DATA MINING ON THE CANDELA CLOUD PLATFORM

Wei Yao[1], Corneliu Octavian Dumitru[1], Jose Lorenzo[2], and Mihai Datcu[1]

[1]Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany
(emails: wei.yao@dlr.de; corneliu.dumitru@dlr.de; mihai.datcu@dlr.de)
[2]ATOS SPAIN SA, Calle Albarracín 25, 28037 Madrid, Spain (email: jose.lorenzo@atos.net)

ABSTRACT

This paper describes the work done with the Data Mining components of the H2020 CANDELA project, mainly the Data Model component on the CANDELA platform as its back end, and the user interaction component of the local user machine as front end.
The Data Mining tool is basically composed of four main sub-modules: the Data Model Generation for Data Mining (DMG-DM), the database management system (DBMS) sub-module that has already been dockerized and deployed on the CANDELA platform, the image search and semantic annotation (KDD) sub-module and the multi-knowledge and query (QE) sub-module. They all require user inputs, and connect directly to the database on the platform; they can be started as a normal GUI (Graphical User Interface) tool.

Index Terms—*Big data, Sentinel data, Earth observation, data mining, semantics*

1. INTRODUCTION

Due to the wealth of open data provided by the Copernicus programme, the CANDELA project intends to build various modules and frameworks on-top of available components from consortium members. Among them is the Data Mining tool.
The aims of the Data Mining tool are to provide a data mining system with a large-scale image interpretation capability for Earth Observation, and to fill the ever-enlarging gap between acquired satellite data and processing tools. The DMG-DM (Data Model Generation for Data Mining) sub-module generates data models to be inserted into a database on the platform; the KDD (Image Search and Semantic Annotation) and QE (Multi-Knowledge and Query) sub-modules query the database to perform semantic annotation and data analytics.
Our data mining system has been developed as an active-learning interactive tool, which is not only able to query the image products, but also to query the semantic annotations. Hence, knowledge discovery and data mining can be carried out in both feature and semantic spaces.

2. RELATED WORK

In the Horizon 2020 EO Big Data Shift call in 2017, CANDELA was one of the five accepted projects of the European Commission [1]. The BETTER project manages Data Pipelines, the CANDELA project builds a platform, the EOPEN project establishes a platform, the OPENEO project writes APIs, and the PerceptiveSentinel project develops the EO-LEARN Python library. Among them, the CANDELA platform manages various algorithmic modules, extracts information and discovers higher-level knowledge out of the big amount of Sentinel data [2], [3], and [5].

3. DATA MINING FLOW

The data flow of the Data Mining tool on the platform and the local user machine are shown in Figure 1.
In the platform layer, Sentinel-1 and Sentinel-2 products can be accessed via these links, and symbolic CreoDIAS [7] links are provided by a platform service. By starting one Data Model Generation for Data Mining (DMG-DM) process, it works for either one Sentinel-1 or one Sentinel-2 product. Afterwards, all extracted image features, metadata information, and quick-look images will be ingested into the 'candela' MonetDB [8] database on the platform. The generated quick-look images will be published on the platform for download by local users via a REpresentational State Transfer (RESTful) service.
In the transfer layer, all services require Internet connections. The CANDELA Jupyterlab (with Python) is the main interface for a local user to interact with the platform, where multiple services are provided. On the platform, users can not only use analytical tools, but also perform programming exercises via Jupyter notebooks. A RESTful service is used to access the quick-look images by local users using the front-end tool. The MonetDB database 'candela' on the platform will be remotely connected for querying product metadata, feature extraction, and for ingesting user-annotated semantics.
In the front-end layer, with a GUI tool, users can perform active learning based semantic annotations for either Sentinel-1 or Sentinel-2 products. Depending on the classes to be extracted and the selected parameter settings; this process usually takes 5-7 iterations, or a bit more, depending also on the complexity of the image content. The semantics annotations will be ingested via Internet into the remote database 'candela' on the platform.

4. DATA MINING SUB-MODULES

This section summarizes the function of the four sub-modules: Data Model Generation for Data Mining (DMG-DM), Database Management System (DBMS), Image Search and Semantic Annotation (KDD), and Multi-Knowledge and Query (QE).

