DOI 10.34185/1562-9945-6-143-2022-01
УДК 004.62

Є.С. Сулема, О.Р. Пеня

ЗАСТОСУВАННЯ КЛАСТЕРИЗАЦІЇ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ ЛІНІЙНИХ НАБЛИЖЕНЬ

Анотація. У статті запропоновано спосіб підвищення точності моделювання об'єкта дослідження на основі набору темпоральних мультимодальних даних за допомогою ліній них наближень з використанням кластеризації. Запропонований підхід може бути засто сований при створенні цифрового двійника досліджуваного об'єкта.
Ключові слова: цифрові двійники, темпоральні мультимодальні дані, аналіз даних.

Вступ. Сучасні технологічні засоби спостереження, моніторингу та вимірювання дозволяють генерувати величезний об'єм гетерогенних даних, що описують стан та поведінку систем та процесів, таких як розумні будинки, автоматичні виробництва та випробувальні стенди, медичні дослідження тощо. Внаслідок цього виникає необхідність аналізу та використання цих даних з великою точністю для отримання розуміння процесів, що спостерігаються, та побудови цифрових моделей складних систем та процесів, зокрема цифрових двійників, з метою дослідження, вивчення та оптимізації прийняття рішень [1–3]. Такі набори даних породжують цілу низку проблем та викликів, які становлять значний інтерес для дослідників у різних областях науки та техніки, наприклад, їх отримання, синхронізація, зберігання і ефективні обробка та аналіз [4].

За своєю природою мультимодальні дані містять більше інформації, ніж значення модальностей, зафіксовані окремо, завдяки наявності зв'язків між ними. Хоча ці зв'язки можуть бути невідомі безпосередньо, про них можна зробити певні висновки з використанням методів поєднання даних (data fusion). Поєднання даних визначається як процес комбінування значень, що надходять з різних джерел, наприклад, сенсорів різних типів, для більш повного опису середовища, процесу або об'єкта, що досліджується, таким чином, щоб при цьому суттєво зросла деяка інформаційна метрика [5]. З практичної точки зору поєднання даних дозволяє отримати більш точні моделі

досліджуваних явищ на основі різних джерел інформації, які використовуються в різних цілях, в тому числі для побудови цифрових двійників – комп'ютерних моделей реальних систем, які досліджуються для аналізу інформації та прийняття рішень в умовах, коли відповідні експерименти з реальною системою не є доцільними з міркувань вартості, безпеки тощо. У таких системах модальності впливають одна на одну, забезпечуючи більш точне представлення об'єкта дослідження.

Постановка проблеми. Задача методів моделювання та аналізу в цілому полягає у створенні цифрової моделі об'єкта дослідження за набором темпоральних мультимодальних даних з попередньо невідомими зв'язками, що дозволятиме передбачення з більшою точністю, ніж одне лінійне наближення. Вхідний набір даних вважатимемо повним та синхронізованим за часом.

Ця стаття присвячена використанню кластеризації для аналізу наборів темпоральних мультимодальних даних, які характеризують об'єкт дослідження з метою підвищення точності моделювання.

Аналіз літературних джерел. У роботі [6] запропоновано базову модель цифрового двійника, яка призначена для аналізу та моделювання виробничих процесів. Вона складається з чотирьох рівнів: рівня фізичного простору виробничого процесу, рівня комунікаційної системи, рівня цифрового двійника та рівня користувацького простору. Програмна система для реалізації цифрового двійника включає модуль керування, модуль симуляції, модуль передбачення і виявлення аномалій, а також модуль збереження даних (хмарне сховище). Недоліком запропонованої моделі є неможливість її використання для інших галузей застосування цифрових двійників.

У роботі [7] представлено метод створення цифрової моделі виробничого пристрою та передбачення дефектів виробництва та зносу робочих частин завдяки збору й обробці опосередкованих даних зовнішніх сенсорів.

Запропонований у роботі [8] метод дозволяє забезпечити зв'язок між різними моделями промислового об'єкта для коректної візуалізації цифрового двійника. Цей метод не може бути застосований для цифрового подання об'єктів іншої природи, наприклад, медико-біологічних об'єктів.

