

ПОШУК ЗНАЧУЩИХ ОЗНАК ДЛЯ ВИЯВЛЕННЯ ЗАХВОРЮВАНЬ ЛІСУ НА ОСНОВІ СУПУТНИКОВИХ ЗНІМКІВ

Є. В. Салій¹, А. М. Лавренюк¹

¹ Навчально-науковий Фізико-технічний інститут

Анотація

В роботі розглянуто метод пошуку значущих ознак за допомогою відстані Бгаттачар'я. Було розглянуто наскільки ефективними, з точки зору сепарабельності класів, є окремі ознаки, а саме: канали супутникових знімків Sentinel-2 та різні вегетаційні індекси, що розраховані за допомогою цих каналів.

Було виявлено, що такі вегетаційні індекси як VARIGreen, NGRDI, NDRE2, NDVI, GLI добре (а в рамках даного експерименту – найкраще) відділяють хворі ліси від здорових.

Ключові слова: Sentinel-2, вегетаційні індекси, відстань Бгаттачар'я

Вступ

Ліси є одним із головних компонент екосистеми нашої планети, яка є домівкою значної частини тваринного світу, допомагає протистояти ерозії ґрунту, забезпечує відчутну частину переробки вуглекислого газу, є важливою складовою економіки.

Тому догляд за лісом є важливим не лише для екології, але й для людства.

Однією із складових частин догляду за лісом є його санітарні рубки. Завдяки яким із лісу вилучаються хворі, згорівші, мертві дерева. Це попереджує розвинення та розповсюдження захворювань лісу та дозволяє йому краще розвиватися.

Таким чином, раннє виявлення аномалій в лісовій масі є запорукою ефективного догляду за ним.

Моделі, які вирішують задачу виявлення аномалій та класифікації земного покриття в цілому, використовують супутникові знімки [1]. Але окрім наявності самих знімків, потрібно розуміти, яку інформацію з них можна отримати та яка частина цієї інформації є значущою для того, щоб з найбільшою ймовірністю виявити, що з лісом щось трапилось.

Разом із цим, задачі такого типу часто вирішуються за допомогою моделей глибинного навчання [2, 3]. Але такі моделі потребують великої кількості грошей, часу та обчислювальних ресурсів для навчання. Тому перед навчанням таких моделей потрібно впевнитися, що вхідні дані дозволяють ефективно вирішувати поставлену задачу.

1. Постановка задачі та опис даних

Перед нами стоїть задача: на основі супутникових знімків Sentinel-2 [4] виявити, чи є якась ділянка лісу хворою/зараженою.

На вхід ми приймаємо супутниковий знімок (рис. 1(а)), який містить 4 канали з 10 метровим,

6 каналів із 20 метровим та 1 канал із 60 метровим просторовим розрізненням, та маску лісу (рис. 1(б)).

У ході роботи в якості маски лісу було взято Forest Type 2018 [5]. У рамках роботи розглядаються лише ділянки хвойного лісу.

Для аналізу розподілів та навчання моделей також надана маска захворювання (рис. 1(в)).

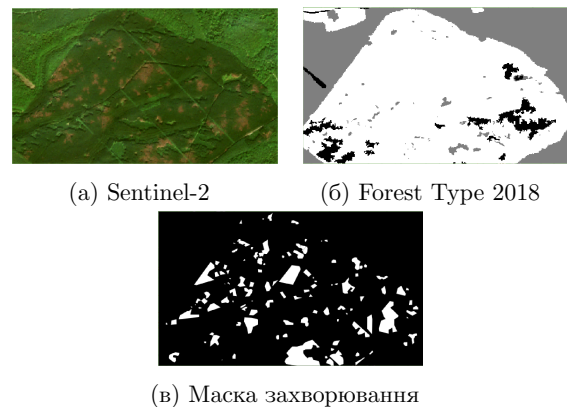


Рис. 1. Приклад вхідних даних. (б) білий - хвойний, сірий - листяний, чорний - не ліс; (в) білий - хворий, чорний - здоровий

Навчальні та тестові діляки розташовані в східній частині регіону Гранд-Ест, Франція. Їхнє місцезнаходження можна побачити на рис. 2.

2. Відстань Бгаттачар'я

Відстань Бгаттачар'я [6] використовується для визначення відносної подібності двох вибірок. Таким чином, дана величина має добре підходити для оцінки роздільності класів класифікації.

