УДК 519.7:004.8

Плісс І. П., Шевякова А. Ю., Шевякова Ю. Ю.

НЕЙРОМЕРЕЖЕВЕ ВІДНОВЛЕННЯ ПРОПУСКІВ У ТАБЛИЦЯХ ДАНИХ

Запропоновано адаптивну нейромережеву систему для відновлення пропусків, що дозволяє відновлювати пропуски в таблицях «об'єкт-властивість» в режимі послідовного надходження даних на обробку. Ця система характеризується високою швидкодією і простотою чисельної реалізації, здатна обробляти інформацію в режимі реального часу.

Ключові слова: *адаптивна нейромережева система, відновлення пропусків.*

Предложена адаптивная нейросетевая система для восстановления пробелов, позволяющая восстанавливать пробелы в таблицах «объект-свойство» в режиме последовательного поступления данных на обработку. Данная система характеризуется високим быстродействием и простотой численной реализации, способна обрабатывать информацию в режиме реального времени.

Ключевые слова: *адаптивная нейросетевая система, восстановление пробелов.*

An adaptive neural network system for restoration gaps that can restore gaps in tables «object-property» in the serial receiving data for processing. This system has high speed and simple numerical realization, can process information in real time.

Key words: *adaptive neural network system, the restoration of gaps.*

Вступ

В багатьох задачах Data Mining, пов'язаних з обробкою емпіричних кількісних даних таблиці «об'єкт – властивість» можуть містити порожні клітини (пропуски), інформація у котрих по тим чи іншим причинам відсутня. Задачею відновлення таких пропущених спостережень приділялось достатньо уваги [1-3], при цьому найефективнішими в даній ситуації виявилися підходи, засновані на математичному апараті і, перш за все, штучних нейронних мережах [4-7]. Разом з тим, описані підходи до відновлення пропусків працездатні лише у випадках, коли вихідна таблиця даних задана апріорно та кількість її рядків або стовпців не може змінюватись в процесі обробки. Існує велика кількість задач, коли дані поступають на обробку послідовно у реальному часі, при цьому завчасно невідомо, котрий з векторів-образів, що оброблюються містять пропуски. Розгляду даної ситуації і присвячена ця робота.

1. Постановка задачі

Нехай задана $N \times n$ таблиця «об'єкт-властивість» виду

Таблиця 1

	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_{N1}	…	x_{Np}	…	x_{Nj}	…	x_{Nn}

що містить інформацію про N об'єктів, кожний з яких описується $(1 \times n)$ – 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 пропуску, а $N_F = N - N_G$ заповнені повністю. 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.

2. Алгоритм вірішення задачі

Представимо таблицю 1 у вигляді $(N \times n)$ – матриці X, в якій відсутній один елемент x_{kj} або в більш загальному випадку відсутні N_G елементів. Припускається [1], 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

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

або

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

де $w_j = (w_{j0}, w_{j1}, ..., w_{jn})^T$ – (n × 1) – вектор параметрів, що підлягають визначенню, $\underline{X}_{kj} = (1, x_{k1}, ...,$ $\underline{X}_{kj} = (1, x_{k1}, ..., x_{k,j-1}, x_{k,j+1}, ..., x_{kn})$ 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*-ий рядок, *j*-ий стовпець, додати зліва стовпець з одиниць та на основі отриманої ((*N* – 1) × *n*) матриці X_j розрахувати оцінки параметрів

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

де $\vec{\mathrm{X}}_j = (x_{1j}, ..., x_{ij}, ..., x_{k-1,j}, x_{k+1,j}, ..., x_{Nj})^T$, $(\bullet)^+$ – 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 × *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 $\hat{x}_{kj}$.