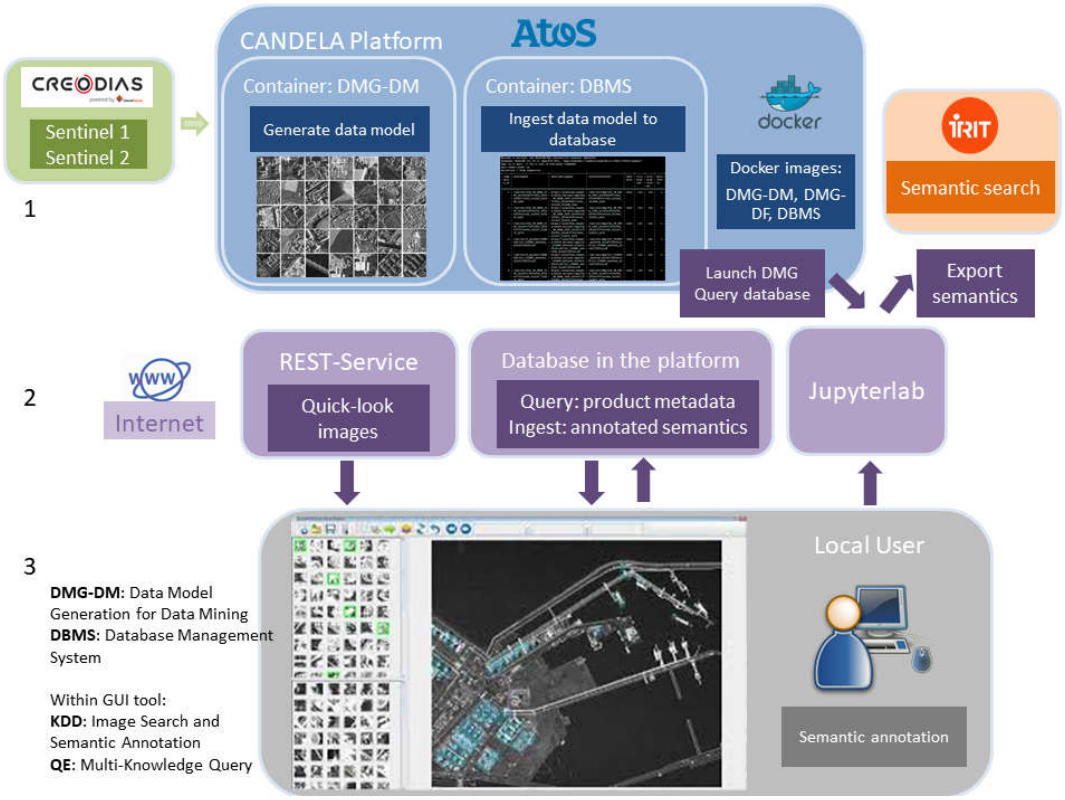

Figure 1: The data mining flowchart.

4.1. Data Model Generation for Data Mining

This sub-module creates the data model: to extract product metadata, to crop image patches, to generate high-resolution quick-look images, and to extract image features from the cropped image patches. To initiate it, users should set the following parameters: the feature selection method, the grid level definition, and the patch size. By searching the CreoDIAS link access library, users can set Sentinel-1 or Sentinel-2 product links as input products to be processed. The output of this sub-module is an XML file with all the extracted and processed information, and the cropped quick-look images.

4.2. Database Management System

This sub-module manages a MonetDB-based relational database, named 'candela', in particular, the ingestion of DMG-DM-generated results. The input of this sub-module is the XML file which is created by DMG-DM. The output of the sub-module is an SQL file which is generated based on the XML file. The extracted product metadata, the quick-look image patch directories, and the image features from the cropped image patches are ingested via the SQL file and stored into the 'candela' database.

In addition, a hierarchical semantic catalogue has been ingested into the 'candela' database. It allows users to select a proper label during semantic annotation by using the GUI tool. In case of an undefined class, after choosing a specific general-level label, users also have the opportunity to define a new detailed-level label using the GUI tool.

4.3. Image Search and Semantic Annotation

This sub-module is embedded in the GUI tool and runs on a local user machine. It allows users to search and query in the database to select products that users are interested in. A cascaded active learning based GUI tool is used to perform semantic annotation. A user picks up positive and negative samples by mouse clicking and trains an SVM classifier; this procedure can be performed iteratively until the user is satisfied with the classification results. Multiple classes will be extracted after multiple operations. This sub-module has been updated to connect directly to the remote 'candela' database on the platform via an Internet port connection, and loads quick-look images from the platform via a RESTful service.
In the toolbar, there is a query button to connect to the 'candela' database on the platform, which allows users to query and load quick-look images from the platform, and an ingestion button to ingest the semantically annotated data directly into the 'candela' database of the platform [4].

4.4. Multi-Knowledge and Query

This sub-module allows users to query images either based on product metadata, semantic annotations, or based on a

combination of different conditions (e.g., metadata and semantics) to search in the database. It is triggered by clicking the 'query the database' button in the GUI tool [4].

