УДК 631.3:528.8:681.518

КЛАСТЕРНА МОДЕЛЬ ОЦІНЮВАННЯ ДАНИХ МОНІТОРИНГУ ВАРІАБЕЛЬНОСТІ ПАРАМЕТРІВ СТАНУ СІЛЬСЬКОГОСПОДАРСЬКИХ УГІДЬ

О. Броварець, канд. техн. наук,
Національний університет біоресурсів і природокористування України

У статті наведено методику оцінювання даних моніторингу параметрів стану сільськогосподарських угідь, отриманих від системи моніторингу.

Ключові слова: *моніторинг, точне землеробство, кластерна модель.*

Актуальність проблеми. Одним з головних питань технології точного землеробства є оцінювання різних алгоритмів для зображення зон управління. Відчутною стала потреба у більш сучасних моделях для оцінювання даних моніторингу варіабельності параметрів стану ґрунтового покриву, оскільки інформація – ключовий елемент процесу прийняття рішення, а кількість різнохарактерної та різного ступеня складної інформації, яку продукують нові технології моніторингу, постійно зростає. Одним з важливих завдань, що виникає у зв'язку із створенням сучасної інформаційної системи, є автоматизація процесу оцінювання даних моніторингу варіабельності параметрів стану сільськогосподарських угідь (образів).

Навіть якщо є загальноприйнятий метод для окреслення зон, то необхідно розробляти методи для їх окреслення. Причому потрібен не статичний фактаж, в якій пропорції знаходяться між собою згадані параметри, а саме динамічні моделі взаємозв'язку в межах конкретного поля, тобто мова йде про необхідність розроблення глобальної моделі виробництва сільськогосподарської продукції, яка базувалась б на закономірностях сумісного розвитку виробництва і природи, сучасних теоріях систем, ефективних методах обґрунтування рішень [1].

Мета дослідження – побудова ефективної моделі обробки результатів досліджень варіабельності стану сільськогосподарських угідь, отриманих від систем моніторингу за допомогою кластерної моделі для підвищення точності складання картограм.

Аналіз останніх досліджень і публікацій. *Кластерний аналіз* – методика, за допомогою якої можна класифікувати дані в різних комбінаціях багатьох змінних в дискретні класи або кластери. Ця структура передбачає дві головні категорії, ієрархічну і неієрархічну. Найголовніше неієрархічне групування – k-засоби (також відомо як c-засоби), де багатовимірні дані класифіковані в k класи (кластери). Середня крапка в кожному класі має мінімальну відстань від кожного пункту даних. Невизначений k-засоби-

продовження k-засобів, що групуються, це рахунки для зв'язаної невпевненості з класовими межами і членством [2, 3].

Ping, J. L., Green, C. J., Bronson, K. F., Zartman, R. E., & Dobermann, A. [4] використовували k-засоби кластерного аналізу, багатовимірний дисперсійний аналіз (MANOVA) і дискретний аналіз в зусилля, щоб окреслити потенційні зони управління у просторі. Вони запропонували кластерний аналіз прибутку залежно від ґрунтових властивостей як підставу для окреслення зон управління.

У Бразилії, ці методи були використані, щоб встановити зони управління полем для забезпечення визначення необхідної грунтобробки для отримання максимальної врожайності [5].

Виклад основного матеріалу. *Кластерний аналіз* – це зручне джерело систематизації додаткового матеріалу. Можливість «розпізнавати» є одним з основних властивостей людських почуттів, як, до речі, й інших живих організмів. Образ являє собою характеристику (опис) об’єктів.

Відповідно з характером розпізнаваних об’єктів акт розпізнання можна розділити на два основні типи: розпізнавання конкретних об’єктів та розпізнавання абстрактних об’єктів. Процес розпізнання можна визначити як «сенсорне» розпізнавання. Процес цього типу забучує ідентифікацію і класифікацію просторових і часових образів.

Розпізнавання образів можна звести до питання оцінювання очевидної ймовірності, що початкові дані відповідають тому або іншому із відомих множин статистичних, що визначаються досвідом, які є орієнтирами і апріорною інформацією для розпізнавання. Таким чином, задачу розпізнавання образів можна розглядати, як задачу встановлення різниці між початковими даними, причому порівнянням з окремими образами, і їх сукупностями (останнє здійснюється при пошуку ознаки (інваріантних властивостей) на багатьох об’єктах, що визначають сукупність.