Аналіз цих та інших літературних джерел, опублікованих нещодавно, дозволяє стверджувати про необхідність удосконалення методів подання та аналізу темпоральних мультимодальних даних для комплексного подання наборів даних та формування цифрових двійників досліджуваних об'єктів.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Подання темпоральних мультимодальних даних. Для формальної специфікації цифрових двійників досліджуваних об'єктів потрібен математичний апарат, який забезпечить логіку подання та оброблення темпоральних мультимодальних даних з урахуванням властивостей цих даних, а саме, мультимодальності (цифровий двійник визначається сукупністю даних різної природи; логіка подання та оброблення даних цифрового двійника залежить від якісного та кількісного складу цієї сукупності даних) та темпоральності (елементи сукупності даних цифрового двійника є впорядкованими за часом їх отримання; порядок слідування окремих елементів послідовності даних впливає на результат оброблення всієї сукупності даних цифрового двійника). Таким математичним апаратом є алгебраїчна система агрегатів [10].

Агрегат A – це впорядкована скінчена сукупність елементів, яка визначається кортежем множин $\{A\}$ та кортежем кортежів елементів $\langle A\rangle$, причому елементи a_i^j кожного кортежу елементів $\langle a_i^j\rangle_{i=1}^{n_j} \in \langle A\rangle$ належать до відповідної множини $M_j \in \{A\}$, $j = [1 \dots N]$, що задає взаємно-однозначний зв'язок між порядком слідування множин та порядком слідування кортежів елементів:

$$A = [\![M_j | \langle a_i^j \rangle_{i=1}^{n_j}]\!]_{j=1}^{N} = [\![\{A\} | \langle A \rangle]\!], \quad (1)$$

де $\{A\}$ – кортеж множин M_j; $\langle A\rangle$ – кортеж кортежів елементів $\langle a_i^j\rangle_{i=1}^{n_j}$; a_i^j – окремий елемент (значення) або складений елемент (кортеж однорідних значень або кортеж різнорідних значень), $a_i^j \in M_j$.

Розглянемо об'єкт спостереження S, який виявляє свою сутність через набір властивостей $F_1, F_2, \dots, F_N$, які можуть бути виміряні. В багатьох випадках ці властивості є взаємопов'язаними, оскільки вони представляють стан одного і того ж об'єкта та, фактично, є різними проявами його поведінки, що змінюється з плином часу. Комплексне подання даних, що відображають характеристики досліджуваного об'єкта, сприяє кращому розумінню загальної картини поведінки цього об'єкта. Виміряні значення властивостей досліджуваного об'єкта S можна представити як агрегат A_S, компонентами якого є кортежі значень $\langle f_i^j \rangle_{i=1}^{n_j}$ властивостей F_j ($j = 1 \dots N$) об'єкта S:

$$A_S = [\![M_1, M_2, \dots, M_N | \langle f_1^1, f_2^1, \dots, f_{n_1}^1 \rangle, \langle f_1^2, f_2^2, \dots, f_{n_2}^2 \rangle, \dots \langle f_1^N, f_2^N, \dots, f_{n_N}^N \rangle]\!] \quad (2)$$

Розглянемо спосіб дослідження зв'язку між даними різних модальностей об'єкта дослідження, визначеного (2), які визначають його властивості.

Використання кластеризації даних для підвищення точності моделювання. У випадку, коли поведінка систем описується нелінійним законом, або сукупністю законів в залежності від певних умов, єдина апроксимація мультимодальних даних, яка будується в більшості методів поєднання даних стає значно менш ефективною через викиди, що сильно відхиляються від лінійного наближення.

Застосування в такому випадку методів факторизації або компонентного аналізу дозволяє описати систему більш точно за рахунок появи більшої кількості ступенів свободи, або введенню принципово інших властивостей, але при цьому все одно будується єдине лінійне наближення, обчислення, аналіз та інтерпретація якого є нетривіальною задачею.

