Rerarch paper

Application of machine learning method–decision tree to classification of oil use sentinel 2

Ha Phong Doan[1*], Thanh Bang Nguyen[1], Dang Hung Tran[1], Dao Xuan Hoang[1], Tran Anh Doan[1]

[1] Vietnam Institute of Meteorology Hydrology and Climate Change (IMHEN); doanhaphong@gmail.com; bangnt29@gmail.com; danghung2261991@gmail.com; hoangdx81@gmail.com; dtrananh2612@gmail.com

*Corresponding author: doanhaphong@gmail.com; Tel.: 84–913212325
Received: 27 February 2021; Accepted: 15 June 2021; Published: 25 August 2021

Abstract: Decision tree classification algorithms have significant potential in classifying remote sensing data. This article's approach method using decision tree technology to classify remote sensing images with the representative object as oil spill. First, this paper discusses the algorithmic structure and algorithmic theory of the decision tree. Second, the build of decision tree classification algorithm with 10 branches for oil spill classification using Sentinel 2 image data based on the JavaScript application's online interface (API) called Code Editor. Decision tree technology has several advantages for remote sensing applications due to their relatively simple, clear and intuitive classification structure.

Keywords: Decision Tree; Classification; Models; Machine Learning; Remote sensing image; Oil Spill.

1. Introduction

Remote sensing techniques in detecting oil spills based on remote sensors are widely used in the world. There are many types of remote sensing materials that have supported in monitoring and detecting polluting oil spill locations. Radar technology can monitor a wide area, day and night regardless of rainy/ cloudy weather due to an infrared device that detects oil spill at night and a UV laser (fluorosensor laser). Therefore, most oil spill maps are used radar technology.

In the Caspian Sea region of Azerbaijan, a stochastic model of oil spills risk leading to deterioration of water quality and coastal ecosystems was established by using satellite imagery. More than 411 satellite images obtained by sensors SENTINEL–1, LANDSAT–8, RADARSAT, ENVISAT and ERS from 1996 to 2017 are used for semi–automatic detection and classification of oil spills and slicks. Natural oil slicks are classified based on object orientation and visual interpretation. Parameters such as size, shape are determined, oil slick location takes into account the interaction between factors such as wind speed and current direction, depth and main source of emissions at the seabed. The pixel values in the combined pollution hazard model are positively correlated with the observed frequency of oil spills during the research period, contributing to improving the reliability of the modeling method [1].

The SANCHI tanker collision accident in January 2018 in the South China Sea shows that traditional techniques using synthetic aperture radar or daylight optics do not provide timely and complete information about the incident. Therefore, a Visible Infrared Imaging Radiometer Suite (VIIRS) and day/ night Automatic Indentification System (AIS) were used. The path and position can also be built as a numerical model, with an average error

from the origin < 15 km. While high–resolution optical images taken 4 days after the tanker sank show a much larger area of oil spill (> 350 km^2). In this case the VIIRS night data is the only satellite data that has proven effective in detecting the orbital drift and daily position of the ship, while all other sensor satellites can not observe or obscured. Together with digital equipment to detect oil tankers and other sensors to re–map the oil spill trajectory [2]. Sentinel–3 satellite with OLCI probe has also been used to track and monitor the Sanchi collision by combining close to the true color of channels 10,5,1 as shown in Figure 1.
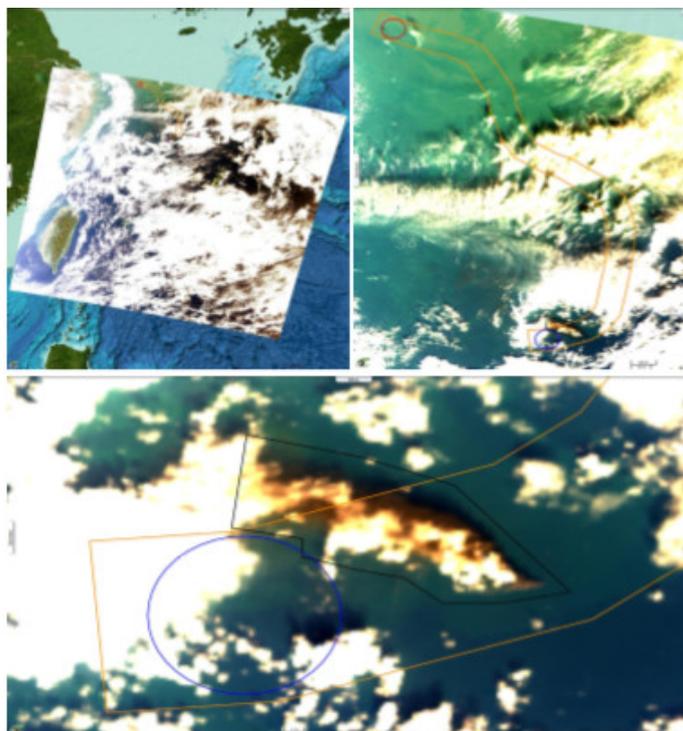

Figure 1. Sentinel–3 OLCI acquired on 14 Jan 2018 at 01:38:58 GMT [2].