В задача розпізнавання образів можна виділити два головних напрямки:

1. Вивчення можливості до розпізнавання, якими володіють живі організми;

2. Розвиток теорії і методів побудови приладів, призначених для розв’язання окремих задач розпізнавання об’єктів.

Предмет розпізнавання образів – елементи, які належать конкретному класу, серед багатьох різних елементів, що відносяться до багатьох класів. Під класом образів розуміють деяку категорію, що визначається рядом властивостей, спільних для всіх елементів.

Образ – це опис будь-якого елемента як представника відповідного класу образів. У випадку, коли багато образів розділяються на неперетинаючі класи, необхідно використовувати для віднесення цих образів до відповідних класів який-небудь автоматичний засіб. Деякі задачі розпізнавання такі, що людина не в змозі їх вирішувати. Очевидно, що логічне рішення задачі розпізнавання об’єктів полягає у виділенні ознак кожного класу. Сукупність цих тестів повинна розрізняти всі допустимі образи з різних класів.

Головні задачі, які виникають при розробленні систем розпізнавання образів. *Перша* з них пов'язана з представленням вихідних даних, отриманих як результат вимірювання для розпізнаваного об'єкта. Це проблема чутливості. Кожна виміряна величина є деякою характеристикою об'єкта.

Друга задача розпізнавання об'єкта пов'язана з виділенням характерних ознак або властивостей з отриманих вихідних даних і зниженням розмірності векторів образів. Цю задачу часто визначають як задачу попередньої обробки і вибору ознак.

Властивості класу образів являють собою характерні властивості, спільні для всіх образів даного класу. Властивості, що характеризують різницю між окремими класами, можна інтерперувати як між класові ознаки. Внутрішньокласові ознаки, загальні для всіх розглядуваних класів, не володіють корисною інформацією з точки зору розпізнавання і можуть не братися до уваги. Вибір ознак вважається однією із головних задач, яка пов'язана з побудовою і розпізнаванням системи. Якщо результати вимірювання дозволяють отримати повний набір різних ознак для всіх класів, тому розпізнавання і класифікація образів не викличе особливої складності. Автоматичне розпізнавання тоді зведеться до простого співставлення або процедури типу перегляду таблиць.

У більшості практичних задач розпізнавання, визначення повного набору різних ознак буде справою дуже важкою, якщо взагалі можливою. Із вихідних даних, як правило, можна виокремити деякі розрізняльні ознаки і використовувати їх для спрощення процесу автоматичного розпізнавання образів. Зокрема, розмірність векторів можна знизити за допомогою перетворення, що забезпечує мінімізацію втрати інформації.

Третя задача, пов'язана з побудовою систем розпізнавання образів, полягає у пошуку оптимальних вирішальних процедур, необхідних для ідентифікації і класифікації. Після того, як зібрані дані, про об'єкти образи яких розпізнаються, представлені точками або векторними вимірами в просторі образів, необхідно за допомогою певного алгоритму з'ясувати до якого класу образів ці дані відповідають.

Розв'язати функції можна кількома способами. У тих випадках, коли про розпізнавані образи наявна повна апріорна інформація, рішення функції може бути визначене точно на основі цієї інформації. Якщо відносно образів є лише якісна інформація, можуть бути висунуті розумні допущення про вид функцій. В останньому випадку межі границь областей рішення можуть суттєво відхилятися від дійсних, тоді необхідно створювати систему, що може привести до позитивного результату за допомогою ряду позитивних коригувань.

Але, як правило, ми володіємо лише незначними апріорними свідченнями про розпізнаваний образ. За цих умов для побудови розпізнаваної системи краще всього використовувати навчальну процедуру. На першому етапі вибирають випадкові функції і потім в процесі виконання інтерактивних кроків ці функції доводять до оптимального виду.

Класифікацію об'єктів за допомогою функцій можна здійснювати найрізноманітнішими способами. Використовуються детерміністичні і статистичні алгоритми знаходження для розв'язання функцій.

