Toc56758819)

[3.2.1 数据清洗 10](#_Toc56758820)

[3.2.2 模型构建分析 12](#_Toc56758821)

[3.3 线路分布 13](#_Toc56758822)

[3.3.1 统计客户历史运单数据的线路分布 13](#_Toc56758823)

[3.3.2 最短路线模型 15](#_Toc56758824)

[3.3.3 专家评议法 16](#_Toc56758825)

[第4章 实现路径 20](#_Toc56758826)

[4.1 运量走势 20](#_Toc56758827)

[4.1.1 预测模型 20](#_Toc56758828)

[4.1.2 实现步骤 21](#_Toc56758829)

[4.2 车货匹配 21](#_Toc56758830)

[4.3 线路分布 23](#_Toc56758831)

[4.3.1 查询满足需求的历史线路分布 23](#_Toc56758832)

[4.3.2 Floyd算法 23](#_Toc56758833)

[4.3.3 最短路线算法的实现 25](#_Toc56758834)

[第5章 分析结果 27](#_Toc56758835)

[5.1 运量走势 27](#_Toc56758836)

[5.1.1 结果展示 27](#_Toc56758837)

[5.1.2 结果分析 27](#_Toc56758838)

[5.2 车货匹配 29](#_Toc56758839)

[5.3 线路分布 30](#_Toc56758840)

[5.3.1 历史上有一条线路但是可能绕路的推荐方案 30](#_Toc56758841)

[5.3.2 历史上有多条线路给出的推荐方案 32](#_Toc56758842)

[5.3.3 历史上没有运输线路给出的推荐方案 33](#_Toc56758843)

[5.4 为网络货运平台提供的参考方案 34](#_Toc56758844)

[第6章 总结 35](#_Toc56758845)

[参考文献 36](#_Toc56758846)

1. 物流领域问题场景

近年来，随着物流行业和车货匹配平台的快速发展，我国公路货运市场规模已超过6万亿。准确把握公路货运量等发展趋势，可以为相关决策提供依据。运量预测是指根据国民经济和社会发展对交通运输的需求，探索未来旅客和货物运量发展的趋势，对未来一定时期内交通运输业所应承担的工作量所做的测算和判断。是国家经济预测的组成部分之一，也是研究分析交通运输发展战略的重要内容和决策的依据。

运量预测的主要内容包括四个方面：社会总运输量的预测；各种运输方式的运输量预测；地区之间的运输量预测；运输企业在运输市场上的占有率预测。在四类预测中，前两类属于宏观预测的范畴，后两类属于微观预测的范畴。由于预测的目的要求不同，因此内容的粗细也不同。一般来讲，宏观预测与长期预测内容要粗一些，微观预测和短期预测的内容则要细一些。例如，列入本企业(或部门)经营的运输量，不仅有客、货运量和周转量，还应包括上行、下行的运输量，淡、旺季的运输量，货物运量中主要货物的分类和比重等。

中国物流的特点是大而复杂。在物流市场中，存在地区性差异和季节性差异，参与其中的角色众多：个体司机，车队老板，物流公司等等。而近年来我国公路运输不断涌现出资源浪费现象：车找不到货、货找不到车的情况在物流市场普遍存在。为了降低物流成本，减少不必要的社会资源浪费，许多企业在互联网的大时代背景下开始探索如何实现车货匹配。研究如何充分整合社会运力资源，实现高效率的车货匹配具有重要的社会意义与研究价值。

物流企业在发展自身时，最先需要分析的就是运输线路，运输线路的测定直接影响着运输工具的规划，间接影响着运输成本的高低，也是决定利润收益的关键。线路的制定一般只有三种，即干线、专线、直线，这三种线路基本上是以区域面积划分的。通俗的讲，干线是省际之间，运输的路线也是国道；专线也是省际之间，通常是城市之间为运输终端，专线基本上建立在一线城市之间；支线则是设立在省内的运输分支，运输的货物量也只是承担省内的运输货物。三者之间是相辅相成的，形成递进式运输。

运量预测是一个复杂问题，能够对运量造成影响的因素很多，包括经营问题、社会问题、环境问题等等。本文采用时间序列分析法预测运量趋势。时间序列分析需要长时间且大量的数据积累，才能达到较为准确预测未来数据的目的。但是，由于比赛数据集只有2018年1月-2019年7月共近一年半时间内的运单数据，数据的时间跨度太短，我们对数据集进行处理后得到的运量月度时间序列只有18个点。面对这样短的时间序列，大多数模型都很难准确预测未来运单运量的发展趋势。

车货匹配中货源是唯一品和非标准品，针对每一次新的订单要求，需要均衡考虑反馈率和环境。本模型中采取简单的将数学建模模型和机器学习XGBoost结果进行加权，从而综合考虑历史与现状，但是这种模型对车辆实时承运情况考虑不足，可能模型给出的最佳匹配车辆当时正在运输其它运单，最终执行运单的车辆不是最佳匹配结果。

面对有多个提货、卸货节点的运单需求时，对节点次序进行调整以满足路线总长最小，以及对节点顺序有特殊要求的优化问题是困难的问题。图论中必须经过某些中间节点的最短路径规划问题是一个NP-hard问题，目前除了穷举之外没有快速的精确算法，有蚁群算法、遗传算法等近似解法可以较高效的求取近似最优解。HUANG等人基于Dijkstra算法和贪心理论提出了给定起终点的情况下必须经过某些指定的中间节点的最短路线算法[1]，可是他们的算法只对固定起终点有效，在物流场景下存在捎单现象和多地点提货卸货的需求，路线的起终点的选择也在路线优化的范围内，此时他们的算法便不适用了。此外模型基于GPS坐标计算两地之间的直线距离作为两地之间的路线长度而忽略具体走的道路，导致路线长度的计算存在本源性误差，本文中提出的最短路线模型只能求出次优解。

货运路线的选择除了考虑路线长度，还要考虑道路状况、车辆状况、司机习惯、客户需求、车辆装载量、道路坡度等其它因素的影响。这些影响因素对于道路选择的影响机制和影响程度难以建模，本文中提出的模型评分项目和权值的指定全凭经验，可能会影响模型优化结果的准确性。

1. 分析目的及思路
   1. 运量走势
      1. 分析目的

运量预测是运输组织工作中规划运能利用和编制日常运输计划的基本依据，也是对运输设备新建和扩能改造提出运营要求的基本依据。因此，运量预测的准确性以及运量发展变化趋势的正确认识与把握，对提高运输组织工作的预见性、改进运输规划工作具有重要意义。

* + 1. 分析思路

运量预测的方法很多，总的来说，可以分为定性预测和定量预测两大类。定性预测方法又叫判断分析法，是依据人们在市场活动中获得的经验和分析能力，通过对影响市场变化的各种因素的分析、判断和推理，来预测未来的发展变化。定性预测方法的特点是简便易行，特别是在不可控因素和不可定量因素比较多时，采用这种方法进行短期判断有其明显的优势。然而，这种方法也有其缺陷，它不能提供以精确数据为依据的预测结果，主观随意性比较大，有时易发生疏忽和失误。定量预测法又称数量分析法或数理统计预测法，它是根据市场调查所取得的数据资料，运用数学模型进行计算，并据此预测未来市场变化的一种预测方法。这种预测方法的一个显著特点是运用数学、统计学和计算机等方法或工具，用数据对未来进行客观描述，因此，其科学性、严密性更强。

本文采用定量预测法中的时间序列分析的三次指数平滑法对运量进行预测。在社会经济发展过程中，很多经济变量的发展变化都表现出与时间呈某种特定关系，运输需求也是如此。基于此，我们可以通过对运量的时间数列外推的方法预测未来的运量变化趋势。时间序列预测法的特点是把预测变量看作时间的函数。当一个序列在每个固定的时间间隔中都出现某种重复的模式，就称之具有季节性特征，而这样的一个时间间隔称为一个季节(理解：比如说在一个周内，销量呈现出重复的模式)。一个季节的长度k为它所包含的序列点个数。

三次指数平滑考虑了一个序列的baseline、趋势和季节分量。比如预测下一个季节第3个点的季节分量时，需要指数平滑地考虑当前季节第3个点的季节分量、上个季节第3个点的季节分量等等。

模型在运量分析的处理过程中，首先设置提货点和卸货点，然后在地址数据中搜索符合条件的运单ID，并记录。最后在运单数据中，搜索运单ID对应的运单，记录该运单的吨数和方数到时间序列中。在完成时间序列的收集后，我们使用holt-winters模型对时间序列进行分析。

* 1. 车货匹配
     1. 分析目的

为了充分利用大数据资源，通过互联网技术提高信息检索能力和匹配效率，减少因信息不对称问题造成的物流资源浪费问题，达到去中介化的目的，尽可能实现最优车货匹配方案。

* + 1. 分析思路

车货匹配算法场景本质可看作推荐场景，尤其需要注意的是其中的反馈率。在一个时间窗口内，进行联系的货源除以总货源叫做反馈率。反馈率和地区（区县一级）的供需关系呈现强烈正相关。反馈率一旦达到一个阈值，就会在地区形成一种新的平衡：用户自然流失等于或小于平台自然流入，地面团队可以把更多精力放在服务用户身上，而非拉新促活。所以对于业务指标来说，完成反馈率提升甚至比提供更有效的用户匹配更重要。