Based on the reflection of the sun on the sea surface, it can be difficult to distinguish an oil stain. However, observed by the 300 m resolution Sentinel–3 image, the position of the boat and the cargo burning fiercely a few hours before sinking, an oil slick away from the sinking area was shown on this complex image.

Research on oil spills in the Deepwater Horizon area, used Synthetic Aperture Radar (SAR) within 10 km of the coast with suitable weather conditions to detect oil. In case of large oil spills occurred, SAR satellites will collect spill location data and provide oil spill feedback from one or more images per day, helping to estimate the trajectory of the oil spill. The combination of oil stains in shallow water will be affected by the current, wind direction and wave direction affecting the oil drift. By collecting and analyzing sequence of time SAR images in the same location, the speed and trajectory of oil drift in coastal areas can be estimated. The simplest oil spill model can be built, simulating the movement of oil on the ocean surface by tracking the motion of a large number of particles moving with velocities that are a combination of ocean surface velocities and speed of winds [3].

The results of satellite monitoring of the oil spill in the Barents Sea 2015–2016 also used SAR images from the European space satellite Sentinel–1A and Sentinel–1B. To efficiently analyze the SAR images, a site–based GeoMixer application–specific geographic approach was used. Dedicated Geoportal with a large volume of oceanographic, geographic and watershed information, including oil and gas infrastructure and coastlines. This tool with data from AIS has shown that the main sources of oil pollution (spill from 0.5 to 90 km^2) are freight transport and fishing vessels from which an overview map of the

oil spill was developed of reportred. Most surface oils are detected in the summer, which is a typical seasonal dynamic. Over time, single cases of large oil spills at sea were observed. Oil spill distribution maps based on SAR images provide assistance in obtaining new information on the extent and sources of oil contamination.

Thus, with the development of Remote Sensing technology in oil pollution detection, in addition to the advantages of Radar sensors, optical sensors (near infrared at night, visible, ultraviolet rays) must still be used. purple) to obtain as much information about oil pollution at sea as possible, combined with the information of the oil decoder [4].

If using a large number of satellites of all kinds, it is possible to build an oil orbit according to the corresponding time (from 15 to 60 days), combined with wave and wind data, which can serve for source tracing oil–based relatively accurately [1–4].

Oil pollution of unknown origin on the coast of Vietnam is usually pollution in two forms: visible to the naked eye: patches, streaks, oil patches, rocks, hydrated oil lumps floating in close waters or rocks, oil lumps that have been ored to varying degrees located onshore , yards; not visible to the naked eye: the form is soluble in water or deposited in sediment with content exceeding the limits allowed in coastal seawater standards [5].

Based on the results of research on the state–level topic [6] two types of ultra–high–frequency satellite imagery materials are ENVISAT ASAR and ALOS PALSAR, with a total of 110 scenes (2007). The number of ENVISAT ASAR images ordered is 10 scenes, at a processing level of 1B. The number of photos ALOS PALSAR ordered was 80 scenes, a processing level of 4.2. In addition, 20 ENVISAT ASAR images collected from the Department of Environmental Protection – Ministry of Natural Resources and Environment are also used to assist in the detection of oil stains at sea [7].
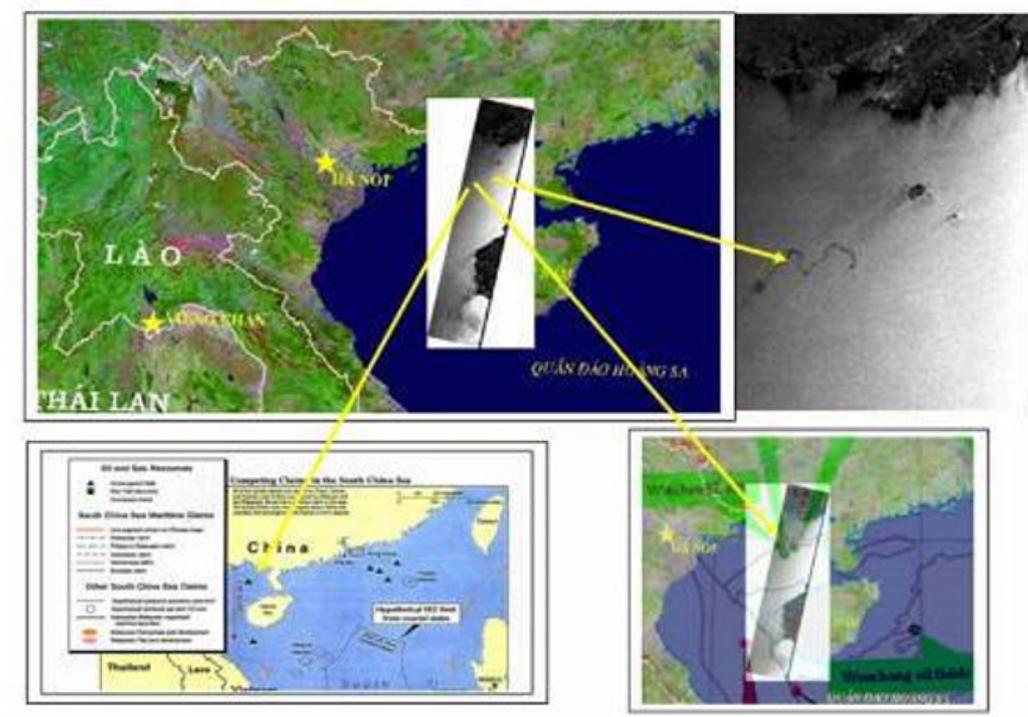