Вирішення задачі попередньої обробки і виділення ознаки і задач отримання оптимального рішення і класифікації, як правило, пов'язане з необхідністю оцінювання і оптимізації ряду параметрів. Це приводить до задачі оцінювання параметрів. Крім того, зрозуміло, що процес виділення ознаки і процес прийняття рішення може бути суттєво удосконалений за рахунок використання інформації заключної в контексті образів. Інформація, що міститься в контексті, може бути виміряна за допомогою умовної ймовірності, лінгвістичних статистик та близьких варіантів. В деяких додатках просто необхідно використовувати певну інформацію для точного розпізнавання.

Об'єкти (образи), що підлягають розпізнаванню і класифікації за допомогою автоматичної системи розпізнавання образів, повинні мати набір вимірюваних характеристик. Коли для цілої групи образів результати відповідних вимірювань виявляються аналогічними, вважається, що ці об'єкти належать до одного класу. Мета роботи розпізнавання об'єктів полягає в тому, щоб на основі зібраної інформації визначити клас об'єктів з характеристиками, аналогічними вимірами у розпізнаваних об'єктів. Правильність розпізнавання залежить від об'єму ідентифікованої інформації, що міститься у вимірюваних характеристиках, і ефективності використання цієї інформації. Якби ми могли виміряти всі можливі характеристики і мати необмежений час для обробки зібраної інформації, то можна було б досягнути цілком адекватного рівня розпізнавання, використовуючи найпримітивніші методи. У звичайній практиці обмеження за часом, простором та затратами вимагають розвитку реалістичних підходів.

Для кластерної моделі оцінювання даних моніторингу варіабельності параметрів стану сільськогосподарських угідь запропоновано два методи її визначення.

Перший метод. Результатом багатовимірного групування у кластерному аналізі є розподіл сукупності спостережень на однорідні групи. Техніка кластерного аналізу базується на поняттях подібності об'єктів. Підбором найбільш схожих одиниць виконується розподіл сукупності на групи (кластери). На відміну від комбінаційних угрупувань, кластерний аналіз потребує поділу на групи з урахуванням відповідних ознак. Чіткі межі кожної групи та їх кількість у досліджуваній сукупності визначаються програмою.

Однорідність сукупності задається правилом обчислення певної метрики, що характеризує ступінь подібності одиниць сукупності. Її вибір є вузловим моментом кластерного аналізу, від якого головним чином залежить кінцевий варіант поділу сукупності на групи у разі даного алгоритму розподілу. Найпоширенішою є Евклідова метрика, за якою відстань між об'єктами обчислюється за формулою [6]:

$$C_{jk} = [\sum_{i=1}^{m} (z_{ij} - z_{ik})^2]^{\frac{1}{2}}, \quad (1)$$

де z_{ij} і z_{ik} – стандартизовані значення *i*-ї в *j*-ї та *k*-ї одиниць сукупності.

Якщо ознаки x_i рівновагомі, то розраховується зважена Евклідова відстань з вагами ω_i:

$$C_{jk} = [\sum_{i=1}^{m} \omega_i (z_{ij} - z_{ik})^2]^{\frac{1}{2}} \quad (2)$$

Оскільки наближеність об'єкта, який підлягає класифікації, до аналогів певного класу буде використовуватися як критерій для її здійснення, то такий підхід називається класифікацією об'єктів за критерієм мінімуму відстані [6].

Таким чином, для дослідження ефективності використання трудових ресурсів сільськогосподарських товариств з обмеженою відповідальністю регіону необхідним є проведення класифікації за сукупністю вищенаведених показників для визначення типових господарств.

Для побудови кластерної моделі оцінювання даних моніторингу варіабельності параметрів стану сільськогосподарських угідь вибрано алгоритм Isodata (Iterative Self-Organizing Data Analysis Techniques) [6]. Він має досить широкий набір допоміжних евристичних процедур, які включені до схеми ітерації.

Для виконання алгоритму необхідно визначити набір N_c вихідних центрів кластерів $z_1, z_2, \ldots, z_{Nc}$. Цей набір, кількість елементів якого не обов'язково повинна дорівнювати кінцевій кількості кластерів, може бути вибіркою образів з даної множини даних.

