МЕТОДИ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ДАНИХ В ЗАДАЧІ ДІАГНОСТИКИ МІТОХОНДРІАЛЬНИХ ЗАХВОРЮВАНЬ

Бойко[1] Д.О., Васильєва[2] О.В., Галкін[1] Д.О., Гречаніна[2] Ю.Б., Поворознюк[1] А.І., Філатова[1] Г.Є.

[1]Національний технічний університет «ХПІ», м. Харків

[2]Харківський спеціалізований медико-генетичний центр, м. Харків

Формування медичних даних є складний та неоднозначний процес тому, що здобуття даних виконується в різний час, різними лікарями з використанням методик, що розрізняються. Це призводить до того, що при стандартизації усієї вибірки можливі пропуски даних у різних пацієнтів. При істотному дефіциті даних в задаче синтезу діагностичних моделей мітохондріальних захворювань (МЗ) видалення даних, які мають пропуски, недопустиме. Тому задача відбудови пропущених даних є актуальною.

Метою даної роботи є аналіз ефективності існуючих методів для відновленням діагностичних ознак МЗ.

На сьогоднішній день немає алгоритму, який дозволяє відновлювати данні без похибок. Для вирішення поставленої задачі пропонується використовувати наступні методи з аналізом їх ефективності для реальних даних: заміна пропуску загальним середнім, заповнення з упередженим підбором, евристичний алгоритм, регресійній аналіз, генетичні алгоритми, методи розпізнавання образів.

Діагностичні ознаки МЗ представлені Харківським спеціалізованим медико-генетичним центром у вигляді таблиці експериментальних даних (ТЕД) типа «об'єкт-ознака» $\mathbf{X} = \left[x_{ij}\right]_{1,1}^{N,m}$, де x_{ij} – j-та ознака i-го об'єкта (пацієнта); N – кількість пацієнтів; m – кількість ознак. Шляхом видалення з ТЕД об'єктів, які мають пропуски даних, отримана навчальна ТЕД (НТЕД). На основі аналізу НТЕД розроблено критерій $J = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_{ij} - x'_{ij}}{x_{ij}}\right|$, де x_{ij}, x'_{ij} – еталонне (з НТЕД) та відновлене (за допомогою метода) значення j-ї ознаки i-го об'єкта відповідно. Таким чином, для відновлення невідомого значення j-ї ознаки необхідно використовувати той метод, для якого виконується умова $p = \arg\min_{k=1,K} J_k$, де $p \in [1, K]$ – номер оптимального методу, J_k – значення критерію для k-го методу; K – кількість методів.

На основі розробленого критерію оцінки методів відновлення даних проаналізовані методи й алгоритми та вибрано кращі методи відновлення ознак, які необхідні при діагностиці МЗ.