УДК 004.6

В.В. Яцишин канд. техн. наук, доцент, Ю.О. Журихін
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ОЦІНЮВАННЯ ЯКОСТІ ДАНИХ ДЛЯ СИСТЕМ МАШИННОГО НАВЧАННЯ

V.V. Yatsyshyn PhD, Assoc. Prof., Y.O. Zhurykhin

DATA QUALITY EVALUATION FOR MACHINE LEARNING SYSTEMS

Сучасні технології проектування інформаційних систем, розробки програмних та програмно-технічних засобів характеризується необхідністю опрацювання великої кількості інформації, що призвело до стрімкого розвитку таких галузей як Big Data, Data Mining, Text Mining, прикладних систем штучного інтелекту. Для ефективного проектування систем штучного інтелекту, систем машинного навчання та інших «smart» систем важливим є забезпечення якості даних, що є фундаментальним аспектом ефективності алгоритмів опрацювання даних та одержання точних і достовірних результатів.

Побудова моделей і розробка методів забезпечення, управління та контролю якості даних є актуальною задачею практично для усіх сфер діяльності. При проектуванні систем машинного навчання характерним є різна природа і походження даних. При цьому дані є слабоструктурованими або апріорі невідомими, присутні дефекти, що призводить до опрацювання недостовірної інформації та як наслідок недостовірних і не точних результатів.

Міжнародний стандарт ISO/IEC 25012 визначає загальну модель якості даних, що зберігаються в структурованому форматі комп'ютерної системи. Він фокусується на якості даних як частини комп'ютерної системи і визначає якісні характеристики для цільових даних, використовуваних людьми і системами.

Цільовими даними є ті, що становлять інтерес для подальшого аналізу і перевірки, шляхом їх представлення у моделях через певні структури. Термін нецільових даних охоплює два випадки: перший відноситься до даних, які не є постійними (наприклад, дані обробляються операційною системою), а другий належить до даних, які можуть бути в рамках стандарту, але організація вирішує не застосовувати стандарт до них.

При використанні цього стандарту з іншими стандартами з ряду SQuaRE, можна визначити вимоги якості даних, визначити метрики якості даних, а також здійснити планування і оцінку якості даних. Атрибути якості даних і відповідні метрики можуть бути класифіковані за характеристиками якості і використані в процесі оцінювання з метою аналізу даних незалежно від інших компонентів комп'ютерної системи. В загальному випадку, якість даних – ступінь, в якій характеристики даних задовольняють зазначені передбачувані потреби при використанні в зазначених обставинах.

Модель якості даних, що визначена у стандарті ISO/IEC 25012, визначає п'ятнадцять характеристик відповідно з невід'ємної і залежної від системи точок зору. Таким, чином для класу систем машинного навчання необхідно визначити атрибути якості даних, здійснити їх класифікацію за характеристиками якості та підібрати метрики для кількісної їх інтерпретації.