МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ
І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ЖУРИХІН ЮРІЙ ОЛЕГОВИЧ

УДК 004.6

МЕТОДИ ЗАБЕЗПЕЧЕННЯ ЯКОСТІ ДАНИХ ПРИ ПРОЕКТУВАННІ СИСТЕМ МАШИННОГО НАВЧАННЯ

123 «Комп'ютерна інженерія»

Автореферат
дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль 2018

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, доцент кафедри комп'ютерних систем та мереж
Яцишин Василь Володимирович,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: кандидат технічних наук, доцент кафедри програмної інженерії
Михалик Дмитро Михайлович,
Тернопільський національний технічний університет імені Івана Пулюя

Захист відбудеться 20 лютого 2018 р. о $9^{.00}$ годині на засіданні екзаменаційної комісії №35 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. 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.

Для підвищення якості даних розроблено ряд вітчизняних та закордонних стандартів (ISO/IEC 25012, ISO/IEC 9126, ДСТУ ISO 9001 – 2001 та ін.), які покликані підвищити якість даних при проектуванні комп’ютерних систем.

Питаннями оцінювання та покращення якості даних присвячено ряд наукових та науково-прикладних публікацій українських (П. Андон, О. Харченко), а також закордонних науковців (B. Boehm, T. Saati, M. Holsted), однак процедур щодо забезпечення, управління, контролю та оцінювання якості даних не наведено. Тому актуальною задачею в галузі інформаційних технологій є побудова моделей якості даних, методів і засобів їх управління та оцінювання.

Мета роботи: дослідження методів забезпечення та оцінювання якості даних при проектуванні систем машинного навчання для підвищення ефективності результатів інтелектуального аналізу.

Об'єкт дослідження – процеси забезпечення та оцінювання якості даних.

Предмет дослідження – моделі якості даних, методи і засоби забезпечення та оцінювання якості даних систем машинного навчання.

Методи дослідження: Для вирішення поставлених задач використано наступні методи: аналіз та узагальнення – при проведенні аналізу існуючих моделей якості даних і методів забезпечення якості; формалізації – при побудові моделі якості даних та розробці методу забезпечення та оцінювання якості даних; проектування та програмування – при розробці програмного засобу формування критеріїв якості даних та оцінювання їх якості; експеримент та вимірювання – для апробації розробленого методу та обґрунтованої моделі..

Наукова новизна отриманих результатів:

– уперше обґрунтовано та формалізовано модель якості даних на основі рекомендацій стандарту ISO/IEC 25012, що дало змогу представити характеристики якості даних у вигляді комплексних критеріїв до складу яких входять атрибути та метрики для кількісного вираження значень якості даних та в перспективі спроектувати засіб для автоматизації процесу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання.

– уперше визначено атрибути якості даних, що дало змогу реалізувати процес забезпечення та управління якістю при проектуванні систем машинного навчання.

– уперше розроблено метод забезпечення та оцінювання якості даних при проектуванні систем машинного навчання, що дало змогу кількісно оцінити відповідність даних критеріям якості та врахувати їх у загальному процесі розробки інтелектуальних систем.

Практичне значення отриманих результатів. Впровадження формалізованої моделі якості даних, методу забезпечення та оцінювання якості даних реалізовано та впроваджено у вигляді програмного засобу, який дає змогу забезпечити якість даних при проектуванні систем машинного навчання.

Апробація. Результати дослідження апробовано на VI Міжнародній науково-технічній конференції молодих учених та студентів «Актуальні задачі сучасних технологій» (16-17 листопада 2017 року) Тернопільського національного технічного університету імені Івана Пулюя та V науково-технічній конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні системи, моделі та технології» (1-2 лютого 2018 р.) у вигляді тез конференцій.

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 6 частин, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка – 117 арк. формату А4, графічна частина – 8 аркушів формату А1.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі визначено актуальність теми дипломної роботи магістра щодо дослідження методів забезпечення якості даних при проектуванні систем машинного навчання, сформульовано мету, задачі і методи дослідження, наведено наукову новизну та практичне значення одержаних результатів.

У першому розділі «Аналіз сучасного стану в галузі проектування систем машинного навчання» проведено аналіз наукових публікацій, методів і засобів побудови систем машинного навчання у результаті якого встановлено необхідність забезпечення якості даних у процесі проектування таких систем, оскільки якість даних впливає на точність і повноту результатів роботи інтелектуальних систем, а також складність алгоритмів при їх побудові. Проведено аналіз моделей даних, які використовуються при побудові систем машинного навчання, встановлено особливості їх застосування і вимоги до якості даних. Обґрунтовано застосування моделі якості даних ISO/IEC 25012 при побудові комп’ютерних систем, в тому числі і машинного навчання, що дало змогу визначити характеристики якості для оцінювання даних при проектуванні систем машинного навчання.

