队伍编号	MC2406053
题号	C

论文标题

摘要

摘要内容

关键词: 随机森林算法、LSTM

目录

1 问题重述

1.1 问题背景

随着网购的流行，电商物流显得更加重要。在电商物流网络中，订单配送的过程包括多个环节，其中核心环节之一是分拣。分拣中心负责根据不同的目的地对包裹进行分类，并将它们发送到下一个目的地，最终交付给顾客。因此，提高分拣中心的管理效率对整个网络的订单交付效率和成本控制至关重要。

货量预测在电商物流网络中扮演着至关重要的角色。准确预测分拣中心的货物量是后续管理和决策的基础。通常，货量预测是根据历史货物量、物流网络配置等信息，来预测每个分拣中心每天和每小时的货物量。

分拣中心的货量预测与网络的运输线路密切相关。通过分析各线路的运输货物量，可以确定各分拣中心之间的网络连接关系。当线路关系发生变化时，可以根据调整信息来提高对各分拣中心货量的准确预测。

基于货量预测的人员排班是下一步需要解决的重要问题。分拣中心的人员包括正式员工和临时工两种类型。合理安排人员旨在完成工作的前提下尽可能降低人力成本。根据物流网络的情况，制定了人员安排的班次和小时人效指标。在确定人员安排时，优先考虑使用正式员工，必要时再增加临时工。

1.2 问题描述

1.2.1 问题一

根据所给的前 3 个月每天的货量数据以及前一个月每一个小时的货量数据建立货量预测模型，对 57 个分拣中心未来 30 天每天及每小时的货量进行预测。

1.2.2 问题二

由于分拣中心之间存在货物运送的关联，在已知过去三个月内的货物运输关联情况下，若运输关系发生变化，预测分拣中心的货量变化。

1.3 总体思路分析

2 符号说明与基本假设

2.1 符号说明

表 1: 符号说明

符号	说明

2.2 基本假设

3 问题一的建模与求解

3.1 分析

问题一给出了分拣中心在过去的四个月内每天的货量以及过去 30 天内每个小时的货量。在这些数据的基础上要对下一个月（30 天）的货量进行预测，我们先对给出的数据进行观察。我们选取，例如 SC6,SC41 来观察。利用 python/matplotlib 作图

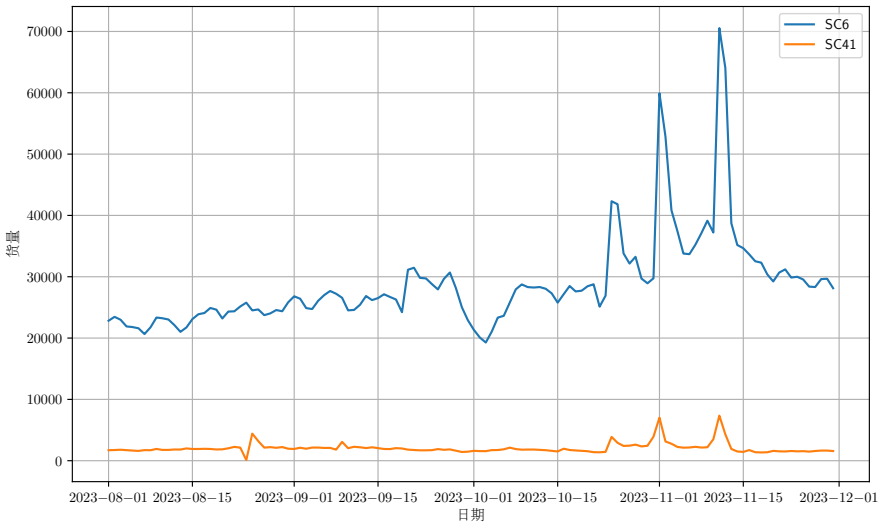

图 1: SC6,SC41 的日期-货量图线

根据观察，我们发现在特定的时间段会出现一些”异常值”，而这些异常值的出现很可能是由于节假日等因素引起的。因此为了预测 30 天后的货物量，我们选取几种深度学习的方法来进行预测。

