УДК 528.06

С.І. АЛЬПЕРТ*

ОЦІНКА ЯКОСТІ КЛАСИФІКАЦІЇ АЕРОКОСМІЧНИХ ЗОБРАЖЕНЬ НА ОСНОВІ МАТРИЦІ ПОМИЛОК ТА КОЕФІЦІЄНТІВ ТОЧНОСТІ

*Науковий Центр аерокосмічних досліджень Землі Інституту геологічних наук НАН України, Київ, Україна

Анотація. *У даній статті розглянуто такі коефіцієнти для оцінки точності тематичних карт, як повна точність класифікації, точність виробника, точність споживача та ризик Байєса. В роботі наведено деякі приклади. Враховуючи результати прикладів, встановлено певні закономірності та взаємозв'язки між точністю виробника та точністю споживача, встановлено взаємозв'язок між помилками omission та помилками commission для кожного класу. Розглянуто ризик Байєса, який використовується для оцінки неточності тематичних карт.*
Ключові слова: *матриця помилок, повна точність класифікації, точність виробника, точність споживача, помилки omission, помилки commission, ризик Байєса.*

Аннотация. *В данной статье рассмотрены такие коэффициенты для оценки точности тематических карт, как полная точность классификации, точность производителя, точность пользователя и риск Байеса. В работе приведены некоторые примеры. Учитывая результаты примеров, установлены определенные закономерности и взаимосвязи между точностью производителя и точностью пользователя, установлена взаимосвязь между ошибками omission и ошибками commission для каждого класса. В статье рассмотрен риск Байеса, который используется для оценки неточности тематических карт.*
Ключевые слова: *матрица ошибок, полная точность классификации, точность производителя, точность пользователя, ошибки omission, ошибки commission, риск Байеса.*

Abstract. *In this article we considered coefficients for accuracy assessment of thematic maps, such as overall accuracy, producer's accuracy, user's accuracy and Bayes risk. In the work we proposed some examples. Taking into account the results of these examples, we established some rules and relationships between producer's accuracy and user's accuracy. The relationship between omission errors and commission errors was established for each class as well. In this article we considered Bayes risk for inaccuracy assessment of thematic maps.*
Keywords: *error matrix, overall accuracy, producer's accuracy, user's accuracy, omission errors, commission errors, Bayes risk.*

1. Вступ

На даний час дистанційне зондування Землі (ДЗЗ) утворює інформаційну основу для дослідження, контролю, спостереження, оцінки та прогнозу змін природного середовища. Методи ДЗЗ з космосу використовуються для розв'язання сільськогосподарських, наукових проблем, для побудови топографічних та тематичних карт, екологічної оцінки територій, класифікації земель. В основі ДЗЗ лежить отримання зображення місцевості за допомогою сенсорів, далі їх обробка та класифікація. Вдале розв'язання задачі ДЗЗ залежить від якості зображень та точності класифікації. Існують різноманітні методи для оцінки точності класифікації аерокосмічних зображень.

У даній статті було розглянуто метод оцінки точності класифікації, який використовує матрицю помилок та коефіцієнти точності класифікації, що отримуються з цієї матриці. Основною метою статті є розгляд таких коефіцієнтів, як повна точність класифікації (overall accuracy), точність виробника (producer's accuracy), точність споживача (user's accuracy) та ризик Байєса. В роботі було проаналізовано два види помилок: помилки omission та помилки commission та встановлені певні взаємозв'язки між коефіцієнтами точності, які отримані з матриці помилок [1–2].

ISSN 1028-9763. Математичні машини і системи, 2014, № 1

Також у даній статті було розглянуто ризик Байєса, що використовується як оцінка неточності класифікації, і який, на відміну від інших коефіцієнтів, застосовується для розпізнавання типів неправильних класифікацій.

2. Матриця помилок і коефіцієнти точності: вихідні положення

Матриця помилок (рис. 1) заповнюється статистичними результатами проведеної класифікації n об'єктів при наявності K класів. Кожен рядок нумерується індексом i, а кожен стовпчик нумерується індексом j. При цьому $i, j = 1, 2, ..., K$. Елемент матриці помилок n_{ij} відображає число об'єктів, що помилково були віднесені при класифікації до класу i, хоча в дійсності вони належать класу j.