本方案中的车货匹配采用了建模与机器学习相结合的方式，车货匹配算法场景本质可看作推荐场景，即可以套在CTR、CVR的模型上，此处选用XGBoost进行加权。车货匹配方案中，考虑起始点终点后，将地址数据、运单数据、整理后的车辆数据进行全连接，可以对同一辆车的历史起终点进行统计，对于同一辆车，考虑到反馈率的影响，历史频次越高则权重更大，即相同需求下，认为同一辆车更倾向选择权重大的起终点进行运输，同时使用XGBoost来预测车-货的基础相关性，实际是一个CTR和CVR混布模型。结合基础相关性和反馈率最后给出合理的车货匹配方案。

* 1. 线路分布
     1. 分析目的

通过统计历史运单的路线分布，对于新的运货需求，规划出一条更多司机走过的，或者距离较短的路线。

* + 1. 分析思路

本方案中的线路分布采用了统计与图论建模相结合的方式。首先对每个客户的所有历史运单信息，统计出走每一条路线的历史运单数，计算每条路线的热门度。之后根据客户新的货运需求，给出提货地点、卸货地点的情况下，在历史线路分布中找到符合需求的订单，可以知道历史上司机们是怎么安排提货、卸货顺序的，走了什么路线。同时系统根据图论中求最短路径的Floyd算法，通过将中间节点排列组合的方式给出不同的途经顺序，求解必须经过某些中间节点的最短路径问题，给出一条最短路径。最后结合路线热门度、流行度和长度三项指标，给出最多三条推荐路线和推荐星级。推荐星级表示路线的可信程度，最小是1，最大是5。

1. 数据分析与建模
   1. 运量走势

我们把2.1.2节中介绍的思路重述如下：统计两地之间一个月所有订单发送货物的吨数和方数之和，得到两地之间运量随时间变化的序列。通过数据处理拟合出前两年的运量预测数据，并使用holt-winters模型对上述时间序列进行分析，预测未来两地之间运量的变化。

* + 1. 整理两地之间运量随月份变化的时间序列

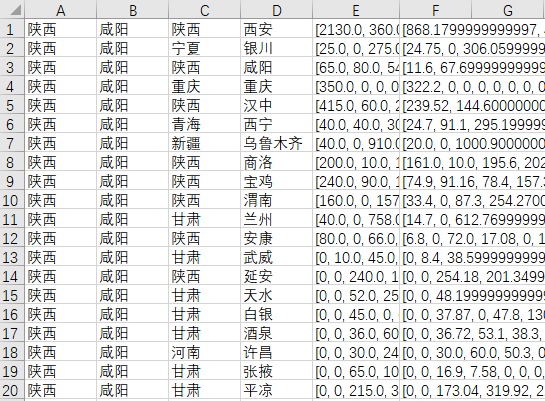
我们使用下面的伪代码来描述整理前18个月运量数据的思路。

|  |
| --- |
| 代码1：整理前18个月运量时间序列  【1】 在地址数据从上到下扫描到第i条。  a) 若为新的运单ID，记录该ID，跳转到【2】  b) 否则继续向下搜索，重复运行【1】  【2】 比对提货地点，判断是否为目标提货点Pos\_A，  a) 若是，跳转到【3】  b) 否则，跳转到【1】  【3】 向下搜索第i+1条数据，判断是否为目标卸货点Pos\_B。  a) 若是目标卸货点，则记录本运单ID。在运单数据中搜索该运单。  i.判断本运单的时间点，将本条数据中包含的吨数和方数，加在预先设置好的吨数/方数时间序列上。  【4】 跳转到【1】 |

经过上述处理，我们可以得到客户（A/B）在2018年1月至2019年7月的所有运单发送货物的吨数和方数的月度时间序列。下面对客户A和客户B的数据处理结果分别进行说明。

**客户A**：未经过整理的数据，地址数据共有21472条，运单数据共有9142条。

经过分析数据，我们观察到客户A所有订单的提货点均位于陕西咸阳，而卸货点有很多城市。但是，我们也观察到地址数据中存在同一运单下多个卸货点在不同城市的情况，为了便于按照前述思路处理数据，我们不考虑这种情况，只统计卸货点在同一城市的运单数据。最终得到的统计结果存储到附件“运量走势\数据整理\客户A\数据\result\_A.csv”文件中，部分结果如图1。



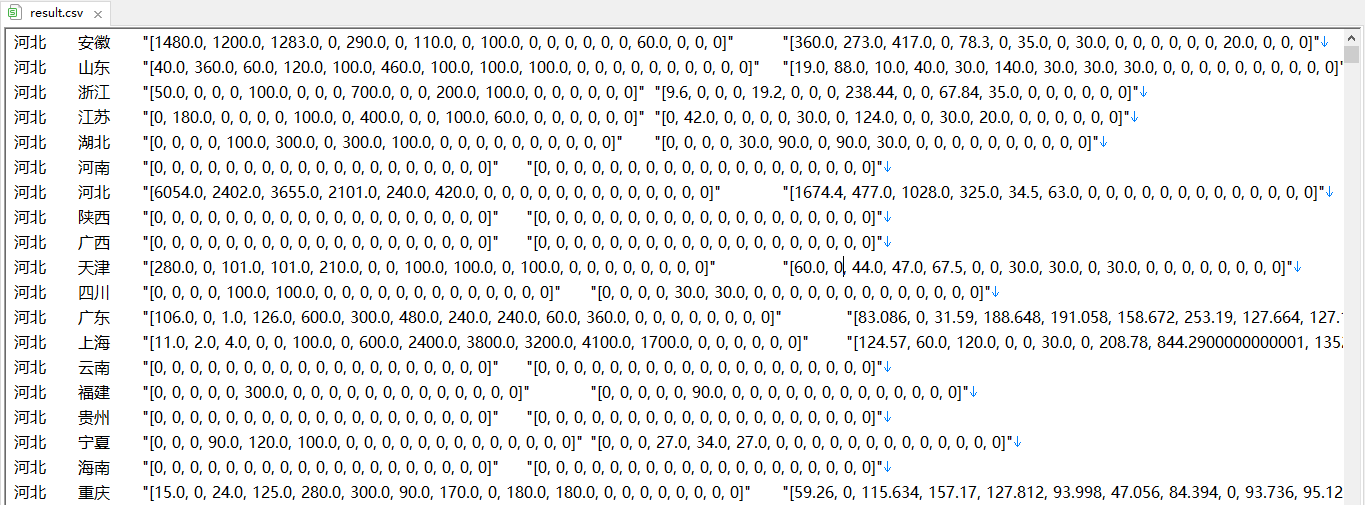
1. 客户A部分运量时间序列

Csv文件中每一行存储的是某提货城市到某卸货城市从2018年1月到2019年7月的运量月度时间序列，数据格式为（提货省，提货市，卸货省，卸货市，方数时间序列，吨数时间序列）。共收集到83条数据。

因为客户A数据集中有19个月的数据，所以统计得到的时间序列长度为19。通过观察方数时间序列和吨数时间序列，我们可以发现大部分的时间序列最后一个点为0。分析数据集我们可以发现，最后一个月的数据只有该月1号的数据。为了便于分析时间序列的规律，我们在分析的时候忽略最后一个点。这就是前文提到“整理前18个月运量数据的思路”而不是“前19个月”的原因。

**客户B**：未经过整理的数据，地址数据共有55757条，运单数据共有26627条。

经过分析数据，我们发现客户B的数据量比客户A的数据量大很多。提货点不只有一个城市，同时也存在一个运单的多个卸货点不在同一个城市的情况。如果以城市为提货点和卸货点的地点单位，计算量相当巨大，所以我们以省为地点单位统计时间序列，同时忽略多个卸货点不在同一个省份的运单。最终得到的统计结果存储到附件“运量走势\数据整理\客户B\数据\result\_B.csv”文件中，部分结果如图2。



1. 客户B部分运量时间序列

Csv文件中每一行存储的是某提货地省份到某卸货地省份从2018年1月到2019年7月运量的月度时间序列，数据格式为（提货省，卸货省，方数时间序列，吨数时间序列）。共收集到868条数据。

和客户A的数据集一样，统计得到客户B的时间序列长度为19。通过观察序列发现大多数序列的最后一个点都是0，原因与客户A相同，所以我们在进行时间序列分析的时候同样忽略最后一个点。

此外，考虑到我们统计了所有省份两两之间的运量时间序列，但存在某提货省到某卸货省没有运输记录的个例，所以部分数据存在运量时间序列全为零的情况。我们在时间序列分析的时候将这些数据视为无效数据，不予分析。