1. LSTM
2. ARIMA
3. 随机森林算法

3.1.1 LSTM

长短期记忆（英语：Long Short-Term Memory，LSTM）是一种时间循环神经网络（RNN）。由于独特的设计结构，LSTM 适合于处理和预测时间序列中间隔和延迟非常长的重要事件。LSTM 通过引入“门”结构来调节信息的流动，这使得它在长序列数据上表现得更好。LSTM 的核心组件有遗忘门（Forget Gate）、输入门（Input Gate）、细胞状态（Cell State）、输出门（Output Gate）。LSTM 的每个时刻会通过上述四个门控制信息的流动。遗忘门决定丢弃哪些过去的信息，输入门和它的候选值共同决定将哪些新信息加入细胞状态。细胞状态随后更新，最后通过输出门确定应输出什么信息。

通过这样的结构，LSTM 能够在处理序列数据时有效地保留长期依赖信息，解决了传统 RNN 在长序列上的梯度消失或爆炸问题。

3.1.2 ARIMA

ARIMA 模型（自回归积分滑动平均模型）是一种广泛应用于时间序列预测的统计模型。ARIMA 模型结合了自回归（AR）、差分（I）和移动平均（MA）三种主要组成部分，适用于分析和预测具有时间依赖性的数据。其核心组件包括自回归（AR）部分、差分（I）部分、移动平均（MA）部分。将这三部分结合起来，ARIMA 模型的完整形式可以表示为：

$$\phi(B)\nabla^d X_t = \theta(B)a_t$$

其中 $\phi(B)X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}$，$p$ 是自回归项的阶数,ϕ_i 是自回归系数，B 是退后算子。$\theta(B)a_t$ 则类似，a_t 是时间序列的误差项。

从式子中可以看出，通过对原始数据进行差分转换，ARIMA 能够将非平稳时间序列转换为平稳时间序列，因此模型能够适用于预测那些显示出趋势或季节性模式的数据。

3.1.3 随机森林算法

随机森林是一种集成学习方法，特别适合用于分类、回归和其他任务，通过构建多棵决策树在训练时并输出模式的类（分类）或平均预测（回归）。随机森林算法的核心思想是通过合并多个决策树的预测结果来提高整体模型的预测准确性和稳定性。

随机森林算法的主要步骤有：自助采样（Bootstrap sampling），从原始数据集中随机（允许重复的）选择 N 个样本构成了一个训练集。N 次采样产生的 N 个训练集被用于训练 N 棵决策树；构建决策树：对于每棵树的每个节点，随机选择 k 个特征（而不是所有特征），然后使用这些特征中的最佳分割方法来分割节点。这种“特征随机选择”的做法增加了树之间的差异性。决策树的生长：每棵树都尽可能地生长而不进行剪枝。每棵树都完全依赖于自助采样得到的训练数据集来建立。聚合预测：对于分类问题，使用多数投票法；对于回归问题，计算所有树的输出值的平均值。

随机森林算法的成功在于它的简单性和在多种数据集上表现出的高效性与准确性。它既能处理分类问题，也能处理回归问题，同时还能进行特征选择，是一种非常强大且灵活的机器学习算法。虽然随机森林通常不如专门的时间序列或图数据模型那样直接适用于处理分拣中心之间的货物流动问题，但它仍然可以在某些情况下提供有价值的见解和预测，特别是随机森林可以提供关于哪些特征（例如历史货物流量、时间因素、分拣中心之间的距离等）对预测货物流量最有影响的洞察。这对于理解货物流量模式和优化物流网络可能非常有用。

3.2 按天的数据模拟与预测

根据选择，我们对其进行数据模拟。我们通过使用 `python` 的 `sklearn` 中提供的模型来进行拟合。首先我们通过分类不同的仓库，获取所有的信息。在分类完成之后，我们利用几种方法来实现。首先是利用 LSTM 来生成预测的曲线。[1]如图2。

