УДК 004.032.26

НЕЧІТКА КЛАСТЕРИЗАЦІЯ МАСИВІВ ДАНИХ З ПРОПУЩЕНИМИ ЗНАЧЕННЯМИ

В.В. Волкова, А.Ю.Шафроненко

Харківський національний університет радіоелектроніки,

val_volkova@ukr.net, alinashafronenko@gmail.com

Запропоновано алгоритм заповнення пропущених значень в масивах даних, що базується на використанні Адаліни. Цей алгоритм дозволяє заповнювати пропуски в режимі on-line. Виконано оцінку якості методів нечіткої кластеризації даних з пропущеними значеннями залежно від алгоритму їх заповнення.
Ключові слова: дані з пропущеними значеннями, кластеризація, алгоритм

The algorithm for filling the missing values in data massive is proposed. This algorithm is based on using the Adaline and allows to close the missing values in on-line mode. An evaluation of quality of fuzzy clustering methods for data with missing values depending on algorithm of they filling is performed.
Keywords: data with missing values, clustering, algorithm

Предложен алгоритм заполнения пропущенных значений в массивах данных, основанный на использовании Адалины. Алгоритм позволяет заполнять пропуски в режиме on-line. Выполнена оценка качества методов нечеткой кластеризации данных с пропущенными значениями в зависимости от алгоритма их заполнения.
Ключевые слова: данные с пропущенными значениями, кластеризация, алгоритм

Вступ

Завдання кластеризації даних є важливим елементом загальної проблеми Data Mining, а для її вирішення сьогодні існує безліч підходів і алгоритмів від суто інтуїтивних і евристичних до строго математичних. Разом з тим, у багатьох задачах Data Mining, включаючи кластеризацію, вихідні таблиці даних «об'єкт-властивість» можуть містити порожні клітини (пропуски), інформація в яких з тих чи інших причин відсутня. Задачі відновлення таких пропущених спостережень приділялося достатньо уваги [1], при тому більш ефективними в даній ситуації опинилися підходи, що базуються на математичному апараті м'яких обчислень (обчислювального інтелекту), і, перш за все, штучних нейронних мережах [2]. Разом з тим, відомі підходи до відновлення пропусків і традиційні алгоритми кластеризації працездатні лише у випадках, коли вихідна таблиця задана апріорно і число її рядків або стовпців не змінюється в процесі обробки.

У той же час існує досить широкий клас задач, коли дані надходять на обробку послідовно в on-line режимі, при цьому заздалегідь невідомо, який з оброблюваних векторів-образів може містити пропуски. При цьому процеси ві-

дновлення даних та їх кластеризація повинні протікати паралельно та в реальному часі.

Традиційний підхід до вирішення таких завдань передбачає, що кожне спостереження може відноситися тільки до одного кластеру, хоча більш природною видається ситуація, коли оброблюваний вектор ознак з різними рівнями належності може відноситись одразу до декількох класів. Ця ситуація є предметом розгляду нечіткого кластерного аналізу, що інтенсивно розвивається в даний час [3,4]. У зв'язку з цим у даній роботі розглядається алгоритм заповнення пропущених даних, що базується на принципах лінійної регресії, а також аналізується залежність якості роботи алгоритмів нечіткої кластеризації даних з пропущеними значеннями залежно від алгоритму заповнення пропусків.

1. Алгоритми заповнення пропущених даних

Алгоритми заповнення пропусків розробляються для емпіричних таблиць типу «об'єкт-властивість». Таблиці типу«об'єкт-властивість» – це таблиці, в яких у рядках перелічені об'єкти (наприклад, підприємства), а в стовпцях – їх властивості. Ці таблиці експериментальних даних можуть містити пропуски.

Завдання алгоритмів заповнення пропущених даних полягає в заповненні пропусків у таблицях найбільш правдоподібними значеннями з найменшою похибкою. Схематично алгоритми заповнення пропусків у таблицях даних ілюструє рис. 1.