Figure 2. The source of oil pollution comes from the area where the oil field is located.

The source of oil pollution comes from the oil field area in southern China (in the Gulf of Tonkin region), the central coordinates of the polluted region about: Latitude 108º45' and latitude 20º50'. Envisat Asar satellite image number: ASA_IM Orbit 25179 Track 75, Swath I1. Collection time: December 24, 2006 at 02:48.64 (UTC). 147 ENVISAT ASAR scenes, 49 ALOS PALSAR scenes (December 2006 to April 2007) were used in the identification of oil pollution sources [8].
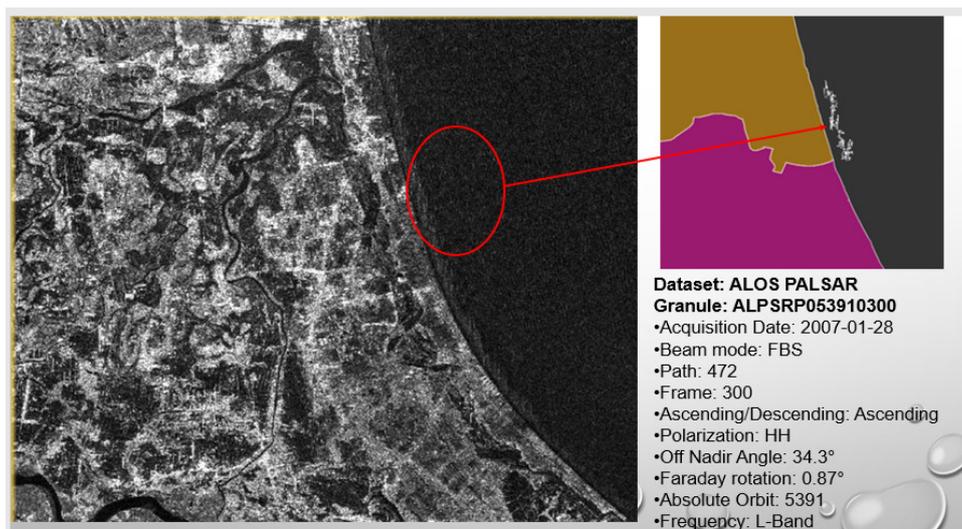
Figure 3. ALOS PALSAR satellite image dated January 28, 2007 Cua Dai Beach – Hoi An (Quang Nam), Non Nuoc (Da Nang). Oil pollution area and interpretation in red circle [8].

On January 30, 2007, the oil clumps and oil carpet stretched nearly 20 km from the sea area of Da Nang to Quang Nam, off the coast, the spreading oil slick was still floating on the water. Many black oil blocks began disintegrated and floated on the sea. Figure 3 shows the oil slick detected after treatment, starting for 6 months of major oil spills occurring in 20 coastal provinces of Vietnam.

Thus, the use of remote reconnaissance in monitoring and monitoring the location of oil spills of unknown origin (has been conducted in Vietnam in case of oil spills with particularly serious consequences, from the time of the oil spill to the detection occurs in 1 month) with a very large number of satellite images (approximately 200 scenes of all kinds have been selected). In particular, mainly using radar images of all kinds (image resolution from 10–30 m, suitable for the establishment of a ratio of 1: 50,000, serving a provincial scale), image processing and data extraction does not keep up with the requirements of location determination, not combined with the back–time model. Optical fars telescopes are almost uns mentioned and used appropriately in determining the trajectory, and specifications of oil spills.

