

A PATTERN ANALYSIS IMAGE VALIDATION TOOL FOR THE GENERATION OF RELIABLE EARTH OBSERVATION IMAGE BENCHMARKS

Wei Yao, Gottfried Schwarz, and Mihai Datcu

EO Data Science, Remote Sensing Technology Institute, German Aerospace Center (DLR)
Münchener Str. 20, 82234 Weßling, Germany
wei.yao@dlr.de, gottfried.schwarz@dlr.de, mihai.datcu@dlr.de.

ABSTRACT

This paper describes an image validation tool for the generation of good-quality Earth Observation (EO) benchmark datasets. We already developed an active-learning-based semantic annotation tool which allows users to fast annotate images with few samples; this tool reaches about 90% accuracy. A subsequent data cleaning tool then helps correct noisy data, thus increasing the number of correctly labeled images to be qualified as benchmark data. However, this work has an annoying bottleneck, namely the manual correction via visual checks still costs a considerable amount of energy. Therefore, this paper aims to discuss about the label relationships in the embedding space, by proposing new metrics to distinguish four different ambiguous cases within a dataset, based on pattern analysis. The interactive visualization then enables users to visualize a dataset and explore unknown patterns.

The benefits are two-fold: firstly, experiments show the proposed metrics greatly help decrease the manual labor whilst keeping the essential data, thus enhancing the degree of automation in the process of generating good-quality benchmark datasets. Secondly, our approach provides possibilities to interactively visualize and explore very large-scale datasets in real time, thus providing help for further data mining.

Index Terms— Benchmarks, interactive visualization, image validation, pattern analysis.

1. INTRODUCTION

As already known, many EO applications exploiting machine learning techniques, especially deep learning tasks, have raised the importance of appropriate training datasets. However, data collection may be quite expensive, as it usually costs lots of time and human labor.

Our motivation is to prepare analysis-ready Earth observation training datasets at minimal cost. We try to tackle the following questions: how do we foster the generation of EO image datasets and how do we guarantee the quality of the generated EO image datasets? The answer to the first question is to use active learning techniques, which adds human

expertise to the data interpretation loop, thus strengthening interactive machine learning. The answer to the second question is to embed the high-dimensional image dataset in a 2D space, and use visualization techniques to enable human experts to validate the generated datasets, correct wrong semantic labels, and enhance the quality of the datasets.

The EO community urgently needs new large-scale benchmark datasets, to support further machine learning and deep learning applications. With the inherent complexity of EO data, more and more complex EO datasets are being generated continuously. A well-known example is the BigEarthNet dataset for optical and SAR applications (using both Sentinel-1 and Sentinel-2 data) [1]. In addition, the booming development of Artificial Intelligence (AI) stresses the importance of good benchmark datasets.

With the emergence of more and more new public EO datasets, when we consider the optical photo datasets collected in the computer vision field, we have to take into account that even the famous ImageNet has 5% label error in its validation set [2].

1.1. Data validation

In our previous work which resulted in an active learning tool for the generation of EO image benchmarks [3], we already tried to develop a data validation strategy. This tool supports manual label corrections by visual inspection. The data validation process helps refine the raw data and purifies the resulting datasets, in order to achieve the required benchmark quality.

However, we noticed a notorious bottleneck: the manual correction led to an ample amount of work for the image analysts.

1.2. Pattern analysis

Pattern analysis concentrates on identifying rules that describe specific patterns within the data, for instance, for anomaly detection. This technique is helpful to analyze the relationships among data points, and to organize the data points that fall into different patterns. Thus, in our case, it

is beneficial to distinguish the assigned labels with different levels of confidence.

1.3. Interactive visualization

The integration of interactive visualization with the benchmark generation tool has shown several benefits [4], [5], mainly in two fields: it enhances the operability of the tool; and it arouses the users' interest in exploring the datasets. Thus, the current work puts some emphasis on the implementation of interactive visualization components within the tool. Moreover, we can visualize four patterns (described in Section 2) in the projected space, using our proposed interactive visualization.

Here, we present a pattern-analysis-based EO image validation tool for the generation of EO benchmarks. This image validation tool is implemented in an interactive visualization manner allowing that a high-dimensional feature space can be re-projected via dimensional reduction techniques.