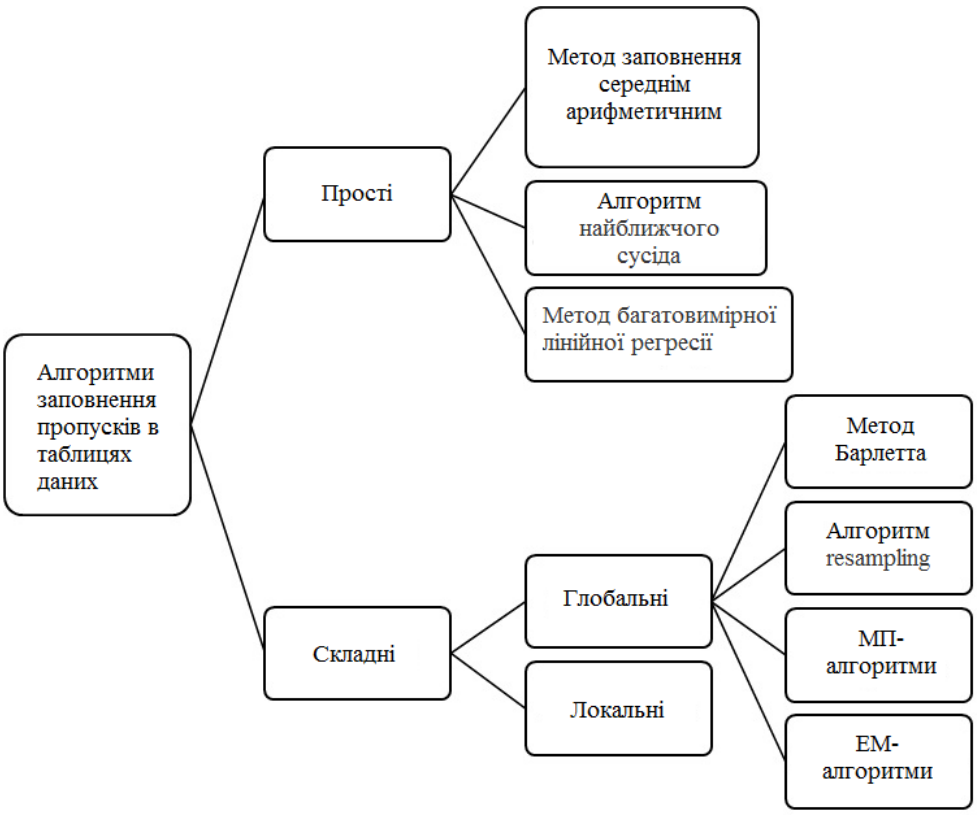

Рис. 1. Алгоритми заповнення пропущених даних

Як видно з рис. 1, алгоритми заповнення пропусків у емпіричних таблицях даних можна умовно розділити на складні та прості. До простих алгоритмів відносяться такі методи заповнення таблиць як: середнє арифметичне, алгоритм найближчого сусіда, метод багатовимірної лінійної регресії. Складні алгоритми, в свою чергу, діляться на дві частини: глобальні та локальні. До глобальних алгоритмів відносяться метод Барлетта, алгоритм resampling, МП-алгоритми, ЕМ-алгоритми, локальний – алгоритм ZET.

2. Постановка задачі

Нехай задана $N \times n$ таблиця «об'єкт-властивість» (табл.1), що містить інформацію про N об'єктів, кожний з яких описується (1×n)-вектором-рядком ознак $\underset{\rightarrow}{x}_i = (x_{i1},...,x_{ip},...,x_{ij},...,x_{in})$. При цьому припускається, що N_G рядків можуть мати по одному пропущеному значенню, а $N_F = N - N_G$ заповнені повністю.

Таблиця 1.

Таблиця виду «об'єкт-властивість».

	1	...	*p*	...	*j*	...	*n*
1	x_{11}	...	x_{1p}	...	x_{1j}	...	x_{1n}
...	...	...	...	...	...	...	...
i	x_{i1}	...	x_{ip}	...	x_{ij}	...	x_{in}
...	...	...	...	...	...	...	...
k	x_{k1}	...	x_{kp}	...	x_{kj}	...	x_{kn}
...	...	...	...	...	...	...	...
N	x_N	...	x_N	...	x_N	...	x_N

,

В процесі обробки таблиці необхідно заповнити пропуски так, щоб відновлені елементи були б у певному сенсі «найбільш правдоподібні» або «близькі» до апріорі невідомих закономірностей, що містяться в таблиці.

3. Алгоритм заповнення пропущених даних на основі Адаліни

Представимо таблицю 1 у вигляді *(N × n)*-матриці *X*, в якій відсутній один елемент x_{kj} або в більш загальному випадку відсутні N_G елементів. Припускається [3], що між стовпцями $\vec{x}_j = (x_{1j},...,x_{ij},...,x_{kj},...,x_{Nj})^T$ існує лінійна кореляція, на підставі якої й проводиться відновлення пропуску за допомогою регресії

$$\hat{x}_{kj} = w_{j0} + w_{j1}x_{k1} + w_{j2}x_{k2} + w_{j,j-1}x_{k,j-1} + w_{j,j+1}x_{k,j+1} + ... + w_{jn}x_{kn}, \qquad (1)$$

або

$$\hat{x}_{kj} = \underset{\rightarrow}{X}_{kj} w_j, \tag{2}$$

де $w_j = (w_{j0}, w_{j1}, ..., w_{jn})^T - (n \times 1)$-вектор параметрів, що підлягають визначенню , $\underset{\rightarrow}{X}_{kj} = (1, x_{k1}, ..., x_{k,j-1}, x_{k,j+1}, ..., x_{kn})$ - $(1 \times n)$-вектор-рядок ознак *k*-того об'єкту без kj-того елементу і з одиницею на першій позиції.

