Обробка інформації в складних технічних системах

УДК 004.032.26

DOI: 10.30748/soi.2020.160.01

С.Г. Удовенко¹, Д.В. Келембет², О.В. Тесленко¹

¹ Харківський національний економічний університет ім. С. Кузнеця, Харків

² Харківський національний університет радіоелектроніки, Харків

КОМБІНОВАНИЙ МЕТОД НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ДАНИХ В СИСТЕМАХ ТЕХНІЧНОЇ ДІАГНОСТИКИ

У статті наведено результати аналізу методів нечіткої кластеризації даних поточних спостережень в технічних системах. Запропоновано і протестовано комбінований підхід до нечіткої кластеризації даних з використанням комбінації алгоритму Густафсона-Кесселя, методу остовного лісу та фільтра Калмана, що може бути використаний в системах технічної діагностики (зокрема, діагностики стану основних параметрів технологічних процесів). Розглянуто можливість подальшого розширення функціональних можливостей наведеного комбінованого методу кластеризації з наданням йому адаптивних властивостей. Наведено результати тестування та визначено переваги запропонованого підходу.

Ключові слова: нечітка кластеризація, діагностика стану параметрів, метод остовних дерев, фільтрація Калмана.

Вступ

Постановка проблеми. В системах технічної діагностики може бути виділений клас завдань, що пов'язані з віднесенням поточного стану технологічного процесу до одного з можливих станів (кластерів), які дозволяють в подальшому приймати рішення (в автоматичному режимі або на рівні оператора) щодо прогнозування розвитку процесу та вибору стратегії керування. Останнім часом для діагностики поточних станів об'єктів діагностування (ОД) за даними спостережень їх входів та виходів, які формують масив спостережень (МС), часто використовують методи нечіткої кластеризації (НК).

Застосування НК дає можливість формування поточних діагностичних параметрів (ДПА(t)), що використовуються для прийняття діагностичних висновків [1–2]. Загальну структуру системи діагностики такого типу наведено на рис. 1.

Рис. 1. Система діагностики стану параметрів з використанням нечіткої кластеризації

Ідентифікація поточних станів ОД (зазвичай, це об'єкти цифрового керування) за експериментальними даними з використанням нечітких моделей є ефективним способом апроксимації нелінійних систем. До найбільш відомих моделей такої ідентифікації слід віднести нечітку модель Такагі-Сугено (ТС) [3]. Ця модель використовує лінеаризацію нечітких областей параметрів ОД в просторі станів. За допомогою наборів нечітких правил ТС можна представити простір входів множиною локальних моделей, що дозволяють описати поведінку системи в нечітких областях [4]. Алгоритми побудови нечітких ТС-моделей ТЗ в загальному випадку передбачають реалізацію двох етапів: формування функцій належності (ФН), що відповідають деяким нечітким правилам; визначення правил нечіткого виводу. Перший з цих етапів може бути реалізований шляхом нечіткої кластеризації в просторі входів-виходів. Так як рівняння нечіткого виведення зазвичай вибираються лінійними, то реалізація другого етапу, як правило, заснована на застосуванні методу найменших квадратів. Чимало алгоритмів кластеризації базуються на оптимізації нечітких функцій с-середнього та використовують евклідову відстань між об'єктами, для яких сформовані кластери мають форму гіперсфери (зокрема, алгоритм Fuzzy C-Means або FCM) [5–6]. Деякі типи алгоритмів використовують адаптивну норму для отримання кластерів різної геометричної форми, що відповідають одній множині даних [7]. Крім того, для нечіткої кластеризації даних може бути додатково використано теорію графів. При цьому комбінація

методів теорії графів та нечіткої логіки дозволяє виділяти кластери довільної форми та вирішувати завдання кластеризації об'єктів з лінгвістичними атрибутами [8]. Перспективним та актуальним науковим напрямком є дослідження можливості комбінованого використання методів нечіткої кластеризації в системах технічної діагностики (зокрема, діагностики стану основних параметрів технологічних процесів).

Аналіз останніх досліджень і публікацій. Нечітка кластеризація в декартовому просторі входів-виходів широко використовується для отримання функцій належності моделей ТС. Застосовуючи нечітку кластеризацію, можна отримати початкові багатовимірні нечіткі множини, які можуть бути використані або безпосередньо в ТС-моделі або за допомогою непрямих змінних (регресорів).

В роботі [9] пропонується метод нечіткого моделювання, де використовуються алгоритм кластеризації Густафсона-Кесселя (ГК) для оцінювання функцій належності та МНК для оцінювання параметрів моделі ТС. В роботі [10] запропоновано адаптивні форми ймовірнісного та можливісного варіантів алгоритму ГК, які відрізняються розрахунковою простотою та забезпечують високу ефективність при роботі в умовах невизначеності, а саме при мінливому у часі характері даних, що обробляються, та при нестачі знань про природу вибірки. В роботі [11] пропонується використовувати алгоритм кластеризації Гата-Гева (ГГ) замість методу ГК. Перевагою такого алгоритму є те, що функції належності можуть бути тут безпосередньо отримані за параметрами кластерів. В роботі [12] наведено гібридний метод нечіткої кластеризації, що поєднує переваги графового методу MST та методу FCM і пропонується для аналізу фактографічних даних. Таким чином, слід відзначити суттєвий інтерес науковців до проведення досліджень в галузі розвитку і комбінованого використання різних методів нечіткої кластеризації даних.

Мета статті. У даній роботі пропонується комбінований підхід до нечіткої кластеризації даних з використанням комбінації алгоритму ГК, методу MST та фільтра Калмана, що може бути використаний в системах технічної діагностики (зокрема, діагностики стану основних параметрів технологічних процесів).

Виклад основного матеріалу

Аналіз проблеми нечіткої кластеризації

Кластерний аналіз використовується для класифікації об'єктів за ознаками подібності між ними і для групування даних. Якщо про внутрішні залежності в даних нічого невідомо, то важливим завданням такого аналізу є завдання кластеризації, тобто виявлення внутрішньої структури в даних, на підста-

ві якої надалі можна буде формувувати більш детальні завдання про пошук залежностей, що впливають на угруповання вхідної множини даних.

