

УДК 528.88.042.4

Морщавка С. В.¹, Шама Є. О.², Піза Д. М.³

¹Канд. техн. наук, доцент Запорізького національного технічного університету

²Аспірант Запорізького національного технічного університету

³Д-р техн. наук, професор Запорізького національного технічного університету

ВПЛИВ ВИДІВ НОРМУВАННЯ НА ЯКІСТЬ РОЗПІЗНАННЯ РОСЛИННИХ ОБ'ЄКТІВ

Розглянуто вплив різних видів нормування на якість розпізнання за результатами дистанційного зондування.

Ключові слова: розпізнавання, хеометрика, нормування, спектральні коефіцієнти яскравості.

ВСТУП

При розпізнанні рослинних об'єктів, необхідно проводити аналіз багатоспектральних потоків даних сукупності спектральних коефіцієнтів яскравості, які описуються деяким масивом чисел [1]. На практиці під час експериментів зручно складати матрицю, в якій стовпці відповідають довжинам хвиль, а рядки вибіркам – значенням спектральних коефіцієнтів яскравості (СКЯ), що отримані з виходу аналогово-цифрового перетворювача (АЦП) спектрометра [2]. Такий підхід є загальновизнаним в хеометриці. Зазвичай значення x в межах однієї строки змінюються від деякого мінімального значення $x_{\min}$ до деякого максимального значення $x_{\max}$. При переході від однієї вибірки до іншої значення СКЯ можуть значно відрізнятися. Їх абсолютні значення не повинні

впливати на подальшу статистичну обробку цих даних. Оскільки інформація про СКЯ міститься не в абсолютних значеннях відбивних характеристик на певних довжинах хвиль, а в його зміні щодо довжини хвилі, то можна провести нормування одержаних в результаті експериментів характеристик, які в подальшому будуть використовуватись для розпізнання об'єктів [3]. Слід зазначити, що значна кількість методів розпізнавання, в особливості параметричних, є самонормуючими, тобто нечутливими до масштабу, оскільки для обчислень використовується обернена ковариційна матриця та виконується центрування. Однак, і у цьому випадку, при практичній реалізації, нормування також може позитивно вплинути на якість розпізнавання за рахунок усунення помилок при обчисленнях у процесорах з фіксованою розрядністю.

Приведення до єдиного масштабу забезпечується нормуванням кожної змінної на діапазон розкиду її значень. Для цього функція перетворення $y=f(x)$ повинна мати наступні властивості:

$$y(x_{\min})=K_{\min}; y(x_{\max})=K_{\max}; dy/dx>0. \quad (1)$$

Будь-яка функція з такими властивостями може бути використана для нормування. Наприклад, якщо $x_{\max} \rightarrow \infty$, то можна обрати функцію:

$$y(x) = 1 - \exp\left(1 - \frac{x}{x_{\min}}\right). \quad (2)$$

Тобто, за рахунок вибору відповідної функції можна врахувати різноманітні ефекти спотворення оцінок. Наприклад, схильність набору даних до крайніх оцінок. При цьому, можливо, слід застосовувати для різних наборів даних і різні функції перетворення, що враховують особливості їх поведінки, статусу і тому подібне. Зразкові графіки таких функцій наведені на рис. 1.

Приведення до єдиного масштабу забезпечується нормуванням кожної змінної на діапазон розкиду її значень. У простому варіанті це – лінійне перетворення:

$$\tilde{x}_i = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}}, \quad (3)$$

де $\tilde{x}_i$, x_i – нормований та виміряний спектральні коефіцієнти яскравості на довжині хвилі λ_i , відповідно; $x_{i,\max}$, $x_{i,\min}$ – максимальний та мінімальний виміряні спектральні коефіцієнти яскравості, відповідно.

Слід зауважити, що лінійне нормування оптимальне, коли значення змінної щільно заповнюють певний інтервал. Але подібний «прямолінійний» підхід можна застосовувати далеко не завжди. Так, якщо при розпізнанні рослинних об'єктів в даних є відносно рідкісні викиди, що перевищують типовий розкид, то саме ці викиди визначають, згідно з попередньою формулою (3), масштаб нормування. Це приведе до того, що основна маса значень нормованої змінної зосередиться поблизу нуля