相同的，我们选取随机森林进行预测可以得到图3。另外由于我们发现数据有较大的波动，而 ARIMA 模型适用于平稳时间序列，而从图中看，原始数据不是平稳的。因此直接使用 ARIMA 模型可能不会得到好的预测效果。我们尝试季节性 ARIMA (SARIMA) 模型。

[1]具体代码在附录中。

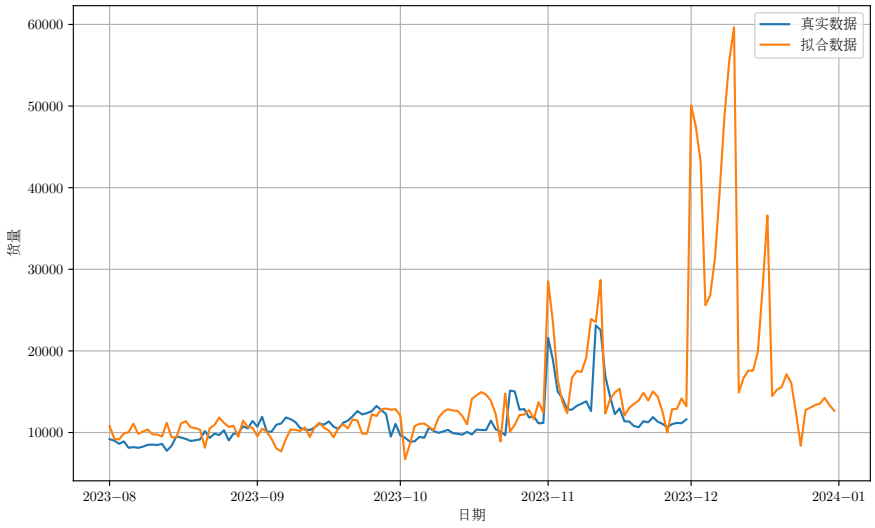

图 2: 利用 LSTM 预测图线示例

经过所输出的平均方差的对比，我们认为使用随机森林算法对于第一题是相对的最优解。因此我们对未来一个月之内每一个小时的货量预测也采取完全相同的方法预测。

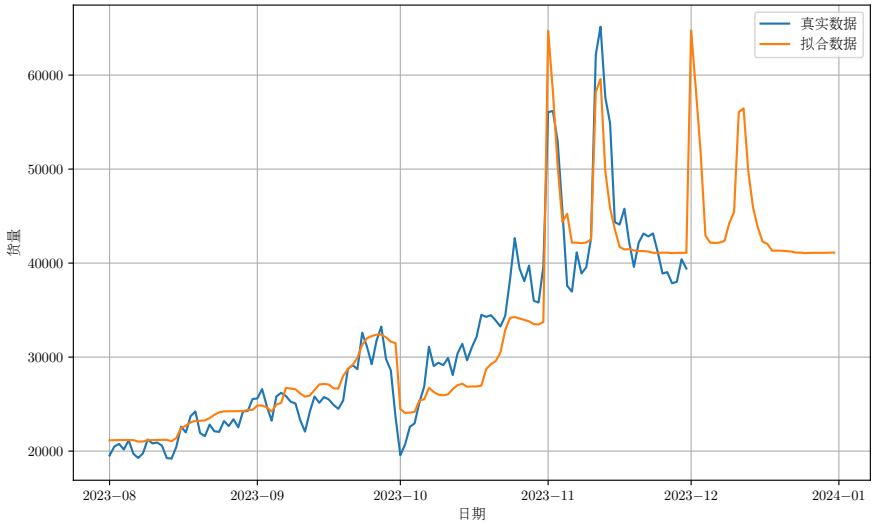

图 3: 利用随机森林预测图线示例（SC10）

3.3 按小时的数据模拟和预测

按照之前的方法，相似的预测之后的数据，如图5。从图中可以看出，拟合效果符合数据波动的周期性。通过计算拟合模型与真实数据之间的平均方差，我们确定使用随机森林算法来拟合数据。

