УДК 004.8 doi: 10.20998/2220-4784.2022.02.04

О. О. БРОВАРНИК, В. В. ОВСЯНІКОВ

ДОСЛІДЖЕННЯ ВЛАСТИВОСТЕЙ СЕРЕДОВИЩА КЕРУВАННЯ ДАНИМИ ТА ОЦІНКА ЧАСУ ПЕРЕДАЧІ ВЕЛИКИХ НАБОРІВ ДАНИХ

У статті розглядається задача оцінювання часу передачі великих наборів даних через розподілене середовище керування даними на основі самостійно створеної моделі нейронної мережі та дослідження властивостей цього середовища за допомогою методів статистичного аналізу. Для початкового аналізу отримано метадані для успішних передач файлів в системі, трансформовано та виділено змінні, які впливають на час передачі файлів. Під час аналізу використані різні вибірки, щоб перевірити, чи схожі результати в усіх наявних даних. Застосовано методи кореляційного, регресійного аналізу для дослідження середовища. Виявлено, що не існує чіткої кореляції між часом передачі та одним з вхідних параметрів. Час передачі файлу залежить від ряду зовнішніх факторів, які неможливо отримати за допомогою метаданих, але можливо частково дослідити середовище використовуючи отримані метадані. Використано модель на основі двох вхідних рівнів для числових та категоріальних змінних, а потім об'єднаних в одну гілку. Для зображення результатів передбачення використовуються показники RMSE та діаграма розсіювання для порівняння цільових та передбачених значень. Проведені розрахунки показують задовільні результати передбачень;

Ключові слова: дослідження властивостей; оцінка часу передачі; методи статистичного аналізу; аналіз даних; регресія; кореляція; нейронна мережа.

Вступ. Середовище керування даними наукових експериментів утворює складну екосистему з динамічною взаємодією між користувачами та центрами обробки даних. Точність прогнозів моделі обмежена кількістю системних даних на момент прогнозування та стохастичними процесами, що відбуваються в окремих частинах системи. Центральна роль Rusio як системи керування даними, а також велика кількість інформації про передачі та життєвий цикл правил даних, яку вона збирає, можуть допомогти створити алгоритм машинного навчання для оцінки часу передачі.

Було розглянуто відомі дослідження команди розробників Rucio, в яких проведені дослідження властивостей середовища та оцінювалася тривалість передачі великої кількості файлів для наукових досліджень [1-5]. У цій статті розглядається проблема ідентифікації факторів, що впливають на процеси в різних системах. Зазвичай такі задачі вирішуються методами кореляційного, регресійного, факторного та компонентного аналізу.

Основним завданням кореляційного аналізу є визначення зв'язку між випадковими величинами, оцінка його інтенсивності та спрямованості [6, с. 56]. Отже, відповідні коефіцієнти кореляції показують величину зв'язків між ознаками.

Завданням регресійного аналізу є створення моделі, яка дозволяє оцінити значення залежної змінної на основі значень незалежних показників. Регресійний аналіз є основним інструментом для вивчення показників взаємозв'язку між різними змінними [6, с. 68].

Кореляційно-регресійний аналіз використовують для розв'язання задач в різноманітних галузях, таких як економіка, соціологія, статистика, географія, демографія та інші[7–10].

Багато робіт, присвячених задачам регресійного аналізу, використовують нейронні мережі, як метод розв'язання [11,12,13].

Метою роботи є проведення дослідження даних подій середовища Rucio, визначення нового підходу в реалізізації моделі нейронної мережі для оцінювання часу передачі великих наборів даних.

Постановка задачі. Основною метою дослідження є аналіз роботи середовища Rucio з використанням інформації про передачу даних та створення нейронної мережі для розрахунку часу передачі файлу методом регресійного аналізу.

Історично задача регресії використовувалася при дослідженні впливу однієї групи безперервних випадкових величин на іншу групу безперервних величин.