Для розподілів ймовірностей p та q відстань Бгаттачар'я визначається як:

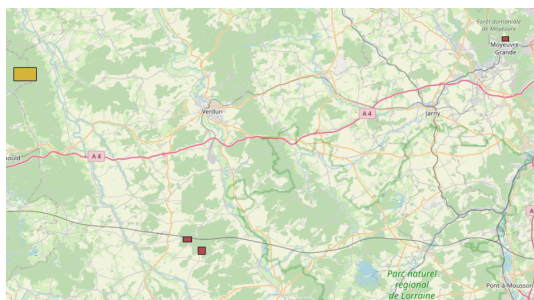


Рис. 2. Зони інтересу: жовтий - тестова ділянка, червоний - навчальні ділянки

$$D_B(p, q) = -\ln BC(p, q),$$

де $BC(p, q)$, також відомий як коефіцієнт Бгаттачар'я [7], у випадку дискретного розподілу з n значень визначається за формулою:

$$BC(p, q) = \sum_{i=1}^n \sqrt{p_i q_i}.$$

З вигляду $BC(p, q)$ зрозуміло, що при перекритті двох вибірок значення коефіцієнта буде зростати. А у разі, якщо вибірки не перетинаються, значення коефіцієнту буде рівним нулю.

У нашому випадку в якості дискретного розподілу буде виступати гістограма значень нашої вибірки з n стовпцями.

Важливо зазначити, що даний коефіцієнт, а відповідно і відстань, є чутливими до кількості стовпців гістограми. При надто малій кількості ми можемо недооцінити розділюванність, а при надто великій навпаки переоцінити її.

3. Аналіз даних

Оскільки у разі повністю розділених класів коефіцієнт Бгаттачар'я рівний нулю, а відповідна йому відстань буде прямувати до нескінченності, то для виявлення найбільш перспективних ознак ми повинні звертати увагу на найбільші значення відстані.

Канали зображення Sentinel-2

На основі маски зараження розіб'ємо наші зображення на 2 множини пікселів: H (healthy, здорові) та S (stressed, хворі). Для кожної компоненти із цих множин обраховуємо гістограми. Зважаючи на незначну кількість елементів в S навчальної вибірки, а саме 717, було вирішено будувати гістограми з $n = 32$ стовпцями.

На основі гістограм обрахували відстань Бгаттачар'я, а для самих множин порахували математичне сподівання (M) та дисперсію (σ). Отримані результати наведені в таблиці 1.

Як можна побачити із таблиці 1, Red-Edge (B5, B6, B7 B8A) та Near-Infrared (B8) компоненти зображення, окрім B5, дуже схожі у заражених та здорових лісів, а тому вони скоріше за все не будуть значущо впливати на результати класифікації.

Таблиця 1. Відстань Бгаттачар'я для каналів супутникових знімків Sentinel-2

Band	M_H	σ_H	M_S	σ_S	D_B
B4	0.021	0.006	0.032	0.008	0.443
B12	0.043	0.014	0.058	0.013	0.269
B9	0.246	0.057	0.245	0.033	0.223
B5	0.054	0.013	0.062	0.01	0.214
B11	0.093	0.027	0.116	0.024	0.211
B2	0.018	0.005	0.022	0.005	0.134
B3	0.033	0.008	0.037	0.007	0.108
B6	0.167	0.044	0.159	0.026	0.106
B8A	0.233	0.059	0.22	0.033	0.096
B7	0.211	0.053	0.195	0.03	0.086
B8	0.222	0.061	0.209	0.041	0.054

Вегетаційні індекси

Оскільки ми маємо справу з лісами, то можемо використовувати добре відомі вегетаційні індекси [8, 2], значна частина яких розроблялася спеціально для моніторингу стану рослинності.

За аналогічними до попередньої підсекції міркуваннями, отримуємо таблицю 2.

Таблиця 2. Відстань Бгаттачар'я для вегетаційних індексів.

Index	M_H	σ_H	M_S	σ_S	D_B
VARIGreen	0.322	0.082	0.113	0.108	0.724
NGRDI	0.215	0.055	0.077	0.074	0.686
NDRE2	0.588	0.044	0.517	0.04	0.478
NDVI	0.829	0.046	0.747	0.052	0.464
GLI	0.254	0.049	0.16	0.061	0.46
NDWI	0.433	0.065	0.313	0.072	0.442
RDI	0.186	0.053	0.265	0.05	0.42
DSWI	2.28	0.389	1.702	0.329	0.385
DRS	0.048	0.014	0.067	0.014	0.344
PBI	7.001	1.629	5.769	1.038	0.174
CIG	6.386	1.802	5.12	1.076	0.155
GNDVI	0.751	0.051	0.713	0.04	0.147
NDRE3	0.05	0.017	0.061	0.02	0.084
CVI	12.992	5.447	10.644	3.153	0.062

Відразу можна помітити, що значна кількість вегетаційних індексів має краще розділювати класи. У нашому випадку, 5 вегетаційних індексів має більшу відстань Бгаттачар'я ніж червоний (B4) канал зображення.

