Є.С. ПРОСКУРКА,
В.Д. КИШЕНЬКО, канд. техн. наук
Національний університет харчових технологій

КЛАСТЕРИЗАЦІЯ ЧАСОВИХ РЯДІВ ТЕХНОЛОГІЧНИХ ЗМІННИХ ЦУКРОВОГО ВИРОБНИЦТВА НА ОСНОВІ ПРЕЦЕДЕНТНОГО ПІДХОДУ

Проведений кластерний аналіз часових рядів технологічних змінних цукрового виробництва, який дозволив виділити характерні стани в поведінці складних об'єктів керування. Отримані результати дозволять реалізувати ефективні алгоритми прецедентного керування технологічними процесами харчових виробництв.
***Ключові слова**: кластерний аналіз, прецедентне керування, часові ряди, технологічні процеси, цукрове виробництво.*

Проведен кластерный анализ часовых рядов технологических переменных сахарного производства, который позволил выделить характерные состояния в поведении сложных объектов управления. Полученные результаты позволят реализовать эффективные алгоритмы прецедентного управления технологическими процессами пищевых производств.
***Ключевые слова**: кластерный анализ, прецедентное управление, часовые ряды, технологические процессы, сахарное производство.*

The cluster analysis of rows of sentinels of technological variables of sugar industry is conducted, which allowed selecting the characteristic consisting of conduct of difficult objects of management. The got results will allow realizing the effective algorithms of precedent control of technological processes of food productions.
***Keywords**: a cluster analysis, precedent management, sentinels, is rows, technological processes, sugar industry.*

В останній час широко використовуються «некласичні» методи управління технологічними процесами, серед них і управління на основі прецедентів — метод ухвалення рішень, в якому використовуються знання про раніше виникаючі ситуації або випадки (прецеденти).

При розгляді нової проблеми (поточного випадку) відшуковується схожий прецедент. Замість того щоб кожного разу шукати рішення спочатку, можна спробувати використовувати рішення, прийняте в схожій ситуації, можливо, адаптувавши його до поточного випадку.

За ситуації, коли відомих параметрів об'єкта управління і навколишнього середовища недостатньо для однозначного визначення поведінки цього об'єкта, управління необхідно здійснювати не за

параметрами об'єкта, а за його станом, який більш повно визначає тенденцію його подальшої поведінки. Виникає завдання ідентифікації стану об'єкта управління за його технологічними параметрами. Для цього потрібно уміти сформувати на основі апріорної інформації узагальнені образи — класи станів об'єкта. Отримати необхідні знання з набору наявних даних можна за допомогою методів знаходження даних — класифікації і кластеризації. Якщо стан об'єкта управління зводиться до присутності в одному з цих класів, то дію, що управляє, можна розглядати як перехід об'єкта одного класу в інший (зокрема, як утримання об'єкта в тому ж класі).

У цьому випадку стан об'єкта управління порівнюється з прецедентами з наперед накопиченої бази даних. На основі вибраної метрики вибирається один із схожих прецедентів. Дія, що управляє, пов'язана з ним, використовується безпосередньо або адаптується до поточного випадку, виходячи із ступеня близькості прецеденту. Результат дії також прогнозується за прецедентами. Підсумок дії заноситься в базу прецедентів для подальшого використання. Одночасно ставиться і більш складніше завдання вибору метрики для визначення схожості керованого об'єкта з прецедентами. Шукана відстань повинна сприяти обмеженню перебору можливих варіантів, а також полегшенню адаптації дії, що управляє, від прецеденту до поточного стану об'єкта управління.

Методи знаходження даних проводять пошук інформації в часових рядах технологічних параметрів, що були отримані під час технологічного процесу на виробництві.

Часовим рядом називається множина упорядкованих часових відліків разом з відповідними їм числовими значеннями:

$$X = \{(x_i, t_i),\ \ i \in N,\ \ x_i \in R, t_i \in T\} \qquad (1)$$

де T — дискретна часова шкала; R — множина дійсних чисел $x_i = x(t_i)$, характерезуючих числові значення часового ряду в і-ті моменти часу.