У другому розділі «Розробка методу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання» формалізовано модель якості даних стандарту ISO/IEC 25012 , що дало змогу представити характеристики якості даних у вигляді комплексних критеріїв до складу яких входять атрибути та

метрики для кількісного вираження значень якості даних, що дало змогу реалізувати процес забезпечення та управління якістю проектування систем машинного навчання. Розроблено метод забезпечення та оцінювання якості даних при проектуванні систем машинного навчання, що дало змогу кількісно оцінити відповідність даних критеріям якості та врахувати їх у загальному процесі розробки інтелектуальних систем. Обґрунтовано застосування методів Сааті, Когера і Ю та простого алгоритму вибору при визначенні вагових коефіцієнтів для атрибутів якості даних, що дало змогу підвищити ефективність проектування систем машинного навчання.

У третьому розділі «Засіб автоматизації процесу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання» спроектовано архітектуру засобу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання. Розроблено засіб забезпечення та оцінювання якості даних у відповідності до характеристик якості стандарту ISO/IEC 25012: точності та послідовності. Проведено експериментальне використання розробленого засобу, що дало змогу покращити першочергові дані і підвищити їх якість при проектуванні систем машинного навчання..

У четвертому розділі «Обґрунтування економічної ефективності» проведено розрахунки щодо доцільності розробки методу і засобу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання.

У п’ятому розділі «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто питання норм і правил охорони праці та техніки безпеки при експлуатації засобу автоматизації процесу забезпечення та оцінювання якості даних, а також проаналізовано планування заходів цивільного захисту на об’єкті у випадку надзвичайних ситуацій.

У шостому розділі «Екологія» проаналізовано роль матеріало- та ресурсозбереження у вирішенні екологічних проблем та розглянуто статистичне оцінювання техногенних впливів.

У загальних висновках до дипломної роботи магістра наведено результати виконання частин дипломної роботи магістра, їх наукове та практичне значення при проектуванні систем машинного навчання.

Додатки до пояснювальної записки містять матеріали конференцій у яких опубліковано основні результати дипломної роботи магістра.

У графічній частині до дипломної роботи магістра проілюстровано основні наукові та практичні результати щодо забезпечення та оцінювання якості даних при проектуванні систем машинного навчання.

ВИСНОВКИ

У дипломній роботі магістра проведено аналіз наукових публікацій, методів і засобів побудови систем машинного навчання у результаті якого встановлено необхідність забезпечення якості даних у процесі проектування таких систем, оскільки якість даних впливає на точність і повноту результатів роботи інтелектуальних систем, а також складність алгоритмів при їх побудові.

Проведено аналітичний огляд щодо класифікації систем машинного навчання і виявлено, що в основі практично усіх таких систем лежать бази даних, які в

подальшому формують бази знань і якість представлення даних є важливою характеристикою цілісної системи машинного навчання.

Обґрунтовано застосування моделі якості даних ISO/IEC 25012 при побудові комп'ютерних систем, в тому числі і машинного навчання, що дало змогу визначити характеристики якості для оцінювання даних при проектуванні систем машинного навчання.

Формалізовано модель якості даних стандарту ISO/IEC 25012 , що дало змогу представити характеристики якості даних у вигляді комплексних критеріїв до складу яких входять атрибути та метрики для кількісного вираження значень якості даних та в перспективі спроектувати засіб для автоматизації процесу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання.

Визначено атрибути якості даних і розроблено метод забезпечення та оцінювання якості даних при проектуванні систем машинного навчання, що дало змогу кількісно оцінити відповідність даних критеріям якості та врахувати їх у загальному процесі інтелектуальних систем, реалізувати процес забезпечення та управління якістю проектування систем машинного навчання.

Досліджено методи підвищення якості даних, що дало змогу інтегрувати процеси профілювання, стандартизації, очищення, збагачення і дедуплікації даних при проектуванні систем машинного навчання.

Спроектовано архітектуру та розроблено засіб забезпечення та оцінювання якості даних при проектуванні систем машинного навчання, що дало змогу оцінити відповідність тестових даних характеристикам точності та послідовності стандарту ISO/IEC 25012.