Методи кластеризації можуть бути застосовані для кількісних (числових) даних, для якісних (категорійних) даних або для змішаних даних. Ідентифікація реальних об'єктів цифрового керування заснована, як правило, на спостереженнях числових даних, що характеризують конкретні фізичні процеси. Кожне спостереження складається з n значень вимірюваних змінних, що утворюють n -мірний вектор-стовпець. Сукупність N спостережень позначимо як $Z = \{Z_k; k = \overline{1, N}\}$. Така сукупність може бути представлена відповідною $(n \times N)$ -мірною матрицею даних. В задачах ідентифікації стовпці матриці Z містять значення сигналів системи в дискретні моменти часу, а рядки відповідають фізичним змінним системи (положенню, швидкості, температурі тощо). Дані можуть групуватися в кластери різної геометричної форми, різних розмірів і щільності. Таким чином, кластери можуть розглядатися як підмножини загальної множини даних. Класичні методи кластеризації базуються на класичній теорії множин і дозволяють визначити, належить чи ні розглянутий елемент до деякого кластеру (тобто підмножини в Z). Методи нечіткої кластеризації допускають приналежність одного і того ж елемента до різних кластерів одночасно, проте з різними оцінками ступеня такої приналежності.

Розглянемо спочатку принципи класичного розбиття множини даних Z на кластери. Таке розбиття може бути визначено як сімейство підмножин $\{A_i / 1 \leq i \leq c\} \subset P(Z)$.

Матриця $U = [\mu_{ik}]$ відповідає класичному розбиттю, якщо її елементи задовольняють таким умовам:

$$\begin{aligned} \mu_{ik} &\in \{0, 1\} & 1 \leq i \leq c, \quad 1 \leq k \leq N; \\ \sum_{i=1}^c \mu_{ik} &= 1 & 1 \leq k \leq N; \\ 0 < \sum_{k=1}^N \mu_{ik} &< N & 1 \leq i \leq c. \end{aligned} \quad (1)$$

Відома велика кількість методів кластеризації, які діляться на ієрархічні і неієрархічні, серед яких найбільшого поширення набули методи розбиття. Спільним для методів розбиття є обов'язкова формалізація відстані як засобу порівняння даних, формування множини центрів кластерів і матриць належності як основних результатів кластеризації, завдання функціоналу, мінімізація якого визначає рішення, а також набір обмежень. До найбільш відомих методів розбиття належать вище згадані методи k -середніх, Fuzzy C-Means, ГК та ГГ. Перший відрі-

зняється тим, що одержувані з його допомогою значення матриці належності є чіткими (приймають значення з множини $\{0,1\}$) [13].

Розглянемо особливості застосування методів розбиття на прикладі методу FCM з використанням прийнятих вище позначень. Метою цього методу є автоматична класифікація множини об'єктів, які задаються векторами в просторі ознак. Інакше, такий алгоритм визначає кластери і відповідно класифікує об'єкти. Кластери представляються нечіткими множинами, і, крім того, кордони між кластерами також є нечіткими.

FCM-алгоритм кластеризації передбачає, що об'єкти належать всім кластерам згідно з певною ФН, а ступінь належності визначається відстанню від об'єкта до відповідних кластерних центрів. Даний алгоритм ітераційно обчислює центри кластерів і нові ступені належності об'єктів.

Базовий FCM-алгоритм заснований на оптимізації функцій, що мають наступну структуру:

$$J(Z;U,V) = \sum_{i=1}^c \sum_{r=1}^N (\mu_{ik})^m \|z_k - v_i\|_A^2, \quad (2)$$

де U – матриця нечіткого розбиття множини Z :

$$U = [\mu_{ik}], \quad (3)$$

V – вектор центрів кластерів:

$$V = [v_1, v_2, \dots, v_c], \quad v \in R^n, \quad (4)$$

m – ваговий показник, що визначає ступінь нечіткості кластерів.

Норма відстані між елементами множини Z та центрами кластерів визначається наступним чином:

$$D_{ikA}^2 = \|z_k - v_i\|_A^2 = (z_k - v_i)^T A (z_k - v_i). \quad (5)$$

З (6) видно, що відповідна відстань зважується величиною ступеня належності елементів $(\mu_{ik})^m$.

Мінімізуючи функцію (2), отримуємо:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{D_{ikA}}{D_{jkA}} \right)^{\frac{2}{m-1}}}; 1 \leq i \leq c, 1 \leq k \leq N, \quad (6)$$

$$v_i = \frac{\sum_{k=1}^N (\mu_{ik})^m z_k}{\sum_{k=1}^N (\mu_{ik})^m}; 1 \leq i \leq c. \quad (7)$$

Відношення (6–7) відповідають принципу нечіткої кластеризації з використанням FCM-алгоритму. Відповідно до цього алгоритмом для вхідної множини даних Z задаються кількість кластерів $c \in (1, N)$, значення показника $m > 1$, значення параметру умови зупинки $\varepsilon > 0$ і матриця норми A . Далі задається матриця розбиття U .

Алгоритм зупиняється згідно з наступною умовою:

$$\|U^{(l)} - U^{(l-1)}\| < \varepsilon. \quad (8)$$

Форма кластерів визначається вибором матриці A в рівнянні (5). Зокрема, якщо $A = I$ (I – одинична матриця), то отримуємо стандартну норму Евкліда:

$$D_{ikA}^2 = (z_k - v_i)^T (z_k - v_i).$$

В цьому випадку кластери, що формуються, мають сферичну форму.

FCM-алгоритм може бути узагальнений з використанням адаптивної норми відстані, що дозволяє виділяти кластери різної геометричної форми в початковій сукупності даних. Припустимо, що елементи матриць A в рівнянні (5) можуть варіюватися, однак їх детермінанти мають залишатися фіксованими, тобто:

$$|A_i| = \rho_i, \quad \rho > 0, \quad (9)$$

де значення констант ρ_i є фіксованими для кожного з кластерів. В цьому випадку, оптимізуючи (2), можна отримати такий вираз для A_i :

$$A_i = [\rho_i \det(F_i)]^{\frac{1}{n}} F_i^{-1}, \quad (10)$$

де F_i – матриця нечіткої коваріації i -го кластера, що визначається таким чином:

$$F_i = \frac{\sum_{k=1}^N (\mu_{ik})^m (z_k - v_i)(z_k - v_i)^T}{\sum_{k=1}^N (\mu_{ik})^m}. \quad (11)$$

Залежності (10–11) лежать в основі деяких алгоритмів розбиття, зокрема, алгоритмів ГК та ГГ. Тут, як і в FCM-алгоритмі, спочатку вибираються такі параметри, як $c \in (1, N)$, $m > 1$, $\varepsilon > 0$, A , а далі випадковим чином ініціюється матриця розбиття U .