Вектор невідомих параметрів w_j може бути знайдений за допомогою стандартного методу найменших квадратів, для чого з матриці X слід вилучити *k*-й рядок, *j*-й стовпець, додати зліва стовпець з одиниць та на основі отриманої $((N-1) \times n)$-матриці X_j розрахувати оцінки параметрів

$$w_j = (X_j^T \mathrm{X}_j)^+ \mathrm{X}_j^T \overrightarrow{\mathrm{X}}_j \tag{3}$$

де $\overrightarrow{\mathrm{X}}_j = (x_{1j}, ..., x_{ij}, ..., x_{k-1,j}, x_{k+1,j}, ..., x_{Nj})^T$, $(\bullet)^+$ – символ псевдообернення по Муру-Пенроузу [8].

Якщо ж пропуски існують в N_G рядках та в інших стовпцях, з матриці X вилучаються усі ці рядки та на підставі отриманої усіченої $(N_F \times n)$-матриці n раз знаходять вектори параметрів (3) для всіх $j = 1,2,...,n$.

Далі за допомогою рівнянь (1) та (3) заповнюються все пропуски отриманими оцінками $\hat{x}_{kj}$.

Цей алгоритм нескладно поширити на випадок, коли дані про об'єкти в таблицю 1 надходять послідовно об'єкт за об'єктом. При появі $(N+1)$ -го спостереження у вигляді цілком заповненого рядка $\underset{\rightarrow}{X}_{N+1}$ оцінка може бути відкоригована за допомогою рекурентного методу найменших квадратів

$$\begin{cases} w_j(N_F+1) = w_j(N_F) + \dfrac{P_j(N_F)(x_{N+1,j} - \underset{\rightarrow}{\mathrm{X}}_{N+1,j} w_j(N_F))}{1 + \underset{\rightarrow}{\mathrm{X}}_{N+1,j} P_j(N_F) \underset{\rightarrow}{\mathrm{X}}^T_{N+1,j}} \underset{\rightarrow}{\mathrm{X}}^T_{N+1,j}, \\ P_j(N_F+1) = P_j(N_F) - \dfrac{P_j(N_F) \underset{\rightarrow}{\mathrm{X}}^T_{N+1,j} \underset{\rightarrow}{\mathrm{X}}_{N+1,j} P_j(N_F)}{1 + \underset{\rightarrow}{\mathrm{X}}_{N+1,j} P_j(N_F) \underset{\rightarrow}{\mathrm{X}}^T_{N+1,j}}, \end{cases} \tag{4}$$

після чого уточнюються відновлені значення $\hat{x}_{kj}$.

Обробку інформації в режимі послідовного надходження даних зручно організувати на основі нейромережевої системи, основними елементами якої є n паралельно працюючих адаптивних лінійних асоціаторів (ALA), які навчаються за допомогою алгоритму (4). На рис.2 наведена схема цієї системи, яка не потребує додаткових пояснень.

