

ВІДГУК

офіційного опонента на дисертаційну роботу Земляного Олексія Дмитровича на тему: «Розроблення методів та програмного забезпечення інтелектуального імпутовання пропусків даних», представлену на здобуття наукового ступеня доктора філософії з галузі знань 12 «Інформаційні технології» за спеціальністю 121 Інженерія програмного забезпечення

Актуальність теми дисертації

У сучасному світі, де дані відіграють ключову роль у прийнятті рішень у різних галузях, проблема неповноти даних стає все більш актуальною. Пропущені значення, які виникають через помилки збору інформації, технічні збої або об'єднання даних з різних джерел, суттєво впливають на якість аналізу. Вони можуть призводити до упереджених результатів, зниження точності моделей машинного навчання та помилкових висновків. Дисертаційна робота Земляного О. Д. спрямована на вирішення цієї проблеми шляхом розробки інтелектуальних методів імпутовання пропусків у даних, які будуть враховувати структуру даних, патерни пропусків, а також кореляційні та ентропійні зв'язки між ознаками, та мають на меті покращити якість подальших аналітичних моделей, що робить дослідження надзвичайно важливим для сучасних систем обробки даних.

Оцінка змісту та завершеності дисертації

Дисертаційна робота має обсяг 169 сторінок, з яких основний текст займає 131 сторінку. Робота містить анотацію, вступ, 4 розділи, висновки, список використаних джерел з 121 найменуванням та додатки. Дисертація є завершеним науковим дослідженням, у якому розроблено методи, алгоритми та програмне забезпечення для інтелектуального імпутовання пропусків у даних.

У вступі обґрунтовано актуальність теми, визначено мету, завдання, об'єкт і предмет дослідження, а також наукову новизну та практичне значення роботи. Наведено відомості щодо апробації результатів та публікацій за темою дисертації.

Перший розділ присвячений аналізу сучасних підходів до обробки даних, зокрема проблеми пропусків, яка є однією з ключових перешкод для ефективного машинного навчання. Розглянуто задачі класифікації на прикладі діагностування ішемічної хвороби серця та прогнозування часових рядів на основі даних гідрологічного моніторингу. Проведено детальний огляд існуючих методів імпутовання, наведено їхню класифікацію за механізмами виникнення та стратегії заповнення.

Другий розділ присвячений розробці алгоритмічного забезпечення для імпутовання пропусків. Описано методи, які враховують структуру даних, патерни пропусків, кореляційні та ентропійні зв'язки між ознаками, а також алгоритми перетворення якісних ознак на кількісні зі збереженням інформації про пропуски. Наведено UML-діаграми діяльності для візуалізації логіки роботи алгоритмів.

Третій розділ присвячений реалізації розроблених алгоритмів на мові програмування Python з інтеграцією в екосистему бібліотеки scikit-learn, а також оптимізації коду для покращення продуктивності обчислень.

Четвертий розділ присвячений тестуванню розроблених методів імпутовання. Описано датасети для задач класифікації та прогнозування часових рядів, проведено тести для порівняння точності імпутовання та швидкодії, аналізу впливу імпутовання на якість моделей у задачах класифікації та прогнозування, та для аналізу змін ентропії до та після імпутовання.

У висновках узагальнено основні результати дослідження. Зазначено, що розроблені методи імпутовання пропусків у даних підвищують точність імпутовання та покращують якість моделей у задачах класифікації та

прогнозування. Наведено практичне значення роботи, зокрема впровадження результатів у освітній процес та потенціал для використання в реальних системах аналізу даних.

У додатках наведено список публікацій здобувача за темою дисертації, коди класів ім'ютерів та кодувальників якісних ознак, а також акт впровадження результатів роботи в освітній процес кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій

При виконанні досліджень використано комплекс сучасних та апробованих методів обробки даних, статистичних підходів та методів машинного навчання. Отримані практичні результати демонструють високу кореляцію з теоретично прогнозованими, а також узгоджуються із результатами інших авторів, доступних з відкритих джерел.

Наукові положення, висновки та рекомендації, представлені у дисертації, обґрунтовані, базуються на аналізі спеціалізованої профільної літератури, інформаційних технологій та підтверджуються результатами експериментальних досліджень.

Основні наукові результати та їх наукова новизна

Серед результатів, які отримані при виконанні досліджень, слід зазначити наступні, які мають наукову новизну:

1. Удосконалено метод імпутування NoNa шляхом розробки методу UnifiedClassRegrImputer, який застосовує комбінацію класифікатора та регресора з урахуванням типу даних ознак, при цьому враховує патерни пропусків у даних та надає змінний порядок обходу ознак залежно від

кількості пропусків, що дозволяє підвищити точність імпутування та покращити якість моделей у задачах класифікації та прогнозування.

2. Вперше розроблено метод імпутування `RegrImputer`, який застосовує регресор з уточненням значення у випадку категоріальних даних, що дозволяє значно підвищити швидкодію та точність імпутування, а також покращити якість моделей у задачах класифікації та прогнозування.

3. Удосконалено ентропійний метод `EntropyImputer` шляхом додавання двох модифікацій: однокрокова та ітераційна процедури зменшення умовної ентропії кожної ознаки у послідовному та паралельному виконанні, що дозволяє зменшити умовну ентропію та час виконання алгоритму.