Розглянутий алгоритм зручно представити у вигляді схеми, наведеної на рис. 1

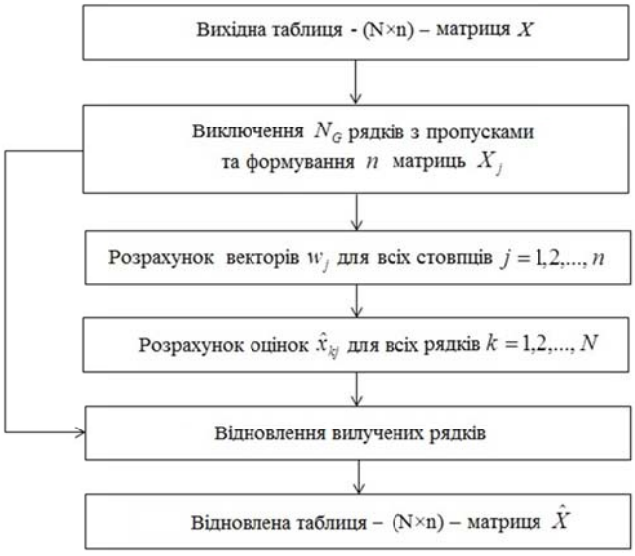

Рис. 1. Пакетний алгоритм заповнення пропусків

Цей алгоритм нескладно поширити на випадок, коли дані про об'єкти в таблицю 1 надходять послідовно об'єкт за об'єктом.

Нехай без втрати спільності від початку задано інформацію у вигляді таблиці 1 з *N* рядків, з яких N_F заповнені повністю. На підставі цих даних може бути побудована оцінка

$$w_j(N_F) = (X_j^T(N_F) X_j(N_F))^+ X_j^T(N_F) \vec{\mathrm{X}}_j(N_F), \tag{4}$$

за допомогою якої відновлюється вихідна таблиця. При появі (*N* + 1)-го спостереження у вигляді цілком заповненого рядка $\underline{x}_{N+1}$ оцінка (4) може бути відкоригована за допомогою рекурентного методу найменших квадратів

$$\begin{cases} w_j(N_F+1) = w_j(N_F) + \dfrac{P_j(N_F)(x_{N+1,j} - \underline{\mathrm{X}}_{N+1,j} w_j(N_F))}{1 + \underline{\mathrm{X}}_{N+1,j} P_j(N_F) \underline{\mathrm{X}}_{N+1,j}^T} \underline{\mathrm{X}}_{N+1,j}^T, \\ P_j(N_F+1) = P_j(N_F) - \dfrac{P_j(N_F) \underline{\mathrm{X}}_{N+1,j}^T \underline{\mathrm{X}}_{N+1,j} P_j(N_F)}{1 + \underline{\mathrm{X}}_{N+1,j} P_j(N_F) \underline{\mathrm{X}}_{N+1,j}^T}, \end{cases} \tag{5}$$

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

Якщо ж (*N* + 1)-й рядок має пропуск, він пропускається і алгоритм (5) очікує появи повністю заповненого рядка, наприклад $\underline{x}_{N+2}$, після чого розраховується $w_j(N_F + 1)$ за допомогою (*N* + 2)-го спостереження і коригуються всі значення $\hat{x}_{kj}$, включно з $\hat{x}_{N+1,j}$.

На початкових етапах оброблення таблиці 1, коли кількість повністю заповнених рядків N_F порівняно з кількістю стовпців *n* є недостатньою, оцінки, отримані за допомогою методу найменших квадратів, характеризуються низькою точністю. У цій ситуації для послідовної обробки більш ефективним є

застосування адаптивних алгоритмів навчання, що мають як фільтруючі, так і слідкуючі (для нестаціонарних ситуацій) властивості [9; 10]. Для розглянутої задачі такий алгоритм може бути записано у вигляді

$$\begin{cases} w_j(N_F+1) = w_j(N_F) + r_j^{-1}(N_F+1)(x_{N+1,j} - \underline{X}_{N+1,j} w_j(N_F))\underline{X}_{N+1,j}^T, \\ r_j(N_F+1) = \alpha r_j(N_F) + \left\|\underline{X}_{N+1,j}\right\|^2, \end{cases} \quad (6)$$

де $0 \le \alpha \le 1$ – параметр згладжування, що задає компроміс між згладжуванням і стеженням за змінними характеристиками об'єктів. Обробку інформації в режимі послідовного надходження даних зручно організувати на основі нейромережевої системи, основними елементами якої є n паралельно працюючих адаптивних лінійних асоціаторів (ALA) [11], які навчаються за допомогою алгоритмів (5) або (6). На рис. 2 наведена схема цієї системи, яка не потребує додаткових пояснень.