* + 1. 数据分析

季节性被用来描述时间序列数据的趋势，它表现出每一个周期重复前面周期的行为，就像任何周期函数一样。“季节”这个词用来表示序列每隔时间段*L*就开始重复自身的特征。在自然界中有不同类型的季节性“累加性”(additive)和“累乘性”(multiplicative)，就像加法和乘法是数学的基本运算。以下举例说明运量序列数据具有“季节性”。

图3中子图的纵坐标是两地之间的运量（方数或吨数），横坐标的单位是季度。描述的是运量随季度的变化曲线，其中蓝线为前4个季度（2018年的一至四季度）的运量，黄色线为后2个季度（2019年一二季度）的运量。可以看出前后两年的曲线趋势近似，所以可得运量走势是一个以年为周期，年与年之间有季节性的“累加性”，且以季度为时间单位时数据随机波动带来的噪声小于数据本身的变化趋势。假若采用更细的粒度（比如天），容易发现数据随机波动大概率会较大影响对数据本身的变化趋势的观测，进而影响运量预测结果。

|  |  |
| --- | --- |
| (1)（客户A）陕西咸阳==>青海海东 | (2)（客户A）陕西咸阳==>河南南阳 |
| 第1年方数序列  第2年方数序列 | 第1年吨数序列  第2年吨数序列 |
| (3)（客户A）陕西咸阳==>青海西宁 | (4)（客户B）河北==>江苏 |
| 第1年吨数序列  第2年吨数序列 | 第1年吨数序列  第2年吨数序列 |
| (5)（客户B）河北==>山西 | (6)（客户B）浙江==>江西 |
| 第1年吨数序列  第2年吨数序列 | 第1年吨数序列  第2年吨数序列 |

1. 运量序列数据具有“季节性”示例

移动平均值永远不可能应用于现有的数据集边缘的数据，因为它们的窗口宽度是有限的。这是一个大问题，因为数据集边缘的变动形态一般都是我们最感兴趣的部分。类似地，移动平均法也不能应用于现有数据集的范围之外。其结果是，边缘数据对预测毫无用处。

幸运的是，有一种很简单的计算方案能够避免所有这些问题。它叫指数平滑法(exponential smoothing)或Holt-Winters法。指数平滑法有几种不同形式：一次指数平滑法针对没有趋势和季节性的序列，二次指数平滑法针对有趋势但没有季节性的序列。Holt-Winters模型为三次指数平滑法，针对既有趋势又有季节性的序列。

所以综上所述我们采用三次指数平滑Holt-Winters模型。

* 1. 车货匹配
     1. 数据清洗

分两步对原始数据进行清洗。

①重点分析运过货的车辆基础数据以及运单数据；

②用历史运单数据中的最大吨数与方数对车辆基础数据中的空数据以及不合理数据（体现在车辆能够承载的最大吨数或最大方数小于货物的吨数或方数）进行修改以及更新。

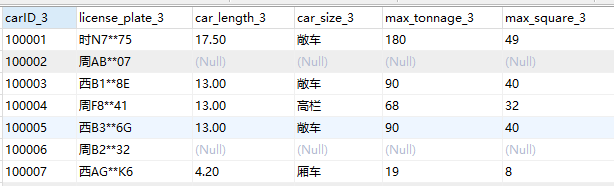
数据清洗的部分结果展示如下，以下对客户A、B的数据分别进行说明。

**客户A**：未经过清洗的数据，车辆基础数据共有106340条数据，运单数据9142条，经过数据清洗后，将车辆基础数据表与运单数据表进行对应连接后得到9142条数据，与运单数据相关的车辆数据共3136条数据。若考虑起始点终点后，将地址数据、运单数据、清洗后的车辆数据进行全连接后，得到21472条数据。

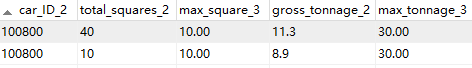
由图4看出，原始数据中存在着空数据的情况。

将运单数据与车辆数据进行全连接操作，得到图5，可以看出，有不合理的数据存在，存在运单数据中的方数大于车辆最大方数的情况，我们假设这种情况是因为车辆数据更新不及时（或车辆数据错误）导致的。

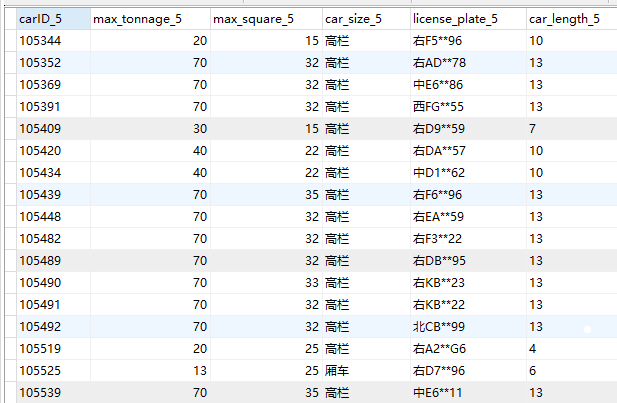
清洗后的车辆基础数据表如图6所示，可以看出对空数据以及不合理数据进行整理更新。



1. 车辆基础数据中存在空数据



1. 原始数据中的不合理数据



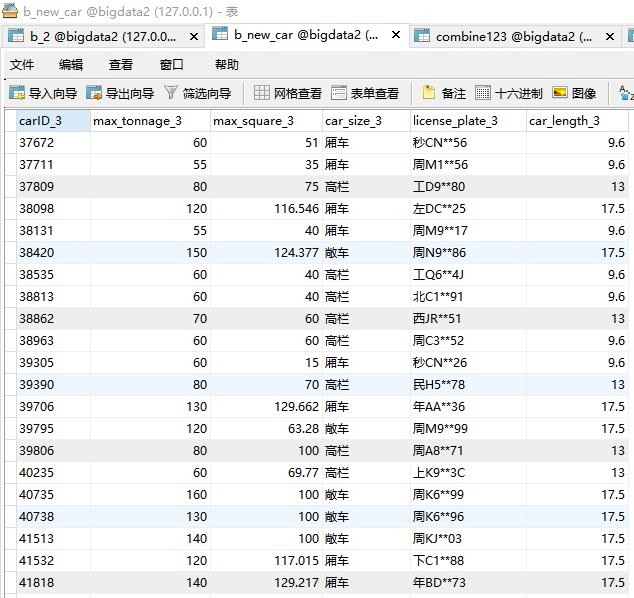
1. 清洗后的承运客户A运单的车辆数据

**客户B**：未经过清洗的数据，车辆基础数据共有106340条数据，运单数据27627条，经过数据清洗后，将车辆基础数据表与运单数据表进行对应连接后得到27627条数据，与运单数据相关的车辆数据共13614条数据。若考虑起始点终点后，将地址数据、运单数据、整理后的车辆数据进行全连接后，得到56757条数据。

同上所述，车辆基础数据源中存在着空数据的情况。

将运单数据与车辆数据进行对应连接操作，可以看出其中有不合理的数据存在，存在运单数据中的吨数大于车辆最大吨数的情况，我们假设这种情况是因为车辆数据更新不及时（或车辆数据错误）导致的。

清洗后的车辆基础数据表如图7所示，可以看出对车辆空数据以及不合理数据进行整理更新。



1. 清洗后的承运客户B运单的车辆数据
   * 1. 模型构建分析

为了考虑全面，还需知道当时车辆的位置坐标，因为是未来数据，此处假设车辆与起点的地理距离为

|  |  |
| --- | --- |
|  |  |

同时还与路径损耗相关，即车辆当前位置到起点的路径情况越好，路径损耗的绝对值越小，此处选用路径系数

|  |  |
| --- | --- |
|  |  |

假若起终点以及吨数方数确定：

①只要车辆最大吨数和方数大于发货需求的吨数和方数，则认为该车可以作为承运车辆；

②考虑到车货匹配的反馈率要求，通过数据分析选出较频繁的备选车辆；

③并通过XGBoost对备选车辆进行预测，得出空闲可用车辆的优先度；

④将同一起终点的所有空闲可用车辆的优先度与路径系数综合考虑，对进行排序，选出最大值即应该匹配的车辆，具体考虑参数可以参考当前的情况，假如物流货运没有很紧张，一般可主要考虑空闲可用车辆的优先度。

* 1. 线路分布
     1. 统计客户历史运单数据的线路分布

这一部分使用MySQL+Python Pandas库的技术路线，python与MySQL数据库的通信使用pymysql库完成。

* + - 1. 解析运单采用的线路

模块对客户A或B，查询运单数据库中所有的运单ID。之后对每一个运单ID，模块从地址数据库中查询该运单的所有途经节点的类型（提货还是卸货）、节点所在的省、市。然后将查询出的所有节点串起来，形成一个途经城市轨迹序列，例如(('提货', '陕西,咸阳'), ('卸货', '陕西,西安'), ('卸货', '陕西,咸阳'), ('卸货', '甘肃,兰州'), ('卸货', '陕西,宝鸡'), ('卸货', '河南,许昌'))。这就是运单采用的线路。对于货车在一个城市不同区域多次卸货的情况，粒度只细分到城市，即只考虑车在这个城市卸过货，只保留在该城市相同操作的节点中的第一个。将运单ID和对应的线路保存在数据帧中，将数据帧的内容保存在csv文件(track\_A或B.csv)中。

* + - 1. 统计每条路线对应的运单数

这里使用pandas库的value\_counts()方法，统计每条线路对应的运单数，按照降序排列，结果保存在数据帧trackCount中。trackCount中部分内容如图9或图11所示。客户A的所有线路和线路对应运单数见附件“线路分布”文件夹下trackCount-A.csv。客户B有对应的trackCount-B.csv.