Remote sensing information classification and exploitation play as an important role in the field of remote sensing technology. In the remote sensing classification application, traditional classification methods such as supervised classification, unsupervised classification and expert classification are all based on spectrum image features. However, since the image itself has the phenomenon of the same object having different spectra, and different objects have the same spectrum, classification methods based only on the ground spectral features always included many false classifications and omit errors. Numerous studies show that a classification combined with image spectrum and other supporting information can greatly improve the accuracy of a classification [9].

A random forest classifier is a group classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables [10]. Due to the accuracy of this classifier has been used in this research. Decision trees classification as a supervised classification method that exploits spatial data, solves previous classification problems and rules, as well as making use of certainty ecological and remote sensing knowledge and the results are always closely related to experience and expertise. It has the classification rules by the decision research process and does not need to satisfy the normal distribution [11]. It can use the knowledge of the earth in the GIS database to help with the classification and improve the accuracy of the classification [12].

Currently, decision tree classification has been applied in the field of land use, land use change as capture and scoping remote sensing images [13]. In the United States, the USGS and EPA, the departments have agreed to put in place the US land cover database plan and decision tree classification technology not only to be applied in soil classification but also in urban density information and urban density class information extraction. Land classification accuracy reached 73–77% and urban density information extraction accuracy is from 83% to 91% and canopy accuracy is 78–93% [14]. The mapping efficiency has been improved by 50% and can meet the requirements of large–scale soil classification data production.

Decision tree research method is one of the data mining methods to find classification problems in reality. It can classify the rules of decision tree expression form. The great advantage of the decision tree is that the research does not require a user with huge background knowledge [15]. As long as the input data examples can be represented by "attribute–result" and use this algorithm to learn. Categorizing the data obtained by the decision tree is easy to show and apply. Currently, foreign scholars have used decision trees to gather knowledge and apply them in spatial research and analysis.

Recently, sentinel 2 optical satellite with a resolution of 10 m is a source of high–resolution free remote images as well as low image repeat (5 days) is a valuable resource for us to monitor and monitor oil spills so it is necessary to develop a method of classification of oil spills from satellite images. Sentinel 2 with new classification algorithms and suitable for this satellite image.

Therefore, the purpose of the study is to test the application of the decision tree machine learning algorithm in oil spill classification using Sentinel 2 satellite images, as well as determine the best number of branches for this classification.

2. Materials and Methods

2.1. Description of study site

In this study, remote sensing data which is Sentinel 2 optical satellite data is used to detect oil spills in the coastal area of Binh Dinh province in May 2020. The reseach site located in Binh Dinh – a coastal province in central Vietnam. The province's territory stretches for 110 km in the North–South direction, with an average width of 55 km (50 km at the narrowest point, 60 km at the widest place). The North borders Quang Ngai province with a common boundary line of 63 km (the northernmost point has coordinates: 14°42'10" North, 108°55'4" East). The South borders Phu Yen province with a common boundary of 50 km (the southernmost point has coordinates: 13°39'10" North, 108°54'00" East). The West borders Gia Lai province with a common boundary line of 130 km (the westernmost point has coordinates: 14°27' North, 108°27' East). The East borders the East Sea with a coastline of 134 km, the easternmost point is Nhon Chau commune (Cu Lao Xanh) in Quy Nhon city (coordinates: 13°36'33" North, 109°21' East). Binh Dinh is considered as one of the gateways to the sea of the Central Highlands provinces and southern of Laos (Figure 4).

2.2. Research data

This study will test the recognition and classification using a decision tree algorithm for an oil spill accident in Binh Dinh Province, Vietnam. Multispectral optical satellite images show a sharp contrast to the surrounding oil–free water, which depends on several factors such as the shape of the sun observed, the type of oil, and the thickness. The Sentinel–2 satellite consists of 13 spectral channels, a near–infrared band at 10 m resolution, six red/short–wave infrared bands at 20 m, and three atmospheric correction bands at 60 m.
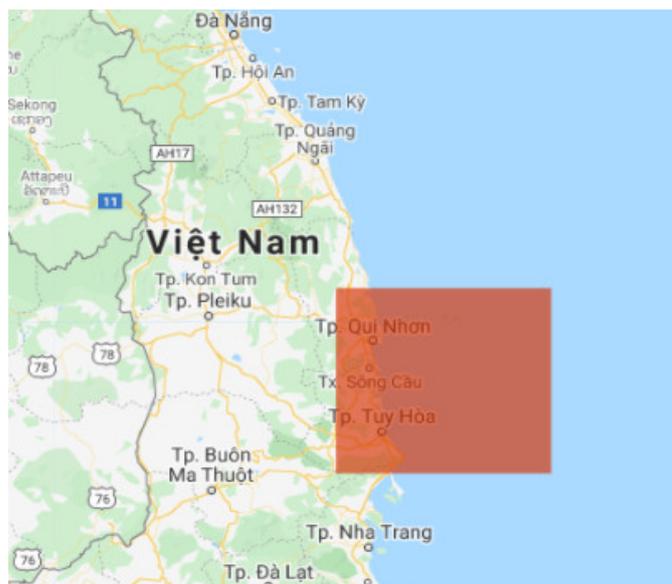
Figure 4. Binh Dinh coastal area.

