

ОЦІНКА СТАНУ ПІДТОПЛЕННЯ ЗА ДАНИМИ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ МЕТОДОМ БАГАТОВИМІРНОЇ СТАТИСТИЧНОЇ КЛАСИФІКАЦІЇ НА ОСНОВІ МОДЕЛІ КОМПОЗИЦІЙНИХ РОЗПОДІЛІВ

(Рекомендовано членом редакційної колегії д-ром геол.- мінералог. наук, проф. В.М. Загнітком)

У статті описано результати експерименту з інтерпретації даних супутникового зондування територій, що знають підтоплення. Запропоновано нову модель для опису статистичного розподілу даних у вигляді композиції локальних розподілів. Для оцінки належності об'єктів до класів використано Бейєсівський критерій. Проведено оцінку спроможності методу шляхом тестування 600 об'єктів. Аналіз помилок класифікації показав високий рівень надійності інтерпретації в умовах нечіткої диференціації розподілів і достатньо великої кількості класів. Метод має перспективу застосування в системах моніторингу стану територій як засіб оперативної оцінки стану підтоплення.

This article is considered with outcome of experiments of remote sensing data interpretation. It is suggested new "compositional model" for describing statistical distribution data. This method was trained for estimation its efficiency. Results show reliable level of interpretation data. Therefore this approach has perspective to use in underflooding monitoring.

Вступ. Зважаючи на зростаючі обсяги дистанційних досліджень Землі, задача підвищення якості інтерпретації результатів супутникової зйомки для оперативної оцінки стану територій стає все більш актуальною. У повній мірі це стосується стану підтоплення. Ведеться дискусія з приводу ефективності тих чи інших моделей та алгоритмів класифікації, переважна більшість яких носить виражений емпіричний характер [3, с. 20-269]. Автори виходять з того, що резерв підвищення якості інтерпретації слід шукати в двох основних напрямках: 1) удосконалення та уніфікація моделей статистичного розподілу показників, по яких здійснюється інтерпретація; 2) реалізація таких моделей у алгоритмах, які забезпечують мінімум помилок класифікації.

Автори дотримуються тієї точки зору, що останню умову може забезпечити лише алгоритми, які побудовані за Бейєсовським принципом [1, с. 391-395; 3, с. 20-45], оскільки саме цей принцип теоретично забезпечує оптимальність в розумінні мінімуму помилок класифікації. Проте потенціал методу обмежується рівнем адекватності моделі багатовимірного розподілу показників яскравості по окремих каналах. Застосування класичного гаусівського закону багатовимірного розподілу до класів об'єктів виглядає штучним і, як показали наші дослідження, пересічно неадекватним у практиці інтерпретації. Такого недоліку позбавлена модель композиції локальних багатовимірних розподілів [2, с. 179-242].

Постановка задачі. Космічний знімок являє собою матрицю, кожний елемент якої (піксел, у подальшому – об'єкт) характеризується вектором ознак $\mathbf{G}_{ij} = \{G_1, G_2, \dots, G_r\}_{ij}$.

Компонентами вектора $\mathbf{G}$ є значення інтенсивності відбитої або випроміненої енергії, зареєстрованої в r спектральних каналах. Кількість r компонент цього вектора визначається кількістю дискретних каналів, у яких проводиться зйомка; i, j – умовні координати пікселя на знімку.

Необхідно провести класифікацію космічного знімку на визначену еталонами кількість класів. Для порівняння проводилося тестування різних методів класифікації на еталонних ділянках.

Метод параметричної класифікації при гіпотезі незалежності ознак.

Позначимо:

$G_1, G_2, \dots, G_r$ – ознаки, обрані для розв'язання задачі класифікації на класи $Q_1, Q_2, \dots, Q_t$;

Y – об'єкт, що підлягає класифікації;

$Y = \{Y_1, Y_2, \dots, Y_r\}$; Y_k – стан ознаки G_k об'єкта Y ;

t – кількість класів;

r – кількість ознак, за якими розв'язується задача класифікації.