Відзначимо недоліки розглянутих алгоритмів розбиття: кількість кластерів є заданим параметром, що призводить до ітераційного проведення досліджень набору даних і може використовуватись для оцінки кожного розбиття на досягнення оптимальної кількості кластерів; чутливість до аномалій в наборі даних, що вимагає використання додаткових інструментів для очищення даних до проведення кластеризації; при використанні критеріїв зупинки циклу розбиття на основі різниці між результатами попередньої та поточної ітерації можливі ситуації, при яких відбувається зациклення обробки даних, що призводить до виникнення невизначеностей; повільна робота на великих обсягах даних, що обмежує застосовність методів; нелінійне збільшення часу аналізу при зростанні обсягів вхідних даних, що призводить до значних тимчасових витрат при динамічній зміні характеристик об'єкту діагностування; неможливість пояснення отриманих результатів розбиття, що знижує довіру до ефективності методів.

За основу первинного поділу об'єктів на кластери може бути використано метод MST (Minimum Spanning Tree), який з допомогою так званих мінімальних остовних дерев виділяє кластери довільної форми. Розглянемо докладніше варіанти цього методу.

Нехай заданий зв'язний неорієнтований граф $G = (V, E)$, де V, E – сукупність вершин та ребер відповідно. Для графа з цілочисельними вагами можливе використання спеціальних прийомів, таких як сортування за розрядами, що призводить до алгоритмів, що вирішують завдання за лінійний час. У розподіленій задачі може бути присутнім додаткова вимога, що обмін даними відбувається тільки уздовж ребер графа.

Рішення завдання залежить не від значень ваг ребер, а від їх порядку. Таким чином, замість завдання ваг можна задати порядок – предикат “менше або дорівнює” на множині пар ребер графа. З цієї причини можна без обмеження спільності вважати, що всі ваги ребер різні: для цього слід упорядкувати ребра спочатку за вагою, а потім за номером. Крім того, до вихідних ваг можна застосувати будь-яку зростаючу функцію, що це не призведе до зміни рішення. Як наслідок, можна вважати, що всі ваги знаходяться в заданому інтервалі, наприклад $[0, 1]$.

Якщо умова зв'язності вихідного графу є не обов'язковою, то набір мінімальних остовних дерев для всіх компонент зв'язності називають мінімальним остовним лісом (Minimum Spanning Forest, MSF). Мінімальний остовний ліс завжди існує, а якщо всі ваги ребер різні, то він єдиний. Як уже зазначалося, завжди можна зробити ваги ребер різними. Якщо умова різних ваг не виконується, може існувати більш одного мінімального остовного дерева.

Вхідну множину даних Z ми представляємо (як для MST, так і для MSF-постановок) як граф, де вершини – елементи множини Z , а дуги – всілякі пари цих елементів, вага яких дорівнює відстані між їх векторними уявленнями. Далі по одному з відомих пошукових алгоритмів будується мінімальне остовне дерево, причому при виборі алгоритму необхідно враховувати велику кількість дуг в графі (для N елементів в Z маємо $N(N-1)$ дуг). Після побудови дерева йде процес видалення ребер з найбільшими довжинами, в результаті чого утворюється ліс більш маленьких дерев, з вузлів яких і генеруються кластери.

Розглянемо модифікований графовий MSF-підхід до проблеми кластеризації з використанням методів “напіввизначених найменших квадратів” – МННК- k [14]. Проблема k -кластера – це NP-проблема жорсткої комбінаторної оптимізації. Нехай в графі $G = (V, E)$, що має n вершин ($1 \leq k \leq n$), ми шукаємо підграф G з k вершинами та максимальною

кількістю ребер. Ця проблема може бути сформульована як задача квадратична програмування для змінних 0-1:

$$(k\text{-cluster}) \begin{cases} \max y^T A y; \\ e^T y = k; \\ y \in \{0, 1\}^n, \end{cases} \quad (12)$$

де $e = (1, \dots, 1) \in R^n$; A – суміжна з G матриця.

Для вирішення цієї проблеми можна застосувати кілька підходів математичного програмування, зокрема лінійне, напіввизначене або опукле квадратичне програмування. Для побудови обчислювальної процедури за методом напіввизначених найменших квадратів використаємо переформулювання двійкових обмежень за допомогою так званого “сферичного” обмеження [15]. Перший крок цього переформулювання полягає у представленні змінної обмеження бівалентності у вигляді $x \in [-1, 1]^n$, потім у гомогенізації квадратичних форм шляхом додавання змінної, щоб отримати чисто квадратичну задачу оптимізації в $x \in R^{n+1}$:

$$(QP) \begin{cases} \max x^T Q_1 x; \\ x^T Q_2 x = 2k - n; \\ x \in \{-1, 1\}^{n+1}, \end{cases} \quad (13)$$

де Q_1, Q_2 залежать від A та e .

Після введення симетричної $(n+1)$ – вимірної матриці $X = xx^T$ рангу 1 задачу кластеризації можна представити наступним еквівалентним чином:

$$(SDP) \begin{cases} \min \|X - Q_1 / \alpha\|^2; \\ \langle Q_2, X \rangle = 2k - n; \\ \text{diag}(X) = e; \\ X \succeq 0. \end{cases} \quad (14)$$

Останнє обмеження означає, що матриця X є напіввизначеною позитивною (SDP). Відзначимо, що в цій ситуації обмеження рангу 1 є еквівалентним сферичному обмеженню $\|X\|^2 = (n+1)^2$. Дуалізація сферичного обмеження приводить до методу напіввизначених найменших квадратів.

Проблему такого типу можна вирішити чисельно ефективними методами навіть для великої кількості дуг графі $G = (V, E)$. Основною перевагою такого підходу є можливе скорочення часу реалізації обчислень в обмін на погіршення якості обмежень, визначеною необхідністю встановлення значення параметра α в (14).

З метою отримання моделі, прийнятної для використання в системі діагностики стану основних параметрів технологічних процесів, функції належності мають визначатися таким чином, щоб було

можливим обчислити ступень належності навіть для даних, які не містяться в вхідному наборі. Для вирішення цього завдання доцільним є використання фільтра Калмана, що дозволяє апроксимувати дискретні функції належності лінійними моделями і отримати їх в трикутних або трапецієподібних формах.

Комбіноване використання розглянутих методів може сприяти подоланню недоліків кожного з них: графові методи дозволяють зробити більш плавною загальну процедуру нечіткої кластеризації, поміщаючи об'єкти в різні кластери з різним ступенем належності; після попереднього використання графових методів з визначенням доцільної кількості кластерів можна здійснити уточнену нечітку кластеризацію за одним з методів розбиття. Такій підхід дозволяє скоротити кількість ітерацій дослідження вхідного набору даних, а отже і знизити і затрати на проведення діагностування.

