STUDY CASES ON FAST COMPRESSION DISTANCE BASED DATA VISUALIZATION

Wei Yao

EO Data Science Department, Institute of Remote Sensing Technology
German Aerospace Center, Oberpfaffenhofen

ABSTRACT

In this paper, we develop a visualization tool to enhance the understanding of up to big data sets. Compared to classic data models which rely on the computing of the features (color, texture, etc.), this tool is fully feature free, as it processes directly on the data file. The Fast Compression Distance (FCD) and t-distributed Stochastic Neighbor Embedding (t-SNE) have been applied to visualize a large TerraSAR-X dataset which are annotated with up to three layers of hierarchical semantic labels, and a Sentinel-1 dataset with 10 annotated classes, in VV and VH polarization modes. We analyze the visualization results in manifold space, and try to understand and interpret them with the available semantic labels. The visualization interpretation is based on a vega-style interactive tool, which allows user zoom in, zoom out for processing large amount of data points.

Index Terms— FCD, t-SNE, TerraSAR-X, Sentinel-1, visualization

1. INTRODUCTION

A wide range of dimensionality reduction methods assume that data lie on a low-dimensional manifold which shows the intrinsic structure properties of the data [1]. Moreover, visualization is critical for data analysis, as the revealed intricate structures which provides us a unique perspective to discover unimagined effects. As known, similarity matrix shows the similarity degree between each data pairs, it actually plays a core role in a number of dimensionality reduction methods, since the objective function builds upon this matrix.

Regarding classification and segmentation tasks, in general, the traditional procedure is to extract from images color, shape or texture features, these features are then used to represent the images and for further applications. In such cases, the preserved image information are dependent on extracted feature characteristics. On the contrary, while compression-based similarity measures are effectively employed in applications on diverse data types as basically parameter free approach, a fast compression distance (FCD) metric has shown its advantages when applied on remote sensing datasets [2], [3], [4]. In such cases, the image information is directly taken account for further processing, without any feature dependency. t-SNE has been a popular tool to reduce data dimension and visualize the data. The idea behind it is to convert similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional manifold and the high-dimensional data. It performs well on preserving the complete local structure and some global structure of the data points. [5]

Hence in this research, we use FCD together with t-SNE to visualize a large semantic annotated TerraSAR-X dataset and a Sentinel-1 dataset as study cases. The visualizations represent the annotated semantic labels in such an intuitive way which help us to better understand the relationships between their annotated semantics and how their actual similarities are in manifold space.

2. METHODS

The FCD clustering method has been proved to be able to achieve similar classification performance comparing with the Normalized Compression Distance (NCD) method, regarding small- to medium- size datasets [4]. Due to its attractive performance, we have chosen the t-SNE method for the visualization.

2.1. Datasets

The TerraSAR-X dataset is obtained via an active learning scheme which was especially designed for high-resolution SAR imageries that covers over all 5 continents. The dataset contain image patches from 288 TerraSAR-X image scenes (41 scenes acquired in Africa, 6 from Antarctica, 59 from Asia, 80 from Europe, 40 from the Middle East, 54 from North and South America and 8 from ocean surfaces), with a total number of over 60000 individual image patches. All TerraSAR-X products are generated via the X-band instrument, using the high-resolution Spotlight mode. The incident angles throughout the scenes varies between 20 and 50 degrees. The resolution of the images scenes is 2.9 meters, with a pixel spacing of 1.25 meters. The polarization mode is horizontal (HH) for all products. Furthermore, for convenience

Thanks to the Helmholtz Association of German Research Centres (HGF) for funding.

we convert all intensity data to 8-bit integer precision. More information regarding this dataset can be found in [6]. In this research, we take into account 7 general classes, with 73 subclasses in total.

The Sentinel-1 dataset contains 33358 image patches of SAR urban targets, covering 21 major cities of China, with 10 categories, 2 polarization modes and 4 formats. [7] In this research, 300 patches are sampled from each class, regarding the 2 polarization mode, respectively.

2.2. Fast Compression Distance

The NCD is a way of measuring the similarity between two arbitrary objects, it is obtained by approximating the notion of Kolmogorov complexity as real-world compressors and normalize the information distance. The FCD is an accelerating version of NCD by avoiding computing the full compression concept. It is claimed as a fast speed without skipping the joint compression step which obtains better performance compared to NCD [8]. The idea behind is: the LZW algorithm extracts a dictionary $D(x)$ from each image patch, and encode into a string x, in ascending order. The definition of FCD is defined as an operation which mainly takes account of the joint number of patterns within two dictionaries $D(x)$ and $D(y)$. It is represented as:

$$FCD(x, y) = \frac{|D(x)| - \cap(D(x), D(y))}{|D(x)|} \qquad (1)$$

2.3. Implementation

Fig. 1 shows the flow chart of the proposed procedure for visualization based data analysis.

