УДК 004.93
Каврін Д. А.[1], Субботін С. О.[2]
[1] асп. ЗНТУ
[2] д-р техн. наук, проф., зав. каф. ЗНТУ

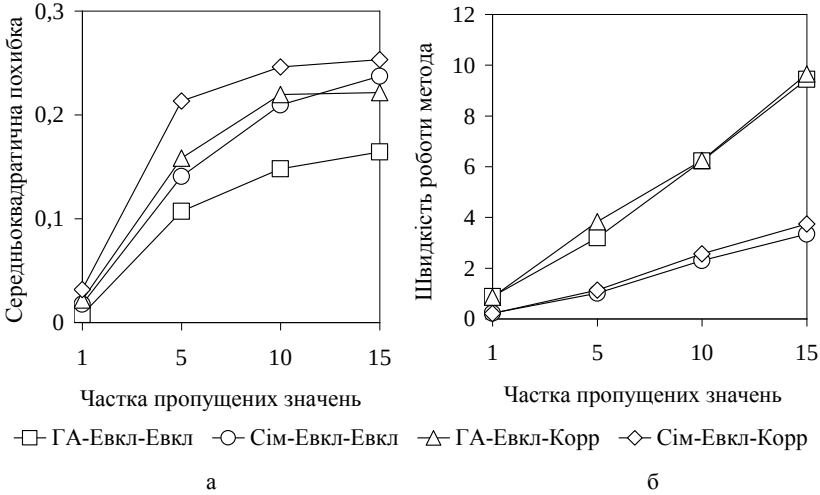
ВІДНОВЛЕННЯ ПРОПУСКІВ У ВИБІРКАХ З ВИКОРИСТАННЯМ МЕТОДІВ ЕВОЛЮЦІЙНОГО ПОШУКУ

Реальні вибірки даних часто містять в собі пропущені значення. Причини виникнення пропусків пов'язані з помилками в роботі обладнання, програмного забезпечення і т.д. Ігнорування подібних дефектів в даних може призводити до зниження продуктивності діагностичних систем і, можливо, до прийняття невірних рішень.

Існують різні методи заповнення (відновлення) пропусків в даних, які відрізняються за своєю природою, галуззю застосування та обчислювальною складністю. Умовно всі методи відновлення можна поділити на глобальні, в яких передбачається оцінка значень пропусків по всіх об'єктах вибірки, і локальні, які оцінюють значення пропусків в деякій локальній близькості [1].

В даній роботі пропонується модифікація методу локального відновлення пропущених даних ZET [1]. В основі методу ZET лежать припущення про надмірність даних у вибірці, їх лінійну залежність і локальну компактність. Одним з основних етапів методу ZET є підбір параметрів компетентності рядків (екземплярів) та стовпців (ознак), для пошуку, яких необхідно вирішити задачу оптимізації. У запропонованій модифікації, задача оптимізації вирішується методами еволюційного пошуку.

Для оцінки та порівняння якості відновлення пропущених значень, базовий метод ZET, в якому застосовується симплекс метод оптимізації, порівнювався з запропонованою модифікацією. В якості вихідної вибірки використовувався набір даних випробувань газотурбінних авіаційних двигунів [2]. Даний набір не мав природних пропусків, тому пропущені значення створювалися штучно, випадковим чином. Такий підхід дозволив оцінити роботу методу відновлення пропущених значень безпосередньо порівнянням відновлених даних з природними значеннями вихідної вибірки і розрахувати середньоквадратичну похибку (RMSE). Також проводилась оцінка швидкості роботи методу для різних обсягів пропущених значень. В якості міри близькості стовпців розглядалися значення кореляції та евклідової метрики. Таким чином, проводився аналіз трьох модифікацій методу, в залежності від способу оптимізації та міри близькості (рис. 1).



а б

Рисунок 1 – Візуалізація залежностей середньоквадратичної похибки (а) і швидкості роботи модифікацій методу ZET (б) від частки пропущених значень у вибірці

Запропонована модифікація методу ZET була програмно реалізована при проведенні обчислювальних експериментів з відновлення пропущених значень у вибірках даних. Експерименти показали, що використання оптимізаційних методів еволюційного пошуку дозволяє більш точно відновлювати пропущені значення при невеликій втраті швидкості, в порівнянні з оптимізацією методом симплекса.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Загоруйко Н. Г. Прикладные методы анализа данных и знаний [Текст] / Н. Г. Загоруйко. – Новосибирск : ИИМ, 1999. – 270 с.

2. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : [монография] / [Богуслаев А. В., Олейник Ал. А., Олейник Ан. А. и др.] ; под ред. Д. В. Павленко, С. А. Субботина. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.