Проведено експериментальне використання розробленого засобу, що дало змогу покращити першочергові дані і підвищити їх якість при проектуванні систем машинного навчання

На основі розрахунку економічних показників обґрунтовано доцільність розробки методу і засобу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання. Визначено, що затрати на проведення НДР становлять 41790,55 грн., а термін окупності – 1,88 року.

Проведено аналіз вимог з охорони праці і техніки безпеки при використанні комп'ютерної техніки та планування заходів цивільного захисту у випадку надзвичайних ситуацій, що дало можливість врахувати їх при експлуатації засобу забезпечення та оцінювання якості даних при проектуванні систем машинного навчання

Проаналізовано роль матеріало- та ресурсозбереження у вирішенні екологічних проблем та розглянуто статистичне оцінювання техногенних впливів.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Журихін Ю.О. Оцінювання якості даних для систем машинного навчання /Ю.О. Журихін, В.В. Яцишин // Матеріали VI Міжнародної науково-технічної конференції молодих учених та студентів «Актуальні задачі сучасних технологій» - Тернопіль – 16 – 17 листопада 2017 р. – с. 196

2. Журихін Ю.О. Процеси забезпечення якості даних при проектуванні систем машинного навчання / Ю.О. Журихін, В.В. Яцишин // Матеріали V науково-технічній конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні системи, моделі та технології» - 1-2 лютого 2018 р. – Тернопіль – с. 68.

АНОТАЦІЯ

Журихін Ю.О. Методи забезпечення якості даних при проектуванні систем машинного навчання

Дипломна робота на здобуття освітнього ступеня магістра 123 – Комп'ютерна інженерія. – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль 2018.

У дипломній роботі магістра досліджено методи забезпечення та оцінювання якості даних при проектуванні систем машинного навчання, що дало змогу підвищити ефективність результатів інтелектуального аналізу.

Основними задачами дипломної роботи є аналіз наукових публікацій та стандартів в галузі забезпечення якості даних для визначення сучасного стану та шляхів удосконалення існуючих моделей якості даних та інтеграції їх у процеси проектування систем машинного навчання, обґрунтування та формалізація моделі якості даних для підвищення якості проектування систем машинного навчання, визначення атрибутів якості даних для систем машинного навчання, розробка методу забезпечення та оцінювання якості даних на основі моделі якості даних ISO/IEC 25012, розробка програмного засобу для формування критеріїв якості даних та проведення відповідного оцінювання їх якості

Розроблено метод забезпечення та оцінювання якості даних на основі моделі якості даних ISO/IEC 25012, досліджено методи підвищення якості даних та методи розрахунку вагових коефіцієнтів для атрибутів якості даних на основі експертних технологій.

У роботі визначено вимоги до програмного засобу підтримки методу забезпечення та оцінювання якості даних, спроектовано архітектуру та реалізовано його на основі технологій PHP та MySQL. Це дало можливість автоматизувати процеси забезпечення та оцінювання якості даних при проектуванні систем машинного навчання.

Ключові слова: МЕТОД, МАШИННЕ НАВЧАННЯ, ДАНІ, ЯКІСТЬ, ПРОЕКТУВАННЯ.

ANNOTATION

Zhurykhin Y.O. Methods of data quality providing at computer-assisted learning systems design

The diploma paper for obtaining the Master's degree 123 – Computer engineering – Ternopil Ivan Puluj National Technical University, Ternopil 2018.

In the diploma paper, the methods of providing and assessing the quality of data in the design of systems of machine learning were investigated, which enabled to increase the efficiency of the results of the intellectual analysis.

The main objectives of the thesis are to analyze scientific publications and standards in the field of data quality assurance in order to determine the current state and ways of improving the existing models of data quality and integrate them into the processes of designing machine learning systems, substantiation and formalization of the data quality model for improving the quality of designing machine learning systems, definition of data quality attributes for machine learning systems, development of data quality assurance and evaluation methodology based on data quality model ISO/IEC 25012 , the development of software for the formation of criteria for the quality of data and conducting an appropriate assessment of their quality.

The method of data quality assurance and evaluation based on the data quality model ISO/IEC 25012 was developed, methods of data quality improvement and methods of calculating weight factors for data quality attributes based on expert technologies were investigated.

The work defines the requirements for software support for the method of data quality assurance and evaluation, architecture is designed and implemented on the basis of PHP and MySQL technologies. This made it possible to automate the processes of ensuring and assessing the quality of data when designing machine learning systems.

Keywords: METHOD, MACHINE LEARNING, DATA, QUALITY, DESIGN.