Матриця помилок вказує на те, як співвідносяться значення збіжних класів, отримані з різних джерел. Як джерела можуть бути дані, які підлягають перевірці, та опорні (еталонні) дані, отримані з більш надійного джерела даних. Також під час інтерпретації результатів припускається, що результат, який перевіряється, є неточним, а перевірочні дані відображають реальну ситуацію. У випадку, коли перевірочні дані також є неточними, ми вже не можемо говорити про "помилку", а слід говорити про "різницю" між двома наборами даних.

Матриця помилок містить назви класів легенди класифікації даних, які підлягають перевірці, та класи легенди даних, що використовуються для перевірки [3].

Головна діагональ матриці вказує на ті випадки, де отримані розрахункові класи та реальні дані співпадають (правильна класифікація). Сума значень діагональних елементів вказує на загальну кількість правильно класифікованих пікселів. Відношення цієї кількості правильно класифікованих пікселів до загальної кількості пікселів у матриці називається загальною точністю класифікації (overall accuracy) і виражається у відсотках.

Для визначення точності певного розрахункового класу необхідно розділити кількість правильно класифікованих пікселів цього класу на загальну кількість пікселів у ньому згідно з перевіреними даними. Даний показник називається точністю виробника (producer's accuracy). Точність виробника показує, наскільки добре результат класифікації для цього класу співпадає з даними, що були перевірені [4].

Також ми можемо обчислити аналогічний показник для реального класу (завіркових даних), якщо розділити кількість правильно класифікованих пікселів класу на загальну кількість пікселів у ньому згідно з даними, що підлягають перевірці. Цей показник називається точністю користувача, оскільки він показує, наскільки ймовірно, що даний клас збігається з результатами класифікації. Недіагональні елементи вказують на випадки розбіжності між розрахунковими та реальними класами (помилки класифікації) [5–7].

Структура матриці показана на рис. 1.

$$n_{i+} = \sum_{j=1}^{K} n_{ij}; \quad n_{+j} = \sum_{i=1}^{K} n_{ij} \cdot \tag{1}$$

Загальна точність класифікації, точність виробника та точність користувача обчислюються за такими формулами [8–9]:

– загальна точність класифікації (Overall Accuracy):

$$OA = \frac{\sum_{i=1}^{K} n_{ii}}{n}; \tag{2}$$

– точність виробника (Producer's Accuracy):

$$PA = \frac{n_{jj}}{n_{+j}}; \tag{3}$$

– точність користувача (User's Accuracy):

$$UA = \frac{n_{ii}}{n_{i+}}. \tag{4}$$

		Завіркові дані					
		Клас 1	*Клас 2*	*Клас j*	*Клас K*	*Сума елементів рядка* n_{i+}	*Точність користувача (UA)*
Класифіковані дані	*Клас 1*	n_{11}	n_{12}	n_{1j}	n_{1K}	n_{1+}	UA_1
	Клас 2	n_{21}	n_{22}	n_{2j}	n_{2K}	n_{2+}	UA_2
	Клас i	n_{i1}	n_{i2}	n_{ij}	n_{iK}	n_{i+}	UA_i
	Клас K	n_{K1}	n_{K2}	n_{Kj}	n_{KK}	n_{K+}	UA_K
	Сума елементів стовпчика n_{+j}	n_{+1}	n_{+2}	n_{+j}	n_{+K}	n	
	Точність виробника (PA)	PA_1	PA_2	PA_j	PA_K		

Рис. 1. Структура матриці помилок (Error Matrix)

Приклад 1. Наведемо приклад класифікації 700 ділянок, якщо ми маємо дві категорії: "Поле" та "Ліс". Припустимо, що у нас є дві класифікації даних ділянок. Одна з цих класифікацій створена на базі даних AVHRR, а друга-MODIS. Результат накладання двох класифікацій буде:

1) два джерела визначили територію як ліс;
2) AVHRR визначив територію як поле, MODIS-як ліс;
3) MODIS визначив територію як поле, AVHRR-як ліс;
4) два джерела визначили територію як поле.