* + - 1. 统计结果

通过3.3.1.1和3.3.1.2节整理，系统统计出客户A和B的所有运单路线，走每一条路线的运单数，按照降序排列可以给出频数直方图，如图8和图10所示。直方图按照路线的运单数大于100，路线数不超过20条进行筛选，画出运单数最多的一些路线。

客户A在陕西咸阳提货、到陕西西安卸货的需求非常大，在总共9142个运单中有3514个走这条路线，占比38.4%。A的部分统计结果见图9，表头track代表路线，freq代表运单数，hot代表路线热门度。

图10是客户B的线路分布直方图，图11是部分数据。可以看到，客户B的业务范围更加广泛，提货城市共94个，卸货城市共264个。前五大线路需求是安徽马鞍山提货、浙江杭州卸货；广东中山提货、广东东莞卸货；陕西宝鸡提货、陕西宝鸡卸货；安徽六安提货、安徽芜湖卸货；安徽马鞍山提货、浙江湖州卸货。

图表, 表格

已生成高可信度的说明

1. 客户A线路分布直方图

表格

已生成高可信度的说明

1. 客户A部分线路的运单数和热门度

图表

已生成高可信度的说明

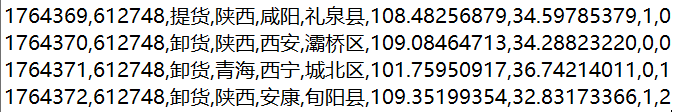
1. 客户B线路分布直方图

表格

已生成极高可信度的说明

1. 客户B部分线路运单数和热门度
   * 1. 最短路线模型
        1. 使用图论F算法求最短路径的出发点

可以看到在运单数据中，当有多个卸货城市时，不同的运单会采取不同的访问顺序，较差的访问顺序会导致时间和车辆资源的浪费。比如客户A的运单612748，途经地点信息如图12所示。该运单从陕西咸阳提货，先后在陕西西安、青海西宁、陕西安康卸货。这条路线不是在这些地点卸货的最短路线，从陕西到青海的路程很长，这会导致绕非常远的路。



1. 运单612748途经地点信息
   * + 1. 模型构建分析

模型假设只需要满足在某些地点提货、在另一些地点卸货的需求，不考虑某个城市是主节点的要求（即必须最先或者最后经过该城市）。在这个假设下，只需要把提货地点、卸货地点全排列，对地点序列中的所有地点两两之间用最短路径连接，每一种可能都会求出一条长度尽量小的路径，所有序列中长度最小的就是最短路线。

使用我们的算法，对运单612748的需求给出的推荐线路是陕西,咸阳(提货)->陕西,西安(卸货)->陕西,安康(卸货)->青海,西宁(卸货)，先在陕西省内卸货完毕之后再去青海省卸货，避免重复走陕西到青海的路程。

* + 1. 专家评议法

专家评议法是出现较早且应用较广的一种评价方法。它是在定量和定性分析的基础上，以打分等方式做出定量评价，其结果具有数理统计特性。其最大的优点在于，能够在缺乏足够统计数据和原始资料的情况下，可以做出定量估计。

在模型中使用“加权评分法”。模型在推荐路线时，先参考历史上满足提货卸货地点要求的线路，一种可能的查询结果如表1，有x个运单采用A路线，y个运单采用B路线，z个运单采用C路线。

1. 查询满足需求的历史路线的一种可能结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index索引 | Track运单路线 | Freq运单数 | Percentage流行度 | Distance路线长度/km |
| 0 | A | *x* | *x/(x+y+z)* |  |
| 1 | B | *y* | *y/(x+y+z)* |  |
| 2 | C | *z* | *z/(x+y+z)* |  |

* + - 1. 评分指标

评分系统考察三项指标，三项得分加权求和得到总评。系统最终推荐总分最高的几条路线（不超过三条）。

1. 热门度

热门度（hot）根据客户的历史线路分布求解。在求出每条历史路线对应的运单数之后，我们得到一张表格，如表2。设表格第2列运单数构成向量

|  |  |
| --- | --- |
|  |  |

式中*N*代表客户历史线路总数。热门度hot定义式如下

|  |  |
| --- | --- |
|  |  |
|  |  |

式中sigmoid函数的表达式是

|  |  |
| --- | --- |
|  |  |

首先定义向量ttt是考虑路线对应的运单数相差极大，比较热门的路线可以有超过一千个运单采用，而很多路线对应运单数小于10，甚至如果求出的最短路线在历史线路分布中不存在，对应运单数是0，所以定义ttt是向量waybill加1取对数。之后根据sigmoid函数的性质，为了使热门度取值分布在[0, 1]之间，将sigmoid函数的输入调整到范围[-0.5\*max(ttt), 0.5\*max(ttt)]，这样热门路线对应的热门度很高，达到0.95以上；冷门路线热门度很低，小于0.05.

模型考虑该指标的原因与反馈率有关，历史上较多司机选择的路线可能具有更加著名，路况更好或者更安全的特点，模型在推荐路线时要考虑司机们的选择的合理性。

1. 客户历史线路分布示例

|  |  |
| --- | --- |
| 线路 | 运单数 |
| A | 3514 |
| B | 188 |
| C | 367 |
| … | … |

1. 流行度

线路X的流行度（Percentage）定义为线路X对应的运单数占满足需求运单总数的比例。。计算示例见表1的Percentage列。模型考虑该指标的原因也是反馈率。

1. 路线长度度量

路线长度度量(short)的定义式如下。

|  |  |
| --- | --- |
|  |  |

式中distance表示如果备选路线多于1条，使用最短路线算法求出每条路线的长度构成的向量。表示distance的平均值，表示distance的标准差。这样做是为了把distance向量标准化（normalization）。之后作为Sigmoid函数的自变量时之所以取相反数是因为让长度短的路线得分更高。如果备选路线只有一条，那么这条路线就是最短的，short指标得分为1.

这样定义路线长度度量，可以让指标得分在[0, 1]之间分布。

模型考虑该指标的原因是减少路线长度，提高运输效率。

* + - 1. 评分方法

路线X的总分score满足公式

|  |  |
| --- | --- |
|  |  |

式中，和是模型的三个参数，其中是路线热门度的权重，模型中设置为0.2；是路线流行度的比重，模型中设置为0.3；表示最短路径指标的权重，模型中设置为0.5.三个参数满足关系。

如果最短路线算法求出的路线不在历史记录中，对应历史运单数为0，，.从公式可以看出，路线的总分是1分。

在所有路线中，取得分最高的几条路线作为算法推荐路线，并且给出推荐星级，路线的星级越高代表路线越好，更推荐司机选择。得分与推荐星级对照表如表3所示。

1. 得分与推荐星级对照表

|  |  |
| --- | --- |
| 得分score | 推荐星级 |
|  | 5 |
|  | 4 |
|  | 3 |
|  | 2 |
|  | 1 |

* + - 1. 关于模型参数a, b和c的补充说明

热门度指标权重*a*设置为0.2，流行度指标权重*b*设置为0.3是考虑让历史线路分布对当下路线选择的影响不至于太弱，而路线长度度量权重*c*设置为0.5是考虑主要由路线长度指导路线选择的过程，但也不会对历史线路选择情况考虑不足。

在初赛答辩中，评审专家提出，和三个参数赋值0.2，0.3，0.5太过主观，建议使用机器学习的方法调整模型参数。我们在经过分析之后认为无法也不必要使用机器学习方法优化。原因有三。

1. 路线的优化往往综合考虑路线长度、到货时间限制和两地之间货量等指标，往往是多约束条件的线性规划或者最优化问题。我们的模型没有用车货匹配、运量走势的结果指导线路规划，所以在线路规划时没有考虑到货时间和两地之间货量两项指标，只考虑路线长度和历史选择。对于只有三个参数的简单模型，我们认为不必使用机器学习这种复杂的方法指定模型参数。
2. 路线的优劣性这一概念的主观色彩非常强。如果我们认为最短的路线就是最优的，那么我们不需要考虑历史线路分布、到货时间和运量的限制。这时也不需要使用机器学习训练，只要考虑长度最小这一项就足够做出选择了。此外，对于这种主观性极强的属性很难对数据集标注有意义的标签，也就无法使用机器学习模型优化调整参数。
3. 输入机器学习模型的数据集的形式不正规。涉及线路选择的分数比较时，只有同样运输需求下的路线才有比较的意义。在已有数据集中，按照需求给路线分组，不同组的路线数量不同，当然最优路线是唯一的。以元素数目不固定的组作为机器学习模型的输入显然是不可行的。