У класичній задачі відтворення регресії навчальна вибірка – це набір незалежних об'єктів $X = \{x_i\}_{i=1}^{n}$, визначених вектором дійсних ознак $x_i = (x_{i,1},...,x_{i,d})$. Потрібно створити алгоритм (регресор), який за вектором ознак x повернув би точкову оцінку значення регресії $\hat{t}$, довірчий інтервал (t_-, t_+) або апостеріорний розподіл на безлічі значень регресійної змінної $p(t \mid x)$.

Підготовка даних. Для дослідження середовища Rucio та навчання нейронної мережі були зібрані метадані про успішні передачі файлів за один місяць (приблизно 20 мільйонів записів), які знаходились у форматі JSON. Ці дані було конвертовано у формат CSV за допомогою методів бібліотек PySpark та Pandas мови програмування Python. Із всіх доступних даних створено вибірки об'ємом від 10 тисяч до 2 мільйонів, щоб перевірити, чи схожі результати в усіх наявних даних та полегшити навантаження на апаратну частину під час дослідження. Побудувавши кореляційну таблицю (табл. 1) було визначено, що залежність між часом передачі файлів та розміром файлів становить лише 36%, тому потрібно визначати додаткові змінні, які впливають на час передачі файлів.

Було виділено 4 змінні в форматі часу, а саме: час створення запиту на трансфер, час підтвердження, час початку трансферу та час закінчення. Використовуючи ці змінні, розраховано час передачі файлів та час перебування файлів у черзі на передачу.

Додатковим кроком, було введено нову змінну (стовпець), яка охарактеризувала розділ файлів за розміром на окремі категорії (<100 Мб, 100-200 Мб, ...) що були ідентифіковані, базуючись на аналізі типових значень розміру файлів з набору даних.

Під час дослідження було з'ясовано, що деякі файли можуть знаходиться в черзі на передачу до 3 днів, а передаватися менше 1 хвилини. Тому введену змінну було використано лише для проведення кореляційного аналізу. Спираючись на графік залежності між часом початку передачі та часом закінчення (рис. 1), можна зробити висновок, що файли передаються не миттєво, а в залежності від розміру та інших характеристик. Також на рисунку можна побачити, як файли передаються в невеликих групах. Провівши перевірку змінних за допомогою кореляційного аналізу та численних експериментів з підстановкою різноманітних наборів змінних до нейронної мережі, залишено лише 7 змінних, які найбільше впливають на точність оцінювання часу передачі. Залишені змінні:

- «Account» - назва аккаунту від якого був запит на трансфер;
- «Activity» - діяльність з якою пов'язаний трансфер;
- «Scope» - область застосування;
- «Dst_Rse» - місце, куди передається файл;
- «Src_Rse» - місце, звідки передається файл;
- «Bytes» - розмір файла в байтах;
- «Transfer_Duration» - час передачі файлів.

Проведена стандартизація числових даних використовуючи метод StandardScaler з бібліотеки машинного навчання scikit-learn.

Кодування категоріальних даних в числовий масив проведено за допомогою методу OneHotEncoding з бібліотеки машинного навчання scikit-learn.

Таблиця 1. Кореляційна таблиця

	Bytes	Transfer_Duration	Sec_Created_At	Sec_Submitted_At	Sec_Started_At	Sec_Transferred_At	Queue_Duration
Bytes	1.0	0.36	0.07	0.07	0.07	0.07	0.09
Transfer_Duration	0.36	1.0	0.04	0.04	0.04	0.04	0.12
Sec_Created_At	0.07	0.04	1.0	0.99	0.99	0.99	0.03
Sec_Submitted_At	0.07	0.04	0.99	1.0	0.99	0.99	0.03
Sec_Started_At	0.07	0.0	0.99	0.99	1.0	1.0	0.05
Sec_Transferred_At	0.07	0.0	0.99	0.99	1.0	1.0	0.05
Queue_Duration	0.09	0.12	0.03	0.03	0.05	0.05	1.0