Результат класифікації визначається порівнянням значень двокомпонентної міри належності пред'явленого об'єкта Y до кожного з класів. Першою компонентою є кількість $L(j)$ тих ознак, по яких об'єкт класифікації (Y) знаходиться у стані, який у відповідному класі (Q_j) має нульову ймовірність, якщо ознака не є статистичною (тобто номінальною) або щільність розподілу дорівнює нулю (якщо ознака статистична). Отже, $L(j)$ є мірою чужорідності об'єкта по відношенню до класу Q_j . Другою компонентою служить бейєсівська ймовірність $P(Y \in Q_j)$ належності об'єкта Y до класу Q_j . В даному разі номінальні ознаки були відсутні. Ця ймовірність обчислюється за формулою:

$$P(Y \in Q_j) = P_j / \sum_{k=1}^t P_k, \quad (1)$$

де

$$P_j = \begin{cases} \prod_{k \in E(j)} P_{jk}(Y_k), & \text{якщо } L(j) = L_{\min}, \\ 0, & \text{якщо } L(j) > L_{\min}, \end{cases} \quad (2)$$

$E(j)$ – множина номерів тих ознак серед $G_1, G_2, \dots, G_r$, для яких $P_{jk}(Y_k)$ не дорівнює нулю (вираз $P_{jk}(Y_k)$ див. нижче);

$$L_{\min} = \min\{L(1), L(2), \dots, L(t)\}. \quad (3)$$

Критерій класифікації складається з таких кроків. Спочатку визначають $L_{\min}$ – найменшу з величин $L(1), L(2), \dots, L(t)$ та встановлюють ті номери $j_1, j_2, \dots, j_m$, при яких $L(j) = L_{\min}$. Для класів, номери яких не входять до цього списку, покладають $P_j = 0$, а решту значень P_j обчислюють за формулою (2). Клас, до якого слід віднести об'єкт класифікації, визначають за найбільшою з бейєсівських ймовірностей (1).

Якщо ознака G_k статистична, то її станом Y_k у об'єкта Y буде локальний розподіл, що описується трійкою значень $(m_{Yk}, \sigma_{Yk}, u_{Yk})$: m_{Yk} – математичне сподівання ознаки G_k об'єкта Y ; σ_{Yk} – її середній квадратичний відхил; u_{Yk} – укажчик закону розподілу (-1 – нормальний; -2 – логнормальний). Згідно з цим

$$P_{jk}(y_k) = \nu_k \cdot \int_R p_{yk}(x) p_{jk}(x) dx, \quad (4)$$

де $p_{jk}(x)$ – щільність розподілу ознаки у класі Q_j , яка розраховується за формулою

$$p_{jk}(x) = \begin{cases} \frac{1}{\sqrt{2\pi} \cdot \sigma_{jk}} \cdot \exp\left(-\frac{(x - m_{jk})^2}{2\sigma_{jk}^2}\right), & \text{якщо } u_{jk} = -1 \\ \frac{1}{\sqrt{2\pi} \cdot \delta_{jk} x} \cdot \exp\left(-\frac{(\ln x - \mu_{jk})^2}{2\delta_{jk}^2}\right), & \text{якщо } u_{jk} = -2 \end{cases} \quad (5)$$

$p_{yk}(x)$ – щільність розподілу тієї ж ознаки у об'єкта Y , для обчислення якої використовується формула (5) із заміною j на y .

Тут μ_{jk} та δ_{jk} – математичне сподівання та середній квадратичний відхил величини $\ln G$ для вказаного об'єкта, значення яких обчислюються за формулами:

$$\delta_{jk} = \sqrt{\ln\left(\left(\sigma_{jk}/m_{jk}\right)^2 + 1\right)}, \quad \mu_{jk} = \ln m_{jk} - 0.5\delta_{jk}^2; \quad (6)$$

ν_k – довжина області Ω_k на числовій вісі, утвореної об'єднанням інтервалів $[a(j, k), b(j, k)]$ ($j = 1, 2, \dots, N$):

$$\Omega_k = \bigcup_{j=1}^N [a(j, k), b(j, k)], \quad (7)$$

$a(j, k) = m_{jk} - K_d \sigma_{jk}$, $b(j, k) = m_{jk} + K_d \sigma_{jk}$ – границі ознаки G_k того ж об'єкта; за умовчанням коефіцієнт інтервалу щільності розподілу $K_d = 3$, що відповідає довірчій ймовірності інтервалу $q = 0.997$.