Для вирішення даної проблеми пропонується виконати розділення області даних на інтервали, в яких лінійні наближення будуть давати задовільну точність – виконати кластеризацію з урахуванням самих значень точок даних та їх статистичних характеристик.

Загальний алгоритм, що пропонується для цього є модифікацією методу DBSCAN, у якому метрикою відстані є відхилення точки від лінійного наближення по поточному кластеру:

1. Обрати мінімальну кількість точок m.

2. Ініціалізувати список відхилень $R = \emptyset$ та список кандидатів $A = \emptyset$, лічильник кластерів $i = 0$, поточний кластер $C_i = \emptyset$.

2. У впорядкованому за незалежною величиною наборі даних побудувати лінійне наближення залежної величини за першими m точками, додати їх у поточний кластер C_i та обчислити середнє відхилення s.

3. Для кожної наступної точки P обчислити відхилення d від наближення на поточному кластері.

3.1. Якщо $d \leq s$:
3.1.1. Якщо $|R| = 0$, $C_i = C_i \cup \{P\}$;
3.1.2. Інакше якщо$|R|+|A| = m$:
3.1.2.1. Якщо $|R| \geq |A|$, то $i = i + 1$, $C_i = R \cup A$;
3.1.2.2. Інакше $C_i = C_i \cup R \cup A$,
3.1.2.3. Переобчислити наближення та s для поточного кластера;
3.1.2.4. $R = \emptyset$, $A = \emptyset$, перейти до п. 3
3.1.3. Інакше якщо $d \leq s$, то $A = A \cup \{P\}$
3.2. Інакше $R = R \cup \{P\}$, перейти до п. 3.1.2.
4. $C_i = C_i \cup R \cup A$

5. Отримано множну кластерів C та відповідні наближення та них.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Запропонований алгоритм по суті є жадібним алгоритмом, який додає точки даних до поточного кластера, поки не накопичить достатню кількість точок, що відхиляються від лінійної моделі даних для поточного кластера, які сигналізують про необхідність виділення нового кластера та віднесення накопчених точок до нього. Таким чином, отримані наближення є чисельно більш точними, ніж застосування єдиного наближення на всій множині даних.

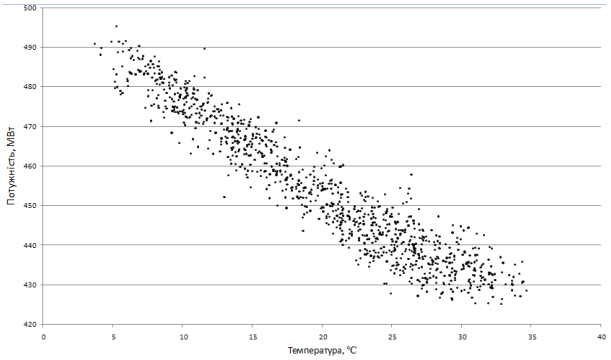
Застосування методу та аналіз результатів. Для дослідження зв'язку між даними різних модальностей розглянемо набір даних містить 9568 точок даних, зібраних з комбінованої електростанції протягом 6 років (2006-2011), коли електростанція була налаштована на роботу з повним навантаженням. Характеристики складаються із середніх годинних значень температури навколишнього середовища (AT), тиску навколишнього середовища (AP), відносної вологості (RH), тиску носія (V) та вихідної потужності електростанції (PE).

Середні значення беруться з різних сенсорів, розташованих на електростанції, які щосекунди реєструють параметри навколишнього середовища. Значення задані без нормування [10].