Table 1. Sentinel 2 used images.

Research area	Date	Sensor Type	Name
Binh Dinh coastal	12/5/2020	Sentinel 2A	S2A_MSIL1C_20200512T030551_N0209_R075_T49PCR_ 20200512T061450
Binh Dinh coastal	12/5/2020	Sentinel 2A	S2A_MSIL2A_20200512T030551_N0214_R075_T49PBS_ 20200512T073119

2.3. Arithmetic Decision Tree

A decision tree is a method that can be inductive by training patterns and building a decision tree or decision rule and then use them to categorize data [16]. A decision tree composed of a root node, a series of inner nodes and leaf nodes. Each node can have only one primary node and two or more sub nodes. Nodes are connected by branches. Each inner node corresponds to an experimental property or property group, and every side corresponds to any possible value of the property. A node corresponds to a class attribute value, and a different node can correspond to the same class attribute value. Decision trees not only can be represented by trees, but also as a group of IF–THEN rules. Each line from the root to the leaf corresponds to a rule, and the rule's condition is, optionally, all of the node's attribute values. And the result of the rule is the class attribute of the leaf node on the road. Compared with decision attributes, the rules are simpler and more convenient to understand, to use, repair and can form the basis of the expert system [17]. Therefore, the rule is being used more and more in practical application.

2.3.1. Classification and Regression Tree (Cart)

Classification and regression tree is a common tree growth algorithm. It was presented by Breiman, etc. and it is a supervised classification method. It trains the sampler to build a binary tree and decode it for classification [18]. This feature is to take advantage of the full binary structure, in other words, the root node includes all patterns. The root node is divided into two subnodes according to defined division rules. This process repeats again in the child node and becomes a regression process until the subnode cannot divide into two subnodes [19]. Building the CART thinking training is based on a whole sampled data, building a multi–level and multi–leaf node tree to reflect the relationship between the nodes and then cutting the tree to create a series of subtree and selecting suitable trees in order to classify the data. In detail, the process involved building a tree and pruning a tree.

a) Tree Development

The distinction of tree nodes is named a branch and it corresponds to a subset of the training patterns. The branch of the root node corresponds to the entire training pattern. Then classification is the partition training sampling process. Therefore, the process of building a tree is to query properties to create partition rules. In this paper, CART adopts an index called "impurity level of the node": i (N) represents the impurity level of node N. When the data is from the same type, i (N) = 0; when the categories of data are evenly distributed, i (N) will be very large. The partition rule is created based on the minimum value of the impurity level function. Here two impurity functions are introduced.

- "Entropy impurity" also known as information impurity:

$$i(N) = -\sum_j P(w_j)\log_2 p(w_j) \tag{1}$$

where P (wj) is the calculation probability calculation for sampling data mode wj that node N belongs to the entire sampling process. According to the entropy parameters, if all data come from the same type then the impurity level is 0; or other impurity level is greater than zero; when all category data appear with the same probability calculation, entropy is the maximum value.

- VAR is the "Gini impurity level". According to nodes sampled from different categories and it is related to the total distribution variance, the below formula is given.

$$-\sum_{i \neq j} P(w_i)P(w_j) = 1 - \sum_j P^2(w_j) \tag{2}$$

The meaning of "Gini impurity level" is to represent the error rate of catalog creation in node N. When given a tree has grown to node N and the node is property queried, a heuristic thinking method has visible is to select the query with the fastest reduction in noise. Reduced impurity levels may be noted:

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L)i(N_R) \tag{3}$$

Include NL and NR as separate left and right buttons; i (NL), i (NR) are the individual impurity level. L P is the probability that when query T is passed, the tree grows from N to NL. And the optimal query value S is the maximum value of Δi (T) [20].

b) Reduce Branches

If we consisted in building the tree until all leaf nodes reach the minimum impurity level. The data will be over–sorted and the tree will degrade into a convenient lookup table. It may not be good for the overall performance of interfering lamps due to the larger Bayes error. Conversely, if the branches stop too early, the sampling training error will not be small enough resulting in a very poor type efficiency. One main method of stopping the branches is pruning, which can also prevent the tree from growing too big. In this paper, the following indicators are adopted to achieve the goal.

$$\cos t = \alpha.size + \sum_{leaf\ node} i(n) \tag{4}$$

where cost is denoted the cost function of the tree weighted error probability and the sum of the complex penalty. Size is expressed from the number of leaf nodes to the weight complexity of the tree classifier, α is denoted as an index of complexity.

