УДК 004.9:004.65:004.71:654.9

ФОРМУВАННЯ БАЗИ ДАНИХ LOGGED VEHICLE DATA ДЛЯ МАШИННОГО НАВЧАННЯ В АВТОМОБІЛЬНОМУ СЕРВІСІ

Грицук Валерій Юрійович, аспірант, Харківський національний автомобільно-дорожній університет, e-mail: valeri.gritsuk@gmail.com, ORCID: https://orcid.org/0000-0002-3780-7815
Грицук Юрій Валерійович, к.т.н., доцент, Національний університет «Острозька академія», e-mail: yuri.gritsuk@gmail.com, ORCID: https://orcid.org/0000-0003-3389-1172

База даних LVD зберігає зведену інформацію про особливості використання транспортних засобів. Дані завантажуються кожного разу, коли автомобіль заїжджає на авторизовану станцію техобслуговування та ремонту. Це буває кілька разів на рік, але з нерегулярними інтервалами, які важко передбачити апріорі. Під час експлуатації автомобіль безперервно збирає та зберігає низку показників, таких як середня швидкість або загальна витрата пального. Загалом, це прості статистичні дані різного роду, оскільки існують дуже жорсткі обмеження на пам'ять і обчислювальну потужність, особливо для старих моделей автомобілів. Більшість параметрів належать до однієї з наступних трьох категорій [1, 2]:

- продуктивність
- експлуатація транспортного засобу
- діагностика або налагоджування.

Для побудови такої бази даних, система повинна надавати попередження про компоненти з підвищеним ризиком виходу з ладу до наступного очікуваного візиту на станцію техобслуговування. Однак, якщо найближче зчитування перед відмовою становить 3-4 місяці, то менш імовірно, що знос мав видимий вплив на дані. Частота зчитування сильно варіюється між транспортними засобами і змінюється з віком транспортного засобу, і може становити лише одне зчитування на рік.

Збирання та обробка даних є найважливішим процесом у системах прогнозування. Дані з датчиків збирають і обробляють за допомогою OBU, а потім передають до VDS за допомогою бездротового модуля для подальшого зберігання у спеціально відведеному місці. Процес діагностування починається, коли транспортний засіб з невідомою несправністю підключається до VDS з метою ідентифікації дефекту. Потім відбувається порівняння фізичних даних, отриманих від датчиків, із відповідними ознаками несправності. Увага приділяється лише тим сеансам, під час яких несправність було успішно діагностовано. Тобто, не розглядається кожна сесія, несправність якої не була успішно визначена. Дані з баз даних LVD та бази технічного обслуговування об'єднані в один набір даних. Дані з LVD доповнюються змінними напрацювання на відмову на основі дат проведення ремонтів у VSR. Дані організовані у

велику матрицю. Ці дані включають інформацію про напрацювання на відмову транспортних засобів, які ще не вийшли з ладу. Показники з напрацюванням нижче заданого граничного значення, що називається горизонтом прогнозування (ГП) (Prediction horizon (PH), отримують мітку «Несправність», тоді як показники з більшим напрацюванням у минулому отримують мітку «Нормальний стан». Це позначення застосовується для розподілу прикладів на дві групи: ті, що неминуче виходять з ладу, і ті, що справно функціонують. Але база даних LVD містить понад 25% пропущених результатів, що занадто багато для інтерполяції, і це спонукає до використання методу на основі фільтрів замість обгорток та інтегрованих методів. В роботі [3] пропонується два метода для заповнення даних параметрів - метод на основі фільтра і метод на основі обгортки. Метод на основі фільтрів виявляє найбільш вагомі параметри, аналізуючи їх індивідуально. Цей метод розбиває дані одного параметра на дві групи на основі міток. Функції щільності розподілу ймовірностей груп порівнюють за допомогою критерію Колмогорова-Смірнова [4]. Обчислюється вірогідність того, що дві групи належать до одного і того ж розподілу, і виражається за допомогою p-значення. Параметри, для яких обидві групи є чітко унікальними (p-значення близьке до нуля), пов'язані зі зносом перед відмовою і включені до набору даних:

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \, e^{-2k^2\lambda^2} \tag{1}$$

$$\lambda = D_n\sqrt{n} \tag{2}$$

$$D_n = max|F_n(x) - F(x)| \tag{3}$$

де $D_n\sqrt{n} \geq \lambda$ при обмеженні збільшення числа незалежних спостережень n

Метод обгортки базується на способі підбору транспортних засобів таким чином, щоб забезпечити послідовність набору даних і уникнути додавання пропущених значень. У даних, що розглядаються, набори ознак відрізняються для різних транспортних засобів. Чистий перелік характеристик - це сукупність доступних показників для всіх транспортних засобів, що входять до складу набору даних.

При внесенні нової характеристики до набору даних усі транспортні засоби, що не мають такої характеристики, вилучаються з набору. Додавання окремих параметрів призводить до того, що набір даних зменшиться до дуже малої частки зразків. Це робить необхідним брати до уваги факт зменшення розміру набору даних при підборі елементів. Запропонований метод в роботі [5] дозволяє використовувати пучковий пошук для знаходження нових ознак для включення, розширюючи лише ті вузли, розмір набору даних яких не перевищує заданий поріг:

$$n_{dataset} = n_{all} * constraintFactor^{n_{params}} \quad (4)$$

де $n_{dataset}$ – позначає необхідний мінімальний розмір набору даних, n_{all} – кількість доступних зчитувань, $constraintFactor$ – коефіцієнт обмеження перебуває в діапазоні від 0 до 1, а n_{params} – кількість параметрів, які входять до обраного масиву даних. Кожен новий параметр може зменшити набір на невеликий відсоток. Це гарантує меншу граничну межу розміру набору даних.

У керованому машинному навчанні (рис 1) індукційний алгоритм представляється набором навчальних даних, де кожен приклад описується за допомогою вектору властивостей (або атрибутів) і мітки класу [6]. Завдання алгоритму індукції, або індуктора, полягає в тому, щоб отримати такий кластер, який буде корисним для класифікації наступних випадків. Алгоритм індукції запускається на наборі даних, розділеному на внутрішню навчальну та тренувальну вибірки, при цьому з даних видаляються різні набори ознак. Набір даних з найвищою оцінкою обирається як остаточний набір, на якому запускається індукційний алгоритм. Потім створений класифікатор оцінюється на незалежному тестовому наборі, який не використовувався під час навчання [7].