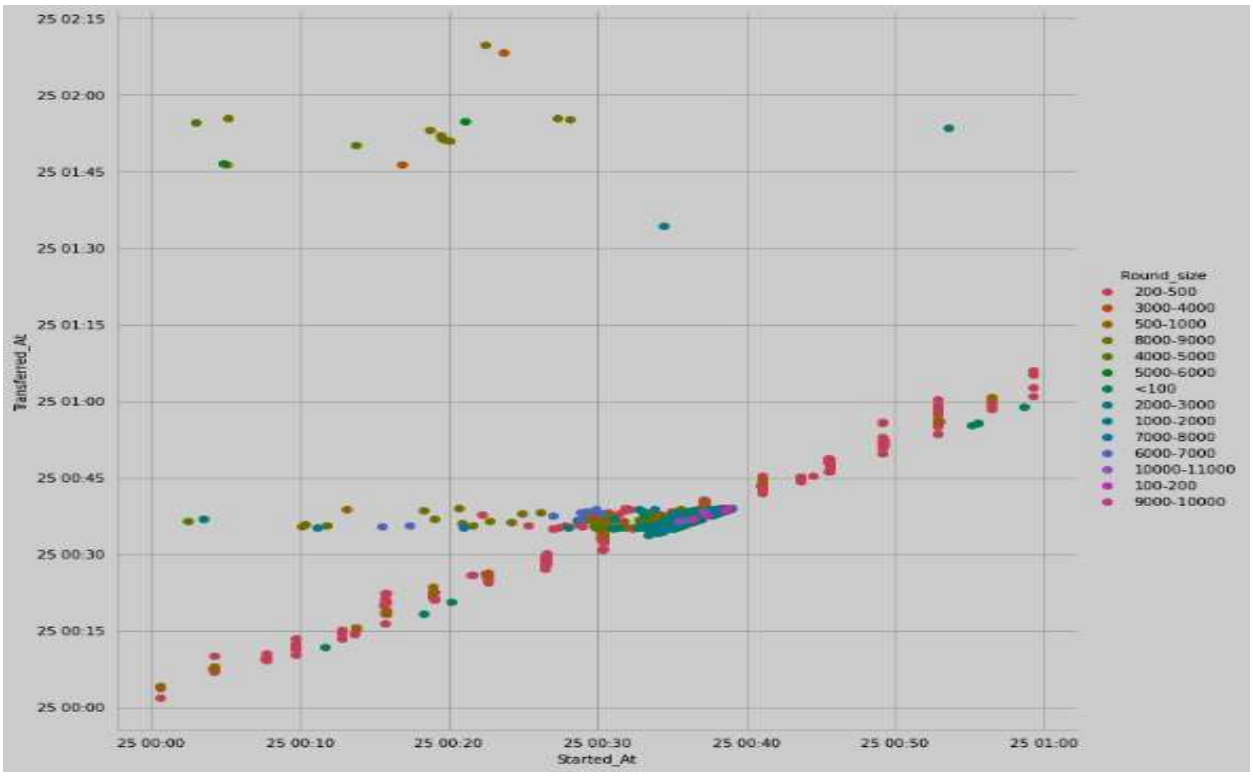

Рис. 1. Залежність між підтвердженням запиту та початком передачі

Програмна реалізація.

Під час вивчення даних було вирішено створити нейронну мережу з 2 вхідними шарами та 1 вихідним типу персептрон. Між ними є кілька шарів, в яких нейрони зменшуються послідовно, і шар, де дві гілки мережі об'єднуються в одну гілку. Схема нейронної мережі зображена на рис. 2.

Використовується функція активації «ReLu», а для останнього шару – функція «Linear».

Використання двох гілок обумовлено використанням числових і категоріальних даних. На вхід лівої гілки подаються категоріальні дані, а на вхід правої подаються числові дані.

