Change Detection and Land Use Classification

1 Overview

Analysing changes in multi-image, multi-date remote sensing data helps us to discover and understand global conditions. This challenge uses satellite imagery-derived geographical features. The data has been processed using computer vision techniques and is ready for exploration using machine learning methods.

The aim of this challenge is to classify a given geographical area into six categories.

2 Classes

This challenge aims to classify a given geographical area into six classes:

- **Demolition**: 0
- **Road**: 1
- **Residential**: 2
- **Commercial**: 3
- **Industrial**: 4
- **Mega Projects**: 5

3 Data Description

The training and test sets are provided in the `train.geojson` and `test.geojson` files, respectively.

3.1 Features

The geographical features are:

1. An irregular polygon (`geometry`).
2. Categorical values describing the status of the polygon on five different dates (e.g., under construction on day 0, completed on the following dates).
3. Neighbourhood urban features (e.g., the polygon is in a dense urban and industrial region).
4. Neighbourhood geographic features (e.g., the polygon is near a river and a hill).

3.2 Dataset Columns

The columns available in the geojson files are:

- `date0` to `date4`: Observation dates (DD-MM-YYYY).
- `change_status_date0` to `change_status_date4`: Status of polygon on each date.
- `urban_type`: Comma-separated multiple values showing neighbourhood urban types.
- `geography_type`: Comma-separated multiple values showing neighbourhood geographic types.
- `geometry`: Vector representation of geographic polygons.
- `change_type`: Label to be classified (training set only).

For each date, scalar mean and standard deviation values of the colour image taken from a 50cm satellite are also provided:

- Means: `img_red_mean_date1` to `img_red_mean_date5`, etc.
- Standard Deviations: `img_red_std_date1` to `img_red_std_date5`, etc.

Important Note: Please note that image statistics columns use suffixes from `date1` to `date5`, while date and status columns use `date0` to `date4`.

4 Pipeline

The proposed pipeline is the one introduced in the first lecture of the ML course:

1. **Data Preprocessing**: You need to preprocess the data and convert it into the appropriate format.
2. **Feature Engineering and Dimensionality Reduction**:
   - Explore One-Hot Encoding for urban and geographic types.
   - Create geometric features (area, perimeter) from the polygons.
   - Calculate the number of days between two consecutive dates.
   - Feature selection or dimensionality reduction could be beneficial.
3. **Learning Algorithm**:
   - A simple baseline (k-NN) is provided in `skeleton_code.py` ($\approx$ 40% performance). The code uses only the **polygon area** as a feature.
   - Test other classifiers: Logistic Regression, SVM, Decision Trees, Neural Networks, Ensemble Learning.
4. **Evaluation**: The evaluation metric is the **Mean F1-Score**.

5 Submission and Grading

Form groups of 3-4 students.

5.1 Deliverables

1. **Kaggle Submission** (40 points):
   - Command to download data:
     `kaggle competitions download -c 2-el-1730-machine-learning-project-2026`
   - **Submission Step-by-Step:**
     (a) Run your trained model on the test data (`test.geojson`).
     (b) Get the predicted class labels for each instance.
     (c) Create a `sample_submission.csv` file containing the header and two columns: `Id` and `change_type`.
     (d) Create an account on Kaggle (one for all team members) and create a new submission by uploading this file.
   - Kaggle will automatically evaluate your predictions:
     – Public Leaderboard (during competition): based on $\approx$ 30% of test data.
     – Private Leaderboard (final results): based on the remaining 70%.
   - Daily limit: 10 submissions.
2. **Report on Edunao** (50 points):
   - PDF file only.
   - Filename: `teamname_student1name_student2name_student3name.pdf`.
   - Must include full names and Kaggle team name.
   - **Section 1: Feature Engineering**: Motivation, intuition, experiments, combinations of features tested.
   - **Section 2: Model Tuning and Comparison**: Compare multiple classifiers, parameter tuning procedure, cross-validation, discussion of discarded models.
3. **Code on Edunao** (10 points - Global Grading):
   - ZIP file: `name_of_your_team.zip` (max 512MB).
   - Reproducible code.
   - The 10 points also cover respecting submission guidelines, organization, and clarity.

Figure 1. Samples from 'test.geojson' showing different **change type**, **change status** on `different dates`, neighborhood label(s), and geography label(s). Latitude-longitude of the change polygon is shown along with city name. First row shows construction of a residential property in suburban area of New York, USA. Second row shows a commercial building in an industrial region which used to be farm lands of a fast growing second tier city in China. Third row shows an industrial construction in desert of Doha, Qatar which went from rural barren desert to a sparse urban area in a time period of 5 years. Fourth row shows construction of a road crossing a river in farm lands of rural China. Fifth row shows special case of urban change, demolition of a farm storage in the fast growing city Changzhou in China. Last row shows construction of a power grid unit which comes under mega project type.

Figure 1: Samples from 'test.geojson' showing different change types, status, dates, and neighbourhood labels.

6 References

The dataset is part of the paper *"QFabric: Multi-Task Change Detection Dataset"* (CVPR 2021W) by Sagar Verma et al.