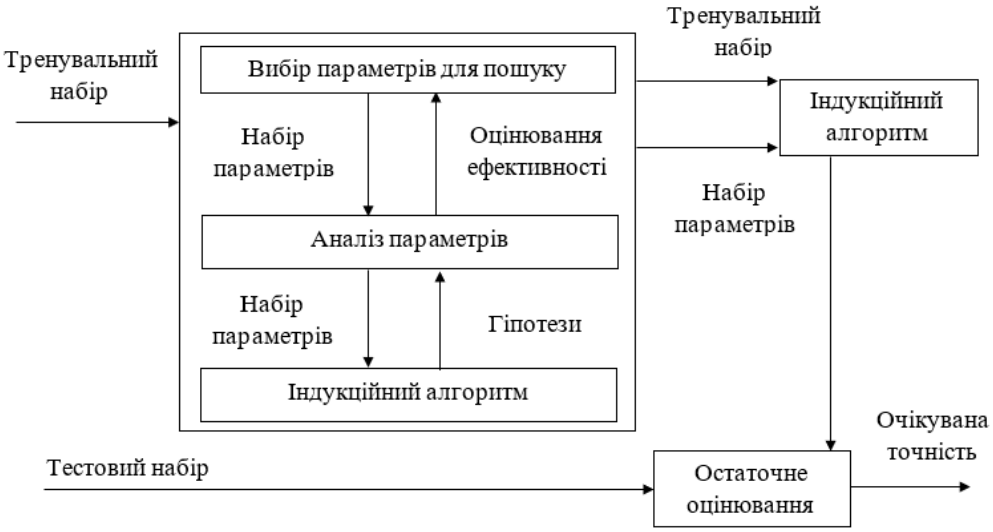

Рис. 1. Обгортковий метод для виділення елементів підмножини параметрів

Однак, чим новіший транспортний засіб, тим більше параметрів LVD доступні, але це також залежить від конфігурації транспортного засобу. Наприклад, детальні параметри коробки передач доступні лише для автомобілів з автоматичною коробкою передач. Це ускладнює отримання послідовних наборів даних для великого потоку транспортних засобів і ускладнює аналіз. Потрібно або вибрати набір даних з невідповідностями і заповнити пропущені значення, або обмежити аналіз лише тими транспортними засобами, які мають необхідні параметри. Такі бази даних містять важливу та цікаву інформацію про несправності транспортних засобів, яка іноді використовується технічним персоналом СТО для діагностики та прогнозування майбутніх несправностей.

Виходячи з цього, маємо два варіанти можливих отриманих даних:

- Неповна вибірка: ці методи спрямовані на збалансування набору даних шляхом усунення зразків класу, що становить більшість.
- Надмірна вибірка: ці методи передбачають відтворення прикладів класу меншини для досягнення більш збалансованого розподілу [8].

Як неповна, так і надмірна вибірка мають певні недоліки. Неповна вибірка може відкинути потенційно корисні дані, а надмірна вибірка може збільшити ймовірність виникнення надмірної пристосованості, оскільки більшість методів надмірної генерації роблять точні копії зразків класу меншості.

Висновки

Таким чином, ефективність систем прогностичного обслуговування безпосередньо залежить від якості попередньої обробки та стратегічного вибору інформативних параметрів. Використання статистичних критеріїв, таких як тест Колмогорова-Смірнова, у поєднанні з оптимізованими методами обгортки дозволяє нівелювати проблему пропущених значень та побудувати релевантні моделі навчання. Запропоновані підходи створюють надійний фундамент для впровадження інтелектуальних систем діагностики, що здатні завчасно ідентифікувати ризики відмов, оптимізуючи графіки технічного обслуговування та підвищуючи надійність транспортних засобів. Розглянутий підхід дозволяє в подальшому рухатися в напрямку інтелектуального сервісу, що базується на аналізі реального експлуатаційного стану транспортних засобів.

Література

1. **Грицук В., Пронін С.** Аналіз та обґрунтування вибору моделі для моніторингу параметрів транспортного засобу та прогнозування технічного обслуговування. *Вісник Приазовського державного технічного університету. Серія: Технічні науки*. 2024. Вип. 49 (1). С. 56–73. DOI: https://doi.org/10.31498/2225-6733.49.1.2024.321206.

2. **Prytz R. et al.** Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering Applications of Artificial Intelligence*. 2015. Vol. 41. P. 139–150. DOI: http://dx.doi.org/10.1016/j.engappai.2015.02.009.

3. **Prytz R.** Machine learning methods for vehicle predictive maintenance using off-board and on-board data : Licentiate thesis. Halmstad : Halmstad University, 2014. 78 p.

4. **Rögnvaldsson T. et al.** Self-monitoring for maintenance of vehicle fleets. *Data Mining and Knowledge Discovery*. 2018. Vol. 32. P. 344–384. DOI: https://doi.org/10.1007/s10618-017-0538-6.

5. **Kimball R., Ross M.** The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence. Hoboken : Wiley, 2010. 744 p.

6. **Kohavi R., John G. H.** Wrappers for feature subset selection. *Artificial Intelligence*. 1997. Vol. 97, no. 1–2. P. 273–324. DOI: https://doi.org/10.1016/S0004-3702(97)00043-X.

7. **Tan P.-N., Steinbach M., Kumar V.** Introduction to Data Mining. Harlow : Pearson Education Limited, 2014. 839 p.

8. **Batista G. E. A. P. A., Bazzan A. L. C., Monard M. C.** Balancing training data for automated annotation of keywords: a case study. *Proceedings of the Workshop on Bioinformatics (WOB)*. 2003. P. 10–18.