Використані гіперпараметри для даної мережі:

- автоматично визначена кількість вхідних сигналів залежно від використовуваного фрейму даних;
- 1 вихідній сигнал;
- кількість епох навчання 200;
- кількість екземплярів тренувальної вибірки, що надходять на вхід за раз, становить 512;
- коефіцієнт навчання дорівнює 0,001;
- оптимізатор - метод оптимізації функції втрат "Adam";
- функція втрат - середня абсолютна похибка.

Підбір значень гіперпараметрів проводився експериментальним шляхом. Компіляція моделі проведена з оптимізатором – «Adam», та функцією втрат – «mean_absolute_error». Мережа навчена за допомогою методу зворотного виклику:

- EarlyStopping (patience = 50) – припинити тренування, коли показник, що відстежується, протягом 50 періодів не покращується;
- ReduceLROnPlateau(factor = 0.5, patience = 20) – зменшення швидкість навчання на коефіцієнт 0.5, коли метрика протягом 20 періодів не покращується.

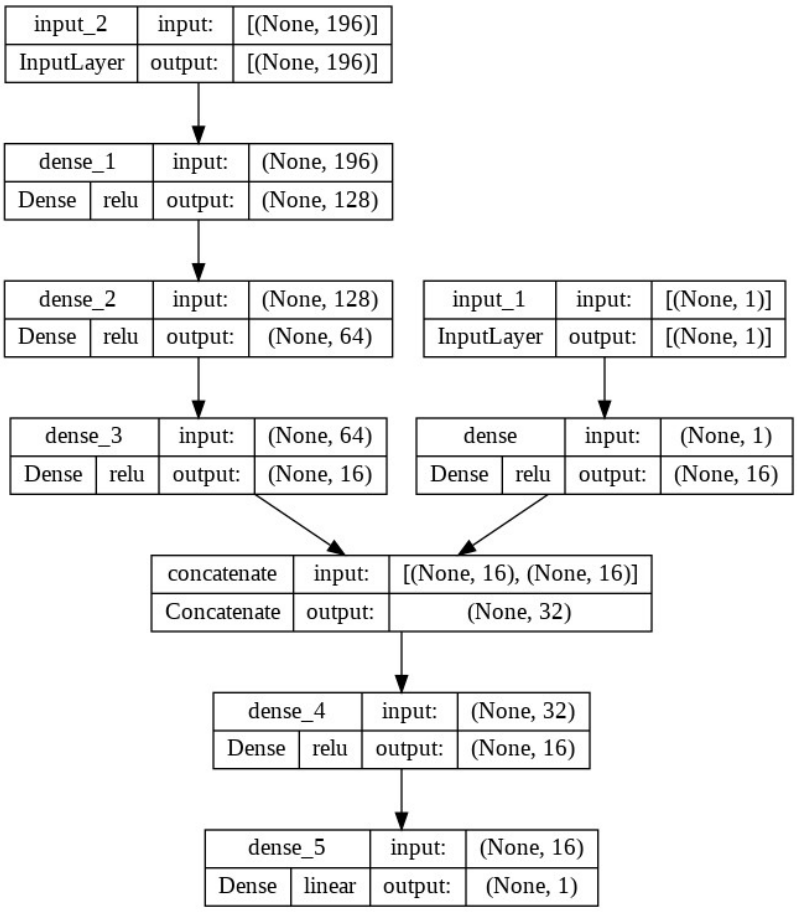

Рис. 2. Схема нейронної мережі

Результати передбачення.

На рис. 3 представлено графіки функції втрат, де графік 1 – це значення валідаційних даних, а графік 2 – значення тренувальних даних. На рис. 4 діаграму розсіювання де крапки – це тренувальні дані, а трикутники - це тестові дані, з яких можна зробити висновок, що алгоритм обробляє тестові дані майже з такою ж точністю, як і тренувальні. Це означає, що під час тренування персептрону вдалося уникнути перенавчання та алгоритм адекватно працює з іншими даними.