Таблиця 1. Результат класифікації 700 ділянок за двома категоріями

	MODIS		
	Поле	Ліс	Σ
AVHRR: Поле	121	87	208
AVHRR: Ліс	17	475	492
AVHRR: Σ	138	562	700

$\text{Overall Accuracy} = \frac{121+475}{700} = 0,85$ – загальна точність класифікації.

У даному прикладі загальна точність становить 85%. Ми отримали таке високе значення загальної точності завдяки територіям, які були класифіковані як ліси обома джерелами.

Тепер підрахуємо точності виробника та точності користувача і встановимо їх взаємозв'язок з помилками omission та помилками commission. Фізичний зміст помилок omission полягає в невірній класифікації, тобто, коли пікселі, які в дійсності мають належати до певного конкретного класу, не були віднесені до цього класу. Фізичний зміст помилок commission полягає в невірній класифікації, коли піксель з одного класу був помилково віднесений до іншого класу, хоча в дійсності він до цього класу не належить.

$\text{Producer's Accuracy} = \frac{121}{138} = 0,88$ – точність виробника для класу полів.

Точність виробника для класу полів становить 88%. Висока точність виробника означає, що під час даної класифікації ми отримали мало помилок omission (omission errors), тобто було пропущено мало пікселів, що відносяться до класу полів. Було встановлено, що невелика кількість пікселів, які насправді відносяться до полів, були помилково віднесені до лісів.

$\text{User's Accuracy} = \frac{121}{208} = 0,59$ – точність користувача для класу полів.

Підраховано, що точність користувача для класу полів складає 59%. Зауважимо, що низька точність користувача означає, що при проведенні даної класифікації ми маємо багато помилок commission (commission errors), тобто маємо багато пікселів, які насправді відносяться до лісу, але були помилково віднесені до полів.

Розрахуємо точність виробника та точність користувача для класу лісів:

$\text{Producer's Accuracy} = \frac{475}{562} = 0,85$ – точність виробника для класу лісів;

$\text{User's Accuracy} = \frac{475}{492} = 0,97$ – точність користувача для класу лісів.

Тепер проаналізуємо отримані результати для класу поля. Для цього класу точність виробника є набагато кращою за точність користувача. Тобто "краще, щоб усі ділянки, які насправді належать до класу поля, були класифіковані як ті, що належать до класу поля", а не "краще, щоб ділянок, що належать до класу поля, було менше, але усі вони точно належали до класу поля".

З даного прикладу випливає, що помилки commission та помилки omission для одного класу є протилежними. Високе значення однієї з них часто пов'язане з низьким значенням іншої. Трактування та інтерпретація якості класифікації залежать від конкретних задач. Також слід зауважити, що однією з найпоширеніших задач є знаходження максимального значення обох типів помилок [10–11].

3. Ризик Байєса та його основні властивості

Розглянемо ризик Байєса, який може бути використаний як оцінка неточності класифікації.

$$R = \sum_{i=1}^{r} \pi_i \sum_{j=1}^{r} \gamma_{ij} p_{j|i}, \qquad (5)$$

де γ_{ij} – величина, яка надається у тому випадку, коли ділянка з категорії C_i віднесена до C_j $(i,\ j = 1, \ldots, r)$;

π_i – апріорні ймовірності категорій C_i;

$p_{j|i}$ – ймовірність того, що ділянка з категорії C_i віднесена до C_j.