According to equation (4), pruning can be done in two steps:

(1) In all sibling leaf nodes, compare t cos after the associated leaf nodes.

(2) Delete the leaf node that has the most t cos. If t cos has not decreased, nothing will be done [21].

Repeat the above process until you cannot continue pruning. During pruning, the training error inferred with leaf nodes increases; The test error deduced at the start and

reached the minimum and then rolled up gradually affected by the training patterns. Use independent data to test and select a subtree with minimum test error as decision tree. This paper applies the heuristics–cross validation technique to select the best tree has 10 cross validations. The training samples are divided into ten subsets that are equal in number and not the same. The classifier will train the data 10 times, and each time the nine subset groups are trained, and one test is the validation set for estimating the test error. The estimated test error is the mean of the ten groups [22].

2.3.2. Enhancement Method – New Applied Techonology in Decision Trees

In decision tree classification design, an enhancement technology was widely used in the mid–1990s in the field of machine learning to improve the accuracy of the classifier. This method can increase the accuracy of the sampled classifier which is difficult to discern. At the same time, this technology can cut down on the sensitivity that the categorization algorithm affects data noise and training sampling error. Reinforcement is a learning method that can increase the accuracy of any learning algorithm, and it can strengthen a weak learning algorithm into a strong learning algorithm. Its theory is probably derived from an approximate learning model. It can leverage a number of mathematical algorithms to generate a wide range of basic classifiers [23]. Every basic classifying training depends on the classification results generated by the old classifier and ends up failing training samples with great weight so that they pay more attention during the next learning process. Finally, the classifier has a vote weighting for all base classifiers and receives the final results and reduces the signal classifier error and improves the classification accuracy. Freund and Schapire came up with the most pragmatic boosting algorithm–AdaBoost followed the basic theory of reinforcement in 1995 and is widely applied.

2.4. Applying Oil Spill Analysis Test Construction Algorithm for an Area

This article uses Google Earth Engine (GEE) software. Google Earth Engine works through a JavaScript Application Online Interface (API) called Code Editor. On this interface, users can write and run scripts/ scripts to share and repeat the processes of analyzing and processing geospatial data. Code Editor helps users to perform all functions available in Earth Engine. The process of constructing oil spill analysis method for Sentinel 2 image is shown in detail in Figure 5.

First, we declare Sentinel 2 data set as input data set for analysis. Thus, we filtered data by boundary, according to the date when the oil spill occurred, as well as filtered the scene with less clouds. After filtering the image, it is necessary to create a data set for training between the oil–free inlet. Conduct a very detailed and accurate sampling as well as the more samples, the better the output. After creating the training data set, the decision tree will be sorted according to Cart algorithm. The classification results are corrected by resample until best results are obtained. This process is also the process of selecting the best number of branches for the classification process. Using the trial and error method, with 10 branches the classification result is the most accurate. Finally, extract oil spill results and make final manual calibration using Arcmap software.
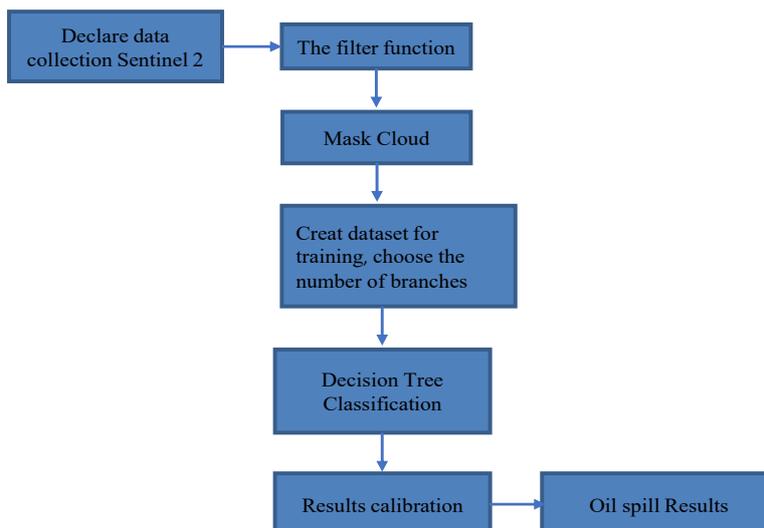
Figure 5. A Method of Constructing a Classification of Decision Trees On GEE.