4. Опис експерименту

Беручи до уваги таблицю 1 та таблицю 2, побудуємо 2 моделі, які будуть приймати на вхід по 6 компонент з найбільшою відстанню Бгаттачар'я.

Таким чином, перша модель (band-based) буде використовувати канали Sentinel-2: B2, B4, B5, B9, B11, B12.

Друга модель (index-based) же буде оперувати і вегетаційними індексами, і каналами: B4, VARIGreen, NGRDI, NDRE2, NDVI, GLI.

В якості моделі візьмемо Random Forest [9] з параметрами:

- Кількість дерев: 300
- Максимальна глибина: 170
- Критерій: Gini

5. Опис результатів

Оскільки маємо справу з бінарною класифікацією, можемо обмежитися лише основними метриками. Детальніше дізнатися про них можна в [10].

Таблиця 3. Тестові результати band-based моделі

	Precision	Recall	F1-Score	Support
Здоровий	0.97	0.95	0.96	113779
Хворий	0.64	0.78	0.70	12881
Accurasy			0.93	126660

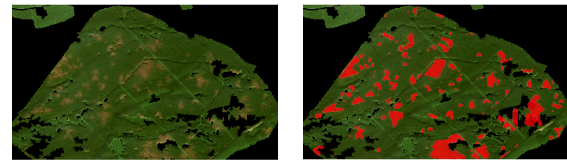
Таблиця 4. Тестові результати index-based моделі

	Precision	Recall	F1-Score	Support
Здоровий	0.99	0.90	0.94	113779
Хворий	0.51	0.92	0.66	12881
Accurasy			0.90	126660

З таблиці 3 та таблиці 4 видно, що band-based модель має на 3% більшу загальну точність (ассурасу = 93%, проти 90%). Але значною мірою це пов'язано з незбалансованістю валідаційного набору даних (114 тисяч здорових та 13 тисяч хворих). Загалом, index-based модель значно краще розпізнає дійсно захворівші ліси (recall = 92% в порівнянні з 78% у band-based), але водночас часто приймає здоровий ліс за хворий.

Але така ситуація може мати місце у випадку, якщо маска зараження недооцінює масштаби захворювання, тому проведемо візуальний огляд ділянки.

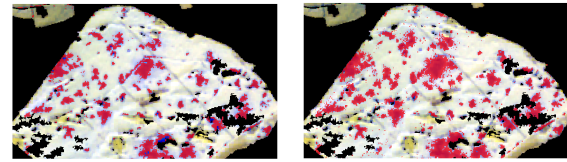
Дивлячись на рис. 3(б) та рис.4(а) стає зрозуміло, що band-based модель дійсно краще відповідає масці. Але якщо поглянути на рис. 5(б) можна побачити, що маска не покриває усієї ділянки хворого лісу. Якщо ж зосередити свою увагу на посинілих ділянках на



(а) Тестова зона

(б) Маска зараження

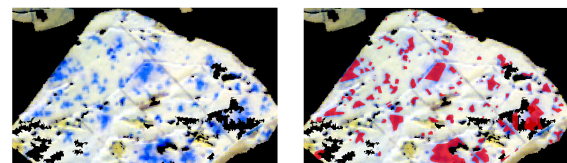
Рис. 3. Тестова зона в RGB (B4, B3, B2) та її маска



(а) Band-based

(б) Index-based

Рис. 4. Результати класифікації



(а) Тестова зона

(б) Маска зараження

Рис. 5. Тестова зона в FalseColor (B12, B11, B9) негативі та її маска

рис. 4, то ми маємо змогу помітити, що index-based модель покрила більшу площу посинілості, ніж band-based.

З огляду на це можемо припустити, що модель, яка навчалася на основі вегетаційних індексів змогла краще виділити ознаки, які говорять про те, що ліс є зараженим. А причиною заниженого значення точності виступає недооцінка маскою ступеня зараження.

Висновки

У ході роботи було проведено дослідження того, чи є відстань Бгаттачар'я придатною для використання в якості показника, що дозволяє виявити значущість ознак в задачі бінарної класифікації.

Як результат проведеного аналізу даних та експерименту, було показано, що відстань Бгаттачар'я ефективно дозволяє визначити відносну значущість ознак. Для визначення ж абсолютної значущості потрібно провести окреме дослідження того, як значення відстані впливає на точність моделей.