Комбінований метод нечіткої кластеризації

За результатами аналізу існуючих методів нечіткої кластеризації, що є прийнятними для вирішення завдання діагностики поточного стану параметрів технологічних процесів, були обрані: MSF-метод з алгоритмічною реалізацією на основі процедури МННК- k і метод ГК. Важливими аргументами на користь такого вибору були здатність при використанні графового підходу та методу ГК виділяти кластери довільної форми і оптимальної структури. При цьому зберігається можливість подальшого розширення функціональних можливостей пропонуваного комбінованого методу кластеризації з наданням йому адаптивних властивостей.

Комбінований метод, що пропонується, передбачає послідовну реалізацію таких кроків: вибір структури діагностичної моделі; збір даних та їх підготовка до кластеризації; визначення кількості кластерів з використанням MSF, МННК- k та фільтрації Калмана; нечітка кластеризація даних з використанням методу ГК; оцінювання якості кластеризації.

Розглянемо докладніше особливості реалізації цих кроків.

Крок 1. Вибір структури діагностичної моделі. Основою цього етапу є вибір інформативної сукупності вхідних і вихідних змінних, що мають використовуватися в діагностичній моделі. Крім того, необхідно вибрати порядок діагностичної моделі. Це дозволить перейти від проблеми нечіткої кластеризації до проблеми створення статичних регресійних залежностей для діагностування стану технологічного процесу.

Крок 2. Збір даних та їх підготовка до кластеризації. Вибір методики реєстрації та збору даних суттєво впливає на кінцевий результат процедури діагностики. Мова може йти як про формування ча-

сової послідовності амплітудних значень вибраних змінних, так і про реєстрацію частотних характеристик об'єкту діагностування (наприклад, в разі використання вібраційної діагностики). У другому випадку на попередньому етапі кластеризації може бути застосовано методи активного експерименту, так як для отримання репрезентативних даних треба впливати на ОД по амплітудній і частотній осях, змінюючи вхідні сигнали в істотному діапазоні. Очевидно, що ці сигнали мають поділятися на значну кількість амплітудних рівнів, а їх частотні властивості мають бути представлені достатньою кількістю гармонік.

Підготовка даних – технологічний етап, що, зокрема, складається з перетворення категоріальних даних (в разі їх наявності) в числові і нормування числових даних в діапазоні $[0,1]$. Перетворення категоріальних даних в числові може проводитися зважуванням або упорядкуванням. Зважування проводиться експертом предметної області шляхом надання числових значень категоріальним атрибутам. Упорядкування не вимагає залучення експерта. При її використанні кожного значення категоріального атрибуту приписується порядковий номер. У разі сумніву в правильності використання стратегії упорядкування і неможливості використовувати знання експерта, можливо, краще виключити атрибут з розгляду. Нормування числових даних необхідно, щоб кожен з атрибутів мав рівну вагу при порівнянні даних. У разі, коли вага атрибутів різний, це необхідно враховувати при нормуванні

Крок 3. Визначення кількості кластерів та первинна кластеризація. На основі характеристик об'єктів кластеризації будується мінімальне остовне дерево та остовний ліс (з використанням процедур MSF і МННК- k) та виділяються первинні кластерні центри. Потім реалізується ітераційна процедура уточнення центрів кластерів і визначається сукупність елементів кластерів (на основі обчислення ступеня належності об'єкта кластеру).

При цьому для кожного елемента вхідного масиву спостережень (X) визначається нормальна міра подібності такого вигляду:

$$\mu_y(x) = 1 - \frac{d(y, x)}{\max_{z \in X}(d(y, z))}, \quad (15)$$

де $x, y, z \in X$.

Це породжує нечіткі множини точок близьких до y (y може бути й центром кластеру). Переваги нормальної міри подібності полягають в тому, що для кожного $x \in X$ існує, принаймні, один елемент даних, який абсолютно схожий з ним, для якого $\mu_y(x) = 1$, і один елемент даних, що максимально відрізняється від x , для якого $\mu_y(x) = 0$.

За допомогою нормальної міри можна визначити відносну міру подібності двох елементів даних відносно третього (3).

$$\xi_y(x, z) = 1 - |\mu_y(x) - \mu_y(z)|, \quad (16)$$

де $x, y, z \in X$, а μ_y – нормальна мера подібності.

Через відносну міру подібності можна визначити міру подібності двох зразків даних на всій множині X , що відповідає такому твердженню: “якщо два елемента подібні щодо y_1 і подібні відносно y_2 , то два елементи даних подібні щодо всієї множини X ”. Таким чином може бути побудовано нечітке відношення толерантності, яке показує подібність між об'єктами з X [8].

Відзначимо, що кластери, які формуються на етапі первинної кластеризації, мають сферичну форму. Це дозволяє прийняти припущення про те, що всередині кожного кластера спостерігається типова щільність об'єктів, яка значно більша за густину об'єктів за межами кластера. Більш того, щільність в областях шуму нижче, ніж щільність в будь-якому з кластерів. З цього випливає, що для кожного об'єкта кластера сусідство заданого радіуса має містити, принаймні, мінімальну кількість об'єктів, тобто щільність в сусідстві повинна перевищувати деяке порогове значення. Таким чином, експериментальним шляхом, задаючи різні пороги та аналізуючи щільність сусідства кожного об'єкта, можна дослідити весь простір об'єктів, а ті об'єкти, які не увійшли до жодного з кластерів, слід вважати шумом.

Для кожного з c кластерів, отриманих під час первинної кластеризації, визначаємо лінійні (або трапецієподібні) функції належності, використовуючи фільтрацію Калмана (ФК). ФК здійснює рекурсивне оцінювання вектора параметрів, що відповідає рівнянню прямої та мінімізує деякий квадратичний критерій. На кожному кроці такого оцінювання використовуються оцінки на попередньому кроці і нові входні дані (в нашому випадку входом є дискретні значення функцій належності). Розглянемо докладніше відповідну процедуру.

Нехай за результатами спостережень параметрів ОД отримано $2cn$ -вимірних наборів даних, кожен з яких представляє лінійну частину деякого нечіткого набору дискретних даних. Лінійна частина виділяється шляхом α -зрізу розглянутого набору. В результаті для кожного набору формуються $2cn$ векторів параметрів. У кожному з цих наборів отримуємо N_j дискретних значень, де j означає належність до j -го набору. При цьому кожен набір в моменті часу k_j може бути промодельований таким рівнянням:

$$\begin{aligned} y_{k_j}^j &= a^j x_{k_j} + b^j + v_{k_j} = \\ &= \begin{bmatrix} x_{k_j} & 1 \end{bmatrix} \begin{bmatrix} a^j \\ b^j \end{bmatrix} + v_{k_j} = C_{k_j}^j \theta_{k_j}^j + v_{k_j}, \\ j &= 1, 2, \dots, 2cn, \quad k_j = 1, 2, \dots, N_j, \end{aligned}$$

де $C_{k_j}^j$ – вектор спостережень; $\theta_{k_j}^j = [a^j b^j]^T$ – вектор параметрів; v_{k_j} – шум вимірювань.