3. Results and discussion

Run time series analysis 2020 in the coast of Binh Dinh Province detects the oil spill on May 12 in detail (Figure 6).

Figure 6. (a) and (b) 2 locations noticed oil spill on Sentinel 2 photo at the coast of Binh Dinh on May 12, 2020.

Create training datasets with 41 elements in google earth engine (Figure 7).

Figure 7. Training datasets information.

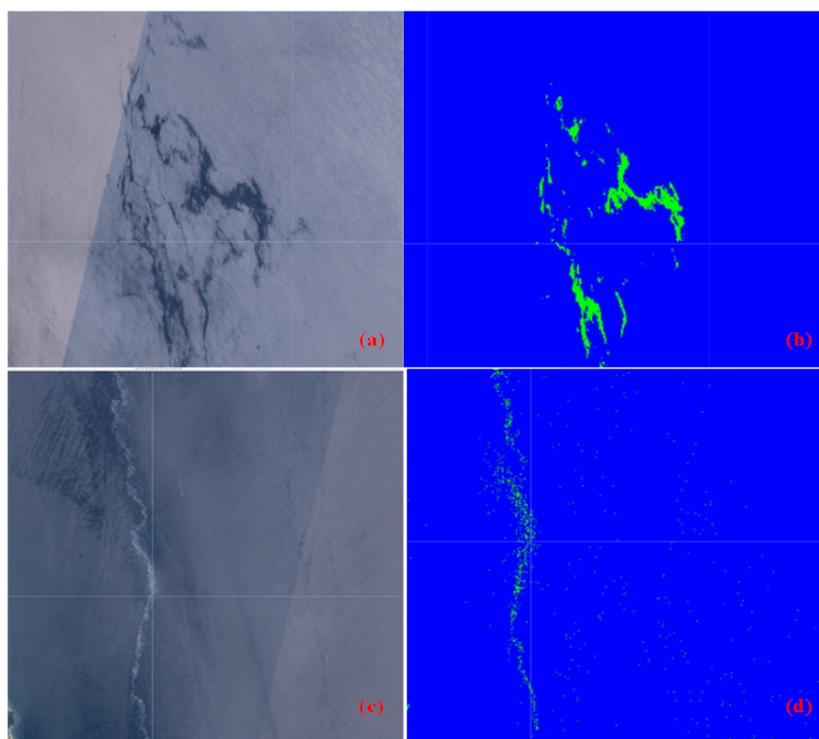The final oil spill result is shown as (Figure 8).

Figure 8. (a–d) The results of classification of two Sentinel 2 images on May 12, 2020

The results of classification were decent, the research was able to detect oil spills using Sentinel 2 images. As can be seen in Figures 8b and 8d, the oil spills areas were highlighted as green in contrast with the blue color of our seawater surface.

4. Conclusion

The advantage of the decision tree algorithm used in remote sensing data classification is that it can show the lack of the MLC algorithm when processing complex distributed data sets. Decision trees offer more flexibility and powerful with data distribution and categorization markup functions. Therefore, evenwhen the distribution of remote sensing image data characteristics is very complex or the source data size has different distributions and statistical ratios, the decision tree classification method can obtain the best classification results.

With the development of artificial intelligence technology and theory, the research of remote sensing image classification has reached to a higher level. Geographic knowledge and support for geographic decision–making can promote the accuracy of remote sensing image classification, extracting information and expert systems is a good approach to solve this problem. The article successfully researched and applied to classify oil spills is a good example to promote in–depth studies on the application of decision tree algorithm in resource and environmental management.

A big problem to be solved in the classification problem using the decision tree algorithm is the optimal number of branches for our problem. In the framework of the new paper, the trial and error method is applied with low accuracy, so it is necessary to develop a new method to give more accurate results.

Competing interest statement: The authors declare no conflict of interest.