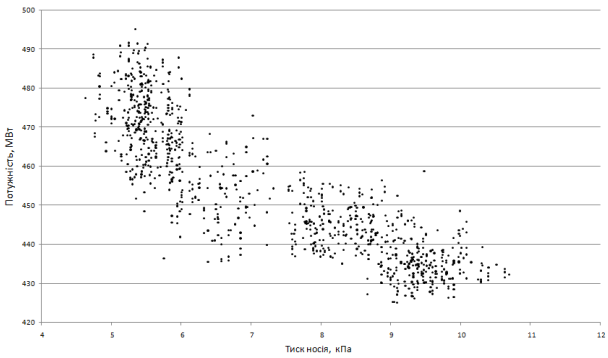
Зв'язки між наведеними модальностями не є очевидними і їх складно повною мірою виявити з наявного набору даних через його розмірність, проте можна застосувати статистичний аналіз та візуалізацію для того, щоб отримати деяке уявлення про зв'язки між модальностями. Їх розуміння дозволяє побудову більш ефективних моделей, що поєднують модальності.

Вважатимемо температуру, тиск, вологість та тиск носія незалежними величинами, оскільки їх визначають приховані особливості процесів установки, а вихідну потужність – залежною, оскільки вона є похідною відповідних процесів, які визначають інші величини. Для дослідження зв'язку між модальностями використаємо графічне представлення та кореляційний аналіз. На рис. 1 (a-e) зображено залежності між деякими модальностями для рівномірної вибірки із тисячі точок даних.

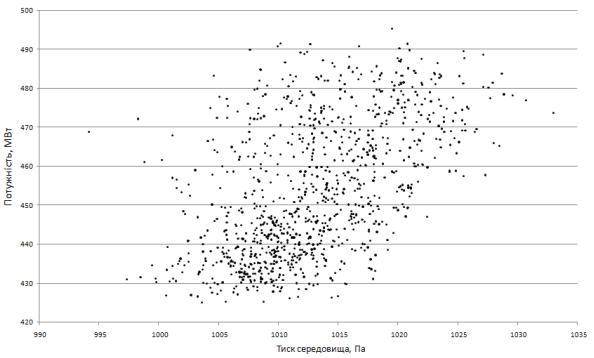
Хоча зображені залежності відображають лише попарний зв'язок між величинами, і не враховують зв'язки між ними, можна зробити висновок, що вихідна потужність станції знаходиться в лінійній залежності від температури середовища, і в експоненціальній залежності від тиску носія. В той час вологість та тиск середовища слабо пов'язані з потужністю, але мають певний зв'язок із великою дисперсією з температурою, що може свідчити про наявність прихованих змінних, які визначають їх співвідношення.