Зауважимо, при вірному розташуванні ділянки величина $\gamma_{ij} = 0$. Квадратна матриця $\Gamma = (\gamma_{ij})$ порядку r називається ціновою матрицею. Оцінка ризику R має вигляд

$$R(X \mid \Gamma) = \sum_{i=1}^{r} \frac{\pi_i}{n_i} \sum_{j=1}^{r} \gamma_{ij} x_{ij}. \tag{6}$$

Розглянемо два випадки: $\pi_i = \frac{1}{r}$ (однакові апріорні ймовірності) та $\pi_i = \frac{n_i}{N}$ (пропорційні апріорні ймовірності). Підставляючи ці значення у вираз (6), маємо

$$\begin{aligned} R_{uni}(X \mid \Gamma) &= \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n_i} \sum_{j=1}^{r} \gamma_{ij} x_{ij}, \\ R_{pro}(X \mid \Gamma) &= \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{r} \gamma_{ij} x_{ij}. \end{aligned} \tag{7}$$

$R(X \mid \Gamma)$ задовольняє таким властивостям:

1) $R(X \mid \Gamma)$ монотонно спадає, коли хоча б один діагональний елемент x_{ii} зростає у незначному діапазоні;

2) $\begin{matrix} R(X \mid \Gamma) \geq 0, \\ R(X \mid \Gamma) = 0 \end{matrix}$ виконується тоді, коли всі ділянки правильно класифіковані (у випадку, якщо $\gamma_{ij} \geq 0$ для $\forall i \neq j$, вираз “тоді” замінюється на “тоді і тільки тоді”);

3) $\begin{matrix} R(X \mid \Gamma) \leq \sum_{i=1}^{r} \pi_i \max\{\gamma_{i1}, \gamma_{i2}, ..., \gamma_{ir}\}, \\ R(X \mid \Gamma) = \sum_{i=1}^{r} \pi_i \max\{\gamma_{i1}, \gamma_{i2}, ..., \gamma_{ir}\} \Leftrightarrow \end{matrix}$ коли всі ділянки невірно віднесені до тих категорій, які при невірній класифікації дають найбільш грубі помилки;

4) $R_{uni}(X \mid \Gamma) = R_{uni}(X^t \mid \Gamma^t)$.

Приклад 2. Надалі, для зручності, будемо використовувати матрицю помилок у такому вигляді:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{r1} & \cdots & x_{rr} \end{pmatrix}.$$

Позначимо через показник x_{ij} кількість об’єктів, які були класифіковані на зображенні як такі, що належать до категорії C_j, хоча в дійсності вони належать до категорії C_i, $i, j = 1, 2, ..., r$, r – загальна кількість класів (категорій).

Наведемо 2 цінові матриці, за допомогою яких оцінюється неточність матриці помилок:

$$X_1 = \begin{pmatrix} 90 & 10 \\ 200 & 1800 \end{pmatrix}: \quad \Gamma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}.$$

$$R_{uni}(X_1 \mid \Gamma_1) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n_i}\sum_{j=1}^{r}\gamma_{ij}x_{ij} =$$

$$= \frac{1}{2}\left[\frac{1}{100}(0\cdot 90 + 1\cdot 10) + \frac{1}{2000}(200\cdot 1 + 1800\cdot 0)\right] = 0,1;$$

$$R_{uni}(X_1 \mid \Gamma_2) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n_i}\sum_{j=1}^{r}\gamma_{ij}x_{ij} =$$

$$= \frac{1}{2}\left[\frac{1}{100}(0\cdot 90 + 10\cdot 10) + \frac{1}{2000}(200\cdot 1 + 1800\cdot 0)\right] = 0,55;$$

$$R_{pro}(X_1 \mid \Gamma_1) = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{r}\gamma_{ij}x_{ij} =$$

$$= \frac{1}{2100}\left[90\cdot 0 + 10\cdot 1 + 1\cdot 200 + 0\cdot 1800\right] = 0,1;$$

$$R_{pro}(X_1 \mid \Gamma_2) = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{r}\gamma_{ij}x_{ij} =$$

$$= \frac{1}{2100}\left[90\cdot 0 + 10\cdot 10 + 1\cdot 200 + 0\cdot 1800\right] \approx 0,1429.$$

4. Висновки

У даній статті ми розглянули метод оцінки точності класифікації, який базується на матриці помилок. Були розглянуті коефіцієнти точності класифікації, які використовуються для обробки аерокосмічних зображень. Було встановлено зв'язок між точністю виробника та помилками omission (omission errors) і зв'язок між точністю користувача та помилками commission (commission errors). У роботі розглянуто приклад класифікації ділянок за двома категоріями: "Поле" та "Ліс", встановлено взаємозв'язок між помилками omission та commission для певного класу. Враховуючи результати прикладів, встановлено, що високе значення помилки commission пов'язане з низьким значенням помилки omission, і навпаки.