Reference

1. Bayramov, E.; Kada, M.; Buchroithner, M. Monitoring oil spill hotspots, contamination probability modelling and assessment of coastal impacts in the Caspian Sea using SENTINEL–1, LANDSAT–8, RADARSAT, ENVISAT and ERS satellite sensors. *J. Oper. Oceanogr.* **2018**, *11*, 27–43.
2. Sun, S.; Lu, Y.; Liu, Y.; Wang, M.; Chuanmin Hu, C. Tracking an Oil Tanker Collision and Spilled Oils in the East China Sea Using Multisensor Day and Night Satellite Imagery. *Geophys. Res. Lett.* **2018**, *45(7)*, 3212–3220. https://doi.org/10.1002/2018GL077433.
3. Garcia–Pineda, O.; Holmes, J.; Rissing, M.; Jones, R.; Wobus, C.; Jan Svejkovsky, J.; Mark Hess, M. Detection of Oil near Shorelines during the Deepwater Horizon Oil Spill Using Synthetic Aperture Radar (SAR). *Remote Sens.* **2017**, *9*, 567. https://doi.org/10.3390/rs9060567.
4. Ivanov, A.Y.; Filimonova, N.A.; Kucheiko, A.Y.; Evtushenko, N.V.; Terleeva, N.V. Oil spills in the Barents Sea based on satellite monitoring using SAR: spatial distribution and main sources. *Int. J. Remote Sens.* **2017**, *39*, 4484–4498. https://doi.org/10.1080/01431161.2017.1371869.
5. Nhân, N.H. và cs. Thử nghiệm mô hình OILSAS – công cụ trợ giúp ứng phó sự cố tràn dầu trên vịnh Vân Phong, tỉnh Khánh Hòa. *Tuyển tập Nghiên cứu Biển* **2014**, *20*, 19–29.
6. Dương, N.Đ. Ô nhiễm dầu trên vùng biển Việt Nam và biển Đông. Đề tài KC.09.22/06–10, 2010.
7. Dương, N.Đ.; Thu, H.L.; Anh, L.V.; Anh, N.K. Ô nhiễm dầu trên vùng biển Việt Nam và kế cận. *Tạp chí các Khoa học về Trái đất* **2013**, *35(4)*, 424–432.
8. Khánh, N.Q. và cs. Ứng dụng công nghệ viễn thám kết hợp với GIS phục vụ giám sát sự cố ô nhiễm dầu ở Việt Nam. *Tạp chí Môi trường* 2014, 7, 13–15.
9. Song, Y.Y.; Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* **2015**, *27*, 130–135.
10. Jenhani, I.; Amor, N.B.; Elouedi, Z. Decision trees as possibilistic classifiers. *Int. J. Approximate Reasoning* **2018**, *48*, 784–807.
11. Fouad, M.M. A Decision Tree Classification Model for University Admission System. *Int. J. Adv. Computer Sci. Appl.* **2012**, *3*, 184–186.
12. Zeng, X.; Yuan, S.; Li, Y.; Zou, Q. Decision Tree Classification Model for Popularity Forecast of Chinese Colleges. *J. Appl. Math.* **2014**, 675806. https://doi.org/10.1155/2014/675806.
13. Dai, Q.Y.; Zhang, C.P.; Wu, H. Research of Decision Tree Classification Algorithm in Data Mining. *Int. J. Database Theory Appl.* **2016**, *9*, 1–8.
14. Patel, H.H.; Prajapati, P. Study and Analysis of Decision Tree Based Classification Algorithms. *Int. J. Computer Sci. Eng.* **2018**, *6(10)*, 74–78.
15. Zhong, Y. The analysis of cases based on decision tree. Proceeding of the 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2016.
16. Lian, H.; Yang, Y. Study on the Test Data Fault Mining Technology Based on Decision Tree. *Procedia Comput. Sci.* **2019**, *154*, 232–237.

17. Ivanov, A. Decision Trees for Evaluation of Mathematical Competencies in the Higher Education: A Case Study. *Mathematics* **2020**, *8(5)*, 748. https://doi.org/10.3390/math8050748.

18. Zimmerman, R.K.; Balasubramani, G.K.; Nowalk, M.P.; Eng, H.; Urbanski, L.; Jackson, M.L.; Jackson, L.A.; McLean, H.Q.; Belongia, E.A.; Monto, A.S.; Malosh, R.E.; Gaglani, M.; Clipper, L.; Flannery, B.; Wisniewski, S.R. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC Infect. Dis.* **2016**, *16(1)*, 503. https://doi.org/10.1186/s12879-016-1839-x.

19. Lemon, S.C.; Roy, J.; Clark, M.A.; Friedmann, P.D.; Rakowski, W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Ann. Behav. Med.* **2003**, *26*, 172–181.

20. Machuca, C.; Vettore, M.V.; Krasuska, M. et al. Using classification and regression tree modelling to investigate response shift patterns in dentine hypersensitivity. *BMC Med. Res. Method.* **2017**, *17*, 120. https://doi.org/10.1186/s12874-017-0396-3.

21. Denisko, D.; Hofman, M.M. Classification and interaction in random forests. *PNAS* **2018**, *115*, 1690–1692.

22. Basu, S.; Kumbier, K.; Brown, J.B.; Yu, B. Iterative random forests to discover predictive and stable high–order interactions. *PNAS* **2018**, *115*, 1943–1948.

23. Riddick, G. Predicting in vitro drug sensitivity using Random Forests. *Bioinf.* **2011**, *27*, 220–224.