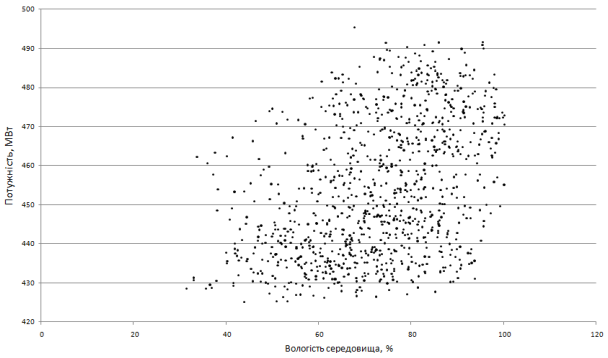
(а) Температура – Потужність



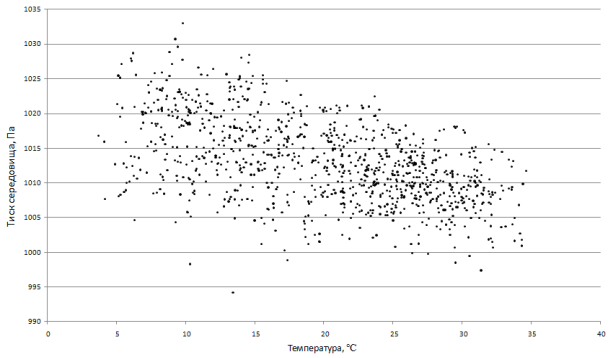
(б) Тиск носія – Потужність



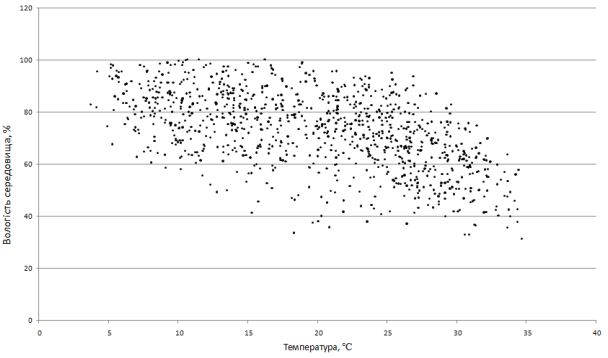
(в) Тиск середовища – Потужність



(г) Вологість середовища – Потужність



(д) Температура – Тиск середовища



(е) Температура – Вологість середовища

Рисунок 1 - Залежності між даними деяких модальностей

Ці висновки підтверджуються кореляційним аналізом (табл. 1, 2). Висока кореляція Пірсона між температурою, тиском носія та потужністю свідчить про наявність залежності між ними, близької до лінійної. Так само, висока кореляція Спірмена означає монотонну (не обов’язково лінійну) залежність між даними величинами.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 1

Кореляційна матриця модальностей (коефіцієнт Пірсона)

	AT	V	AP	RH	PE
AT	1.000	0.844	-0.508	-0.543	-0.948
V	0.844	1.000	-0.414	-0.312	-0.870
AP	-0.508	-0.414	1.000	0.010	0.518
RH	-0.543	-0.312	0.010	1.000	0.370
PE	-0.948	-0.870	0.518	0.390	1.000

Таблиця 2

Кореляційна матриця модальностей (коефіцієнт Спірмена)

	AT	V	AP	RH	PE
AT	1.000	0.851	-0.519	-0.543	-0.944
V	0.851	1.000	-0.426	-0.305	-0.884
AP	-0.519	-0.426	1.000	0.087	0.543
RH	-0.543	-0.305	0.087	1.000	0.390
PE	-0.944	-0.884	0.543	0.390	1.000

З урахуванням отриманих висновків, застосуємо запропонований алгоритм для будови лінійної моделі для обчислення вихідної потужності за температурою та тиском.

Результат використання алгоритму із m = 1435 (15% від загальної кількості даних щоб забезпечити стійкість до відхилень та уникнути виділення великої кількості кластерів) дозволяє отримати один кластер для температури та два кластери для тиску (рис. 2).

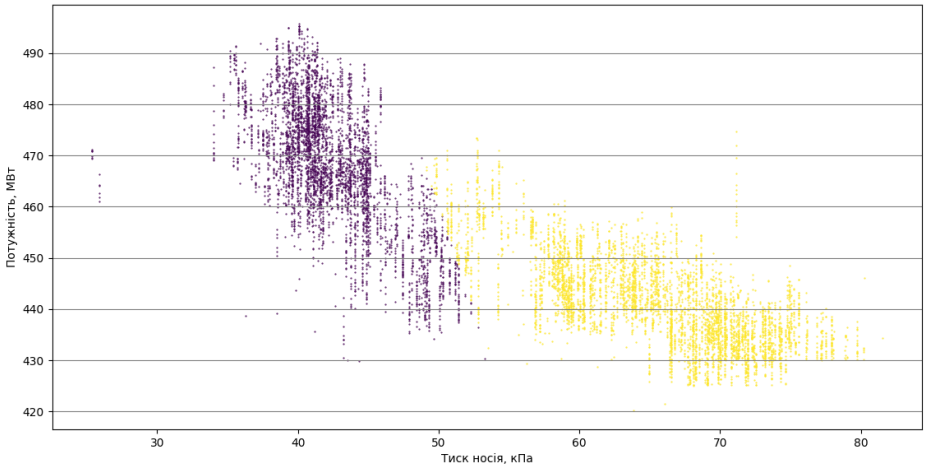


Рисунок 2 - Результат кластеризації за тиском

Для використання отриманого поділу і побудови наближень необхідно скоригувати отримані результати кластеризації таким чином, щоб приналежність вхідної точки до певного інтервалу однозначно визначалась набором залежних змінних. Для цього, межа поділу за тиском обрана як точка, рівновіддалена від крайніх точок відповідних кластерів (рис. 3).