Метод непараметричної класифікації для залежних та незалежних ознак. Наведений нижче алгоритм обіцяє найвищу ефективність з трьох причин. Першою є оптимальність критерію Бейеса, який тут реалізовано. Друга – у вживанні толерантної композиційної моделі багатовимірного розподілу, яку обходять традиційні труднощі параметричної класифікації за статистично залежними компонентами. Третя – працездатність в умовах браку інформації, навіть тоді, коли для класичного бейєсівського критерію виникає ситуація невизначеності через потрапляння значень окремих компонент у множину нульової ймовірності для всіх класів.

Так само, як і в попередньому алгоритмі, результат класифікації визначається порівнянням значень двокомпонентної міри належності пред'явленого об'єкта Y до кожного з класів. Першою компонентою є кількість $L(j)$ тих ознак, по яких об'єкт класифікації (Y) знаходиться у стані, який у відповідному класі (Q_j) має нульову ймовірність. $L(j)$ служить мірою чужорідності об'єкту по відношенню до класу Q_j . Другою компонентою служить бейєсівська ймовірність $P(Y \in Q_j)$ належності об'єкта Y до класу Q_j . Ця ймовірність обчислюється за формулою:

$$P(Y \in Q_j) = P_j(y) / \sum_{i=1}^t P_i(y), \quad (8)$$

де

$$P_i(y) = \sum_{j=1}^{b(i)} P_j^{(i)} / b(i) - \quad (9)$$

умовна ймовірність g -вимірного стану y в класі Q_j ;

$b(i)$ – кількість об'єктів-еталонів у класі Q_j ;

$$P_j^{(i)} = \begin{cases} \prod_{k \in E(i, j)} P_{jk}^{(i)}, & \text{якщо } L^{(i)}(j) = L_{\min}^{(i)} \\ 0, & \text{якщо } L^{(i)}(j) > L_{\min}^{(i)} \end{cases}, \quad (10)$$

$L^{(i)}(j)$ – кількість тих ознак, для яких $P_{jk}^{(i)}(y_k)$ дорівнює нулю;

$$L_{\min}^{(i)} = \min\{L^{(i)}(1), L^{(i)}(j), \dots, L^{(i)}(b(i))\}, \quad (11)$$

де $E(i, j)$ – множина номерів тих ознак серед $G_1, G_2, \dots, G_r$, для яких $P_{jk}^{(i)}(y_k^{(i)})$ не дорівнює нулю (вираз $P_{jk}(y_k)$ для j -го об'єкта з i -го класу Q_i див. нижче);

$$L(i) = L_{\min}^{(i)} \quad (12)$$

$P_{jk}^{(i)}(y_k)$ – ймовірність стану y_k ознаки G_k у j -го еталонного об'єкта з класу Q_i . Якщо ознака нестатистична, і її стан у заданого об'єкта y_k , то ймовірність стану дорівнює 1. Якщо стан її інший, то ймовірність стану дорівнює 0. Вираз ймовірності стану для статистичної ознаки має вигляд:

$$P_{jk}^{(i)}(y_j) = \int_R p_{yk}(x) p_{jk}^{(i)}(x) dx, \quad (13)$$

де $p_{jk}^{(i)}(x)$ – щільність розподілу ознаки G_k у j -го еталонного об'єкта з класу Q_i ; $p_{yk}(x)$ – щільність розподілу ознаки G_k в стані y_k . Якщо $p_{jk}^{(i)}(x)$ невідома через відсутність інформації, то

$$P_{jk}^{(i)}(y_j) = 1/\nu_k. \quad (14)$$

Висновок про внутрішню належність до класу Q_i , якому відповідає найбільша ймовірність належності (10), робиться при $L(i) = 0$. В протилежному разі належність зовнішня.