Для спрощення будемо позначати далі k_j як k .

В рівнянні ТС представимо вектор $\theta_{k_j}^j$ як змінну станів, що дозволяє перейти до такого рівняння станів:

$$\theta_{k_j}^j = A^j \theta_{k-1}^j + w_{k-1}^j, \quad j = 1, 2, \dots, 2cn, \quad (17)$$

де A^j – перехідна матриця станів розмірності (2×2) ; w_{k-1}^j – шум станів; $\theta_{k_j}^j$ – значення змінної станів в момент k .

Шуми станів і вимірювань передбачаються статистично незалежними і можуть бути розглянуті як білий шум з нульовим математичним очікуванням і обмеженими дисперсіями. Рекурентні процедури перерахунку змінних стану і прогнозу виходу мають в цьому разі такий вигляд:

$$\begin{aligned} \hat{\theta}_{k/k-1}^j &= A^j \hat{\theta}_{k-1}^j, \\ \hat{y}_k^j &= C_k^j \hat{\theta}_{k/k-1}^j. \end{aligned} \quad (18)$$

Крок 4. Нечітка кластеризація даних з використанням метода ГК. На цьому кроці за допомогою алгоритму ГК здійснюється уточнення форми та напрямів c кластерів, сформованих на попередньому кроці. Обчислювальна схема алгоритму ГК полягає у поетапному виконанні операцій, описаних нижче.

Етап 1. Визначаємо центри кластерів:

$$v_i^l = \frac{\sum_{k=1}^N (\mu_{ik}^{(l-1)})^m z_k}{\sum_{k=1}^N (\mu_{ik}^{(l-1)})^m} \quad 1 \leq i \leq c.$$

Етап 2. Визначаємо відстані між елементами множини Z та центрами кластерів відповідно до рівняння (5).

Етап 3. Формуємо матрицю розбиття:

якщо $D_{ikA_i} > 0$ для $1 \leq i \leq c$, $1 \leq k \leq N$, то

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^c \left(\frac{D_{ikA_i}}{D_{jkA_j}} \right)^{\frac{2}{m-1}}};$$

якщо $D_{ikA_i} < 0$, то $\mu_{ik}^{(l)} = 0$.

У загальному випадку $\mu_{ik}^{(l)} \in [0, 1]$ та $\sum_{i=1}^c \mu_{ik}^{(l)} = 1$.

Умовою зупинки алгоритму є виконання наступної нерівності:

$$\|U^{(l)} - U^{(l-1)}\| < \varepsilon.$$

Структура коваріаційної матриці F_i остаточно сформованого i -го кластера містить інформацію про його форму і орієнтацію.

Позначимо j -ті власні числа та j -ті власні вектори матриці F_i як λ_{ij} і Φ_{ij} відповідно. Проранжируємо власні числа:

$$\lambda_{i_1} \geq \lambda_{i_2} \geq \dots \geq \lambda_{i_n}.$$

Найменший з власних векторів цієї матриці відповідає перпендикуляру до напрямку кластера, в той час, як найбільший власний вектор визначає напрямки кластера. Лінеаризовані підмножини вхідної сукупності даних можуть бути представлені гіперповерхнями. Власний вектор, що відповідає найменшому власному числу, визначає перпендикуляр до гіперповерхні та може використовуватися для локальної лінійної моделі.

Власні вектори $\Phi_{i_1}, \Phi_{i_2}, \dots, \Phi_{i_{(n-1)}}$ генерують лінійний підпростір X і Y , що відповідає i -му кластеру. Власний n -й вектор Φ_{i_n} буде нормальним (перпендикулярним) по відношенню до цього лінійного підпростору. Так як λ_{i_n} є найменшим власним числом, то Φ_{i_n} є найменшим власним вектором. При цьому дані розмірності n можуть бути представлені комбінацією $(n-1)$ або меншої кількості змінних. У разі, коли всі власні числа мають приблизно рівні значення, важко знайти відношення між входом і виходом регресійної моделі.

Відзначимо, що первинні кластери сферичної форми, сформовані на кроці 3, уточнюються на кроці 4 за допомогою алгоритму ГК та можуть приймати різні форми і напрямки.

Крок 5. Оцінювання якості кластеризації. Для оцінювання якості нечіткої кластеризації в запропонованому методі можуть бути використані критерії, наведені в [8]. Перший з цих критеріїв — коефіцієнт розбиття:

$$PC = \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|C|} u_{ij}^2}{|X|}, \quad (19)$$

де u_{ij} — відповідний елемент матриці належності, X — вхідна множина, C — множина кластерів. Цей критерій приймає значення з $[|C|^{-1}, 1]$, причому значенням $(|C|^{-1})$ відповідає найгірший випадок розбиття (максимальна невизначеність), а значенням 1 — максимально чітке розбиття. Відзначимо, що на малих значеннях кількості кластерів коефіцієнт розбиття може давати помилкові результати, що пов'язано з його областю значень. Не змінюючи характеру критерію, його область значень доцільно змістити таким чином, щоб залежність від кількості кластерів була пов'язана не з початком зазначеного відрізка, а з його закінченням. Це може бути зроблено шляхом

вирахування з коефіцієнта розбиття величини $|C|^{-1}$. При цьому отримується модифікований критерій розбиття з областю визначення $[0, (|C| - 1)/|C|]$:

$$PC_M = \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|C|} u_{ij}^2}{|X|} - \frac{1}{|C|}. \quad (20)$$

Більш універсальним та традиційним є такий критерій ефективності методів розбиття:

$$PI = \sum_{j=1}^{|C|} \sum_{i=1}^{|X|} u_{ij}^2 \left(d^2(c_j, \bar{x}) - d^2(x_i, c_j) \right) = \sum_{j=1}^{|C|} \sum_{i=1}^{|X|} u_{ij}^2 d^2(c_j, \bar{x}) - \sum_{j=1}^{|C|} \sum_{i=1}^{|X|} u_{ij}^2 d^2(x_i, c_j), \quad (21)$$

де $\bar{x}$ — середнє значення елементів вхідної множини, c_j — центр кластера j .

