Defence Research and Development Canada | Recherche et développement pour la défense Canada

Using Shapley values and game theory to measure the effectiveness of different satellite image products in hybrid constellations

BACKGROUND & MOTIVATION

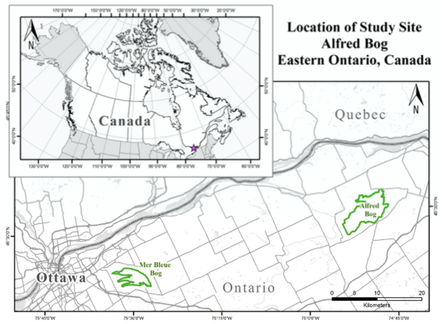
- The advent of small satellites and hybrid constellations have made multiple types of sensors and image products available.
- In classification problems, these diverse data sources can be used as inputs (i.e., variables) to perform categorization tasks.
- Using the optimal number of variables is key because:
  - Too little (under-fitting) may result in poor accuracy.
  - Too many (over-fitting) may increase computation time and yield a classifier that is too specific to the dataset.
- Common techniques for selecting variables include:
  - Metrics such as Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (MDG).
  - Methods such as Principal Component Analysis (PCA), Support Vector Machines (SVM), and genetic algorithms.
- However, the aforementioned variable selection techniques have some limitations, including:
  - Inability to define arbitrary groups of variables and determine the importance of each group as a unit.
  - No indication provided as to why a particular variable or group of variables is ranked as more or less important than another variable or group of variables.
  - No consideration of interactions between variables (for example, correlated or dependent variables).
  - For MDA and MDG, difficulty interpreting the importance values (both provide scaled numbers).

OBJECTIVE

Develop and demonstrate a metric of variable importance that:

- Is suited to classifiers that make use of imagery from different sensor types and from hybrid constellations;
- Can be applied to arbitrary groups of variables;
- Accommodates correlated and dependent variables; and
- Provides an easily interpreted measure of variable importance.

Figure 1. The location of the Alfred Bog temperate peatland complex near Alfred, ON, Canada. The bog spans an area of over 10,000 acres (40 km²). Source: Millard and Richardson, 2005.



METHODOLOGY

- The proposed method treats variables (or groups thereof) as players in a cooperative game where the goal is to maximize classification accuracy.
- Importance is measured using the Shapley value.
  - This metric quantifies the contribution of each individual player to the overall game outcome.
  - Proposed by Lloyd Shapley in 1951 as a way of determining fair wages in economics.
  - Defined as the weighted average of a player's contributions over all coalitions that the player can contribute to:

$$\frac{1}{\text{number of players}} \sum_{\text{coalitions excluding player k}} \frac{\text{marginal contribution of player k to coalition}}{\text{the number of coalitions of this size that exclude player k}}$$

- Properties of the Shapley value:
  - Non-discrimination (players with identical contributions have identical Shapley values).
  - Marginality (players that contribute more to the outcome have higher Shapley values).
  - Efficiency (sum of all Shapley values equals the score when every player participates).
- Example scenario: land-cover classification in Alfred Bog (Figure 1).
  - Classes: agriculture, forest, wetland.
  - Variables grouped into: SAR, optical, LiDAR.
  - Random forest classifier with 1000 trees per model (see Millard and Richardson, 2015).

Figure 2. Accuracy of the classifier by land-cover class and as a function of which sensor types are used as input variables (S = Synthetic Aperture Radar, O = Optical, and L = LiDAR).

Land-cover type	Accuracy type	Accuracy (%) by input data used (S = SAR, O = optical, L = LiDAR)						
		L	O	L+O	S	S+L	S+O	S+O+L
Agriculture	User's	83.6	48.5	86.2	92.0	97.1	92.8	97.1
	Producer's	88.2	65.1	87.1	91.3	94.8	90.4	95.1
Forest	User's	70.1	46.3	69.4	55.2	78.4	55.2	75.4
	Producer's	76.4	58.5	79.5	66.7	84.7	67.9	83.5
Wetland	User's	87.5	77.6	87.1	94.9	94.7	94.5	95.1
	Producer's	82.2	61.9	83.5	91.1	94.5	91.6	94.1
All	Overall	83.7	62.4	84.4	88.5	93.4	88.6	93.2

Figure 3. Flowchart illustrating the use of Shapley values to quantify the contribution each sensor type makes to the accuracy of the classifier.