5. RESULTS AND CONCLUSIONS

This section starts with a first example of the CANDELA vineyard use case, by annotating and processing a Sentinel-2 image acquired on April 19, 2017 in the area of Bordeaux, France.
A second example is the cyclone around Beira, Mozambique acquired by Sentinel-1 in March, 2019. It is chosen from a Big Data demonstration for urban areas in the CANDELA project. Up to now, we have processed data which covers an area of 800,000 km^2 including Sentinel-1 and Sentinel-2 images, and we generated a semantic catalogue which will be migrated to the database on the cloud platform. The images cover 15 locations from all over the world.
Figure 3 and 4 show the outputs of these two examples as classification maps. The validation results demonstrate the data retrieval and mining capabilities of the Data Mining tool.
Further in the CANDELA project, Sentinel-1 and Sentinel-2 data will be fused, so that data mining and analysis can be performed by taking account of both products.

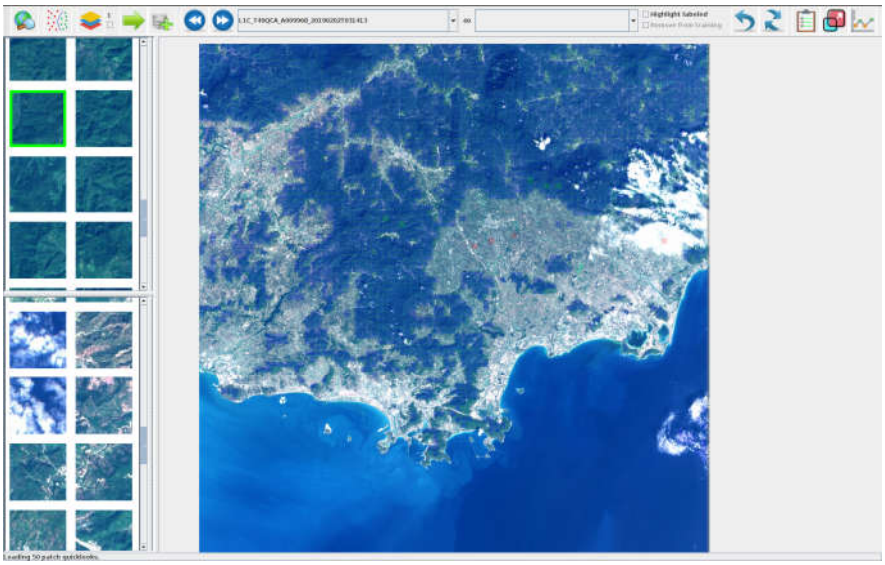

Figure 2: Graphical User Interface of the Data Mining tool.

6. ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776193.

7. REFERENCES

[1] EO-2-2017-EO Big Data Shift, 2017, [Online]. Available: https://cordis.europa.eu/programme/rcn/701821/en. [Accessed: 15-Jan-2020].

[2] M. Datcu, C.O. Dumitru, G. Schwarz, F. Castel, and J. Lorenzo, "Data Science Workflows for the CANDELA Project", Big Data from Space, Munich, Germany, 19-21 February 2019. Available: https://www.bigdatafromspace2019.org/QuickEventWebsitePortal/2019-conference-on-big-data-from-space-bids19/bids-2019/ExtraContent/ContentSubPage?page=4&subPage=2.

[3] J-F. Rolland, A. Haugommard, F. Castel, M. Aubrun, W. Yao, C. O. Dumitru, M. Datcu, M. Bylicki, B-H. Tran, N. Aussenac-Gillles, C. Comparot, C. Trojahn, CANDELA: A cloud platform for Copernicus Earth observation data analytics, IGARSS 2020.

[4] CANDELA Data mining deliverable v2, 2019, [Online]. Available: http://www.candela-h2020.eu/content/data-mining-v2.

[5] C.O. Dumitru, G. Schwarz, F. Castel, J. Lorenzo, M. Datcu, "Artificial Intelligence Data Science Methodology for Earth Observation," in Advanced Analytics and Artificial Intelligence Applications, InTech Publishing, 2019. Available: https://www.intechopen.com/books/advanced-analytics-and-artificial-intelligence-applications/artificial-intelligence-data-science-methodology-for-earth-observation.

[6] G. Dax, "Supervised and Unsupervised Methods in Data Mining", Master's thesis, Salzburg University of Applied Sciences, August 2019.

[7] CREODIAS, [Online], Available: https://creodias.eu/. [Accessed: 15-Jan-2020].

[8] MonetDB documentation, [Online], Available: https://www.monetdb.org/Documentation/. [Accessed: 15-Jan-2020].