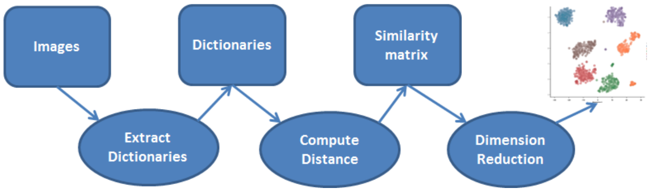


Fig. 1. Flow chart of proposed method.

3. RESULTS

The results in Figure 4 have been shown using t-SNE method to reduce the dimensionality. In our semantically annotated TerraSAR-X dataset, at the moment we have considered 6 general classes which contain 48 subclasses in total. Here different colors represent different semantic labels.

3.1. Discussions

- Figure (a) shows the visualization result of TerraSAR-X dataset, Class Residential. It contains 2789 patches, 11 subclasses: administrative, cemetery, commercial, high density, house, industrial, low density, medium density, mixed, skyscraper, sport.

- Figure (b) shows the visualization result of TerraSAR-X dataset, Class Industrial. It contains 3332 patches, 5 subclasses: car storage, industrial areas, mineral, pipe plant and storage tanks. The ’industrial areas’ subclass dominates and covers almost all the manifold space. The ’car storage’ subclass is shown as difficult to be classified.

- Figure (c) shows the visualization result of TerraSAR-X dataset, Class Public Transportation. It contains 11451 patches, 11 subclasses: airport, boat, bridge, docks, military, ports, public, railways, roads, shipping and train. There are outliers of subclass ’mixed forest’. The ’roads’ subclass has a very compact group outside of the main manifold space, however the rest spreads out in the space.

- Figure (d) shows the visualization result of TerraSAR-X dataset, Class Agriculture. It contains 7575 patches, 6 subclasses: agriculture, greenhouse, pasture, rice, stubble and vineyard. The ’agriculture’ and ’stubble’ subclasses cover all over the manifold space, showing their difficulties of being correctly classified.

- Figure (e) shows the visualization result of TerraSAR-X dataset, Class Natural Vegetation. It contains 2848 patches, 6 subclasses: broad leaf, coniferous, exotic, forest, mixed forest and natural.

- Figure (f) shows the visualization result of TerraSAR-X dataset, Class Bare Ground. It contains 3338 patches, 8 subclasses: brush, cliff, desert, hill, mountain, sand, shadow and soil. The ’desert’ subclass has a very compact area outside of the main manifold space, also some samples spread over the space.

- Figure (g) shows the visualization result of TerraSAR-X dataset, Class Water Bodies. It contains 4129 patches, 12 subclasses: beach, breaking waves, buoy, channels, delta, firth, flooded, ice, lake, ocean, rivers and sea. The subclass ”Lake” shows a very compact grouping outside of the main manifold space.

- Figure (h) shows the visualization result of Sentinel-1 dataset, with VV polarization mode. It contains 3000 patches, 10 subclasses: airport, dense low, general residential, high building, highway, railway, skyscraper, storage area, vegetation, villa.

- Figure (i) shows the visualization result of Sentinel-1 dataset, with VH polarization mode. It contains the same number of patches and subclasses as Figure (h).

Comparing to VV polarization mode, in VH polarization mode, each class shows more compactness in the manifold space.

4. CONCLUSION

In the paper, we have proposed a framework of a tool for unbiased data analysis, which can be applied to any kind of data, for visual data mining, understanding semantics and data clusters, etc. The FCD based similarity matrix effectively provides us a fast yet performance preserved insights in high-dimensional datasets with a non-parametric distance metric. Via the visualization on a TerraSAR-X dataset and a Sentinel-1 dataset, we have gained quick intuition and better understanding of the connections between the annotated semantics and the relationships within the data which is revealed as similarities in manifold space.

5. REFERENCES

[1] M.I. Jordan and T.M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, pp. 255–260, 2015.

[2] R. Cilibrasi and P.M.B Vitanyi, "Clustering by compression," *Journal of IEEE Transaction on Information Theory*, vol. 51, pp. 1523–1545, April 2005.

[3] R. Cilibrasi, *Statistical inference through data compression*, December 2006.