Кластерний аналіз дозволяє робити розбиття об'єктів не по одному параметру, а за цілими набором ознак. Крім цього, кластерний аналіз, на відміну від більшості математико-статистичних методів, не накладає ніяких обмежень на вид розглянутих об'єктів і дозволяє розглядати множину вихідних даних практично довільної природи. Кластерний аналіз дозволяє розглядати досить великий обсяг інформації і різко скорочувати, стискати великі масиви інформації, робити їх компактними і наочними.

Задача кластеризації складається в поділі досліджуваної множини об'єктів на групи «схожих» об'єктів, називаних кластерами. Слово кластер англійського походження (cluster), перекладаєся як пучок, група. Рішення задачі розбиття елементів на кластери називають кластерним аналізом.

Задача кластеризації (або навчання без вчителя) полягає у наступному. Є навчальна вибірка $X = \{x_1, x_2, ..., x_m\}$ та функція відстані між елементами цієї вибірки $p(x_i, x_j)$. Потрібно розбити вибірку на підмножини — кластери, так, щоб кожний кластер складався з об'єктів, близьких за метрикою p, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X$ приписується мітка (номер) кластеру y_i.

Алгоритм кластеризації це функція $a: X \to Y$, яка будь-якому об'єкту $x_i \in X$ ставиться у відповідність мітку $y_i \in Y$.

Відстань (метрика) між об'єктами в просторі параметрів визначається величиною d_{ab}, що задовольняє таким вимогам [1]:

$$d_{ab} > 0;\ \ d_{aa} = 0;$$

$$d_{ab} = d_{ba}.$$

В таблиці представлені найбільш поширені відстані між об'єктами.

Варто зробити два зауваження. По-перше, розрізняють задачі кластерного аналізу (і відповідно алгоритми) із заданим (або відомим) числом кластерів, а також із незаданим (невідомим) числом кластерів. В останньому випадку оптимальна кластеризація і число кластерів перебувають у результаті розв'язання єдиного завдання. По-друге, крім звичайної постановки задачі кластеризації, як задачі пошуку розбиття, існують постановки, як задачі пошуку покриття і структур на заданій множині прецедентів.

Найбільш поширені відстані між об'єктами

Показник	Формула
Лінійна відстань	$d_1(x_i, x_j) = \sum_{t=1}^{m} \lvert x_{it} - x_{jt} \rvert$
Евклідова відстань	$d_E(x_i, x_j) = \sqrt{\sum_{t=1}^{m} (x_{it} - x_{jt})^2}$
Відстань Махаланобіса	$d_M(x_i, x_j) = (x_i - x_j)^t \cdot S^{-1}(x_i - x)$ S^{-1} — обернена коваріаційна матриця векторів x_i та x_j; t — позначає операцію транспонування.

Розглянемо ці задачі детальніше [2]:

1. **Задача пошуку оптимального розбиття.** Розглядається задача кластеризації на l кластерів. Вибірку ознак об'єкта позначимо, як $X = \{x_1, x_2, ..., x_m\}$, $x_i = \{x_1, x_2, ..., x_m\}$. Розбиттям $K = \{K_1, K_2, ..., K_m\}$ вибірки $X = \{x_1, x_2, ..., x_m\}$ на l груп являє собою довільну сукупність непересинаючих підмножин множини X, які покривають всі об'єкти вибірки

$$K_i \subseteq X,\ i = 1, 2, ..., l,\ \bigcup_{i=1}^{l} K_i = X,\ K_i \cap K_j = \varnothing,\ i \neq j.$$

Нехай задано деякий критерій якості F(K) розбиття K. Тоді задача кластеризації полягає в знаходженні екстремуму критерію $F(K^*) = \underset{K \in \{K\}}{extr}\, F(K)$.

Наведемо приклади таких критеріїв [3]:

Сума дисперсій в середині класу чи квадратів помилок:

$$F(K) = \sum_{j=1}^{l} \sum_{x_i \in K_j} p^2(x_i, y_j),$$

де $y_j = \frac{1}{n} \sum_{x_i \in K_j} x_i$, $n_j = K_j$ — число об'єктів в K_j. Вирішенням задачі вважається таке розбиття K^* при якому $F(K)$ досягає мінімуму.

Критерій на базі матриць розсіювання.

Матриця розсіювання групи K_j визначається, як $\Sigma_j = \sum_{x_i \in K_j} (x_i - y_j)(x_i - y_j)^T$, а матриця розсіювання в середині класу $\Sigma = \sum_{i=1}^{l} \Sigma_j$; T — позначає операцію транспонування.

