Data Challenge 2, GEO881 FS22

Cheng Fu

1. Introduction

In this data challenge, we will further practice clustering and classification in the context of geographical information science. Particularly, we will solve a problem regarding land use labeling with a data set partially belonging to the data set used in Fu et al. (2019).

In land use modeling, an important but challenging task is to identify different land use types in urban. With remotely sensed imagery data, we can only sense the physical characteristics of a land parcel, such as the spectrum of the roof, the area, and the periphery of the land parcel. However, how human beings actually use the parcel for residential, commercial, or other uses is the key to defining the use of the land. Human mobility data, such as geotagged tweets in this case, shed light on modeling human activities that happened on the land parcels. From the raw data, we can extract the temporal variance of activity volume, activity type, activity diversity, etc., using different technical tools.

This case study takes the metropolitan of Washington D.C.-Baltimore as the study area. Washington is the capital of the US, and Baltimore is the largest city in Maryland State. Since 1985, the metropolitan area has experienced a boost in land development due to the investment in bio-techniques and IT, as well as real estate. It is thus critical to know where the new land development projects are and which type, i.e., residential vs. non-residential, takes a larger proportion.

With the help of impervious surface modeling on a series of Landsat imagery between 1986 and 2008, we have identified the land parcel polygons of the new developments. As you can imagine, the new land development has a noticeable increase in impervious surface coverage. We then associated the identified land parcel polygons with geotagged tweets collected from the same study area and calculated several groups of aggregated features. The complete feature set includes over 1000 features belonging to two groups: physical and activity features. However, we only use about half of the features in this data challenge, mainly from the activity feature group. To better understand how the data were processed and aggregated, please read the referred article.

2. Description of the data set

The data set has 507 columns and 2520 records in total. The features are the columns afterward Column tw_0. I will use the **feature set** to refer to these columns. You can find the detailed description below. All rows have been labeled. There are no missing data (missing values have been assigned as 0).

Feature name	Index	Type
id	0	integer
parcel_name	1	text, unique identity of a land parcel
label	2	categorical R: residential, NR: non-residential
tw_0 … tw_167	3-170	normalized hourly tweet count
user_0 … user_167	171-338	normalized hourly user count
entropy_0 … entropy_167	339-506	entropy of tweets regarding to the users

3. Tasks

3.1. Explore the data

Before going to the detailed tasks, it is highly recommended to explore the data and understand its nature of the data, including exploring the value distribution of the numerical columns and counts of categorical columns and checking the basic statistics of the data.

3.2. Clustering

*3.2.1. Exercise with k-means (**10 pts**)*

Using the raw feature set, apply k-means clustering to **all** records. You need to decide which metric to use for evaluating the clustering results and determine which k is the best, given the metric you choose. Justify your choice of metric and the best k value in the report.

Assign labels to each of the k groups of your best k. Here we can assume that we know some of the truth of a small sample. In the real world, we select a few samples from each group and check their labels in the real world to determine the label of a group. In this data challenge, we check the values of the **label** column. You make your own choice regarding the sample size and the strategy to determine the label of a group. Justify your strategies for sample selection and label assignment in the report.

Then compare your labels with the true labels using the confusion matrix and evaluate the final performance of the k-means method. Discuss the results you have in the report with the metrics you choose for the evaluation.

*3.2.2. Exercise with dimension reduction (**10 pts**)*

You need to apply principal component analysis (PCA) to reduce the raw feature set into a smaller number. It is recommended to keep principal components that can retain over 90% or higher variance.

Following the same procedure of model selection as we have in the previous section, you will have a different clustering result. Compare the new result with the confusion matrix of Section 3.2.1. Moreover, discuss if the clustering result is improved.

*3.2.3. Comparison of k-means and agglomerative hierarchical clustering (**10 pts**)*

Apply the AGNES hierarchical clustering algorithm to the raw feature set with Euclidean distance as the metric and set the number of groups as the same as the best k value in Section 3.2.1. Please try at least one linkage type. Then follow the same workflow as Section 3.2.1 on labeling the group. Compare the results of this method and the result in Section 3.2.1.

3.3. Classification (**60 pts**)

Use the raw feature set and reserve 20% of all records as the test set and use the rest as the training set. Please note that preprocessing on the features might still be needed for certain classification algorithms. Apply classification algorithms **k nearest neighbor (kNN)**, **decision tree (DT)**, **random forest (RF)**, and **artificial neural network (ANN)** to the training set.

For each algorithm, use 10-fold cross-validation for model selection. Choose the metric(s) for model evaluation and find the 'best' hyperparameter set for each algorithm. Please try at least 3 hyperparameter sets for each algorithm. Report your selected model for each algorithm and justify your selection. Discuss if any overfitting/underfitting is observed in any of the models during the model selection. Compare the performances of the selected models. Discuss the pros and cons of the algorithms and suggest possible improvements for better model performances.

4. Submission requirements (**10 pts**)

You should submit at least two documents: 1) an executable rmd or r file with well-documented comments; 2) a text document as a Word docx file or pdf file that describes the major steps for coding design, the results, and discussion to accomplish each task in Section 3. It is also recommended to upload the HTML file that is generated by the rmd.

The comments in the code file do not have been line by line. However, you shall have comments for functional sections to explain the attempt. In addition, the variable names should suggest their meaning. For justifying the model selection during the k-fold cross-validation, you do not have to report the result of every validation but report the mean or median of the k results.

The submitted document names should be formatted as Geo881_YourLastName_FirstName.rmd/r/docx/pdf

Please finish this data challenge independently and submit the documents before 8:00 am, 5.5.2022 (Thursday).

Reference

Fu, C., Song, X.-P., & Stewart, K. (2019). Integrating Activity-Based Geographic Information and Long-Term Remote Sensing to Characterize Urban Land Use Change. *Remote Sensing*, *11*(24), 2965. https://doi.org/10.3390/rs11242965