综上所述，目前我们没有使用机器学习模型调整参数。调参需要结合物流公司的需求，根据他们对指标的重视程度进行调整。

1. 实现路径
   1. 运量走势
      1. 预测模型

运量走势部分采用Holt-Winters模型对运单的方数和吨数月度数据进行研究预测。

霍尔特-温特（Holt-Winters）方法是一种时间序列分析和预报方法。该方法适用于趋势线性且周期固定的非平稳序列，利用指数平滑法（EMA）让模型参数不断适应非平稳序列的变化，并对未来趋势进行短期预报。Holt-Winters 方法在 Holt模型基础上引入了 Winters 周期项（也叫做季节项），可以用来处理月度数据（周期 12）、季度数据（周期 4）、星期数据（周期 7）等时间序列中的固定周期的波动行为。引入多个 Winters 项还可以处理多种周期并存的情况[2]。

Holt-Winters分为加法模型和乘法模型，此处我们采用加法模型。累加性公式如下：

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

其中 α是数据平滑因子， 0 < α < 1；β是趋势平滑因子，0 < β < 1； γ是季节改变平滑因子0 < γ < 1。

初始化趋势估计b0的公式为：

|  |  |
| --- | --- |
|  |  |

对三次指数平滑法而言，我们必须初始化一个完整的“季节”Ci的值，不过我们可以简单地设置为全1(针对累乘式)或全0(针对累加式)。只有当序列的长度较短时，我们才需要慎重考虑初始值的选取[3]。

Holt-Winters模型就是三次指数平滑法。所有的指数平滑法都要更新上一时间步长的计算结果，并使用当前时间步长的数据中包含的新信息。它们通过“混合”新信息和旧信息来实现，而相关的新旧信息的权重由一个可调整的参数来控制。各种指数平滑法的不同之处在于它们跟踪的量的个数和对应的参数的个数。其中三次指数平滑法的功能最强大，既能体现趋势性又能体现季节性，体现在三次指数平滑法的参数最多，有三个[3]。

Holt-Winters模型有三个可调参数，我们的目的就是训练出有效的α，β，γ。有两种方法，一种是手动赋值，一种是采用数值优化的思想，比如使用最小二乘法最小化误差来求参数。此处根据固定的起始周期数计算初始参数值。由于数据量过小，采用全部数据进行参数值计算。

* + 1. 实现步骤

在运量走势方面，我们将数据进行以下处理并进行操作。

* 1. 整理出两地之间前18个月运量随月份变化的序列。
  2. 通过计算前18个月中每年前6个月的趋势和差值，首先基于第7～12个月的曲线人工拟合出第19～24个月的曲线。因为要获得正确合理的holt-winters中年与年之间的变化趋势，至少需要两个周期的数据，而前面分析得此时间序列以年为周期，而给的数据过少只有18个月，所以只能人为先拟合后六个月的数据。
  3. 将前两年（2018和2019年）的数据作为输入，通过holt-winters模型进行三次指数平滑，通过所有数据（前两年）训练出有效的α，β，γ，即根据固定的起始周期数计算初始参数值。并最后通过加法模型获得预测的序列。
  4. 车货匹配

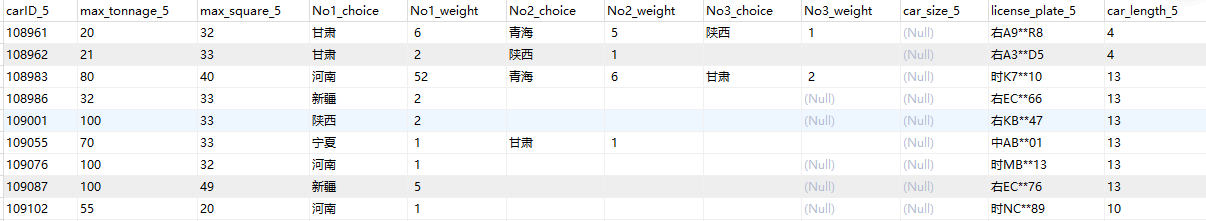
本车货匹配方案综合考虑供需匹配与地点，需要考虑起始点终点后，将地址数据、运单数据、整理后的车辆数据进行全连接后进行分析，可以看出其实对于一辆车来说，它的运单起终点选择是相对固定的，结合mysql以及python，使用数学建模与机器学习方法XGBoost，可以对同一辆车的历史起终点进行统计，对于同一辆车，历史频次越高则权重更大，即相同需求下，认为同一辆车更倾向选择权重大的起终点进行运输。

这里有一点需要特别说明的，通过数据分析，可以看出客户A的发货点只有一个地方，而客户B的发货点和收货点都相对随机，在以下所有车货匹配方案中的推荐车辆备选中，比如客户A仅仅需要考虑收货地点与车辆的关系，而客户B则需要综合考虑起点和终点与车辆的关系。

假若起终点以及吨数方数确定，具体车货匹配如下：

①只要车辆最大吨数和方数大于运单需求的吨数和方数，则认为该车可以作为承运车辆；

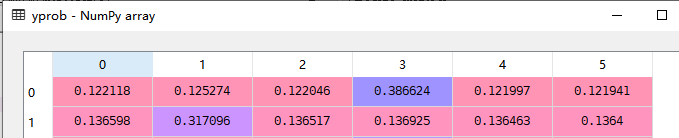
②考虑到车货匹配的反馈率要求，通过数据分析选出较频繁的备选车辆（即最好起终点均在推荐方案中，可以简单理解为起点和终点的频次和），此处每辆车我们给出三种程度的推荐，如图13所示。



1. 推荐结果

由图13可以看出，No1\_choice,No2\_choice,No3\_choice分别对应三种程度的优先度，No1\_weight,No2\_weight,No3\_weight分别对应三种程度的权重（与历史运单频次相关，两地间运单越频繁，相对应的权重越大），其实这里的权重也会在后面的XGBoost的预测中体现出来。

③通过XGBoost进行预测，得出空闲可用车辆的优先度。比如假如通过数据分析选出较频繁的6辆备选车辆，将此六辆车与之前一年半的订单的供需关系进行训练得到XGBoost模型，并通过XGBoost模型进行概率预测（类似于多分类问题），并将预测出的匹配概率进行排序，得到预测优先度。比如如图14所示，每行是输入的订单，每列是XGBoost模型选择对应的概率~，一定程度上反映了历史的记录的影响，并具有一定的泛化能力。



1. XGBoost模型预测6辆车对应的概率

④将同一起终点的所有空闲可用车辆的优先度与路径系数综合考虑，将  进行排序，选出最大值即应该匹配的车辆，具体考虑方案可以参考当前的情况，假如物流货运没有很紧张，一般可主要考虑空闲可用车辆的优先度，即将设为0.85左右。

* 1. 线路分布
     1. 查询满足需求的历史线路分布

模块查找历史路线中经过需求城市的记录，对于不同路线的运单数，统计占经过需求城市的运单总数的比例。以在陕西咸阳提货，在甘肃庆阳和宁夏银川卸货的需求为例，查询结果如图15所示，历史运单中满足需求的线路共有2条，有3个运单选择第一条路线，占满足需求的运单总数5的60%，所以流行度是0.6；2个运单选择第二条路线，占满足需求的运单总数5的40%，所以流行度是0.4.热门度已经提前计算，结果如表格第3列。

文本

已生成高可信度的说明

1. 查询满足需求的线路记录示例
   * 1. Floyd算法
2. 定义城市与F算法矩阵索引的对应关系。

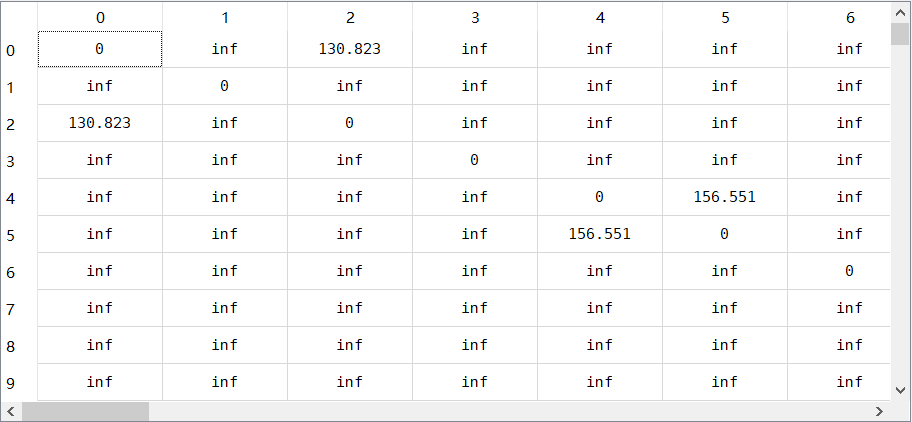
给全国720个城市（含县级市）编号。部分映射关系如表4所示。

1. 城市对应的矩阵索引

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 城市 | 北京 | 上海 | 天津 | 重庆 | 呼和浩特 | 包头 | 乌海 | 赤峰 | 通辽 |
| 编号 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 城市 | 鄂尔多斯 | 呼伦贝尔 | 巴彦淖尔 | 乌兰察布 | 兴安盟 | 锡林郭勒盟 | 阿拉善盟 | 丰镇市 | … |
| 编号 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | … |

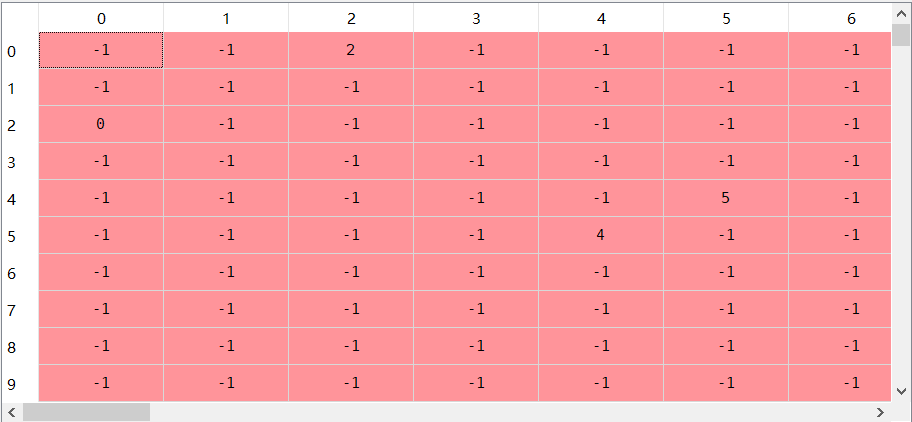
1. 初始化图的权值矩阵和路由矩阵

初始化权值矩阵如图16所示。主对角线全是0，表示每个城市到自身的距离是0.两城市如果接壤，矩阵对应位置是有限值，表示城市之间的距离，单位千米；如果不接壤，则对应位置是无穷。计算方法是使用geopy库查询出每个城市的GPS坐标，基于GPS坐标计算两城市的直线距离。



1. 初始化权值矩阵

初始化路由矩阵如图17所示。和矩阵的值存储在newDistance.npz文件中.



1. 初始化路由矩阵
2. 运行Floyd算法，迭代求任意两节点之间的最短路径。
3. 运行结束后，得到任意两节点之间最短路径的权值矩阵和路由矩阵。和存储在newDistance\_shortest.npz文件中。

权值矩阵如图18所示，矩阵中每个元素代表两城市之间最短路径的长度。

矩阵中的无穷表示对应两城市之间不可达。出现不可达的原因有：①两城市确实没有可达路线。比如海南省的城市与广东省的城市因为隔海而不接壤，所以求最短路径会返回不可达。②地图数据库数据不准确，或者说数据不是我们希望的结果。最短路线模块判断两节点之间是否有边（城市接壤与否）是基于OpenStreetMap平台的地图数据库，数据库中部分城市边界的坐标并不与城市实际的边界完全对应，例如经查询发现数据库中重庆市的边界坐标大概在重庆市郊，而重庆下辖很多区县，实际范围比数据库中的边界大。所以最终结果中出现重庆周边城市到重庆都不可达的现象。解决这个问题需要更符合我们需求的数据库，限于比赛时间我们暂时搁置这个问题。

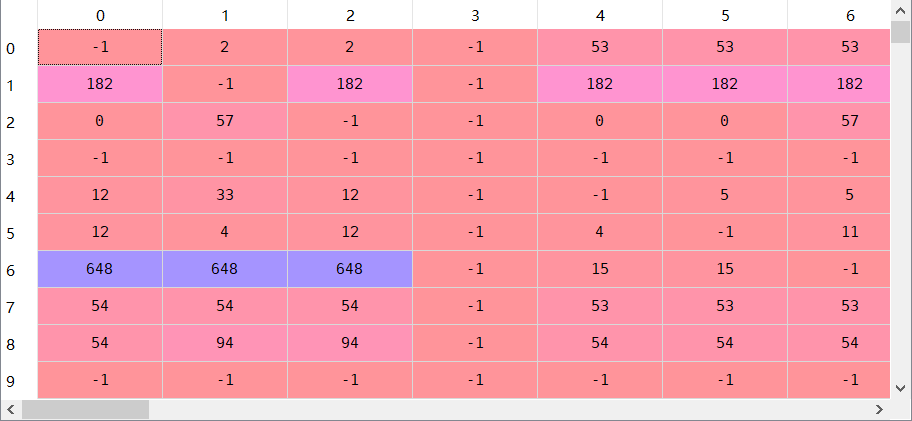
图片包含 表格

已生成极高可信度的说明

1. 最终结果权值矩阵

路由矩阵如图19所示，采用前向路由，由矩阵可以得到任意两节点之间最短路径的路由。

例如，由路由矩阵得到陕西咸阳到青海西宁的最短路径是陕西,咸阳市->陕西,宝鸡市->甘肃,天水市->甘肃,甘南藏族自治州->青海,海东地区->青海,西宁市，由权值矩阵得到路线近似总长度是834.69千米。



1. 最终结果路由矩阵
   * 1. 最短路线算法的实现

最短路线算法的伪代码如下。

|  |
| --- |
| 代码2：最短路线算法  Step1: 设所有的提货城市组成集合*S*，卸货城市组成集合*T*。设*m*是集合*S*的元素个数，*n*是集合*T*的元素个数。  Step2: 分情况讨论  Step2.1: 如果提货城市和卸货城市的数量分别都是1，由于图论优化算法以城市作为图的节点，所以对于两个节点的情况直接输出最短路径，给出最短路线的长度和路由。  Step2.2: 对于有多个提货城市或者多个卸货城市的情况，这时候最短路线问题转化为图论中必须经过某些中间节点的最短路问题。经过查阅资料，我们发现这是一个NP-hard 问题，即没有好的多项式时间算法能解决这个问题，想要获得全局最优解只能靠穷举。考虑到必须先在所有的提货点提货完毕后才能进行卸货的约束条件，问题转化为必须先依次经过集合*S*中的所有节点之后再依次经过集合*T*中所有节点的最短路径。  模型采用遍历算法，即对于*S*中元素的每一种全排列和*T*中元素的每一种全排列，将提货点集合中两两节点之间用最短路径连接（形成的路径称为提货路径），卸货点集合中两两节点之间用最短路径连接（形成的路径称为卸货路径），将提货路径终点与卸货路径起点用一条最短的路径连接起来。对于所有路径，找出长度最短的就是全局最优解，即满足需求的最短路线。  //遍历算法  Step 2.2.1: 对*S*中元素，两两之间求最短路径，保存路径的长度和路由信息  Step 2.2.2: 对*T*中元素，两两之间求最短路径，保存路径的长度和路由信息  Step 2.2.3: 对*S*中任意一个元素*x*和*T*中任意一个元素*y*，求*x*和*y*之间的最短路径，保存路径的长度和路由信息  Step 2.2.4: 对于*S*中元素的一种全排列（，就是一种顺序）和*T*中元素的一种全排列（），将两个节点序列拼接起来得到，求序列中相邻两节点之间的最短路径的长度和路由，求和得到整个序列的长度，拼接得到整条路线的路由。  Step 2.2.5: 求出所有序列中长度最小的，该城市序列就是最优的提货、卸货地点顺序*Si*和*Tj*。  Step 2.2.6: 将*Si*和*Tj*拼接起来，得到最优地点访问顺序。对应的路线路由和路线长度都可以得到  Step3: 输出近似最短路线的地点访问顺序、路线的路由和路线的长度 |

实际代码实现中，最后会输出长度最短的三条路线给评分模型。评分模型会推荐得分最高的三条路线。

对最短路线算法的可行性进行简要说明。遍历算法的时间复杂度和空间复杂度都很高，均为O(*m*!*n*!)级。如果提货或卸货城市数量非常多，算法求解的时间将非常长。不过客户A的数据集中涉及到的城市数（即图的节点数）最多为8，客户B的数据集中涉及到的城市数最多为5，使用该算法可以做到计算时间是毫秒量级或者秒级，所以针对本问题的场景是可以使用遍历算法。

1. 分析结果
   1. 运量走势
      1. 结果展示

以下以客户A咸阳到西安两地为例，对运量走势的分析结果进行说明。图20是运量预测结果曲线图。

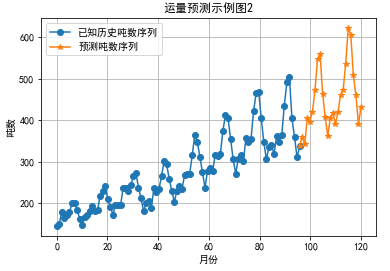
* 1. 整理出两地之间前18个月运量随月份变化的时间序列后（如图中蓝线所示），通过计算前18个月中每年前6个月的趋势和差值，基于第7～12个月的曲线人工拟合出第19～24个月的曲线（如图中黄线所示）。
  2. 将前24个月的数据作为模型的输入，通过holt-winters模型进行三次指数平滑，通过所有数据（前两年）训练出有效的α，β，γ，即根据固定的起始周期数计算初始参数值。并最后通过加法模型获得预测的序列。预测出的后两年的运量时间序列如图中绿线所示。
     1. 结果分析