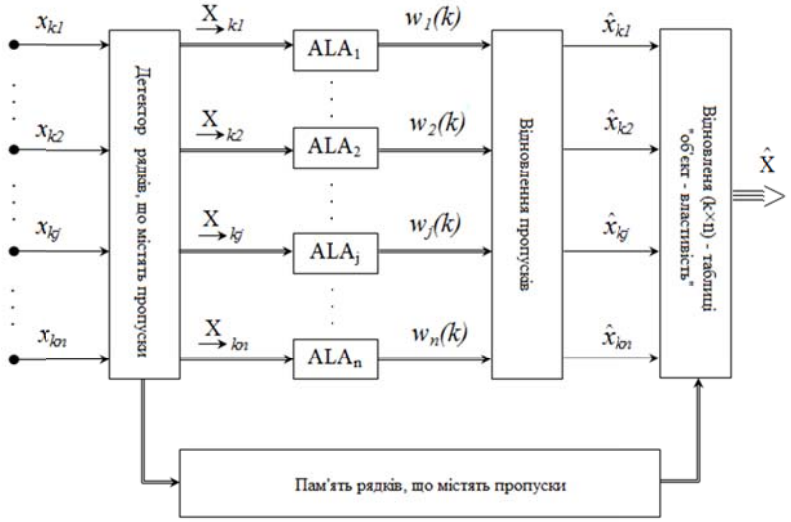

Рис. 2. Адаптивна нейромережева система для відновлення пропусків

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

Висновок

Розглянуто задачу відновлення пропусків в таблицях «об'єкт-властивість» в режимі послідовного надходження даних на обробку. Запропоновано адаптивну нейромережеву систему, що дозволяє вирішувати цю задачу в on-line режимі з постійним коригуванням відновлених елементів таблиці. Ця система характеризується високою швидкодією і простотою чисельної реалізації.

ЛІТЕРАТУРА

1. Загоруйко Н. Г. Эмпирические предсказания. – Новосибирск : Наука, 1979. – 120 с.
2. Han J. Data Mining: Concepts and Techniques / Han J., Kamber M. – Amsterdam : Morgan Kaufman Publ., 2006. – 743 p.
3. Gorban A. Principal Manifolds for Data Visualization and Dimension Reduction / A. Gorban, B. Kegl, B. Wunsch, A. Zinovyev (Eds.) // Lecture Notes in Computational Science and Engineering. – Berlin ; Heidelberg ; New York : Springer, 2007. – Vol. 58. – 330 p.
4. Bishop C. M. Neural Networks for Pattern Recognition. – Oxford : Clarendon Press, 1995. – 482 p.
5. Gorban A. N. Neural network modeling of data with gaps / A. N. Gorban, A. A. Rossiev, D. C. Wunsch II // Радіоелектроніка. Інформатика. Управління. – 2000. – № 1 (3). – С. 47–55.
6. Tkacz M. Artificial neural networks in incomplete data sets processing // In: Eds. Klopotek M. A., Wierzchon S. T., Trojanowski K. Intelligent Information Processing and Web Mining. – Berlin ; Heidelberg : Springer – Verlag, 2005. – P. 577–583.
7. Marwala T. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. – Hershey ; New York : Information Science Reference, 2009. – 303 p.
8. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. – Москва : Наука, 1977. – 224 с.
9. Бодянский Е. В. Многошаговые оптимальные учредители многомерных нестационарных стохастических процессов / Е. В. Бодянский, И. П. Плисс, Т. В. Соловьева // Доклады АН УССР. – 1986. – № 12. – С. 47–49.
10. BodyanskiyYe. An adaptive learning algorithm for a neuro – fuzzy network / BodyanskiyYe., Kolodyazhniy V., Stephan A. ; [ed. by B. Reusch] // «Computational Intelligence. Theory and Applications». – Berlin ; Heidelberg ; New York : Springer, 2001. – P. 68–75.
11. Haykin S. Neural Networks. A Comprehensive Foundation. – Upper Saddle River ; N. J. : Prentice Hall, 1999. – 842 p.

Стаття надійшла до редколегії 15.04.2011 р.