Перша частина критерію (21) відображує міжкластерні відмінності (чим вони більше, тим вище якість кластеризації), а друга — внутрішньокластерні відмінності (чим вони менше, тим вище якість кластеризації). Таким чином, чим більше сумарне значення критерію (21), тим краще виконано кластеризацію.

Результати моделювання та перспективи продовження досліджень

Розглянемо приклад моделювання вхідної нелінійної тестової функції за допомогою лінійних підмоделей, використовуючи запропонований підхід до нечіткої кластеризації. Нехай нелінійна функція має вигляд:

$$y = \begin{cases} 70 \sin(0.04x) + 4v, & \text{якщо } x \leq 152, \\ 4v, & \text{якщо } x > 152, \end{cases} \quad (22)$$

де v — гаусівський білий шум.

Пари даних для ідентифікації, $i = 1, \dots, 500$ зведемо в матрицю даних виду:

$$Z = \begin{bmatrix} x_1 & x_2 & \dots & x_{500} \\ y_1 & y_2 & \dots & y_{500} \end{bmatrix}, \quad (23)$$

а потім здійснимо їх кластеризацію за розглянутим вище алгоритмом. На попередньому етапі кластеризації (крок 3) було визначено 6 сферичних кластерів. Алгоритм ГК (крок 4) зійшовся після 15 ітерацій і дозволив виділити остаточні кластери, форма та орієнтація яких визначається отриманими значеннями власних чисел та векторів (табл. 1).

Максимальна дисперсія відповідає кластеру C_4 , значить, найменше власне число цього кластера має бути найбільшим серед найменших власних чисел інших кластерів, що підтверджується даними табл. 1 ($\lambda_{42} = 28$; $\lambda_{42} > \lambda_{32} > \lambda_{12} > \lambda_{22}$).

Таблиця 1
Центри та власні числа сформованих кластерів

Кластери	Центри	Власні числа	Власні вектори
C1	(21.4552 53.3003)	11.2 746	(-0.8960 0.4441)
C2	(78.4121 0.4971)	2.9 2804.7	(-0.9382 0.3461)
C3	(140.8354 45.3985)	18.9 1176.9	(-0.8716 0.4902)
C4	(239.6839 0.1439)	1297.3 28	(0.0000 1.0000)

Орієнтація кластерів задана напрямками власних векторів. У розглянутому прикладі найменшим власним вектором кластера C4 є вектор $\Phi_{42} = (0.0 \quad 1.0)^T$. Відзначимо, що цей вектор перпендикулярний осі абсцис x , тобто кластер паралельний цій осі.

Результати моделювання підтверджують можливість ефективного здійснення нечіткої кластеризації даних, що відповідають нелінійній залежності “вхід – вихід”, за розглянутим методом.

Також було розглянуто можливість використання запропонованого методу в системі нечіткого діагностування стану параметрів газоперекачувальних агрегатів (ГПА) компресорних станцій газотранспортної системи.

Для організації ефективного функціонування систем керування компресорною станцією доцільно застосовувати системи діагностики стану параметрів з елементами штучного інтелекту. Визначення нелінійної моделі такої діагностики може бути достатньо складним завданням (зокрема, для ТП КС). Альтернатива вирішення цієї проблеми полягає у використанні апроксимуючих можливостей нечітких систем для подання нелінійних динамічних процесів. Модель нечіткого діагностування та прогнозування стану параметрів ГПА, як і будь-яка модель з нечітким описом параметрів, будується на основі формалізації суб'єктивних знань експертів.

Після зміни завдання з вихідного тиску оператор для підвищення тиску поступово збільшує частоти обертання агрегатів та оцінює після закінчення перехідних процесів для кожного ГПА запаси регулювання за параметрами системи керування ГПА, які контролюється, і стан яких має діагностуватися та прогнозуватися. При практичному обнулінні сигналу неузгодженості за будь-яким параметром ГПА, подальше навантаження цього ГПА неможливо, тобто необхідно перерозподіляти навантаження між агрегатами.

Відомо, що для різних діапазонів вихідного тиску (ступеня стиснення газу) динамічні властивості ГПА можуть змінюватися (через нелінійність стати-

чної залежності по каналу “швидкість обертів валу ГПА – вихідний тиск”) [16]. У зв'язку з цим було здійснено поділ тестової сукупності реєстраційних даних зміни вихідного тиску на 5 кластерів, що дозволило після вибору функцій належності сформувати 5 лінійних правил діагностування. Порівняння результатів використання FCM-алгоритму та розглянутого в статті гібридного методу для такої кластеризації свідчить про безумовні переваги останнього (значення критерію якості кластеризації (21) дорівнюють 0.71 та 0.88 відповідно).

Слід відзначити перспективи подальшого розвитку досліджень наведеного комбінованого методу наведеного комбінованого методу кластеризації з метою надання йому адаптивних властивостей. Розробка методології адаптивної кластеризації, основою якої може бути запропонований метод, дозволить здійснити автоматичний вибір локальних критеріїв оцінки якості формування кластерів, стабілізацію процедури пошуку їх центрів тощо (за умов можливої зміни властивостей ОД).

Крім того, доцільно розглянути ефективність застосування адаптивних варіантів алгоритмів нечіткої кластеризації ГК (імовірнісного і можливісного) в системі діагностування (з автоматичним налаштуванням коку навчання), які були б здатні обробляти викиди без попередньої обробки даних і адаптуватися до мінливих характеристик ОД.

Висновки

У статті досліджено можливість комбінованого використання методів нечіткої кластеризації в системах технічної діагностики (зокрема, діагностики стану основних параметрів технологічних процесів).

За результатами аналізу сучасного стану цієї проблеми запропоновано комбінований метод нечіткої кластеризації, що передбачає послідовну реалізацію таких кроків: вибір структури діагностичної моделі; збір даних та їх підготовка до кластеризації; визначення кількості кластерів з використанням MSF, МННК- k та фільтрації Калмана; нечітка кластеризація даних з використанням метода ГК; оцінювання якості кластеризації.

Комбінований підхід дозволяє після попереднього використання графових методів з визначенням доцільної кількості кластерів здійснювати уточнену нечітку кластеризацію за одним з базових методів розбиття. Такий підхід сприяє скороченню процедури дослідження вхідного набору даних та знижує загальний час проведення діагностування.

Результати моделювання підтверджують можливість практичного використання запропонованого методу (зокрема в системі нечіткого діагностування стану параметрів газоперекачувальних агрегатів компресорних станцій газотранспортної системи).

Перспективним напрямком продовження проведених досліджень є розробка та аналіз адаптивних варіантів алгоритмів нечіткої кластеризації, які були

б здатні адаптуватися до мінливих характеристик ОД в реальному часі.