Результати експериментальної оцінки ефективності методу. Для практичної оцінки був використаний знімок, зроблений 21 серпня 2000 р. спектрорадіометром "ETM+" із супутника "Landsat-7". У якості ознак було використано знімки, зроблені на різних спектральних каналах – пп10 (0.45-0.52 мкм), пп20 (0.52-0.60 мкм), пп30 (0.63-0.69 мкм), пп40 (0.76-0.9 мкм), пп50 (1.55-1.75 мкм), пп70 (2.08-2.35 мкм), пп61 (10.4-12.5 мкм), пп62 (10.4-12.5 мкм).

На космоснімку виділено еталонну область, в межах якої проведено тематичне районування на 8 класів: 1) "не задернована підтоплена поверхня", 2) "відкрита водна поверхня", 3) "сівозміна-1 не підтоплена", 4) "сівозміна-1 підтоплена", 5) "сівозміна-2 не підтоплена", 6) "сівозміна-2 підтоплена", 7) "урбанізована територія", 8) "не задернована суха поверхня". Термінами "сі-

возміна-1" і "сівозіміна-2" позначені два класи відбиваючих поверхонь, що фізично представлені сільськогосподарськими угіддями, засадженими зерновими та незерновими культурами відповідно. До експерименту залучено 600 об'єктів, по 75 з кожного класу.

За наведеними вище алгоритмами проводилася тестова класифікація. Результати параметричної класифікації подані у табл.1, непараметричної – у табл. 2.

Таблиця 1. Результати тестової параметричної класифікації

Клас, до якого віднесено об'єкт	Частота належності об'єктів до класів							
	Клас №1	Клас №2	Клас №3	Клас №4	Клас №5	Клас №6	Клас №7	Клас №8
1	0.800	0.000	0.000	0.000	0.000	0.000	0.027	0.000
2	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.067	0.073
4	0.080	0.000	0.000	1.000	0.000	0.000	0.307	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
6	0.080	0.000	0.000	0.000	0.000	1.000	0.107	0.000
7	0.040	0.000	0.000	0.000	0.000	0.000	0.480	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.927

Таблиця 2. Результати тестової непараметричної класифікації

Клас, до якого віднесено об'єкт	Частота належності об'єктів до класів							
	Клас №1	Клас №2	Клас №3	Клас №4	Клас №5	Клас №6	Клас №7	Клас №8
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.073
4	0.000	0.000	0.000	1.000	0.000	0.000	0.107	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	1.000	0.027	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.867	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.927

Результати тестової класифікації показали:

1. Помітну перевагу методу непараметричної класифікації, побудованому на композиційній моделі еталонних розподілів.

2. Високу достовірність інтерпретації по всіх класах за винятком 7-го класу – урбанізовані території. Зниження достовірності саме по цьому класу викликане тим, що він складно побудований і містить у собі різні види відбиваючих поверхонь (табл. 3).

Таблиця 3. Класифікація об'єктів класу "урбанізовані території" на інші еталонні класи

Параметрична класифікація		Непараметрична класифікація	
Клас, до якого віднесено об'єкт	Частота належності	Клас, до якого віднесено об'єкт	Частота належності
1	0.200	1	0.133
2	0.000	2	0.000
3	0.067	3	0.053
4	0.560	4	0.530
5	0.000	5	0.000
6	0.147	6	0.213
8	0.027	8	0.067

Висновок. Модель композиції багатовимірною розподілу показників яскравості по каналах супутникової зйомки забезпечує більш ніж задовільну якість інтерпретації при оцінці ступеня підтоплення територій в разі застосування цієї моделі в алгоритмі класифікації, побудованому за принципом Бейеса. За результатами проведеного експерименту зі знімком, зробленим спектро радіометром "ETM+" із супутника "Landsat-7", по 6 класах оці-

нка ймовірності помилки становила 0, по класу "урбанізовані території" – 0.13, по класу "незадернована суха поверхня" – 0.07. Середня ймовірність помилки – 0.025.

1. Жуков М. Н. Математична статистика та обробка геологічних даних: підручник. – К., 2008. 2. Жуков Н. Н. Вероятностно-статистические методы анализа геолого-геофизической информации. – К., 1975. 3. Р. Дуда, П. Харт. Распознавание образов и анализ сцен. – М., 1976.

Надійшла до редколегії 25.12.09