1. 由图20可以看出，我们预测的结果中，可以较明显的看出序列以年为周期变化，同时模型较好的提取和保存了序列周期内的变化特征。所以可以看出运量走势方面预测结果较好。
2. 需要注意到，应用时间序列平滑法的前提条件是：（1）所预测的客观事物渐进式发展，无跳跃性的变化；（2）过去和现在影响客观事物发展的因素也决定着客观事物未来的发展。由于客观事物的发展变动受多种因素的影响，时间序列平滑法在客观影响因素发生较大变化时可能产生较大的预测误差。为降低这些可能的预测误差，必须充分研究客观影响因素可能的发展与变动，将定性分析和定量研究结合起来，这样才能提高预测的精度。而如果季节性需求不平稳，序列变化趋势没有明显的季节性特征，无法与随机变化区分开来，就很难预测下一期的需求走势。因为运量数据集本身时间跨度的限制，所以为了避免随机性（即噪声）过大，预测的时间粒度不能过小，否则会导致捕获不到正确的特征。我们预测的粒度是月，序列的周期是年。
3. 为了更好的说明本预测模型的合理性，我们假设有较长时间跨度的输入数据，如图21中蓝线所示，是96个月的运量时间序列，则可以得到如黄线所示的预测结果。从图中可以更清晰地看出，此模型在正确保持了序列周期性的同时，也较好地学习到了周期之间的变动趋势和周期内的共同特征。同时关于累加性与累乘性模型的选择取决于数据本身，此处我们根据数据的特征选择了累加性模型进行预测。需要注意的是，预测是不可避免地存在一定的估算误差，短期预测比长期预测更准确。运量预测是基于历史数据的，考虑到均衡补货和缺货，我们的运量预测模型尝试给出一个尽可能准确的未来运量的估值。

图表, 折线图

已生成极高可信度的说明

1. 客户A咸阳到西安运量走势预测结果



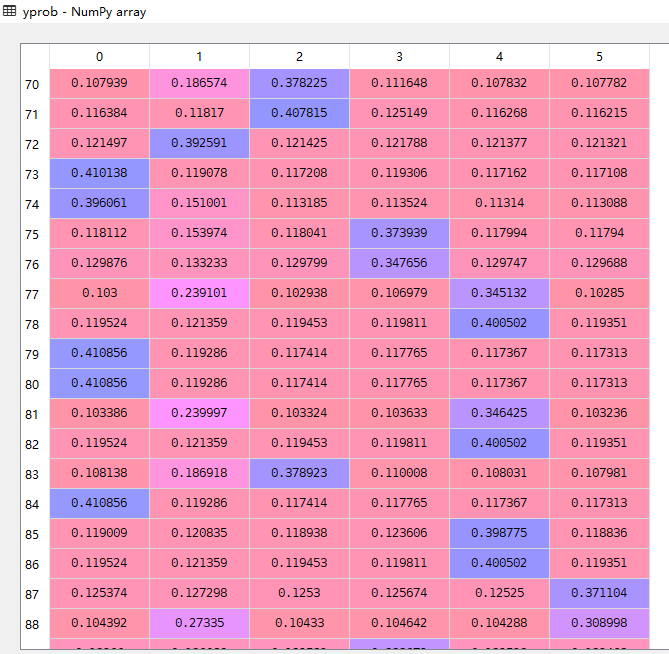
1. 输入较长时间跨度的运量时间序列时模型的预测结果示例
   1. 车货匹配

本车货匹配方案综合考虑供需匹配与地点，一共可对3838辆车进行进一步带有优先级权重的推荐，即将同一起终点的所有空闲可用车辆的优先度与路径系数β综合考虑并按比例进行排序，选出最大值即应该匹配的车辆，假如物流货运没有很紧张，一般可主要考虑空闲可用车辆的优先度α。

XGBoost批量预测结果如图22所示，每行是某时刻同时输入的订单（以订单为单位），每列是XGBoost模型选择对应的概率α，一定程度上反映了历史的记录的影响，并具有一定的泛化能力。

需要注意的是，此处我们假定这些同起终点的订单是并发的，所以可以看出有多订单同时偏好同一辆车的情况，面对这种情况我们有以下解释：第一现实中不会有这么多同起终点运单并发；第二最终还要结合路径系数β综合考虑并按比例进行排序，选出最大值即应该匹配的车辆；第三假如发现选出来的车已经占用，则再进行一次预测（此时之前匹配的车已经不可用，因为它已不再空闲，所以不会进入备选名单）。

最终本车货匹配方案结合了数学建模与机器学习XGBoost的方法，对每一个订单给出最匹配的车辆与基于车辆参数表表5给出对应的车型推荐（因为车辆基础数据中的车型与车辆参数的相关性很低，所以使用表5的统一标准）。



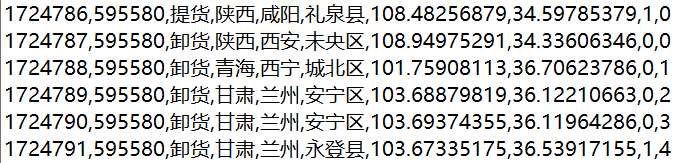
1. XGBoost批量预测结果
2. 常用车型基本参数



* 1. 线路分布
     1. 历史上有一条线路但是可能绕路的推荐方案

客户A有在陕西咸阳提货，在陕西西安、青海西宁、甘肃兰州卸货的需求。从图23看到，查询到满足需求的历史运单（ID 595580）采用的路线是在陕西咸阳提货，依次在陕西西安、青海西宁、甘肃兰州卸货。但是通过查看地图可知，从陕西西安去往青海西宁的途中会经过甘肃兰州，所以这项记录所使用的路线绕路。

从图24看到，模型用最短路线算法推荐的路线是依次在陕西西安、甘肃兰州、青海西宁卸货，具体路由是陕西,咸阳市->陕西,西安市->陕西,咸阳市->甘肃,平凉市->甘肃,定西市->甘肃,兰州市->甘肃,白银市->甘肃,武威市->青海,西宁市。从表6看到模型的评分结果。模型最终推荐的最高分路线是历史运单的路线1，而且给出了路线的近似总长度是1384.486千米和推荐星级是3.第二高分路线是最短路线，近似总长度是1042.47 km和推荐星级是2.从推荐结果看出，模型在推荐路线时综合考虑历史流行度和路线长度两方面因素，推荐多条合适路线给司机选择。



1. 运单595580经过的节点信息

1. 案例5.3.1程序运行结果

算法推荐的最优地点访问顺序:

陕西,咸阳(提货)->陕西,西安(卸货)->甘肃,兰州(卸货)->青海,西宁(卸货)

历史上符合查询要求的运单的路线和选择这条路线的人数

track freq distance

0 陕西,咸阳(提货)->陕西,西安(卸货)->青海,西宁(卸货)->甘肃,兰州(卸货) 1 1384.486368

系统推荐路线：

路线1：

节点序列：陕西,咸阳(提货)->陕西,西安(卸货)->青海,西宁(卸货)->甘肃,兰州(卸货)

具体路线：陕西,咸阳市->陕西,西安市->陕西,宝鸡市->甘肃,天水市->甘肃,甘南藏族自治州->青海,海东地区->青海,西宁市->甘肃,武威市->甘肃,白银市->甘肃,兰州市

推荐星级：3

近似总长度:1384.486 km

路线2：

节点序列：陕西,咸阳(提货)->陕西,西安(卸货)->甘肃,兰州(卸货)->青海,西宁(卸货)

具体路线：陕西,咸阳市->陕西,西安市->陕西,咸阳市->甘肃,平凉市->甘肃,定西市->甘肃,兰州市->甘肃,白银市->甘肃,武威市->青海,西宁市

推荐星级：2

近似总长度:1042.47 km

路线3：

节点序列：陕西,咸阳(提货)->甘肃,兰州(卸货)->青海,西宁(卸货)->陕西,西安(卸货)

具体路线：陕西,咸阳市->甘肃,平凉市->甘肃,定西市->甘肃,兰州市->甘肃,白银市->甘肃,武威市->青海,西宁市->青海,海东地区->甘肃,甘南藏族自治州->甘肃,天水市->陕西,宝鸡市->陕西,西安市

推荐星级：1

近似总长度:1856.416 km

1. 专家评议法对路线评分情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 路线 | 运单数 | 热门度 | 流行度 | 长度 | 路线长度  度量 | 总分 |
| 陕西,咸阳(提货)->陕西,西安(卸货)->青海,西宁(卸货)->甘肃,兰州(卸货) | 1 | 0.0326331 | 1 | 1384.49 | 0.526465 | 0.569759 |
| 陕西,咸阳(提货)->陕西,西安(卸货)->甘肃,兰州(卸货)->青海,西宁(卸货) | 0 | 0.0165872 | 0 | 1042.47 | 0.719665 | 0.36315 |
| 陕西,咸阳(提货)->甘肃,兰州(卸货)->青海,西宁(卸货)->陕西,西安(卸货) | 0 | 0.0165872 | 0 | 1856.42 | 0.259463 | 0.133049 |