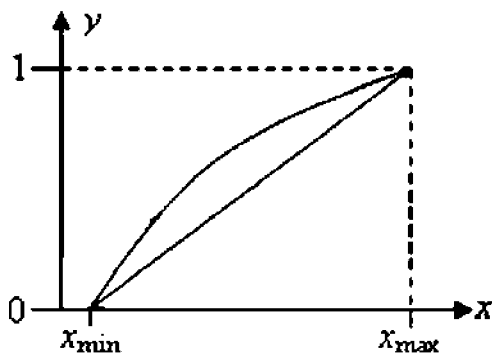


Рис. 1. Графіки функцій нормування

$|\tilde{x}_i| \ll 1$, або, навпаки, поблизу одиниці (крива «ранжована» на рис. 2 має мало значень в середині діапазону). На рис. 2 під назвою «вибірка» наведена для прикладу одна вибірка реальних даних – СКЯ амброзії у логарифмічному масштабі, у відносних одиницях виходу 12 розрядного АЦП спектрометра, в залежності від умовної довжини хвилі i , що відповідає λ_i . Для рангової статистики цієї вибірки (крива «ранжована») на рис. 2 позначені пунктиром межі $P_0, P_5, P_{10}, P_{20}, P_{30}, P_{50}, P_{70}, P_{80}, P_{90}, P_{100}$, що є процентилями відповідного рівня.

Ще одним фактором є обмеження, що виникають при цифруванні СКЯ за рахунок обмеження розрядної сітки АЦП, що ілюструє крива «вибірка» на рис. 2 в діапазоні від 7 до 16 умовної довжини хвилі.

Для усунення впливу обмежень можна використувати не діапазон від мінімального значення до максимального, а від $N-i$ до $1-N$ процентилі. При цьому N може мати значення від 0 до 0,5 (0,5 відповідає 50-й процентилю, або медіані). Узагальнення формули (3) матиме вигляд:

$$\tilde{x}_i = (1 - 2 \cdot N) \frac{x_i - x_{P_N}}{x_{P_{(1-N)}} - x_{P_N}} + N, \quad (4)$$

де x_{P_N} – значення N -процентилю.

Іноді при нормуванні набагато надійніше орієнтуватися не на екстремальні значення, а на статистичні характеристики даних, такі як середнє значення $\bar{x}_i$ і дисперсія σ_i :

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\sigma_i}. \quad (5)$$

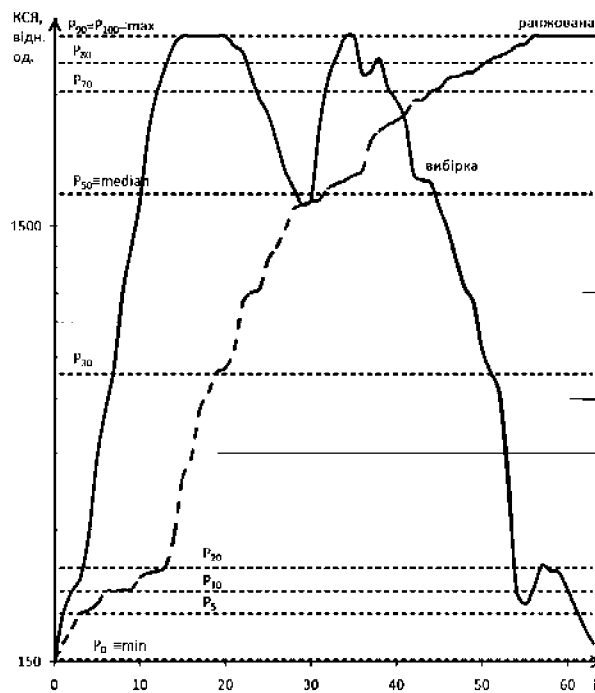


Рис. 2. Використання рангових статистик для нормування

В цьому випадку основна маса даних матиме одиничний масштаб, тобто типові значення усіх змінних будуть порівняними. Перетворення (5) в хемометриці має назву автошкалювання.

Іноді для обмеження діапазону можливих значень вхідних даних використовуються нелінійні гладкі функції, наприклад, – логістичну

$$f(\tilde{x}_i) = (1 + e^{-\tilde{x}_i})^{-1}. \quad (6)$$

Однак до «такого кроку слід вдаватися з обережністю оскільки таке перетворення змінює структуру даних. А у випадку розпізнавання за СКЯ у цьому не має потреби, оскільки максимальні значення вхідних даних і так обмежені розрядністю АЦП спектрометра.

Для порівняння впливу видів нормування на якість розпізнавання рослинних об'єктів, порівнюємо ті види нормування, де масив даних нормується за допомогою лінійного перетворення (4) і за допомогою автошкалювання (5).

УМОВИ І ПРОВЕДЕННЯ ЕКСПЕРИМЕНТУ

Експериментальні дослідження проводилися в два етапи:

- польові дослідження спектрів відбиття від рослин;
- обробка результатів досліджень спектрів.

Польові дослідження проводилися на тестовому полігоні Інституту олійних культур Національної академії аграрних наук, який розташований біля м. Запоріжжя. З отриманих експериментальних даних було відібрано 1000 наборів з 2048 спектральними точками. Критерієм відбору була можливість візуального визначення виду рослин по зображеннях і формі спектральної кривої (відсутність обмежень і надмірної зашумленості). Кількість вибірок різних видів рослин, що були відібрані для подальшого дослідження, та їх назви наведені в табл. 1.