2. BUILDING BLOCKS

In the following, we present our proposed image benchmarking tool. Fig. 1 shows its framework, there are different modules included, mainly, an active-learning-based image annotation tool, and an image validation tool (providing data visualization and cleaning). The image annotation tool is inherently implemented with active learning components which allow the users to interact with it to generate trustworthy and reliable benchmarks by using only few samples. One can achieve (on average) a classification accuracy of more than 90 percent (still to be improved). More details of this tool can be found in two papers [3], [6]. In order to recognize patterns (i.e., relationships within a dataset), and to validate and clean the generated datasets, an image validation tool has been developed, too. It provides interactive visualization, equipped with real-time rendering. The bottleneck is this real-time rendering that requires lots of resources, as well as the handling of labels with uncertainties, for instance, points on the classification boundaries. Our proposed optimization is to perform statistical neighborhood pattern analyses, and a neighborhood decision tree was designed to distinguish four kinds of neighborhood patterns.

Thanks to the selected active learning techniques, the image annotation tool can increase the efficiency of generating EO image benchmark datasets - due to the facts that it's fast in speed, robust in processing different types of products, and smart to train with user interactions. Furthermore, the image validation tool enables users to visualize a dataset, and perform an exploration of unknown patterns or classes, and conduct visual data mining.

Moreover, compared with some state-of-the-art benchmark datasets which were mainly generated manually with a fixed range of semantic labels, this tool extends our bench-

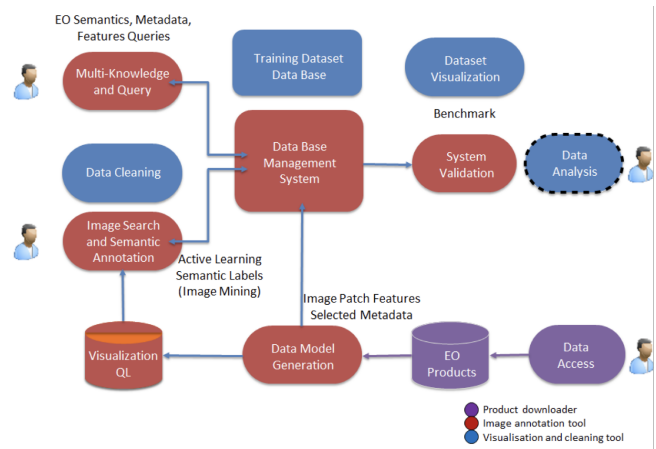


Fig. 1: Framework of the benchmarking tool. In this paper, two components are discussed: the image annotation tool which is presented in [7], and the visualization and cleaning tool.

mark robustness as it allows for user-defined labels, rather than just applying already existing labels to maps.

2.1. Pattern analysis

With the already generated labeled datasets, some classes are more homogeneous and prone to be correct, while some other classes are more ambiguous and are prone to be wrong. These cases are often linked to false positive and false negative errors. Thus, a big question for us is how to quantify the label uncertainty.

In the projected feature space, we observed that there are data points located on a decision border, and data points that mix with a totally different cluster than they semantically belong to; These cases make the classification an intricate task. When we analyze the neighborhood relationships, four patterns are normally observed (cf. Fig.2)

- An outlier is a point which is far away from the main body of data points, and with few neighbors.
- A border case is a point with a varying number of identical or different labels.
- A heterogeneous point is one surrounded by many aliens.
- A homogeneous point is one surrounded by its own group.

For a specific data point, the uncertainty of labels can be represented by means of a Coefficient of Variation (CV). In probability theory and statistics, CV is a standardized measure of dispersion of a probability distribution or frequency distribution. It indicates how "spread out" the members of a sample or population are relative to the mean. The Coefficient

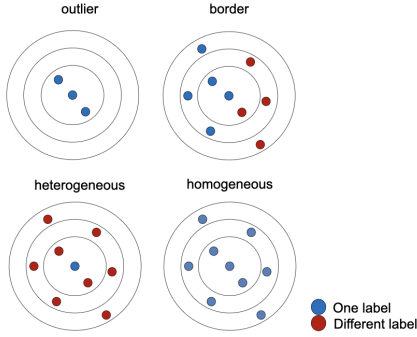


Fig. 2: Neighborhood pattern analysis in a re-projected feature space. Four patterns of labels in relation with their neighbor labels are shown.

of Variation is defined as: $CV = sd/mv$ with standard deviation (sd) and mean value (mv) referring to the distance from its neighbors.