* + 1. 历史上有多条线路给出的推荐方案

客户A有在陕西咸阳提货，在甘肃庆阳和宁夏银川卸货的需求。通过检索历史运单，如图25所示，发现有3个运单先在甘肃庆阳，后在宁夏银川卸货；有2个运单顺序相反。最短路线算法也同时推荐了这两条路线，给出了两条路线具体路由情况，发现路线2绕路，长度比路线1多了337km。两条路线的评分如表7所示，最后推荐的得分较高的路线1是陕西,咸阳(提货)->甘肃,庆阳(卸货)->宁夏,银川(卸货)，近似总长度567千米，推荐星级是3；路线2是陕西,咸阳(提货)->宁夏,银川(卸货)->甘肃,庆阳(卸货)，近似总长度904千米，推荐星级是2。从推荐结果看出，模型根据路线长度给流行度高的路线加分，推荐的较高分数的路线1更加科学。

算法推荐的最优地点访问顺序:

陕西,咸阳(提货)->甘肃,庆阳(卸货)->宁夏,银川(卸货)

历史上符合查询要求的运单的路线和选择这条路线的人数

track freq distance

0 陕西,咸阳(提货)->甘肃,庆阳(卸货)->宁夏,银川(卸货) 3 567.485280

1 陕西,咸阳(提货)->宁夏,银川(卸货)->甘肃,庆阳(卸货) 2 904.063803

系统推荐路线：

路线 1：

节点序列：陕西,咸阳(提货)->甘肃,庆阳(卸货)->宁夏,银川(卸货)

具体路线：陕西,咸阳市->甘肃,庆阳市->宁夏,吴忠市->宁夏,灵武市->宁夏,银川市

推荐星级：3

近似总长度:567.485 km

路线 2：

节点序列：陕西,咸阳(提货)->宁夏,银川(卸货)->甘肃,庆阳(卸货)

具体路线：陕西,咸阳市->甘肃,庆阳市->宁夏,吴忠市->宁夏,灵武市->宁夏,银川市->宁夏,灵武市->宁夏,吴忠市->甘肃,庆阳市

推荐星级：2

近似总长度:904.064 km

1. 案例5.3.2程序运行结果
2. 专家评议法对路线评分情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 路线 | 运单数 | 热门度 | 流行度 | 路线长度 | 路线长度  度量 | 总分 |
| 陕西,咸阳(提货)->甘肃,庆阳(卸货)->宁夏,银川(卸货) | 3 | 0.0632037 | 0.6 | 567.485 | 0.669762 | 0.527522 |
| 陕西,咸阳(提货)->宁夏,银川(卸货)->甘肃,庆阳(卸货) | 2 | 0.0481638 | 0.4 | 904.064 | 0.330238 | 0.294752 |

* + 1. 历史上没有运输线路给出的推荐方案

客户B有在浙江宁波提货，在河南郑州和广东广州卸货的需求。通过查询历史运单，发现没有符合需求的运单。系统推荐两条长度最短的路线，线路1是浙江宁波(提货)->河南郑州(卸货)->广东广州(卸货)，近似总长度2372.609 km，推荐星级是2。线路2是浙江,宁波(提货)->广东,广州(卸货)->河南,郑州(卸货)，近似总长度2639.757 km，推荐星级是1。程序运行结果如图26所示。路线评分结果见表8。模型根据路线长度度量的分数优先推荐长度更短的路线。

1. 案例5.3.3程序运行结果

算法推荐的最优地点访问顺序:

浙江,宁波(提货)->河南,郑州(卸货)->广东,广州(卸货)

历史上没有符合要求的运单

系统推荐路线：

路线 1：

节点序列：浙江,宁波(提货)->河南,郑州(卸货)->广东,广州(卸货)

具体路线：浙江,宁波市->浙江,嘉兴市->浙江,湖州市->江苏,溧阳市->江苏,南京市->安徽,明光市->安徽,蚌埠市->安徽,亳州市->河南,周口市->河南,郑州市->河南,许昌市->河南,漯河市->河南,驻马店市->河南,信阳市->湖北,黄冈市->江西,九江市->江西,南昌市->江西,丰城市->江西,吉安市->江西,赣州市->广东,英德市->广东,广州市

推荐星级：2

近似总长度:2372.609 km

路线 2：

节点序列：浙江,宁波(提货)->广东,广州(卸货)->河南,郑州(卸货)

具体路线：浙江,宁波市->浙江,台州市->浙江,温州市->福建,宁德市->福建,三明市->福建,龙岩市->广东,梅州市->广东,惠州市->广东,广州市->广东,英德市->江西,赣州市->江西,吉安市->江西,丰城市->江西,南昌市->江西,九江市->湖北,黄冈市->河南,信阳市->河南,驻马店市->河南,漯河市->河南,许昌市->河南,郑州市

推荐星级：1

近似总长度:2639.757 km

1. 专家评议法对路线评分情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 路线 | 运单数 | 热门度 | 流行度 | 路线长度 | 路线长度  度量 | 总分 |
| 浙江,宁波(提货)->河南,郑州(卸货)->广东,广州(卸货) | 0 | 0.0240339 | 0 | 2372.61 | 0.669762 | 0.339688 |
| 浙江,宁波(提货)->广东,广州(卸货)->河南,郑州(卸货) | 0 | 0.0240339 | 0 | 2639.76 | 0.330238 | 0.169926 |

* 1. 为网络货运平台提供的参考方案

通过Holt-Winters模型我们成功预测客户未来在任意两地之间提货、卸货运量需求情况。例如我们预测出客户A有在陕西咸阳提货，陕西西安、甘肃兰州、宁夏银川卸货的需求，要求提货时间2020年12月10日8：00，要求卸货时间2020年12月11日23：59，方数28，吨数9.6。

我们将预测的运单需求输入车货匹配模型，模型会根据运单参数推荐出合适的车辆1、2、3。根据反馈率等指标对车辆的优先度进行排序，再根据承运时间内车辆预计的空闲情况给车辆分配运输任务，或者让车辆在运单的要求提货时间之前保持空闲，并到预测的提货地点附近等待接单。这可以大大减少车辆的空闲等待时间，提高车辆利用率。例如，车辆1的优先度0.1，车辆2的优先度0.3，车辆3的优先度0.2。车辆2在2020年12月8日12：00到2020年12月10日23：59承担运单，无法参加预测运单的运输任务，将车辆2从推荐车辆列表中删除。这时推荐优先度较高的车辆3，平台提前给车辆3的司机提供建议，建议其在2020年12月10日8：00之前到陕西咸阳客户A的提货地附近等待接单，并在运单要求提货时间之前不要接受其它运单请求。

此外根据预测的运单需求，线路分布模块通过查找历史上满足需求的运单走的线路，给出历史上较多司机选择的运输线路。结合使用最短路线算法得到的提货、卸货地点顺序，帮助货车司机更合理的安排运输路线，节约时间和车辆成本。例如，前述预测需求在客户A的数据库中不存在，模块根据最短路线算法的结果向车辆3的司机推荐的顺序是陕西咸阳(提货)->陕西西安(卸货)->甘肃兰州(卸货)->宁夏银川(卸货)，建议司机不要先去甘肃兰州卸货、再去陕西西安卸货、最后去宁夏银川卸货，帮助司机少走回头路。

1. 总结

在运量走势方面，本数据分析方案给出两地之间未来运量的预测方法。使用python进行数据处理，得到了运量吨数和方数的时间序列。通过基于Holt-Winters模型的时间序列分析方法，较好地实现运量预测功能。

在车货匹配方面，本数据分析方案将数学建模与机器学习相结合。通过XGBoost模型预测车辆可用性度量，并考虑到车货匹配场景下的反馈率的需求以及车辆与提货地点之间的距离、路况，最后选择匹配推荐度最高的车辆作为订单匹配车辆。

在线路分布方面，本数据分析方案首先统计客户的历史线路分布，计算路线热门度；其次根据历史上符合需求的路线计算路线流行度，再结合路线长度度量，最后通过对三项指标评分，按照评分从高到低依次推荐最多三条路线。

我们将三个方面分析的结果串联起来，可以为网络货运平台提供参考方案。5.4节给出一个参考方案的示例。

综上所述，模型使用时下火热的数据挖掘和人工智能技术对客户物流数据集进行分析。通过时间序列分析预测客户的货运需求；根据车辆可用性和路况推荐车辆，帮助车辆匹配货运需求，实现物流资源的科学调度和合理配置；根据历史热门和最短路线原则优化节点访问顺序，缩短运输路线长度，为货运平台降本增效提供了解决方案。

参考文献

1. HUANG Shuli,HU Dasha,JIANG Yuming. Algorithm for finding shortest path which must go through specified intermediate node set[J]. CEA, 2015, 51(11): 41-46.

1. <https://baike.baidu.com/item/Holt-Winters%20%E6%96%B9%E6%B3%95/24137738>
2. https://www.cnblogs.com/bonelee/p/9441934.html