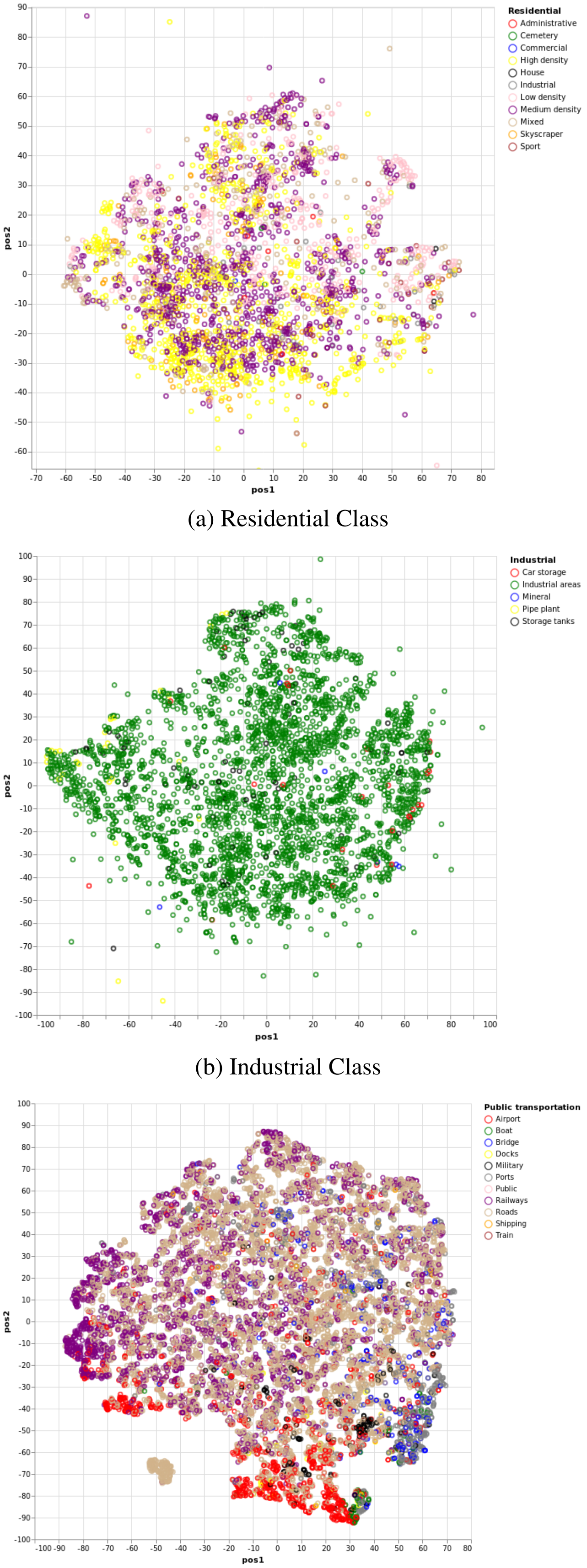
[4] C. Daniele and M. Datcu, "A fast compression-based similarity measure with applications to content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, pp. 293–302, February 2012.

[5] L. van der Maaten and G. Hinten, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, pp. 2579–2605, November 2008.

[6] C.O. Dumitru, G. Schwarz, and M. Datcu, "Land cover semantic annotation derived from high-resolution sar images," *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, vol. 9, pp. 2215 – 2232, June 2016.

[7] J. Zhao, Z. Zhang, W. Yao, M. Datcu, H. Xiong, and W. Yu, "Opensarurban: A sentinel-1 sar image dataset dedicated to urban class interpretation," *In draft.*

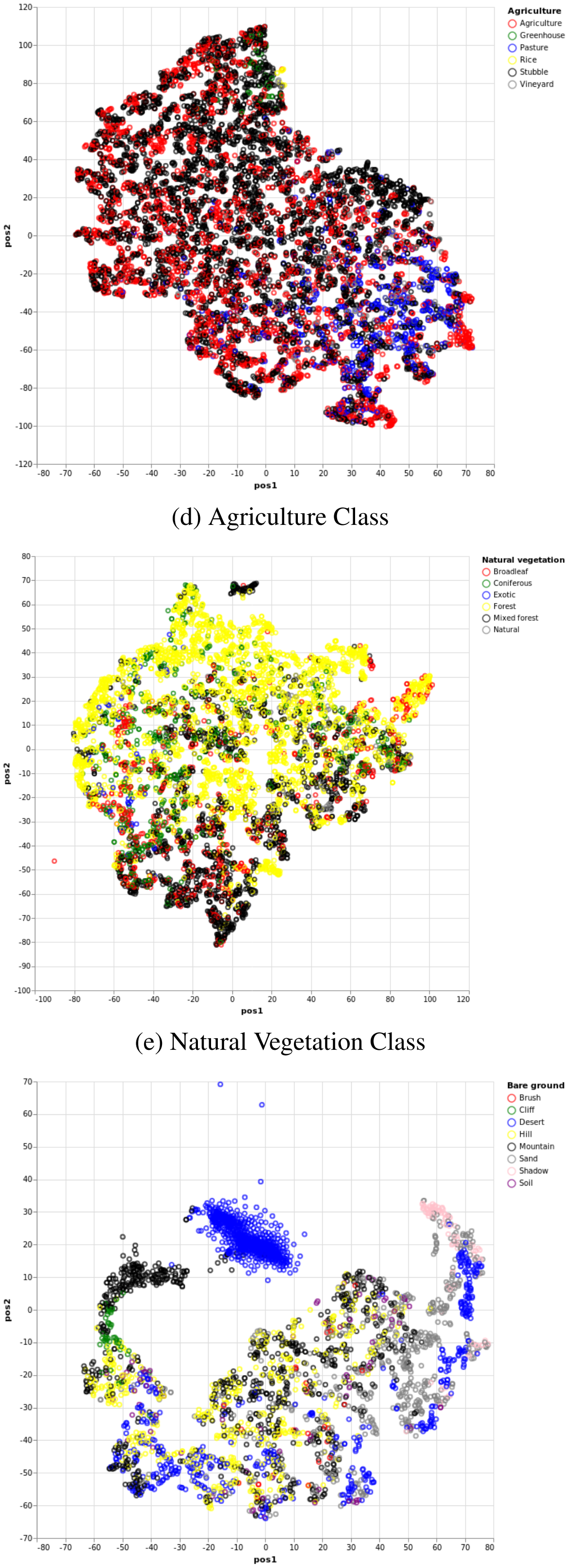
[8] M. Li, X. Chen, X. Li, Ma. B., and P.M.B. Vitányi, "The similarity metric," *IEEE Transaction of Information Theory*, vol. 50, pp. 3250–3264, 2004.