2. Задача пошуку оптимального покриття. Розглянемо один з алгоритмів вирішення даної задачі для заданого числа l кластерів. Вважаємо також попередньо заданими l векторів $y_1, y_2 \ldots y_l \in R^n$, які є центрами «згущення» вибірки й інтерпретуються, як центри кластерів. Вони можуть бути отримані, наприклад, як центри гіперкуль після попередньої кластеризації вибірки за допомогою алгоритму FOREL або за допомогою методу k-means [4].

Введемо в розгляд спеціальні сімейства вкладених гіперкуль. Нехай p — деяка метрика в просторі R^n. Упорядкувавши за зростанням значення відстаней від y_i до всіх елементів з $X = \{x_1, x_2, \ldots, x_m\}$, одержимо монотонно зростаючі послідовності $r_1^i < r_2^i < \ldots < r_{k_i}^i$, $k_i \le m$, $i = 1, 2, \ldots, l$.

Через $\{R_j^i = \{x : p(x, y_j) \le r_j^i\}, j = 1, 2, \ldots, k\}$, позначимо множину вкладених гіперкуль з центрами в y_i. Множина даних гіперкуль покриває всі об'єкти вибірки. Позначимо $n_{ij} = \{x_t : x_t \in R_j^i\}$ — число об'єктів вибірки, що належать гіперкулі R_j^i.

Тепер можна поставити наступне завдання: «Знайти l гіперкуль $R_{t1}^i, R_{t2}^i, \ldots, R_{tl}^i$, що покриває всі об'єкти з $X = \{x_1, x_2, \ldots, x_m\}$ і маючи мінімальне число загальних об'єктів $\sum_{i=1}^{l} n_{it} \to \min$.

Дана відома задача пошуку покриття мінімальної ваги формулюється у вигляді наступного завдання цілочислового лінійного програмування.

$$\sum_{i=1}^{l} \sum_{j=1}^{k_i} n_{ij} z_{ij} \to \min \qquad (2)$$

$$\sum_{i=1}^{l} \sum_{j=1}^{k_i} c_{ij}^t z_{ij} \ge 1, \; t = 1, 2, \ldots, m \qquad (3)$$

$$\sum_{j=1}^{k_i} z_{ij} \ge 1, \; i = 1, 2, \ldots, l \qquad (4)$$

де $z_{ij} \in \{0,1\}$, $c_{ij}^t = \begin{cases} 1, x_t \in R_j^i \\ 0, \text{інакше} \end{cases}$

Обмеження (3) виражає вимогу покриття кожного об'єкта хоча б однією кулею, а (4) — присутність у розв'язку основного завдання хоча б однієї кулі із центром в y_j для кожного $j = 1, 2, \ldots, l$.

3. Задача пошуку структур в даних. Класичним прикладом даної задачі є ієрархічне групування. Тут завдання кластеризації на l кластерів вирішуються послідовним об'єднанням найближчих груп (існують і підходи, засновані на послідовній розбивці груп об'єктів). На першому кроці вважається, що кожний об'єкт утворить «одноточковий» кластер: $K_i^1 = \{x_i\}$, $i = 1, 2, \ldots, m$. На другому кроці два найближчих кластери поєднуються в один. Процес об'єднання повторюється до знаходження l кластерів. Приведемо найпоширеніший алгоритм ієрархічної кластеризації [4].

Нехай фіксована метрика $p(x, y)$ на множині векторів $x = (x_1, x_2, x_n)$, x_i — дійсні числа. Уведемо також міру відстані між групами об'єктів $d(T_i, T_j)$. Прикладами подібних функцій є функції (5) – (6).

$$d_{min} \to \min_{x_v \in T_i, x_u \in T_j} p(x_v, x_u) \qquad (5)$$

$$d_{max} \to \max_{x_v \in T_i, x_u \in T_j} p(x_v, x_u) \qquad (6)$$

Нехай задана функція $p(x, y)$ і функції $d(T_i, T_j)$ для груп об'єктів T_i і T_j. Розглянемо завдання кластеризації на задане число кластерів. Нехай до кроку k, $k = 1,2,\ldots,m$ отримані кластери $T_1^k, T_2^k, \ldots, T_{m-k+1}^k$ (при $k = 1$, $T_v^1 = \{x_v\}$, $v = 1, 2, \ldots, m$). Від даних $(m - k + 1)$ кластерів здійснюється перехід до $(m - k)$ кластерів шляхом об'єднання в однієї тієї пари T_v, T_u для якої:

$$d(T_v^k, T_u^k) = \min_{t,\sigma} d(T_t^1, T_\sigma^1) \qquad (7)$$

Для знаходження розбиття вихідної вибірки на l кластерів потрібне виконання $m - l$ кроків. Отримані кластери є розбиттям вихідної вибірки, причому кожний із кластерів має структурне представлення в термінах відстаней об'єктів і кластерів.