За допомогою пристрою для дослідження спектрів рослин вимірювалися відносні спектральні характеристики в автоматичному режимі з періодом 1 с зі збереженням як поточної спектрограми, так і додаткової інформації (місце розташування і зовнішній вигляд рослини). Характеристики рослин в рядку знімалися послідовно, об'єкти приймача (спектрометричного блоку) переміщався уздовж рядка плавно, без ривків і зупи-

нок. Пристрій для дослідження спектрів складався з: спектрометра Red Tide 650 з волоконно-оптичним кабелем P200-2-UV-VIS і лінзою 74-VIS (фірма Ocean Optics); програмні компоненти, що входять в базовий комплект програмного забезпечення спектрометра: SpectraSuite і OOI_base32; Web-камера Genius VideoCamEye для реєстрації зображення рослин; датчик місцезнаходження BT-Q920 фірми QSatarz.

ОБРОБКА РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТУ

Після усереднення 2048 СКЯ, що забезпечуються спектрометром за один замір, до 64, були здобуті 3 масиви чисел: 587×64 – матриця коефіцієнтів відбиття для соняшника, 189×64 – матриця коефіцієнтів відбиття для мишію і 224×64 – матриця коефіцієнтів відбиття для амброзії. Наступним кроком було нормування – кожна з трьох матриць піддавалася нормуванню по лінійному закону нормування згідно формули (4) з різними значеннями N і автошкалювання згідно (5).

Для перевірки якості розпізнавання рослинних об'єктів за експериментальними даними був використаний блок дискримінаційного аналізу General Discriminant Analysis (GDA) програми Statistica 8.0 компанії StatSoft Inc. Він дозволяє ефективно здійснювати статистичний аналіз і обробку експериментальних даних, зокрема визначити статистичні характеристики вибірових сукупностей, підгонку розподілів, здійснювати варіаційний, регресійний і дискримінаційний аналіз, будувати гістограми і графіки різної складності. Проведення дискримінаційного аналізу потрібне для того, щоб визначити можливість класифікації рослинних об'єктів за спектральними характеристиками, а також оцінити якість розпізнавання [4]. На даному етапі, для цього аналізу, вибір кількості ознак не брався до уваги, тому, що головною метою було порівняти вплив різних видів нормування на якість розпізнавання при однакових умовах, а не як збільшити якість розпізнавання за рахунок відбору найбільш інформативних ознак.

Були отримані наступні результати. Табл. 2 демонструє показники розпізнавання без проведення нормування взагалі. Табл. 3 дозволяє виявити залежність якості розпізнавання від номеру використаної для нормування процентилю.

Залежність вірогідності вірної класифікації від рівнів (процентилів), що використовуються для нормування, проілюстровано на рис. 3. Для порівняння на рис. 3 наведені вірогідності для випадків використання автошкалювання та розпізнавання без попереднього нормування у вигляді горизонтальних ліній. Згідно з цим рисунком нормування може як покращити якість подальшого розпізнавання, так і погіршити її, але в незначних межах.

Однією з умов застосування методів дискримінаційного аналізу є нормальність розподілу вибірок. Методика перевірки на нормальність розподілу описана в міжнародному стандарті ISO 5479-97 [5]. Дозволяється деяке відхилення від нормального розподілу, при якому досить успішно класифікуються вибірки. В цілому вибірки відповідали нормальному розподілу.

Таблиця 1. Видовий склад досліджуваних рослин

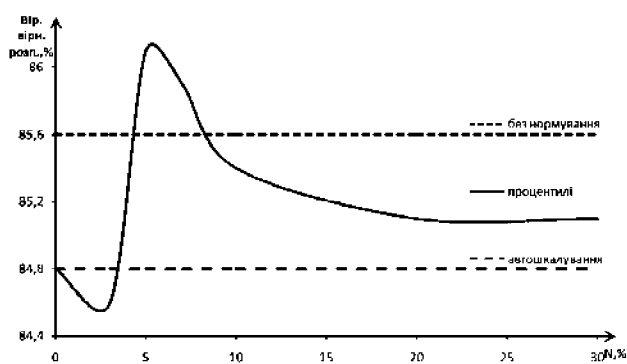
Вид рослини	Латинська назва	Клас	Кількість вибірок
соняшник	<i>Helianthus annuus</i>	1	587
амброзія	<i>Ambrósia artemisiifolia</i>	2	224
мишій	<i>Setaria viridis</i>	3	189
всього			1000