Аналіз результатів експерименту також показав, що при вирішенні таких задач потрібно проводити візуальний огляд отриманої класифікації, оскільки чисельна оцінка точності класифікації може виявитися неточною через недостатньо якісну розмітку даних. Разом із цим потрібно пам'ятати, що ми маємо справу з незбалансованою класифікацією.

Було показано, що вибір ознак з найбільшою відстанню Бгаттачар'я серед каналів супутникових зображень та вегетаційних індексів призводить до покращення результатів класифікації.

У подальшому можна провести дослідження сепарабельності не лише для відомих вегетативних індексів, а й для створених самостійно. Багато вегетативних індексів своєю формулою відповідають таким шаблонам як: $\frac{A-B}{A+B}$, $\frac{A}{B}$, $\sqrt{A^2 + B^2}$. Тому підставивши на місце A та B свій канал чи взагалі інший індекс можна створити новий. Додавши до шаблону коефіцієнти можливо вдалось би ще покращити сепарабельність.

Дивлячись на рис. 3(a) та рис. 5(a), можна візуально виділити фрагменти з кольором, різуче відмінним від кольору здорового лісу. Тому, можливо, що індекс, побудований на основі перетворень кольорових просторів, що добре розділяють колір та інтенсивність, був би ефективним.

Перелік використаних джерел

1. Parcel based classification for agricultural mapping and monitoring using multi-temporal satellite image sequences / N. Kussul, G. Lemoine, J. Gallego, S. Skakun, M. Lavreniuk // 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). — IEEE, 07.2015. — DOI: [10.1109/igarss.2015.7325725](https://doi.org/10.1109/igarss.2015.7325725). — URL: <https://doi.org/10.1109/igarss.2015.7325725>.
2. Detecting Pest-Infested Forest Damage through Multispectral Satellite Imagery and Improved UNet++ / J. Zhang, S. Cong, G. Zhang, Y. Ma, Y. Zhang, J. Huang // Sensors. — 2022. — Т. 22, вип. 19. — ISSN 14248220. — DOI: [10.3390/s22197440](https://doi.org/10.3390/s22197440).
3. Land Cover Changes Analysis Based on Deep Machine Learning Technique / N. N. Kussul, N. S. Lavreniuk, A. Y. Shelestov, B. Y. Yailymov, I. N. Butko // Journal of Automation and Information Sciences. — 2016. — Т. 48, № 5. — С. 42—54. — DOI: [10.1615/jautomatinfscien.v48.i5.40](https://doi.org/10.1615/jautomatinfscien.v48.i5.40). — URL: <https://doi.org/10.1615/jautomatinfscien.v48.i5.40>.
4. Sentinel-2 data for land cover/use mapping: A review / D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, M. Ranagalage. — 2020. — DOI: [10.3390/rs12142291](https://doi.org/10.3390/rs12142291).
5. Forest Type 2018. — URL: <https://land.copernicus.eu/pan-european/high-resolution-layers/forests/forest-type-1/status-maps/forest-type-2018> ; [Accessed 07-Apr-2023].
6. Ilnitskiy A., Burba O. Statistical criteria for assessing the informativity of the sources of radio emission of telecommunication networks and systems in their recognition // Cybersecurity: Education, Science, Technique. — 2019. — Вип. 5. — DOI: [10.28925/2663-4023.2019.5.8394](https://doi.org/10.28925/2663-4023.2019.5.8394).
7. Bhattacharyya A. On a measure of divergence between two statistical populations defined by probability distributions // The Indian Journal of Statistics (1933-1960). — 1946. — Т. 7, вип. 4.
8. Huo L., Jan Persson H., Lindberg E. Early detection of forest stress from European spruce bark beetle attack, and a new vegetation index: Normalized distance red & SWIR (NDRS) // Remote Sensing of Environment. — 2021. — Т. 255. — ISSN 00344257. — DOI: [10.1016/j.rse.2020.112240](https://doi.org/10.1016/j.rse.2020.112240).
9. Belgiu M., Drăguț L. Random forest in remote sensing: A review of applications and future directions // ISPRS Journal of Photogrammetry and Remote Sensing. — 2016. — Apr. — Vol. 114. — P. 24–31. — DOI: [10.1016/j.isprsjprs.2016.01.011](https://doi.org/10.1016/j.isprsjprs.2016.01.011).
10. Tharwat A. Classification assessment methods // Applied Computing and Informatics. — 2018. — Т. 17, вип. 1. — ISSN 22108327. — DOI: [10.1016/j.aci.2018.08.003](https://doi.org/10.1016/j.aci.2018.08.003).