Значення метрик оцінки показані в табл. 2. Було прийнято рішення взяти RMSE як основну міру, оскільки вона має ті самі одиниці, що й вихідні значення (на відміну від MSE), і її легко

інтерпретувати. Метрика також працює з малими абсолютними значеннями, що корисно для комп'ютерних розрахунків.

Таблиця 2. Значення метрик не згрупованих даних

R2_val	0.656
R2_train	0.753
R2_test	0.799
Max_error	1521.351
MAE	47.897
RMSE	142.526

Під час дослідження було визначено, що дані групуються перед передачею, тому було прийнято рішення згрупувати файли по середньому значенню і перевірити відгук нейронної мережі.

Після перевірки, знову було побудовано графік функції втрат (рис. 5) і діаграму розсіювання (рис. 6). Функції втрат для тренувальних та тестових даних майже збігаються (рис. 5), а значення на діаграмі розсіювання близькі до діагоналі, що означає, що передбачені значення близькі до цільових. Метрика RMSE має значення 59.9, MAE – 29.7, а R2 для тестових даних – 93% (табл. 3).

Таблиця 3. Значення метрик згрупованих даних

R2_val	0.841
R2_train	0.915
R2_test	0.932
Max_error	250.334
MAE	29.716
RMSE	59.906

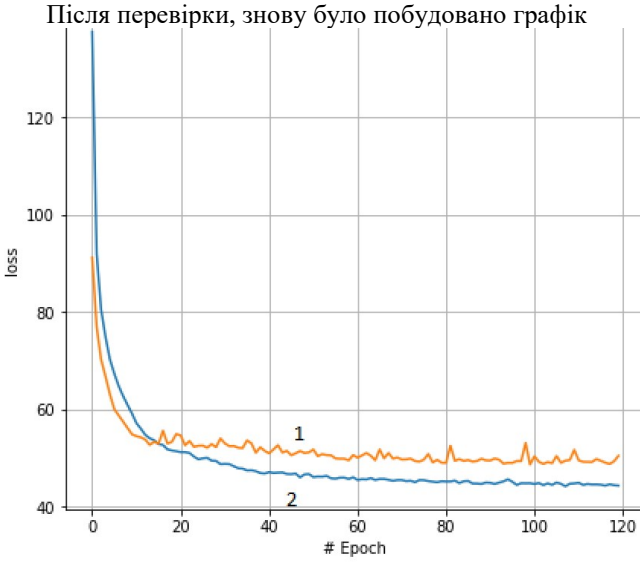

Рис. 3. Графік функції втрат

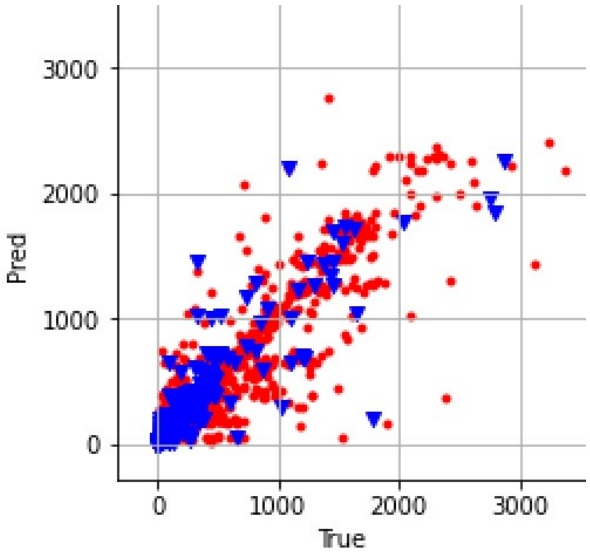