4. Вперше розроблено гібридний метод імпутування `HybridRegrEntropyImputer`, який поєднує ентропійний та регресійний підходи в залежності від типу даних ознаки та патернів пропусків, що дозволяє підвищити точність імпутування та швидкодію у порівнянні з `EntropyImputer`.

5. Вперше розроблено метод імпутування `CorrelationImputer` на основі виявленого кореляційного зв'язку між ознаками, який автоматизує вибір найкращої моделі залежності серед лінійних та квазілінійних моделей, що дозволяє підвищити точність імпутування, а також точність моделей у задачі прогнозування часових рядів.

6. Вперше запропоновано метод імпутування пропусків у часових рядах з сезонною складовою на основі виявленого кореляційного зв'язку між ознаками (метод імпутування `CorrelationImputer`), але дія виконується для кожного окремого місяця року. Це дозволяє підвищити точність імпутування і якість моделей прогнозування у порівнянні з аналогом без розбиття по місяцям.

7. Удосконалено два методи перетворення якісних ознак на кількісні шляхом збереження інформації про пропуски та можливістю виконувати

зворотне перетворення (IgnoreNaNLabelEncoder, IgnoreNaNFrequentEncoder).

Практичне значення одержаних результатів

Практична цінність дисертації полягає у розробці алгоритмів та програмного забезпечення для імпутування пропусків у даних, які реалізують запропоновані методи (UnifiedClassRegrImputer, RegrImputer, EntropyImputer, HybridRegrEntropyImputer, CorrelationImputer, IgnoreNaNLabelEncoder, IgnoreNaNFrequentEncoder). Всі алгоритми розроблені із дотриманням принципів архітектури бібліотеки scikit-learn, що дозволяє використовувати їх в аналітичних конвеєрах (pipeline).

Розроблені методи підвищують точність імпутування та покращують якість моделей класифікації та прогнозування для даних медичного та гідрологічного моніторингу. Результати впроваджено в освітній процес кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара. Розроблені методи можуть бути впроваджені в реальні системи аналізу даних.

Зауваження до викладеного у дисертації

Незважаючи на високий рівень дослідження, є декілька зауважень, які не знижують наукової та практичної цінності роботи, але можуть бути враховані для подальшого вдосконалення:

1. У тексті дисертації потрібно було б навести результати перевірки адекватності моделей, покладених в основу імпутерів, особливо в контексті їхнього застосування до різних типів даних (медичних, гідрологічних тощо).

2. Недостатньо розкрито процес проєктування програмного забезпечення для нових методів імпутування. Додатковий опис етапів розробки, вибору архітектурних рішень та обґрунтування вибору мови

програмування Python (наприклад, порівняно з іншими мовами, такими як C++ або R) допомогло б краще зрозуміти технічні рішення, прийняті автором.

3. Бажано було б детальніше описати метрики, використані для оцінки моделей. Наприклад, які саме метрики (окрім точності) застосовувалися для оцінки ефективності імпутовання та їхнього впливу на кінцеві результати.

4. Не вказано, які інструменти використовувалися для проведення досліджень. Наприклад, можна було б описати, які інтегровані середовища чи спеціалізовані бібліотеки застосовувалися.

5. Не зрозуміло, який із рисунків дисертації відображає інформаційну технологію, яка розроблялась у цій роботі.

Проте висловлені зауваження не знижують загальної позитивної оцінки виконаної автором роботи, її наукової новизни, обґрунтованості та практичної цінності.

Повнота викладу результатів дисертації в публікаціях

За темою дисертації опубліковано 11 робіт: 4 статті у наукових фахових виданнях України категорії Б, 2 статті у інших збірниках наукових праць та 5 тез доповідей у збірниках матеріалів наукових конференцій. Результати досліджень автора пройшли апробацію і дістали позитивну оцінку на конференціях та інших науково-практичних заходах.

Загальний висновок

Дисертаційна робота Земляного Олексія Дмитровича є завершеним науковим дослідженням, у якому отримано нові результати, які мають ознаки наукової новизни та практичної значимості, що ефективно вирішують актуальну науково-прикладну задачу інтелектуального імпутовання пропусків у даних.

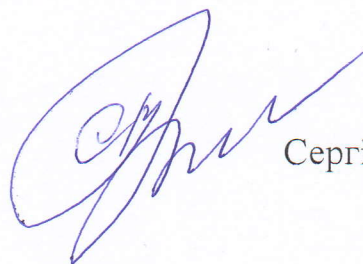
Робота відповідає спеціальності 121 Інженерія програмного забезпечення, оскільки присвячена розробці методів, алгоритмів та

програмного забезпечення, які забезпечують точність та ефективність у задачах моніторингу.

Враховуючи обсяг виконаних досліджень, наукову новизну, обґрунтованість і практичну цінність отриманих результатів, дисертація відповідає вимогам Наказу Міністерства освіти і науки України № 40 від 12 січня 2017 р. та Постанови Кабінету Міністрів України № 44 від 12 січня 2022 р. зі змінами. Земляний Олексій Дмитрович заслуговує присудження йому ступеня доктора філософії за спеціальністю 121 Інженерія програмного забезпечення.

Офіційний опонент

доктор технічних наук, професор,
професор кафедри програмного
забезпечення автоматизованих
систем Черкаського державного
технологічного університету МОН
України



Сергій ГОЛУБ

Підпис проф. С. В. Голуба засвідчую.

Перший проректор Черкаського
державного технологічного
університету, к.т.н., проф.



Артем ГОНЧАРОВ