Була проведена кластеризація часових рядів основних технологічних змінних цукрового заводу (загальна кількість 24 змінних). Як приклад, наведені результати кластеризації часового ряду витрати дифузійного соку (рис. 1), які зображені на рис. 2.

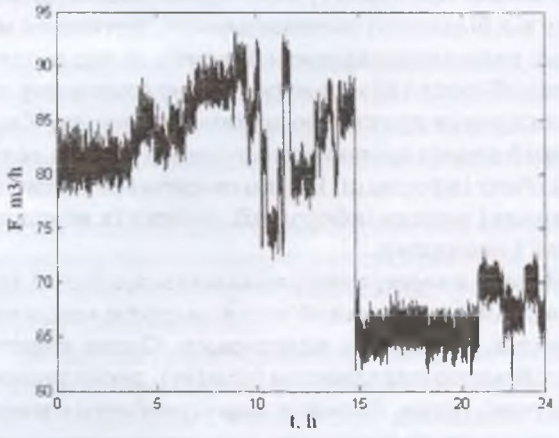

Рис. 1. Часовий ряд витрати дифузійного соку

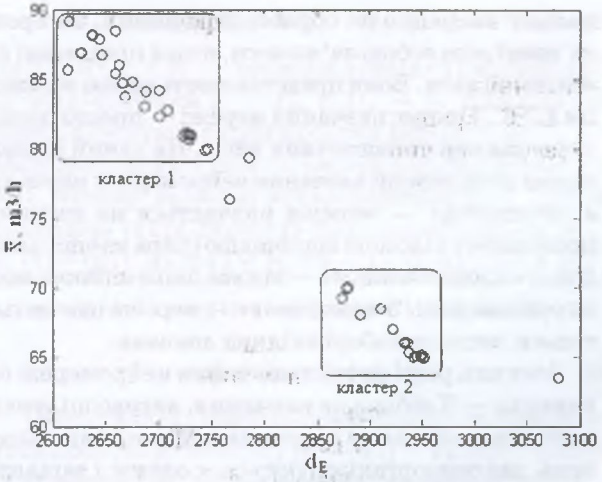

Рис. 2. Виділені кластери в часовому ряді витрати дифузійного соку

Часові ряди розбили на інтервали тривалістю 30 хв. Для кожного ряду знайдемо його середнє значення витрати дифузійного соку за формулою (8):

$$\overline{x}_j = \frac{1}{n}\sum_{t=1}^{m} x_{jt} \qquad (8)$$

та відстань Евкліда d_E, в якій x_{it} — значення витрати дифузійного соку, x_{jt} — вектор часу.

Після того, як знайшли відстані виконуємо перевірку за формулою (9):

$$\left|d_{E1} - d_{E2}\right| < \varepsilon \qquad (9)$$

Якщо різниця відстаней обох векторів менше ε, то відносимо ці два вектори до одного кластеру, в протилежному випадку створюємо множину кластерів.

З рис. 2 видно, що було виділено два кластери. Кожен кластер характеризує певний стан системи.

Висновки. Проведений кластерний аналіз часових рядів дозволив визначити стани об'єктів управління, що забезпечить розробку ефективних алгоритмів прецедентного управління технологічними процесами цукрового виробництва.

ЛІТЕРАТУРА

1. *Мандель И.Д.* Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.

2. *Журавлев Ю.И., Рязанов В.В., Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2005. — 159 с.

3. *Дуда Р., Харт П.,* Распознавание образов и анализ сцен. — М.: Мир, 1976. — 511 с.

4. *Барсегян А.А., Куприянов М.С., Степаненко В.В.,* Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004. — 336с.

Одержана редколегією 09.03.2010