Рис. 4. Графік розсіювання

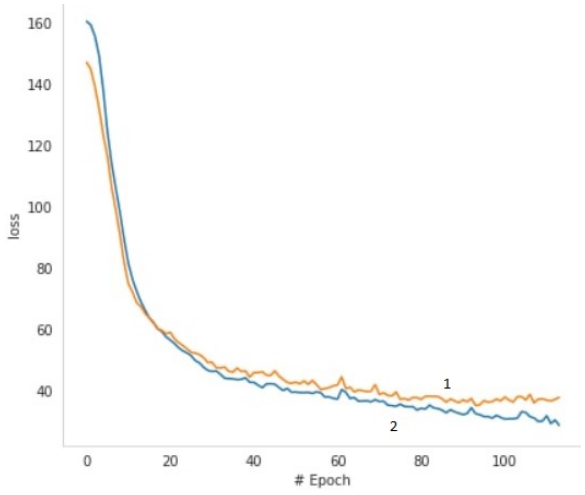

Рис. 5. Графік функції втрат для згрупованих даних

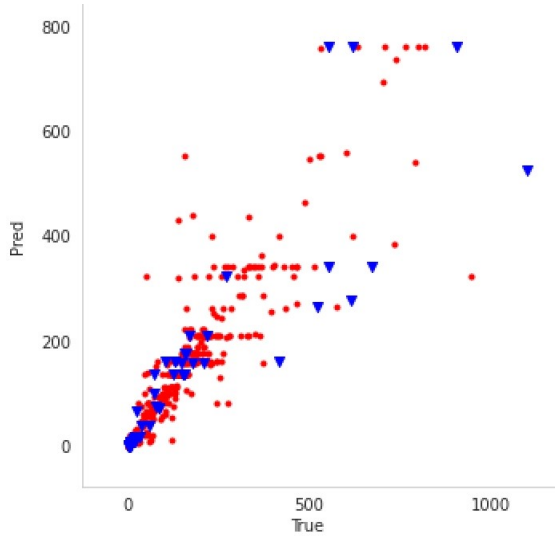

Рис. 6. Графік розсіювання для згрупованих даних

Висновки з даного дослідження і перспективи подальших розробок у даному напрямку.

Було проведено дослідження даних подій середовища Rucio, вивчено властивості та поведінку

середовища Rucio та створено програмне забезпечення для розрахунку часу передачі великих наборів даних за допомогою нейронних мереж.

Проаналізувавши дані стосовно часу передачі файлів, виявлено, що не існує чіткої кореляції між часом передачі та одним з вхідних параметрів. Час передачі файлу залежить від ряду зовнішніх факторів, які неможливо отримати за допомогою метаданих, але за допомогою отриманих даних можливо частково дослідити середовище керування розподіленою інформацією.

Була побудована модель нейронної мережі для прогнозування часу передачі файлів на основі метаданих передачі. Визначено, що при використанні згрупованих файлів під час навчання нейронної мережі значно покращуються результати оцінювання, показники R2_score збільшується до 93%, а RMSE зменшується до 59.9. Отриманий показник є цілком задовільним.