Під час роботи з набором $\{X_1, X_2, \ldots, X_N\}$, утвореним із *N* елементів, алгоритм Isodata виконує такі основні етапи:

Eman 1. Визначаються параметри процесу кластеризації:

K – необхідна кількість кластерів;

θ_N – параметр, з яким порівнюється кількість вибіркових образів, включених у кластер;

θ_s – параметр, який характеризує середньоквадратичне відхилення;

θ_c – параметр, який характеризує компактність;

L – максимальна кількість пар центрів кластерів, які можна об'єднати;

I – необхідна кількість ітерацій.

Eman 2. Задані *N* образи розподіляються по кластерах, які відповідають вибраним початковим центрам. За правилом X належить до класу S_j , якщо

$\|X - Z_j\| < \|X - Z_i\|$, $i = 1,2, \ldots N_c$; $i \neq j$, яке застосовується до всіх образів X вибірки; через S_j позначимо підмножину образів вибірки, які включені до кластера з центром Z_j .

Eman 3. Якщо для деякого j виконується умова $N_j < \theta_N$, то множина S_j виключається з подальшого розгляду і значення N_c зменшується на одиницю.

Етап 4. Кожний центр кластера z_j , j = 1,2, … N_c локалізується і коригується через порівняння його з вибірковою середньою, яка обчислюється за відповідною підмножиною S_j , тобто

$$Z_j = \frac{1}{N_j} \sum_{x \in S_j} x \tag{3}$$

j = 1, 2, …, N_c , де N_j – кількість об'єктів, які утворили множину S_j.

Етап 5. Розраховується середня відстань $\overline{D}_j$ між об'єктами, які входять в підмножину S_j, та відповідним центром кластера за формулою:

$$\overline{D}_j = \frac{1}{N_j} \sum_{x \in S_j} || x - Z_{j|} ||, \quad j = 1, 2, \ldots, N_c. \tag{4}$$

Етап 6. Обчислюють узагальнену середню відстань між об'єктами, які входять в окремі кластери, і відповідними центрами кластерів за формулою:

$$\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \overline{D}_j . \tag{5}$$

Етап 7. Передбачає наявність: а) якщо поточний цикл ітерації останній, то задається $\theta_c = 0$; перехід до етапу 11; б) якщо умова $N_c \le K/2$ виконується, то відбувається перехід до етапу 8; в) якщо поточний цикл ітерації має парний порядковий номер або виконується умова $N_c \ge 2K$, то переходимо до етапу 11, а в іншому випадку процес ітерації продовжується.

Етап 8. Для кожної підмножини вибіркових образів за допомогою формули

$$\sigma_{ij} = \sqrt{\frac{1}{N_j} \sum_{x \in S_j} (x_{ik} - z_{ij})^2} , \tag{6}$$

i = 1, 2, …, *n*; j = 1, 2, …, N_c обчислюють вектор середньоквадратичного відхилення $\sigma_j = (\sigma_{1j}, \sigma_{2j}, \ldots, \sigma_{nj})$, де n – розмірність образу, x_{ik} є i-а компонента k-го об'єкта в підмножині S_j, z_{ij} є i-а компонента вектора, яка відображає центр кластера z_j і N_j – кількість вибіркових образів, які увійшли в підмножину S_j. Кожна компонента вектора середньоквадратичного відхилення σ_j характеризує середньоквадратичне відхилення образу, який входить в підмножину S_j по одній із головних осей координат.

Етап 9. У кожному векторі середньоквадратичного відхилення σ_j, j = 1, 2, …, N_c знаходиться максимальна компонента σ_{jmax}.

Етап 10. Якщо для будь-якого σ_{jmax}, j = 1, 2, …, N_c виконується умова $\sigma_{jmax} > \theta_s$ і 1) $\overline{D}_i > \overline{D}$ і $N_j > 2(\theta_N + 1)$ або 2) $N_c \le K/2$, то кластер із центром z_j поділяється на два нових кластери відповідно з центрами Z_j^+ і Z_j^-. Кластер з центром Z_j ліквідовується, а значення N_c збільшується на одиницю. Для визначення центра кластера Z_j^+ до компоненти вектора, яка відповідає максимальній компоненті вектора σ_j, додається величина γ_j. Центр кластера

Z_j^- визначають відніманням величини γ_j із компоненти вектора Z_j. Величину γ_j визначають із співвідношення:

$$\gamma_j = k\sigma_{jmax}, \quad (7)$$

де $0< k \leq 1$.