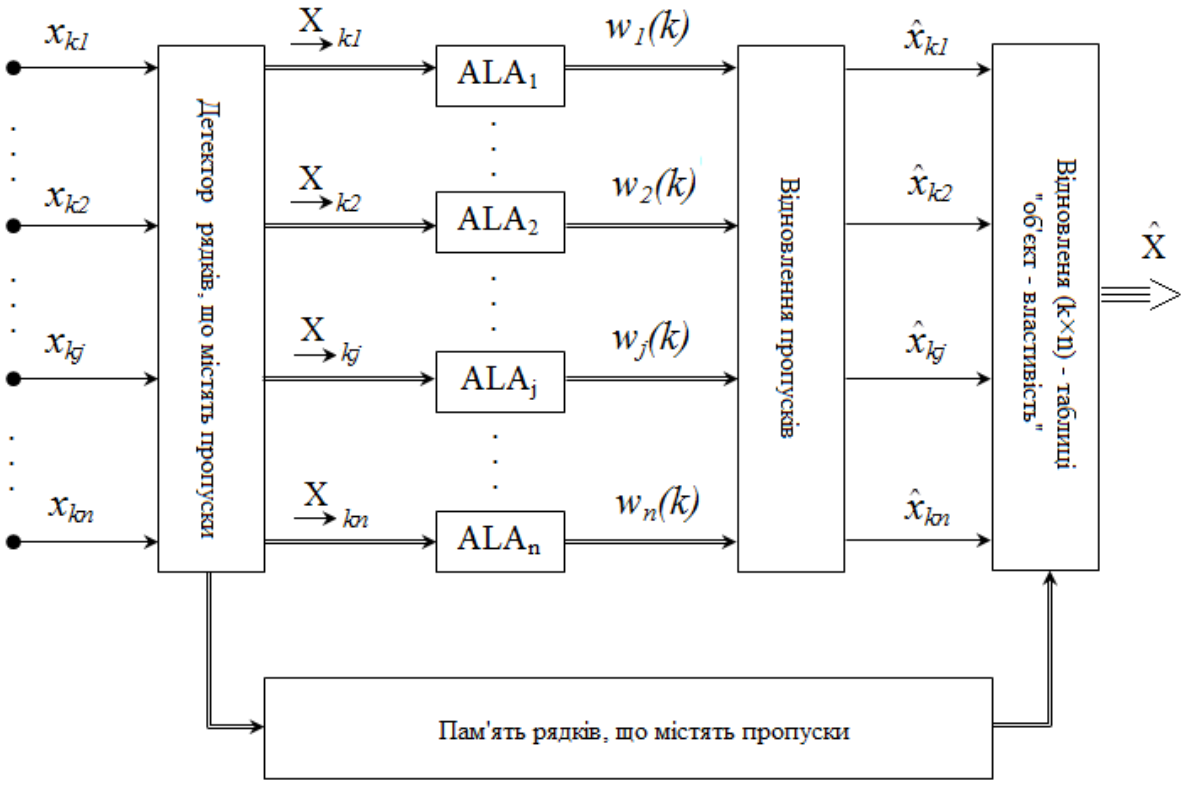

Рис. 2. Адаптивна нейромережева система для відновлення пропусків

Зауважимо тільки, що індекс *k* тут позначає номер об'єкту, характеристики якого в поточний момент часу подаються на обробку.

4. Нечітка кластеризація даних з пропущеними значеннями

Для оцінки якості нечіткої кластеризації даних з пропущеними значеннями було обрано алгоритми fuzzy c-means, Gustafson-Kessel та Gath-Geva [3, 4]. Якість їх роботи перевірялась на вибірках даних UCI-репозиторію, зокрема Iris, Wines та Breast-Cancer-Wisconsin.

У таблиці 2 наведено результати тестувань для вказаних алгоритмів для вибірки даних Wines з відновленими 100 пропущеними значеннями.

Таблиця 2.

Оцінка якості роботи алгоритмів нечіткої кластеризації.

Алгоритм кластеризації	PC		CE		SC	
	AdA	Mean	AdA	Mean	AdA	Mean
Fuzzy c-means	0.5036	0.5025	0.8541	0.8559	1.6255	1.6374
Gustafson-Kessel	0.2755	0.2726	1.3386	1.3441	44.7481	51.0859
Gath-Geva	*0.2542*	*0.2542*	*1.3785*	*1.3785*	*4.4064e-005*	4.4134e-005

Продовження табл. 2

Алгоритм кластеризації	S		XB		DI	
	AdA	Mean	AdA	Mean	AdA	Mean
Fuzzy c-means	0.0119	0.0120	0.0119	0.9692	0.1423	0.1423
Gustafson-Kessel	0.3350	0.3793	0.5717	0.5137	*0.0782*	*0.0782*
Gath-Geva	3.2490e-007	*3.2378e-007*	*53.3548*	52.2223	0.1145	0.1145

У якості критеріїв оцінки якості кластеризації було використано такі характеристики: Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index (S), Xie and Beni's Undex (XB) та Dunn's Index (DI). Розглянуті критерії використовуються для оцінки роботи саме алгоритмів нечіткої кластеризації та показують якість розбиття даних по кластерах.

Як можна бачити з табл.2, запропонований підхід до відновлення пропущених значень має найкращі показники якості розбиття даних на кластери.

Висновки

Розглянуто задачу відновлення пропущених значень у таблицях типу «об'єкт-властивість» у режимі послідовного надходження їх на обробку. Запропоновано алгоритм вирішення цієї задачі на основі нейромережевої системи, що складається з адаптивних лінійних асоціаторів, та дозволяє обробляти дані у on-line режимі. Виконано оцінку якості роботи деяких алгоритмів нечіткої кластеризації даних з пропущеними значеннями в залежності від алгоритмів їх відновлення. Експериментальні дослідження показали, що найбільш ефективною є кластеризація даних, пропущені значення яких були відновлені за допомогою запропонованого алгоритму.

Література

1. Gorban A., Kegl B., Wunsch B., Zinovyev A. (Eds.) Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering, Vol. 58. – Berlin –Heidelberg– New York: Springer, 2007, – 330 p.
2. Bishop C.M. Neural Networks for Pattern Recognition. – Oxford: Clarendon Press, 1995, – 482 p.
3. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. – N.Y.: Plenum Press, 1981, – 272 p.
4. Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. – Chichester, Wiley, 1999, – 300 p.