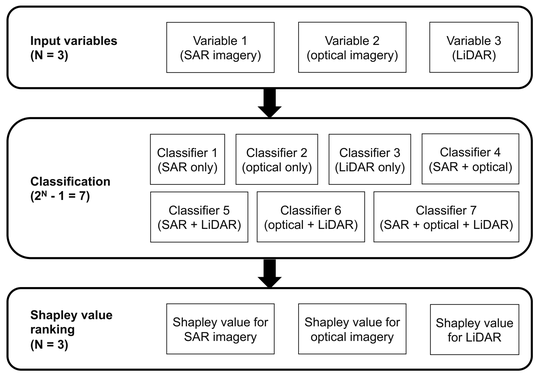


Figure 4. Shapley values for each sensor type, listed by land-cover class, with the highest value (i.e., most important group) for each in bold. The values quantify each group's contribution to overall accuracy. Per the efficiency property, the sum of the Shapley values on each row equals the rightmost column in Figure 2.

Land-cover type	Accuracy type	Shapley value (%)		
		SAR	Optical	LiDAR
Agriculture	User's	18.7	**40.1**	38.3
	Producer's	22.9	35.6	**36.6**
Forest	User's	9.6	32.7	**33.1**
	Producer's	15.5	**35.0**	32.9
Wetland	User's	28.5	33.1	**33.5**
	Producer's	23.3	**35.8**	34.9
All	Overall	22.2	**35.7**	35.3

RESULTS & DISCUSSION

- The accuracy of the land-cover classifier for different combinations of variable groups (sensor types) in the example scenario is shown in Figure 2. At first glance, it is not clear which group contributes the most to overall accuracy.
- By computing Shapley values per the flowchart in Figure 3, the importance of each group can be obtained (Figure 4). Here, the Shapley values represent the individual contribution each group of variables makes to the accuracy of the classifier.
- Per the efficiency property, on each row of Figure 4, the sum of all three Shapley values equals the value in the last column of Figure 2 (i.e., the accuracy achieved using all three groups).
- One limitation of this method is that the number of classifications scales exponentially with the number of variables (or groups of variables). Two ways to address this are:
  - Grouping the variables by a known characteristic and retaining only the group with the largest Shapley value.
  - Using a more conventional metric (such as MDA or MDG) to filter the list of variables before applying Shapley.

CONCLUSIONS & FUTURE WORK

This work shows that the Shapley value can be used as a metric of variable importance for classifiers that make use of imagery from different sensor types and from hybrid constellations. This metric:

- Can be applied to arbitrary groups of variables;
- Accommodates correlated and dependent variables (or groups of variables); and
- Provides an easily interpreted measure of variable importance.

Future applications may include other scenarios that can be modelled as cooperative games, such as the coverage of large geographical areas using a constellation of small satellites.

REFERENCES

Nandlall, S. D. and Millard, K. (2020), IEEE Geoscience and Remote Sensing Letters, 17(1), 42-46.

Banks, S. N. et al. (2017), Remote Sensing, 9(12), 1–27.

Millard, K. and Richardson, M. (2015), Remote Sensing, 7(7), 8489–8515.

Shapley, L. S. (1953), Annals of Mathematical Studies, 28, 307–317.

White, L. et al. (2017), Remote Sensing, 9(6), 1–29.

DRDC | RDDC

AUTHORS **Sacha D. Nandlall[1], Koreen Millard[2]**
[1]*DRDC – Ottawa Research Centre;* [2]*Carleton University*

VENUE **SmallSat 2020 Conference (August 1-6)**
Poster SSC20-P2-10

Canada