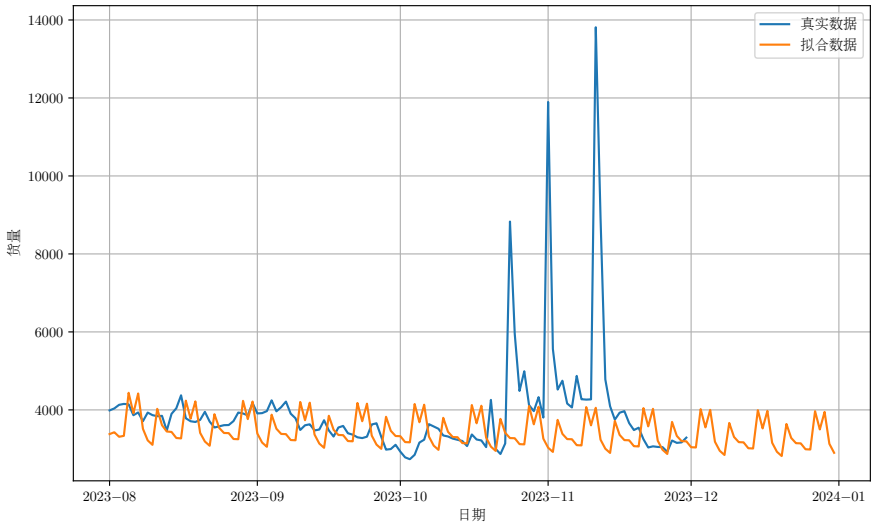

图 4: 利用 SARIMA 预测图线示例（SC55）

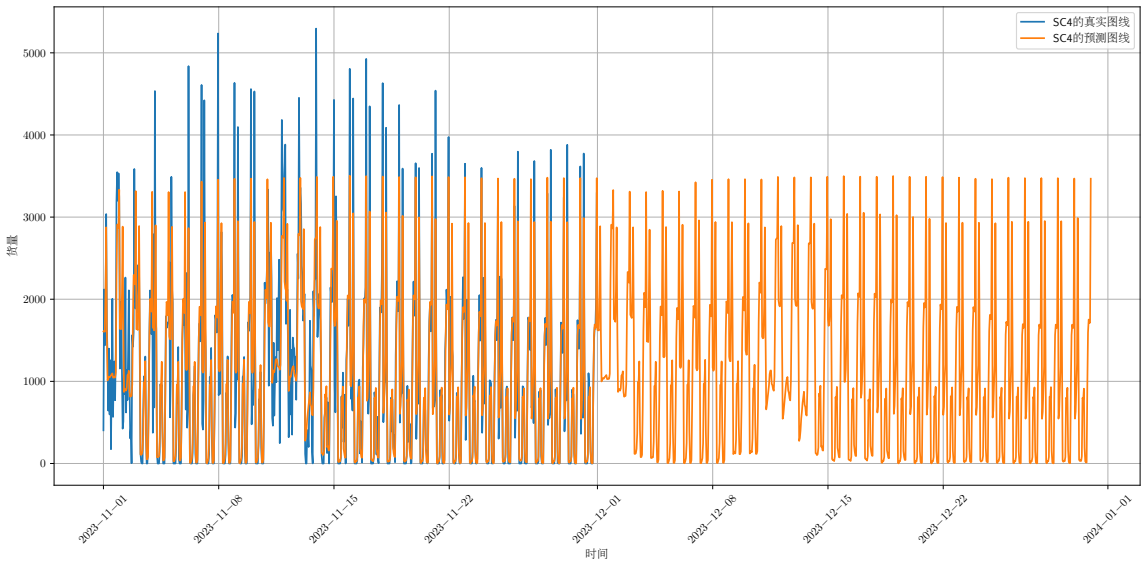

图 5: 利用随机森林模型预测小时图线示例（SC4）

4 问题二的建模与求解

4.1 分析

我们需要对已知的关系进行分析。因此我们首先先将仓库之间的关联进行可视化，我们选用弦图的形式，如图6。在现有的运输关系模型上发生了变动，此时我们需要考虑

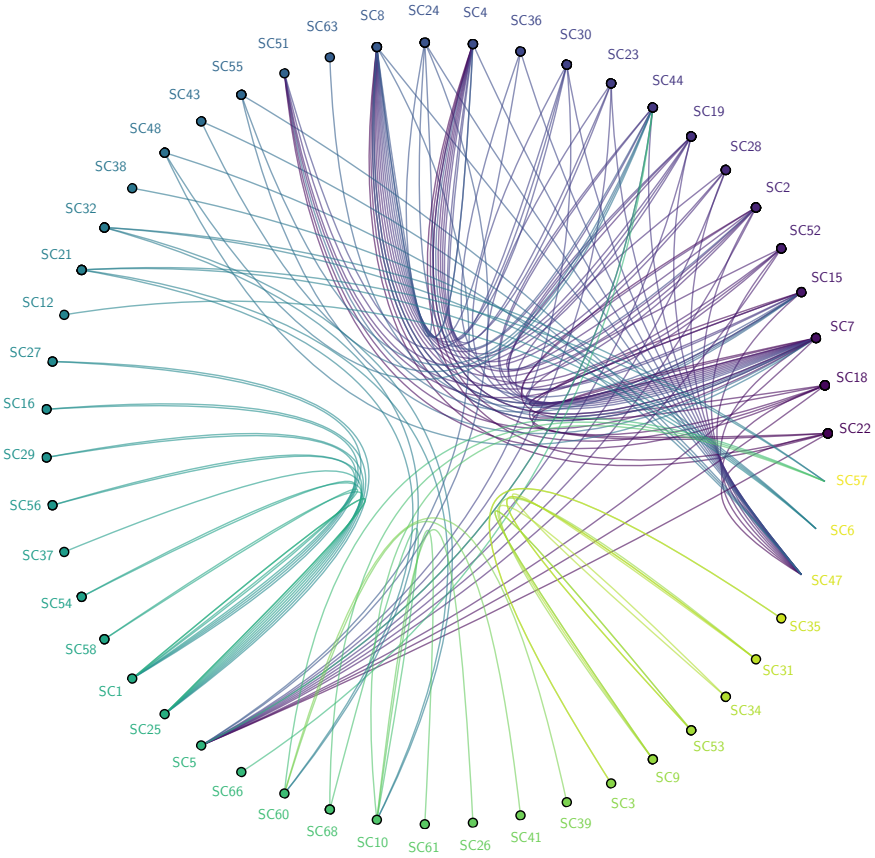

图 6: 分配中心之间的关系弦图

附录

A 问题一代码

随机森林调参代码

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler

# 假设存在的 SCid 列表
data_for_sc = pd.read_csv("../../附件/附件1.csv", encoding="GB2312")
ALL_SC = list(set(data_for_sc["分拣中心"]))
existing_scs = list(map(lambda SC_: int(SC_[2:]), ALL_SC))
existing_scs.sort()


# 加载数据
def load_data(existing_scs):
    all_data = []
    for sc_id in existing_scs:
        try:
            data = pd.read_csv(f"SC{sc_id}.csv")
            data["center_id"] = sc_id
            all_data.append(data)
        except FileNotFoundError:
            print(f"File for center {sc_id} not found.")
    return pd.concat(all_data, ignore_index=True) if all_data else pd.DataFrame()


# 数据清洗和预处理
def preprocess_data(data):
    data["date"] = pd.to_datetime(data["date"], errors="coerce")
    data.dropna(subset=["date", "value"], inplace=True)

    data["year"] = data["date"].dt.year
    data["month"] = data["date"].dt.month
    data["day"] = data["date"].dt.day
    data["weekday"] = data["date"].dt.weekday

    scaler = StandardScaler()
    data[["year", "month", "day", "weekday"]] = scaler.fit_transform(
        data[["year", "month", "day", "weekday"]]
    )

```