Список літератури

1. Barisits, M., Beermann, T., Berghaus, F. (2019), “Rucio: Scientific Data Management, Springer CSBS”, doi: https://doi.org/10.1007/s41781-019-0026-3
2. Bogado J., Lassnig M., Monticelli F., Diaz J., Beermann T. (2020), Zenodo, doi: https://doi.org/10.5281/zenodo.4320937
3. Lassnig M., Toler W., Vamosi R., Bogado J. (2017), Journal of Physics: Conference Series 898, 062009, doi: https://doi.org/10.1088/1742-6596/898/6/062009
4. Begy V., Barisits M., Lassnig M., Schikuta E. (2020), Journal of Computational Science 44, 101158, doi: https://doi.org/10.1016/j.jocs.2020.101158
5. Bogado J., Monticelli F., Diaz J., Lassnig M., Vukotic I IEEE 14th International Conference on e-Science. 2018, pp. 334–335.
6. Руденко В. М., Математична статистика: навч. посіб. – К.: Центр учбової літератури, 2012. – 304 с.
7. Бараз В.Р., Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel / Бараз В.Р. – Екатеринбург, 2005. – 103с
8. Гайдаєнко О.М., Факторний аналіз ефективності використання основних засобів на прикладі ПАТ «Одескабель» [Електронний ресурс] – 2016. – Режим доступу до ресурсу: http://dspace.oneu.edu.ua/jspui/handle/123456789/4685
9. Харченко Ю.А., Кореляційно-регресійний аналіз обсягів збуту продукції промислового підприємства / Ю.А. Харченко // Економічний простір. – 2014. –№ 86. – С. 214–223.
10. Настенко Є. А., Якимчук В. С., Носовець О. К., Інтелектуальний аналіз даних: методичні вказівки до виконання комп’ютерних практикумів з навчальної дисципліни «Інтелектуальний аналіз даних». Частина-1. «Кореляційний та регресійний аналіз медичних даних». – К.: НТУУ «КПІ ім. І. Сікорського», 2017. – 51 с.
11. Енгалыч Г. А. (2019), Методы решения задачи регрессии. [Електронний ресурс]. – Режим доступу до ресурсу: https://dspace.spbu.ru/bitstream/11701/25903/1/document.pdf
12. A Walk-through of Regression Analysis Using Artificial Neural Networks in Tensorflow, available at: https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/
13. Asadujjaman, Md & Supto, Mushfiqur. Implementation of Artificial Neural Network on Regression Analysis (2021), doi: https://doi.org/10.1109/SMC53803.2021.9569881

References (transliterated)

1. Barisits, M., Beermann, T., Berghaus, F. (2019), “Rucio: Scientific Data Management, Springer CSBS”, doi: https://doi.org/10.1007/s41781-019-0026-3
2. Bogado J., Lassnig M., Monticelli F., Diaz J., Beermann T. (2020), Zenodo. doi: https://doi.org/10.5281/zenodo.4320937
3. Lassnig M., Toler W., Vamosi R., Bogado J. (2017), Journal of Physics: Conference Series 898, 062009, doi: https://doi.org/10.1088/1742-6596/898/6/062009
4. Begy V., Barisits M., Lassnig M., Schikuta E. (2020), Journal of Computational Science 44, 101158, doi: https://doi.org/10.1016/j.jocs.2020.101158
5. Bogado J., Monticelli F., Diaz J., Lassnig M., Vukotic I. IEEE 14th International Conference on e-Science. 2018, pp. 334–335.
6. Rudenko, V. M. Matematychna statystyka: navch. posib. – K.: Tsentr uchbovoyi literatury, 2012. – 304 p.
7. Baraz V.R. Korrelyatsionno-regressionnyy analiz svyazi pokazateley kommercheskoy deyatel'nosti s ispol'zovaniyem programmy Excel / Baraz V.R. – Yekaterinburg, 2005. – 103 p.
8. Haydayenko O.M. (2016), Faktornyy analiz efektyvnosti vykorystannya osnovnykh zasobiv na prykladi PAT «Odeskabel′» available at: http://dspace.oneu.edu.ua/jspui/handle/123456789/4685
9. Kharchenko YU.A. Korelyatsiyno-rehresiynyy analiz obsyahiv zbutu produktsiyi promyslovoho pidpryyemstva / YU.A. Kharchenko // Ekonomichnyy prostir.– 2014. –№ 86. – pp. 214–223.
10. Nastenko YE. A., Yakymchuk V. S., Nosovets′ O. K., Intelektual′nyy analiz danykh: metodychni vkazivky do vykonannya komp″yuternykh praktykumiv z navchal′noyi dystsypliny «Intelektual′nyy analiz danykh». Chastyna-1. «Korelyatsiynyy ta rehresiynyy analiz medychnykh danykh». – K.: NTU «KPI im. I. Sikors′koho», 2017. 51 p.
11. Yenhalych H.A. (2019), Metody resheniya zadachi regressii. available at: https://dspace.spbu.ru/bitstream/11711/25903/1/document.pdf
12. A Walk-through of Regression Analysis Using Artificial Neural Networks in Tensorflow, available at: https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/
13. Asadujjaman, Md & Supto, Mushfiqur. Implementation of Artificial Neural Network on Regression Analysis (2021), doi: https://doi.org/10.1109/SMC53803.2021.9569881