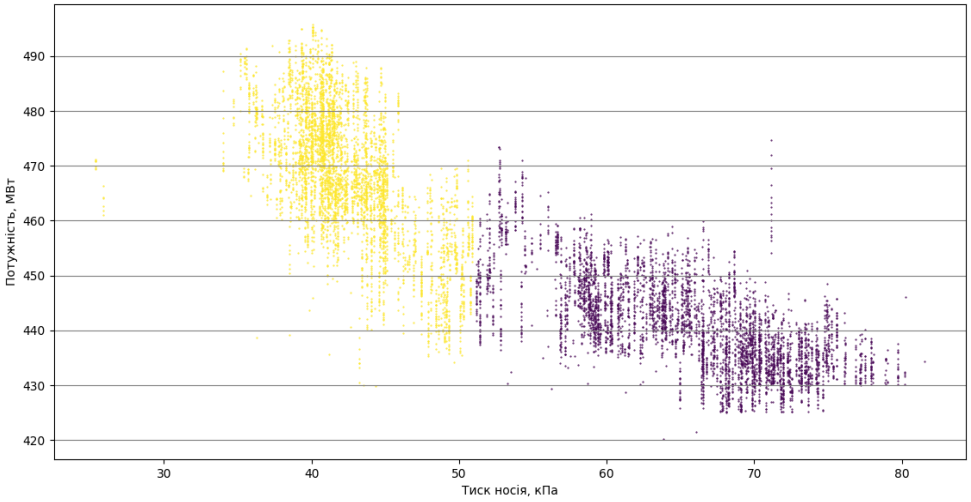


Рисунок 3 - Поділ, однозначний за незалежними змінними

Для виділених інтервалів будуються та використовуються окремі наближення. Для побудови моделі використовувались 90% наявних у наборі даних, і відповідно 10% використовувались для її оцінки на кожному інтервалі. Наближення побудовані та оцінені з використанням програмного пакету Scikit Learn мови python для множинної лінійної регресії та оцінки моделей. В якості параметрів оцінки моделі обрано середньоквадратичну помилку та коефіцієнт детермінації.

Використання єдиного наближення дозволяє отримати модель із середньоквадратичною помилкою 24.13 та коефіцієнтом детермінації 0.91. Модель із окремими наближеннями на кластерах, побудована з використанням запропонованого методу, відповідно має середньоквадратичне відхилення 21.49 та коефіцієнтом детермінації 0.93, тобто помилка моделі зменшилась на 11% і вона краще пояснює вплив незалежних змінних на залежну.

Також, у рамках дослідження за набором даних були побудовані та оцінені інші можливі наближення, результати наведено в табл. 3.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 3

Побудова та оцінка інших наближень

Залежність	Середньоквадратичне відхилення	Коефіцієнт детермінації
Лінійна від тиску	72.20	0.75
Лінійна від температури	29.28	0.93
Лінійна від усіх параметрів	33.72	0.91
Лінійна від температури та квадратична від тиску	66.51	0.82
Лінійна від температури та експоненціальна від тиску	2049.39	0.76

Висновки. У роботі запропоновано метод побудови моделей об’єктів за набором мультимодальних даних з використанням кластеризації для підвищення точності. Так, запропонований підхід можна використовувати в моделюванні при створенні цифрових двійників.

Запропонований алгоритм кластеризації може бути модифікований в залежності від особливостей задачі та набору даних. Наприклад, кількість точок у початковій множині, що формує перший кластер може бути більша, ніж мінімальна кількість точок у кластері, щоб зменшити вплив граничних значень, або умова виділення нового кластеру може залежати від співвідношення наборів відхилень та кандидатів на додавання в поточний кластер.

Було розглянуто його застосування запропонованого методу на наборі даних, що описує характеристики електростанції комбінованого циклу з метою передбачення вихідної потужності за характеристиками середовища. В результаті, досліджено зв’язки між модальностями та проведено низку експериментів з моделювання системи з урахуванням виявлених залежностей.