У роботі ми розглянули ризик Байєса та його основні властивості. Були наведені формули ризику Байєса при апріорних та апостеріорних імовірностях і проілюстровано їх застосування на прикладі класифікації об'єктів.

Також у статті були окреслені перспективні напрямки у сфері обробки зображень, отримані за допомогою дистанційного зондування Землі, а саме переваги застосування матриці помилок, наведених коефіцієнтів точності та ризику Байєса. В залежності від значення ризику Байєса ми в подальшому можемо проводити оцінку точності класифікації об'єктів і зробити висновок, наскільки точно була проведена класифікація.

В перспективі наведені коефіцієнти точності класифікації, ризик Байєса та його модифікації можуть бути застосовані при розв'язанні задач пошуку нафти та газу, а також при проведенні класифікації сільськогосподарських та урбанізованих територій [12–14].

СПИСОК ЛІТЕРАТУРИ

1. Congalton R.G. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices / R.G. Congalton, K. Green. – CRC Press: Taylor & Francis Group, 1999. – P. 130 – 137.

2. Janssen L.L.F. Accuracy assessment of satellite derived land-cover data: A review / L.L.F. Janssen, F.J.M. van derWel // Photogramm. Eng. Remote Sensing. – 1994. – Vol. 60. – P. 419 – 426.
3. Альперт С. Оцінка точності класифікації космічних зображень на основі теорії Демпстера-Шафера / С. Альперт // Зб. праць XI Міжнар. молодіжної наук.-практ. конф. "Історія розвитку науки, техніки та освіти" за темою "Розбудова дослідницького університету". – Київ, 2013. – С. 242 – 245.
4. Cochran W.G. Sampling Techniques / Cochran W.G. – New York: John Wiley and Sons, 1977. – P. 421 – 428.
5. Story M. Accuracy assessment: A user's perspective/ M. Story, R.G. Congalton // Photogramm. Eng. Remote Sensing. – 1986. – Vol. 52. – P. 397 – 399.
6. Hardin P.J. Statistical significance and normalized confusion matrices/ P.J. Hardin, J.M. Shumway // Photogramm. Eng. Remote Sensing. – 1997. – Vol. 63. – P. 735 – 740.
7. Hegarat-Mascle S. Application of Dempster-Shafer Evidence Theory to Unsupervised Classification in Multisource Remote Sensing / S. Hegarat-Mascle, I. Bloch, D. Vidal-Madjar // IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. – 1997. – Vol. 35, N 4. – P. 1018 – 1031.
8. Rosenfield G.H. A coefficient of agreement as a measure of thematic classification accuracy / G.H. Rosenfield, K. Fitzpartrick-Lins // Photogramm. Eng. Remote Sensing. – 1986. – Vol. 52. – P. 223 – 227.
9. McCoy R.M. Fields Methods in Remote Sensing / McCoy R.M. – New York: Guilford Press, 2005. – P. 150 – 160.
10. Abidi M.A. Data Fusion in Robotics and Machine Intelligence / M.A. Abidi, R.C. Gonzalez. – New York: Academic, 1992. – P. 562 – 569.
11. Brownlee K.A. Statistical theory and methodology in science and engineering / Brownlee K.A. – New York: John Wiley and Sons, 1965. – P. 580 – 590.
12. Shafer G. A Mathematical Theory of Evidence / Shafer G. – Princeton, NY: Princeton University Press, 1976. – P. 875 – 883.
13. Попов М. Методология оценки точности классификации объектов на космических изображениях / М. Попов // Проблемы управления и информатики. – 2007. – № 1. – С. 97 – 103.
14. Альперт С. Сучасні критерії оцінки точності класифікації аерокосмічних зображень / С. Альперт // Математичні машини і системи. – 2013. – № 4. – С. 187 – 197.

Стаття надійшла до редакції 22.08.2013