Список літератури

1. Eltyshv D. Influence of fuzzy clustering on the accuracy of electrical equipment diagnostic models / D. Eltyshv, K. Gnutova // *Proceedings of the 6th International Conference on Applied Innovations in IT*. – 2018. – P. 23-28.
2. Алексеев А.А. Алгоритмы нечеткой кластеризации в задачах диагностики технических систем / А.А. Алексеев, Ю.Б. Попова, М.Ю. Шестопалов // *Известия вузов. Технические науки*. – 2012. – № 3. – С. 3-7.
3. Shtovba S.D. Analyzing the Criteria for Fuzzy Classifier Learning / S.D. Shtovba, O.D. Pankevich, A.V. Nagorna // *Automatic Control and Computer Sciences*. – 2015. – Vol. 49. – № 3. – P. 123-132. <https://doi.org/10.3103/S0146411615030098>.
4. Назаров А.О. Модель и метод концептуальной кластеризации объектов, характеризующихся нечеткими параметрами / А.О. Назаров // *Фундаментальные исследования*. – 2014. – № 9. – С. 993-999.
5. Lee S. Various Validity Indices for Fuzzy K-means Clustering / S. Lee, J. Kim, Y. Jeong // *Korean Management Review*. – 2017. – 46(4). – P. 1201-1226.
6. Kumar P. Probabilistic query generation and fuzzy c-means clustering for energy-efficient operation in wireless sensor networks. / P. Kumar, A. Chaturvedi // *International Journal of Communication Systems*. – 2016. – 29 (8). – P. 1439-1450.
7. Нечеткая кластеризация временных рядов с неравномерными и асинхронными тактами квантования / Е.В. Бодянский, Е.А. Винокурова, Д.Д. Пелешко, И.О. Кобылин, О.А. Кобылин // *Системи обробки інформації*. – 2017. – № 5(151). – С. 47-54. <https://doi.org/10.30748/soi.2017.151.06>.
8. Елизаров С.И. Разработка и исследование методов и алгоритмов кластеризации для систем анализа данных [Электронный ресурс] / С.И. Елизаров. – Режим доступа: <http://www.eltech.ru/education/aspir/SIElizarov.doc> (дата обращения 11.09.2016).
9. Babichev S. Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise / S. Babichev, V. Lytvynenko, M. Taif // *Radio Electronics, Computer Science, Control*. – 2016. – № 4. – С. 54-60. <https://doi.org/10.15588/1607-3274-2016-4-7>.
10. Адаптивная нечеткая кластеризация данных на основе метода Густафсона–Кесселя УСМ / Е.В. Бодянский, Б.В. Колчигин, В.В. Волкова, И.П. Плисс // *Управляющие системы и машины*. – 2013. – № 2. – С. 40-46.
11. Удовенко С.Г. Метод идентификации нелинейных объектов с использованием нечеткой кластеризации и фильтрации Калмана / С.Г. Удовенко, Р. Альхайек // *Автоматизированные системы управления и приборы автоматики*. – 2010. – Вып. 152. – С. 65-74.
12. Нейский И.М. Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-mean / И.М. Нейский, А.Ю. Филиппович // *Известия вузов. Проблемы полиграфии и издательского дела*. – 2009. – №3. – С. 48-61.
13. Bodyanskiy Ye. An evolving neuro-fuzzy system for online fuzzy clustering / Ye. Bodyanskiy, O. Tyshchenko, D. Kopaliani // *Proc. Xth Int. Scientific and Technical Conf. "Computer Sciences and Information Technologies (CSIT'2015)"*. – P. 158-161.
14. Malick J. Solving k-cluster problems to optimality with semidefinite programming / J. Malick, F. Roupin // *Mathematical programming*. – 2012. – № 136(2). – P. 279-300.
15. Krislock N. Computational results of a semidefinite branch-and-bound algorithm for k-cluster / N. Krislock, J. Malick, F. Roupin // *Computers & Operations Research*. – 2016. – № 66. – P. 153-159.
16. Удовенко С.Г. Керування агрегатами компресорної станції з використанням нечітких інверсних моделей / С.Г. Удовенко, Д.В. Келембет // *Біоніка інтелекту*. – 2018. – Вып. 2(91). – С. 25-30.

References

1. Eltyshv, D. and Gnutova, K. (2018), Influence of fuzzy clustering on the accuracy of electrical equipment diagnostic models, *Proceedings of the 6th International Conference on Applied Innovations in IT*, pp. 23-28.
2. Alekseyev, A.A., Popova, Yu.B. and Shestopalov, M.Yu. (2012), "Algoritmy nechetkoy klasterizatsii v zadachakh diagnostiki tekhnicheskikh sistem" [Fuzzy clustering algorithms in the problems of diagnosing technical systems], *University proceedings. Technical science*, No. 3, pp. 3-7.
3. Shtovba, S.D., Pankevich, O.D. and Nagorna, A.V. (2015), Analyzing the Criteria for Fuzzy Classifier Learning, *Automatic Control and Computer Sciences*, Vol. 49, No. 3, pp. 123-132. <https://doi.org/10.3103/s0146411615030098>.
4. Nazarov, A.O. (2014), "Model' i metod kontseptual'noy klasterizatsii ob'yektov, kharakterizuyemykh nechetkimi parametrami" [Model and method of conceptual clustering of objects characterized by fuzzy parameters], *Fundamental research*, No. 9, pp. 993-999.
5. Lee, S., Kim, J. and Jeong, Y. (2017), Various Validity Indices for Fuzzy K-means Clustering, *Korean Management Review*, No. 46(4), pp. 1201-1226.
6. Kumar, P. and Chaturvedi, A. (2016), Probabilistic query generation and fuzzy c-means clustering for energy-efficient operation in wireless sensor networks, *International Journal of Communication Systems*, No. 29(8), pp. 1439-1450.