Надійшла (received) 21.10.2022

Відомості про авторів / Сведения об авторах / About the Authors

Броварник Олексій Олексійович (Броварник Алексей Алексеевич, Brovarnyk Oleksii Oleksiiovych) – студент 6 курсу НТУ «ХПІ», м. Харків, Україна;

Овсяніков Владислав Валерійович (Овсяников Владислав Валерьевич, Ovsianikov Vladyslav Valerievich) – аспірант кафедри комп'ютерного моделювання процесів та систем, Національний технічний університет «Харківський політехнічний інститут»

О. О. BROVARNYK, V. V. OVSIANIKOV

DATA MANAGEMENT ENVIRONMENT PROPERTIES INVESTIGATION AND TIME ESTIMATION OF LARGE DATA SET TRANSFER

The article considers the task of estimating the time of transmission of large data sets through a distributed data management environment based on a self-created neural network model and investigating the properties of this environment using statistical analysis methods. For the initial analysis, metadata for successful file transfers in the system was obtained, variables that affect file transfer time were transformed and highlighted. Different samples were used in the analysis to check whether the results were similar across the available data. The methods of correlation and regression analysis are applied for the study of the environment. It was found that there is no clear correlation between the transmission time and one of the input parameters. The file transfer time depends on a number of external factors that cannot be obtained using metadata, but it is possible to partially investigate the environment using the obtained metadata. A model based on two input levels for numerical and categorical variables was used and then combined into one branch. RMSE metric value and a scatter plot are used to display the prediction results to compare the target and predicted values. The performed calculations show satisfactory prediction results;

Keywords: research of properties, transmission time estimation, methods of statistical analysis, data analysis, regression, correlation, neural network.

О. О. БРОВАРНИК, В. В. ОВСЯНІКОВ

ИССЛЕДОВАНИЕ СВОЙСТВ СРЕДЫ УПРАВЛЕНИЯ ДАННЫМИ И ОЦЕНКА ВРЕМЕНИ ПЕРЕДАЧИ БОЛЬШИХ НАБОРОВ ДАННЫХ

В статье рассматривается задача оценки времени передачи больших наборов данных через распределенную среду управления данными на основе самостоятельно созданной модели нейронной сети и исследования свойств этой среды с помощью методов статистического анализа. Для начального анализа получены метаданные для успешной передачи файлов в системе, трансформированы и выделены переменные, которые влияют на время передачи файлов. При анализе использованы различные выборки для проверки, похожи ли результаты во всех имеющихся данных. Использованы методы корреляционного, регрессионного анализа для исследования среды. Обнаружено, что нет четкой корреляции между временем передачи и одним из входных параметров. Время передачи файла зависит от ряда внешних факторов, которые невозможно получить с помощью метаданных, но можно частично исследовать среду используя полученные метаданные. Использована модель на основе двух входных уровней для числовых и категориальных переменных, затем объединенных в одну ветвь. Для изображения результатов предсказания используются RMSE и диаграмма рассеяния для сравнения целевых и предсказанных значений. Произведенные расчеты показывают удовлетворительные результаты предсказаний;

Ключевые слова: исследование свойств, оценка времени передачи, методы статистического анализа, анализ данных, регрессия, корреляция, нейронная сеть.