Data-Preprocessing

Code Analysis Paper: Lambda Function for Data Preprocessing in Pathway and Orthology Datasets

1. Introduction

This paper provides an analysis of a Lambda function developed to preprocess two different types of biological datasets: **Pathway** and **Orthology**. The function downloads the data from AWS S3, preprocesses it by removing missing values, and uploads the cleaned data and metadata back to S3. The function operates in a flexible manner, handling both datasets with a shared structure but different content.

The preprocessing task is critical for ensuring clean and ready-to-use data, which is essential for further statistical or machine learning analyses. This paper discusses the input and output data formats and breaks down the function's logic step by step, supported by pseudocode.

2. Input and Output Data

Input Data

The Lambda function accepts the following parameters in the `event` object:

- `bucket`: The S3 bucket where the input and output files are stored.
- `dataset_type`: The type of dataset to preprocess. Can either be `Pathway` or `Orthology`. If no dataset is specified, the function defaults to processing `Pathway`.
- **Pathway Dataset**:
  - Label File: `PC_Pathway/pathway_label_list.csv`
  - Data File: `PC_Pathway/PC Pathway.csv`
- **Orthology Dataset**:
  - Label File: `PC_Orthology/orthology_label_list.csv`
  - Data File: `PC_Orthology/PC Orthology.csv`

Output Data

The function produces two output files for both datasets, saved back to the S3 bucket:

1. **Preprocessed Data File**: A pickle file containing cleaned data with missing values removed.
2. **Missing Columns File**: A JSON file listing columns that were removed due to missing values.

For example:

- **Pathway Dataset Output**:
  - Preprocessed Data: `PC_Pathway/intermediate/preprocessed_data_pathway.pkl`
  - Missing Columns Info: `PC_Pathway/intermediate/missing_columns_pathway.json`
- **Orthology Dataset Output**:
  - Preprocessed Data: `PC_Orthology/intermediate/preprocessed_data_orthology.pkl`

- Missing Columns Info: `PC_Orthology/intermediate/missing_columns_orthology.json`

3. High-Level Algorithm Design

The Lambda function follows a structured flow to preprocess either Pathway or Orthology data. It ensures that the data is clean, removing columns with missing values, and logs important metadata to track data changes. The key steps of the function can be summarized as follows:

1. **S3 Initialization**: Establish connection to AWS S3 to download and upload files.
2. **Dataset Type Determination**: Determine whether the dataset is Pathway or Orthology, and set corresponding file paths.
3. **Download Data**: Download the required files from S3 to the Lambda `/tmp` directory.
4. **Data Preprocessing**:
   - Ensure all values are numeric.
   - Identify and remove columns containing missing values.
   - Log the shape of the dataset before and after cleaning.
5. **Upload Preprocessed Data and Missing Columns**: Save the cleaned data and removed column information to S3.
6. **Return Success or Error**: Return the status of the function execution.

4. Pseudocode Representation

The following pseudocode provides an overview of the function's steps:

```
pseudo
코드 복사
FUNCTION lambda_handler(event, context):
    INITIALIZE s3 client
    INITIALIZE lambda client

    # Retrieve bucket and dataset type from the event
    SET bucket = event['bucket']
    SET dataset_type = event['dataset_type'] (DEFAULT to 'Pathway')

    # Determine file paths based on dataset type
    IF dataset_type == 'Pathway':
        SET label_file_key = 'PC_Pathway/pathway_label_list.csv'
        SET function_file_key = 'PC_Pathway/PC Pathway.csv'
        SET preprocessed_data_key = 'PC_Pathway/intermediate/preprocessed_d
ata_pathway.pkl'
        SET missing_columns_key = 'PC_Pathway/intermediate/missing_columns_
pathway.json'
    ELSE:
        SET label_file_key = 'PC_Orthology/orthology_label_list.csv'
        SET function_file_key = 'PC_Orthology/PC Orthology.csv'
        SET preprocessed_data_key = 'PC_Orthology/intermediate/preprocessed
```

```
_data_orthology.pkl'
        SET missing_columns_key = 'PC_Orthology/intermediate/missing_column
s_orthology.json'

    TRY:
        # Download files from S3 to local /tmp directory
        DOWNLOAD label_file_key FROM bucket TO /tmp
        DOWNLOAD function_file_key FROM bucket TO /tmp

        # Read the CSV files
        READ label_data FROM label_file_path
        READ function_data FROM function_file_path

        # Preprocess the data: Extract feature columns and ensure numeric t
ypes
        SET X_dat = function_data[:, 3:]
        CONVERT X_dat to numeric

        # Log initial shape of the data
        PRINT initial shape of X_dat

        # Identify columns with missing values
        SET missing_columns = columns in X_dat WITH missing values

        # Remove columns with missing values
        SET X_cleaned = X_dat WITHOUT columns with missing values

        # Log the cleaned data shape
        PRINT cleaned shape of X_cleaned

        # Concatenate target variable y and cleaned data X_cleaned
        SET data = CONCATENATE function_data['group_2'] AND X_cleaned

        # Save preprocessed data as pickle file
        SAVE data AS preprocessed_data_path

        # Upload preprocessed data and missing columns info to S3
        UPLOAD preprocessed_data_path TO bucket
        SAVE missing_columns AS JSON FILE
        UPLOAD missing_columns TO bucket

        RETURN success message WITH paths to preprocessed data and missing
columns
    EXCEPT Exception AS e:
        PRINT error
        RETURN error message
```

5. Detailed Step-by-Step Analysis

Step 1: S3 Initialization

The function initializes AWS S3 and Lambda clients using the `boto3` library. This is crucial for downloading input data from S3 and uploading processed data back to S3.

Step 2: Dataset Type Determination

The function inspects the event object to determine whether it is processing the **Pathway** or **Orthology** dataset. Depending on the type, it sets the appropriate file paths for the data files, preprocessed output, and missing columns metadata.

Step 3: Download Data

Using the `s3.download_file` method, the function downloads the CSV files for both the label and data (Pathway or Orthology) from S3 to the local `/tmp` directory, where Lambda functions can temporarily store data.

Step 4: Data Preprocessing

- **Initial Logging**: The function logs the initial shape of the dataset (`X_dat`) before cleaning.
- **Numeric Conversion**: The columns from the dataset are converted to numeric values to ensure consistency in the data type.
- **Missing Value Identification**: Columns with missing values are identified.
- **Data Cleaning**: Columns containing missing values are removed, and the function logs the new shape of the cleaned data.
- **Concatenation**: The cleaned data (`X_cleaned`) is concatenated with the target variable (`y`), resulting in a complete dataset that is ready for further analysis.

Step 5: Save and Upload Results

- **Preprocessed Data**: The cleaned dataset is saved as a pickle file in the `/tmp/` directory.
- **Missing Columns**: The list of removed columns is saved as a JSON file.
- Both files are uploaded to the specified S3 bucket for later use.

Step 6: Return Status

After processing, the function returns a success message along with the paths to the uploaded preprocessed data and missing columns. In the event of an error, the function catches the exception and returns an error message.