7. Bodyanskiy, Ye.V., Vinokurova, Ye.A., Peleshko, D.D., Kobylin, I.O. and Kobylin, O.A. (2017), "Nechetskaya klasterizatsiya vremennykh ryadov s neravnomernymi i asinkhronnymi taktami kvantovaniya" [Fuzzy clustering of time series with non-uniform and asynchronous quantization clocks], *Information Processing Systems*, No. 5(151), pp. 47-54. <https://doi.org/10.30748/soi.2017.151.06>.
8. Yelizarov, S.I. (2016), "Razrabotka i issledovaniye metodov i algoritmov klasterizatsii dlya sistem analiza dannykh" [Development and research of clustering methods and algorithms for data analysis systems], available at: www.eltech.ru/education/aspir/SIElizarov.doc (accessed 09.11.2016).
9. Babichev, S., Lytvynenko, V. and Taif, M. (2016), Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise, *Radio Electronics, Computer Science, Control*, No. 4, pp. 54-60. <https://doi.org/10.15588/1607-3274-2016-4-7>.
10. Bodyanskiy, Ye.V., Kolchigin, B.V., Volkova, V.V. and Pliss, I.P. (2013), "Adaptivnaya nechetkaya klasterizatsiya dannykh na osnove metoda Gustafsona–Kesselya" [Adaptive fuzzy data clustering based on the Gustafson – Kessel], *Control systems and machines*, No. 2, pp. 40-46.
11. Udovenko, S.G. and Al'khayek, R. (2010), "Metod identifikatsii nelineynykh ob'yektov s ispol'zovaniyem nechetkoy klasterizatsii i fil'tratsii Kalmana" [Method for identification of nonlinear objects using fuzzy clustering and Kalman filtering], *Automated control systems and automation devices*, No. 152, pp. 65-74.
12. Neyskiy, I.M. and Filippovich, A.Yu. (2009), "Metodika adaptivnoy klasterizatsii faktograficheskikh dannykh na osnove integratsii algoritmov MST i Fuzzy C-mean" [Adaptive clustering of factual data based on the integration of MST and Fuzzy C-mean algorithms], *University proceedings. Problems of printing and publishing*, No. 3, pp. 48-61.
13. Bodyanskiy, Ye., Tyshchenko, O. and Kopaliani, D. (2015), An evolving neuro-fuzzy system for online fuzzy clustering, *Proc. Xth Int. Scientific and Technical Conf. "Computer Sciences and Information Technologies (CSIT'2015)"*, pp. 158-161.
14. Malick, J. and Roupin, F. (2012), Solving k-cluster problems to optimality with semidefinite programming, *Mathematical programming*, No. 136, pp. 279-300.
15. Krislock, N., Malick, J. and Roupin, F. (2016), Computational results of a semidefinite branch-and-bound algorithm for k-cluster, *Computers & Operations Research*, No. 66, pp. 153-159.
16. Udovenko, S.G. and Kelembet, D.V. (2018), "Keruvannya ahrehatamy kompresornoyi stantsiyi z vykorystanniam nechitkykh inversnykh modeley" [Control of compressor station units using fuzzy inverse models], *Bionics of intelligence*, No. 2(91), pp. 25-30.

Надійшла до редколегії 13.01.2020

Схвалена до друку 11.02.2020

Відомості про авторів:

Удовенко Сергій Григорович

доктор технічних наук професор
завідувач кафедри
Харківського національного
економічного університету ім. С. Кузнеця,
Харків, Україна
<https://orcid.org/0000-0001-5945-8647>

Келембет Даніл Вадимович

аспірант кафедри
Харківського національного університету
радіоелектроніки,
Харків, Україна
<https://orcid.org/0000-0002-5853-4166>

Тесленко Олег Володимирович

кандидат технічних наук доцент
доцент кафедри Харківського національного
економічного університету ім. С. Кузнеця,
Харків, Україна
<https://orcid.org/0000-0003-3105-9323>

Information about the authors:

Serhii Udovenko

Doctor of Technical Sciences Professor
Head of the Department
of Simon Kuznets Kharkiv
National University of Economics,
Kharkiv, Ukraine
<https://orcid.org/0000-0001-5945-8647>

Danil Kelembet

Doctoral Student
of Kharkiv National University
of Radio Electronics,
Kharkiv, Ukraine
<https://orcid.org/0000-0002-5853-4166>

Oleg Teslenko

Candidate of Technical Sciences
Associate Professor Senior Lecturer of Simon Kuznets
Kharkiv National University of Economics,
Kharkiv, Ukraine
<https://orcid.org/0000-0003-3105-9323>

**КОМБИНИРОВАННЫЙ МЕТОД НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ
В СИСТЕМАХ ТЕХНИЧЕСКОЙ ДИАГНОСТИКИ**

С.Г. Удовенко, Д.В. Келембет, О.В. Тесленко

В статье приведены результаты анализа методов нечеткой кластеризации данных текущих наблюдений в технических системах. Предложен и протестирован комбинированный подход к нечеткой кластеризации данных с использованием комбинации алгоритма Густафсона-Кесселя, метода остовного леса и фильтра Калмана, который может быть использован в системах технической диагностики (в частности, диагностики состояния основных параметров технологических процессов). Рассмотрена возможность дальнейшего расширения функциональных возможностей приведенного комбинированного метода кластеризации с приданием ему адаптивных свойств. Приведены результаты тестирования и определены преимущества предлагаемого подхода.

Ключевые слова: нечеткая кластеризация, диагностика состояния параметров, метод остовных деревьев, фильтрация Калмана.

**COMBINED METHOD OF FUZZY DATA CLUSTERING
IN TECHNICAL DIAGNOSTIC SYSTEMS**

S. Udovenko, D. Kelembet, O. Teslenko

The article investigates the possibility of combined use of fuzzy clustering methods in systems of technical diagnostics (in particular, diagnostics of the state of the basic parameters of technological processes). According to the results of the analysis of the current state of this problem, a combined method of some clustering is proposed, which implies a consistent implementation of the following steps: choice of the structure of the diagnostic model; data collection and preparation for clustering; determining the number of clusters using MSF (Minimum Spanning Forest,), solving k-Cluster problems to optimality with semidefinite programming and Kalman filtering; fuzzy classification of data using the Gustafson - Kessel method; evaluation of the quality of clustering.

The combined approach allows, after using graph methods beforehand, to determine the appropriate number of clusters, to perform refined fuzzy clustering using one of the basic partitioning methods. This approach shortens the procedure for examining the incoming data set and reduces the overall time for diagnosis.

The simulation results confirm the possibility of practical use of the proposed method (in particular, in the system of fuzzy diagnostics of the parameters of gas pumping units of compressor stations of the gas transmission system).

It is worth noting the prospects for further development of the studies of the above combined method of the given combined method of clustering in order to give it adaptive properties. The development of an adaptive clustering methodology, the basis of which may be a proposed method, will allow the automatic selection of local criteria for assessing the quality of cluster formation, stabilization of the procedure for finding their centers, etc. (subject to possible change in diagnostic object properties).

A promising direction for continued research is the development and analysis of adaptive variants of fuzzy clustering algorithms that can adapt to the changing characteristics of diagnostic object in real time.

Keywords: fuzzy clustering, parameter state diagnostics, spanning tree method, Kalman filtering.