In this work, the proposed two metrics are: $CV(d)$ and $CV(d*I)$ with distance as weight (same label: 1, different label: -1). Then the four patterns can be distinguished by a decision tree model. Outlier: $CV(d)$ large, when $CV(d)$ is small, we look at the $CV(d*I)$ metric; homogeneous: positive small; heterogeneous: negative small; border: large.

As shown in Fig. 3, $CV(d)$ and $CV(d*I)$ metrics are applied when we discriminate against these four patterns. On the upper right corner is the embedding space showing the raw dataset; on the bottom and on the left side are the corresponding embedding spaces for each pattern. In the homogeneous case, we notice very smooth pattern changes among the labeled image patches. The other three patterns are ambiguous cases that need further processing.

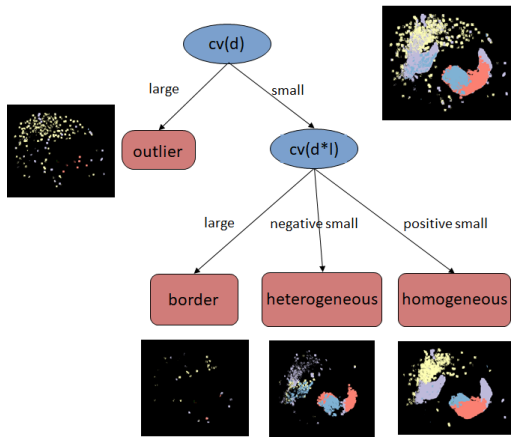


Fig. 3: Neighborhood decision tree. An example of the corresponding visualizations for the four patterns are shown close to the branch nodes.

However, for the generation of good-quality benchmark datasets, it's not necessary to analyze all patterns, a simple solution would be to concentrate on the homogeneous classes.

2.2. Interactive visualization-based data validation tool

With the interactive visualization-based image validation tool, users can perform visual data mining. We can see the low-dimensional embedding feature space, where similar patterns are grouped together, as shown on the right side of Fig. 4.

Here we present an example dataset which was generated within the framework of the H2020 CANDELA project¹. The dataset contains 30,813 non-overlapping image patches cropped from Sentinel-1 products, each of which has a size of 120x120 pixels with 20 m resolution. For this, we calculated the SAR-adapted Weber features[8] with a length of 144 dimensions. There are four classes shown in different colors: yellow represents mixed forest, purple represents mixed urban, red represents cropland, and blue represents sea.

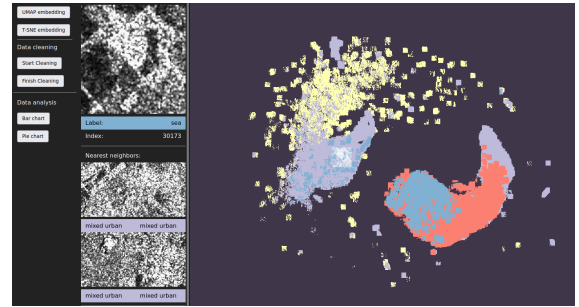


Fig. 4: Data validation tool: Sentinel-1 dataset with a four-class annotation made by the annotation tool [7]; different colors represent different classes.

Fig. 4 shows the original "messy" embedding space, if only the homogeneous data points are shown. In contrast, Fig. 6 (c) presents a much cleaner embedding space with 25,000 points out of the 30,000 points in total, where more than 80% of the original data are preserved.

3. DISCUSSIONS

In this section, some intermediate results will be presented and discussed. Fig. 5 shows the decision tree results of the Sentinel-1 dataset for four patterns. A larger absolute value means a higher discrepancy of a data point between its neighborhood. With four scales of absolute values, the four patterns are clearly depicted in red, green, magenta and blue.

Fig. 6 illustrates how these four patterns look like when visualizing them in the embedding space. While the outlier pattern (a) extracts the outlier layer of the embedding space; the border pattern (b) extracts the 'skeleton' of the embedding

¹<https://candela-h2020.eu/>