Таблиця 2. Класифікаційна матриця без нормування

		Результати класифікації для кожного класу			
		1(соняшник)	2(амброзія)	3(мишій)	Всього
Апріорні дані про клас	1(соняшник)	557	26	4	587
	2(амброзія)	53	155	16	224
	3(мишій)	36	9	144	189
	Всього	646	190	164	1000
Вірогідність вірної класифікації		94,89 %	69,20 %	76,19 %	85,6 %

Таблиця 3. Класифікаційна матриця при використанні різних видів лінійного нормування

Рівні для нормування	N	0(min)	0,03	0,05	0,07	0,1	0,2	0,3	авто- шкал.	без норм.
	1-N	1(max)	0,97	0,95	0,93	0,9	0,8	0,7		
Кількість вірно класифікованих вибірок різних класів	2(сон.)	156	156	156	156	154	152	155	154	155
	3 кл.	136	135	147	143	142	143	142	135	144
	1 кл.	556	555	558	560	558	556	554	559	557
Вірогідність вірної класифікації для всіх класів, %		84,8	84,6	86,1	85,9	85,4	85,1	85,1	84,8	85,6

**Рис. 3.** Залежність вірогідності вірного розпізнавання від рівнів для нормування

ВИСНОВКИ

Перевірка впливу різних видів нормування на якість розпізнавання рослинних об'єктів за результатами дистанційного зондування виявила, що види нормування не значно, але впливають на кінцеву якість розпізнавання. Обробка експериментальних результатів за допомогою дискримінаційного аналізу показала, що для розпізнавання рослин найкраще використовувати лінійне нормування зі значеннями 5 та 95 процентиля замість мінімальних та максимальних, відповідно. При цьому виграш відносно розпізнавання без нормування склав 0,5 %, а відносно автошкалювання 1,3 %. Виграш – незначний, але при вірогідностях вірного розпізнавання близько 90 % кожен додатковий відсоток має значення та досягається значним ускладненням процедури розпізнавання. Крім того, як вже зауважувалося вище, використання розпізнавання без нормування не завжди можливе на практиці. Отримані результати слід віднести до випадків, практичне значення яких не слід узагальнювати без перевірки. На користь цього говорить те, що найбільший вплив нормування створює на розпізнавання лише одного класу – третього, що відповідає мишій – бур'яну з невеликою площею листя, і тому меншим інтегральним коефіцієнтом відбиття. Але результати цієї статті показують, що при вирішенні задач розпізнавання образів

при необхідності нормування початкових даних, слід враховувати їх вплив та обирати їх оптимальні види для досягнення якісного розпізнавання.

СПИСОК ЛІТЕРАТУРИ

1. Пиза, Д. М. Выбор эффективного метода распознавания растений по коэффициентам спектральной яркости / Д. М. Пиза, С. В. Морщавка, Ю. В. Скоробогатов // Радіоелектроніка, інформатика, управління. – 1999. – № 1. – С. 85–88.
2. Дистанционное зондирование: количественный подход / [Ш. М. Дейвис, Д. А. Ландгребе, Т. Л. Филлипс и др.]; под ред. Ф. Свейна и Ш. Дейвис. Пер. с англ. М.: Недра. – 1983. – С. 415. (Пер. изд. США, 1978, 396 с.)
3. Морщавка, С. В. Информационные технологии обработки пропашных культур. Радиометрические аспекты / С. В. Морщавка, Д. М. Пиза, Е. А. Белоусов // Радіоелектроніка, інформатика, управління. – 2008. – № 2. – С. 45–52.
4. Боровиков, В. П. Популярное введение в программу STATISTICA/ В. П. Боровиков. – М.: Компьютер Пресс, 1998. – 238 с.
5. Статистическое представление данных. Проверка отклонения распределения вероятностей от нормального распределения (ISO 5479-97): ГОСТ Р ИСО 5479-2002. – [Действителен с 2002-01-22] – М.: Госстандарт России, 2002. – 27 с. – (Технический комитет по стандартизации ТК125)

Стаття надійшла до редакції 11.05.2011.

Морщавка С. В., Шама Е. А., Пиза Д. М.

ВЛИЯНИЕ ВИДОВ НОРМИРОВАНИЯ НА КАЧЕСТВО РАСПОЗНАВАНИЯ РАСТИТЕЛЬНЫХ ОБЪЕКТОВ

Рассмотрено влияние разных видов нормирования выборок на качество распознавания по результатам дистанционного зондирования.

Ключевые слова: распознавание, хеометрика, нормирование, спектральные коэффициенты яркости.

Morshchavka S. V., Shama E. O., Piza D. M.

EFFECT OF NORMALIZATION ON THE QUALITY OF RECOGNITION OF PLANT OBJECTS

Influence of different types of samples norm is considered on quality of recognition on results of the remote sensing.

Key words: recognition, xemometrics, normalization, the spectral brightness coefficients.