При виборі γ_j потрібно керуватися в основному тим, щоб її величина була досить великою для того, щоб визначити різницю у відстані довільного образу до нових центрів кластерів, але досить малою, щоб загальна структура кластеризації суттєво не змінилась.

Якщо розщеплення відбувається на цьому етапі, то потрібно перейти до етапу 2, в іншому випадку продовжити виконання алгоритму.

Eman 11. Обчислюють відстань D_{ij} між усіма парами центрів кластерів:

$$D_{ij} = \|z_i - z_j\|, \quad (8)$$

$i = 1, 2, \ldots, N_c\text{-}1$; $j = i+1, \ldots, N_c$.

Eman 12. Відстань D_{ij} порівнюється з параметром θ_c. Відстані L , що виявилися меншими θ_c, групуються за збільшенням:

$$[D_{i1j1}, D_{i2j2}, \ldots, D_{iLjL}], \quad (9)$$

до того ж $D_{i1\ j1} < D_{i2\ j2} < \ldots < D_{iL\ jL}$, а L – максимальна кількість пар центрів кластерів, які можна об'єднати. Наступний етап передбачає процес об'єднання кластерів.

Eman 13. Кожну відстань $D_{iL\ jL}$ розраховано для визначеної пари кластерів із центрами z_{il} і z_{jl}. До цих пар у послідовності за порядком їх збільшення, яка відповідає зростанню відстані між центрами, застосовують процедуру об'єднання, яка виконується на основі такого правила: кластери з центрами z_{il} і z_{jl}, $l = 1,2, \ldots, L$ об'єднуються за умови, що в поточному циклі процедура об'єднання не застосовувалася ні до того, ні до другого кластера. Новий центр кластера визначають за формулою:

$$Z_l^* \frac{1}{N_{il} + N_{jl}}[N_{il}(Z_{il}) + N_{jl}(Z_{jl})] \cdot \quad (10)$$

Центри кластерів z_{il} і z_{jl} ліквідуються, а значення N_c зменшується на одиницю.

Слід зазначити, що допускається тільки попарне об'єднання кластерів і центр отриманого в результаті кластера розраховується, виходячи з позицій, які займали центри об'єднаних кластерів, пропорційно кількості вибіркових образів у відповідних кластерах.

Eman 14. Якщо поточний цикл ітерації – останній, то виконання алгоритму завершується. В іншому випадку слід повернутися до етапу 1, якщо необхідно змінити параметри процесу кластеризації або до етапу 2, якщо параметри не змінюються. Завершенням циклу ітерації вважається кожний перехід до етапів 1 або 2.

Було визначено такі параметри класифікації: початкову кількість кластерів

К = 4; необхідна кількість ітерацій I = 50. Параметри визначалися, виходячи з обсягу вибірки, середніх значень показників у всій сукупності об'єктів.

Другий метод. Результатом багатовимірного групування у кластерному аналізі є розподіл сукупності спостережень на однорідні групи. Техніка кластерного аналізу базується на поняттях подібності об'єктів. Підбором найбільш схожих одиниць (елементів) виконується розподіл сукупності на групи (кластери). На відміну від комбінаційних угрупувань, кластерний аналіз потребує розбивки на групи з урахуванням ознак, які групуються. Чіткі межі кожної групи і їх кількість у досліджуваній сукупності визначаються програмою.

Однорідність сукупності задається правилом обчислення певної метрики, що характеризує ступінь подібності одиниць сукупності. Вибір метрики є вузловим моментом кластерного аналізу, від якого головним чином залежить кінцевий варіант розмежування сукупності на групи у разі даного алгоритму розподілу.

Для проведення кластерної моделі оцінювання даних моніторингу варіабельності параметрів стан сільськогосподарських угідь вибрано алгоритм, представлений нижче, який мінімізує показник якості, визначений як сума квадратів відстаней всіх крапок, що входять в кластерну область, до центру кластера. Ця процедура, яку часто називають алгоритмом, що грунтується на обчисленні *K* внутрішньогрупових середніх, складається з наступних кроків [6].