```
    return data


# 模型训练和参数调整
def train_and_optimize_model(X_train, y_train):
    param_grid = {
        "n_estimators": [100 * i for i in range(1, 10)],
        "max_depth": [10 * i for i in range(1, 10)],
        "min_samples_split": [10 * i for i in range(1, 10)],
    }
    model = RandomForestRegressor(random_state=42)
    grid_search = GridSearchCV(
        model, param_grid, cv=5, scoring="neg_mean_squared_error", verbose=2
    )
    grid_search.fit(X_train, y_train)

    print("Best parameters:", grid_search.best_params_)
    return grid_search.best_estimator_


if __name__ == "__main__":
    data = load_data(existing_scs)
    if not data.empty:
        data = preprocess_data(data)
        features = data[["center_id", "year", "month", "day", "weekday"]]
        target = data["value"]

        X_train, X_test, y_train, y_test = train_test_split(
            features, target, test_size=0.2, random_state=42
        )

        best_model = train_and_optimize_model(X_train, y_train)

        y_pred = best_model.predict(X_test)
        mse = mean_squared_error(y_test, y_pred)
        print(f"Test MSE: {mse}")
    else:
        print("No data loaded, please check the data files.")
```

随机森林训练代码

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# 假设数据已经按照分拣中心编号和日期排列好
# 每个文件名格式为'SCi.csv', 其中 i 是分拣中心编号
```

```python
# 假设存在的 SCid 列表
data_for_sc = pd.read_csv("../../附件/附件1.csv", encoding="GB2312")
ALL_SC = list(set(data_for_sc["分拣中心"]))
existing_scs = list(map(lambda SC_: int(SC_[2:]), ALL_SC))
existing_scs.sort()

# 加载数据
def load_data(existing_scs):
    all_data = []
    for i in existing_scs:
        data = pd.read_csv(f"SC{i}.csv")
        data["SC_id"] = i
        all_data.append(data)
    return pd.concat(all_data, ignore_index=True)


# 数据预处理
def preprocess(data):
    # 假设数据包含日期和货量两列，日期列名为'date'，货量列名为'value'
    data["date"] = pd.to_datetime(data["date"])
    data["year"] = (pd.to_datetime(data["date"])).dt.year
    data["month"] = (pd.to_datetime(data["date"])).dt.month
    data["day"] = (pd.to_datetime(data["date"])).dt.day
    data["weekday"] = (pd.to_datetime(data["date"])).dt.weekday
    return data


# 训练模型
def train_model(data):
    features = data[["SC_id", "year", "month", "day", "weekday"]]
    target = data["value"]
    X_train, X_test, y_train, y_test = train_test_split(
        features,
        target,
        test_size=0.2,
        random_state=50,

    )

    model = RandomForestRegressor(n_estimators=300, random_state=42, min_samples_split=20)
    model.fit(X_train, y_train)

    # 预测和评估
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    print(f"Mean Squared Error: {mse}")
    return model

```

```python

# 预测未来的货量
def predict_future(model, start_date, num_days, existing_scs):
    future_dates = pd.date_range(start_date, periods=num_days)
    future_data = pd.DataFrame(
        {
            "date": np.repeat(future_dates, len(existing_scs)),
            "SC_id": np.tile(existing_scs, num_days),
        }
    )
    future_data = preprocess(future_data)
    features = future_data[["SC_id", "year", "month", "day", "weekday"]]
    predictions = model.predict(features)
    future_data["predicted_volume"] = predictions
    return future_data


# 保存结果到 CSV
def save_predictions_to_csv(predictions, file_name):
    predictions["date"] = predictions["date"].dt.strftime("%Y/%m/%d")
    predictions.to_csv(file_name, index=False)
    print(f"Saved predictions to {file_name}")


# 主函数
def main():
    data = load_data(existing_scs)
    data = preprocess(data)
    model = train_model(data)
    future_predictions = predict_future(model, "2023-08-01", 153, existing_scs)
    save_predictions_to_csv(future_predictions, "predicted_volumes.csv")


if __name__ == "__main__":
    main()
```

ARIMA 和 LSTM 的代码未被采用，因此仅放在支撑材料中。