(a) Residential Class

(b) Industrial Class

(c) Public Transportation Class

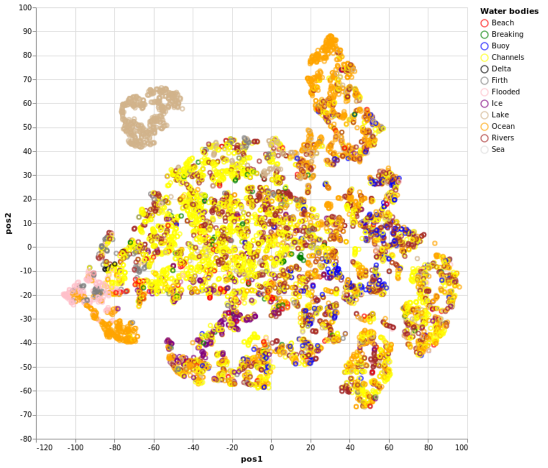


(d) Agriculture Class

(e) Natural Vegetation Class

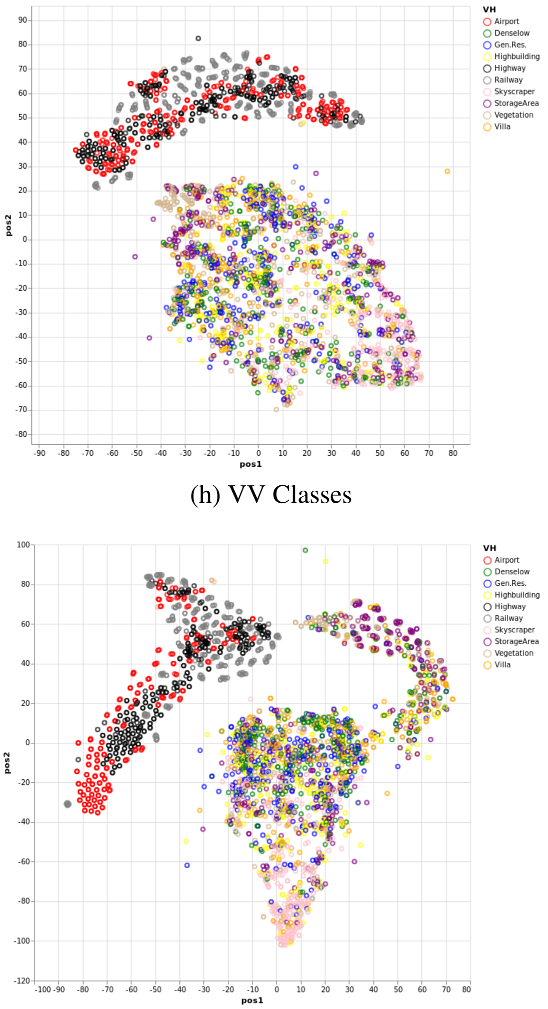
(f) Bare Ground Class

Fig. 2. Class Visualizations of TerraSAR-X Dataset.



g) Water Bodies Class

Fig. 3. Class Visualizations of TerraSAR-X Dataset.



(h) VV Classes

(i) VH Classes

Fig. 4. Class Visualizations of OpenSARUrban Dataset.