Крок 1. Вибирають *K* вихідних центрів кластерів $z_1(1)$, $z_2(1)$, …, $z_k(1)$. Цей вибір роблять довільно і зазвичай як вихідні центри використовують перші *K* результатів вибірки із заданої безлічі образів.

Крок 2. На *k*-му кроці ітерації задана безліч образів {x} розподіляють по *K* кластерів за правилом (11):

$$x \in S_j(k), \text{якщо } \| x - z_j(k) \| < \| x - z_i(k) \| \qquad (11)$$

для всіх $i = 1, 2, \ldots, K, i \neq j$, де $S_j(k)$ – множина образів, що входять до кластера з центром $z_j(k)$. У разі рівності в (1.1) рішення приймається довільним чином.

Крок 3. На основі результатів кроку 2 визначаються нові центри кластерів $z_j(k+1)$, $j = 1, 2, \ldots, K$, виходячи з умови, що сума квадратів відстаней між всіма образами, що належать множині $S_j(k)$, і новим центром кластера має бути мінімальною. Іншими словами, нові центри кластерів $z_j(k+1)$ вибираються таким чином, щоб мінімізувати показник якості (12):

$$J_j = \sum_{x \in S_j(k)} \| x - z_j(k+1) \|^2, \; j = 1, 2, \ldots, K. \qquad (12)$$

Центр $z_j(k+1)$, що забезпечує мінімізацію показника якості, є, по суті, вибірковим середнім, визначеним за множиною $S_j(k)$. Отже, нові центри кластерів визначаються як (13):

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \; j = 1, 2, \ldots, K, \qquad (13)$$

де N_j – число вибіркових образів, що входять до множини S_j (k). Очевидно, що назва алгоритму «K внутрішньогрупових середніх» визначають способом, прийнятим для послідовної корекції призначення центрів кластерів.

Крок 4. Рівність z_j (k +1) = z_j(k) при j= 1, 2, …, K є умовою подібності алгоритму, і при його досягненні виконання алгоритму закінчується. Інакше алгоритм повторюється від кроку 2.

Якість роботи алгоритмів, що грунтуються на обчисленні K внутрішньогрупових середніх, залежить від числа обраних центрів кластерів, від вибору вихідних центрів кластерів, від послідовності огляду образів і, звичайно, від геометричних особливостей даних. У більшості випадків практичне вживання цього алгоритму потребує проведення експериментів, пов'язаних з вибором різних значень параметра K і вихідного розташування центрів кластерів.

Висновки. Запропоновано до використання удосконалену кластерну модель оцінювання даних моніторингу варіабельності параметрів стану сільськогосподарських угідь, що дає можливість підвищити точність складання картограм.

Література

1. Войтюк Д.Г., Гаврилюк Г.Р. Сільськогосподарські машини, Підручник. Видавництво «Каравела», - 2004. – 552с.

2. Dobermann, A., Ping, J. L., Adamchuk, V. I., Simbahan, G. C., & Ferguson, R. B. (2003). Classification of crop yield variability in irrigated production fields. Agronomy Journal, 95, 1105–1120;

3. Guastaferro, F., Castrignano, A., De Benedetto, D., Sollitto, D., Troccoli, A., & Cafarelli, B. (2010). A comparison of different algorithms for the delineation of management zones. Precision Agriculture, 11, 600–620.

4. Ping, J. L., Green, C. J., Bronson, K. F., Zartman, R. E., & Dobermann, A. (2005). Delineating potential management zones for cotton based on yields and soil properties. Soil Science, 170(5), 371–385.

5. Molin, J. P., & Castro, C. N. (2008). Establishing management zones using soil electrical conductivity and other soil properties by the fuzzy clustering technique. Scientia Agricola, 65(6), 567–573.

6. Дж. Ту. Принципы распознавания образов : [пер. с англ.: И. Б. Гуревича; под ред. Ю.И. Журавлева] / Дж. Ту, Р. Гонсалес. – М. : Мир, 1978. – 411 с.

Аннотация

В статье приведена методика оценки данных мониторинга параметров состояния сельскохозяйственных угодий, полученных от системы мониторинга.

Summary

The article presents a method of agricultural land monitoring data parameters evaluating obtained from the monitoring system.