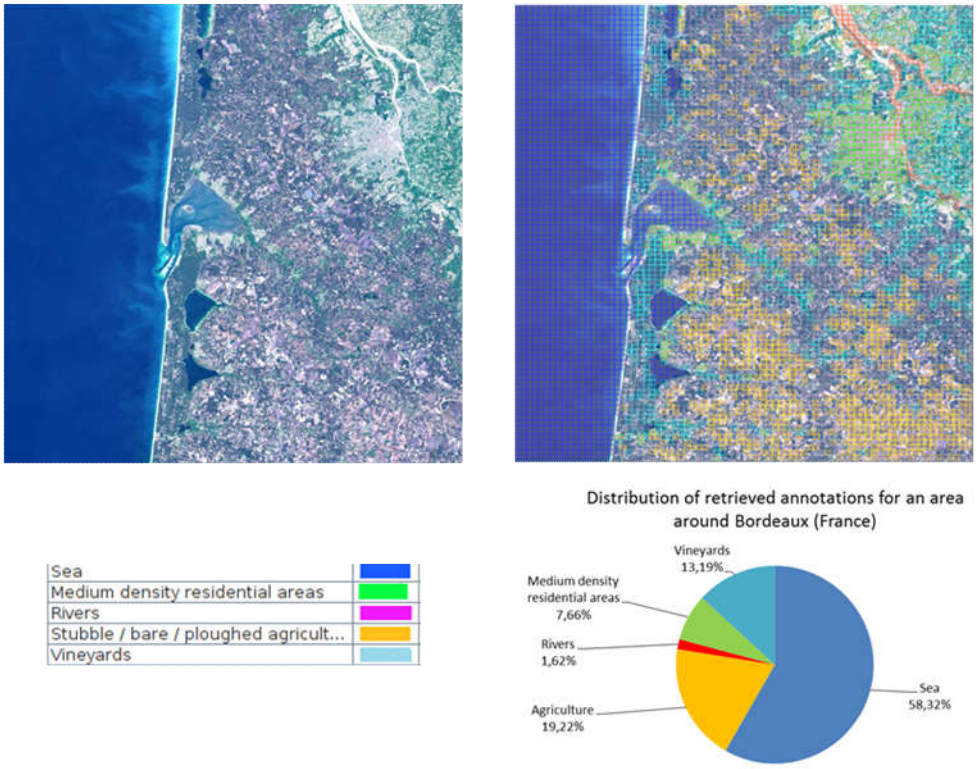

Figure 3: Sentinel-2 quick-look view (top-left), classification map (top-right), and the diversity of categories identified from a single image of Bordeaux, France acquired by Sentinel-2 (bottom-right).

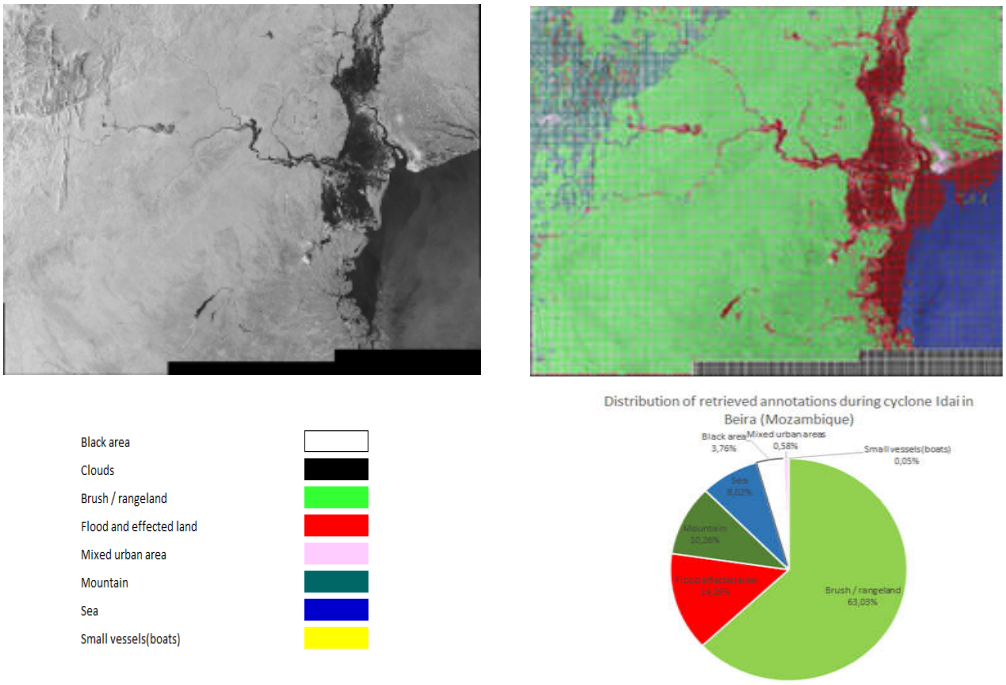

Figure 4: Sentinel-1 quick-look view (top-left), classification map (top-right), and the diversity of categories identified from a single image of Beira, Mozambique acquired by Sentinel-1 (bottom-right) [6].