Застосування запропонованого підходу призвело до підвищення точності моделювання (значення середньоквадратичної помилки, що використовується як метрика точності, зменшилось на 11%). При цьому лінійні моделі мають значно кращу точність завдяки алгоритмам обчислення параметрів, що менш схильні до перенавчання та є чисельно стійкішими у порівнянні з іншими видами апроксимації.

Недоліком методу є обчислювальна складність через необхідність виконання кластеризації, обчислення параметрів та зберігання більшої

кількості моделей, що описують систему. За наявності більшої кількості даних, модальностей та варіацій в поведінці системи їх кількість може бути значно більшою і може призвести до зменшення продуктивності та точності моделі.

ЛІТЕРАТУРА

1. Norris S. Systematically working with multimodal data: Research methods in multimodal discourse analysis. John Wiley & Sons, 2019.

2. Gao J. et al. A survey on deep learning for multimodal data fusion. Neural Computation. 2020. Т. 32. №. 5. С. 829–864.

3. Worsley M. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. Proceedings of the 14th ACM international conference on Multimodal interaction. 2012. С. 353–356.

4. Lahat D., Adali T., Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE. 2015. Т. 103. №. 9. С. 1449–1477.

5. Raol J. R. Data fusion mathematics: theory and practice. CRC Press, 2015.

6. Bevilacqua M. et al. Digital Twin Reference Model Development to Prevent Operators' Risk in Process Plants. Sustainability, 2020, Issue 12, Paper 1088, 17 p.

7. Cai Y. et al. Sensor data and information fusion to construct digital-twins virtual machine tools for cyber-physical manufacturing //Procedia manufacturing. – 2017. – Т. 10. – С. 1031-1042.

8. Talkhestania B.A., Jazdib N., Schlöglc W., Weyrich M. A concept in synchronization of virtual production system with real factory based on anchor-point method. Procedia CIRP, 2018. Vol. 67, P. 13–17.

9. Sulema Ye., Kerre E., et al. Mathematical Methods in Interdisciplinary Sciences. Wiley, USA, 2020. 464 p.

10. Tüfekci P. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. Intern. Journal of Electrical Power & Energy Systems. 2014. Т. 60. С. 126–140.

REFERENCES

1. Norris S. Systematically working with multimodal data: Research methods in multimodal discourse analysis. John Wiley & Sons, 2019.

2. Gao J. et al. A survey on deep learning for multimodal data fusion. Neural Computation. 2020. Т. 32. №. 5. С. 829–864.

3. Worsley M. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. Proceedings of the 14th ACM international conference on Multimodal interaction. 2012. С. 353–356.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

4. Lahat D., Adali T., Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE. 2015. T. 103. №. 9. C. 1449–1477.
5. Raol J. R. Data fusion mathematics: theory and practice. CRC Press, 2015.
6. Bevilacqua M. et al. Digital Twin Reference Model Development to Prevent Operators' Risk in Process Plants. Sustainability, 2020, Issue 12, Paper 1088, 17 p.
7. Cai Y. et al. Sensor data and information fusion to construct digital-twins virtual machine tools for cyber-physical manufacturing //Procedia manufacturing. – 2017. – T. 10. – C. 1031-1042.
8. Talkhestania B.A., Jazdib N., Schlöglc W., Weyrich M. A concept in synchronization of virtual production system with real factory based on anchor-point method. Procedia CIRP, 2018. Vol. 67, P. 13–17.
9. Sulema Ye., Kerre E., et al. Mathematical Methods in Interdisciplinary Sciences. Wiley, USA, 2020. 464 p.
10. Tüfekci P. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. Intern. Journal of Electrical Power & Energy Systems. 2014. T. 60. C. 126–140.

Received 01.11.2022.
Accepted 07.11.2022.

Application of clustering to improve the accuracy of linear approximations

The paper presents an approach to increase the accuracy of modelling an object of research based on a temporal multimodal data set with linear approximations using clustering. The proposed approach can be applied for creating digital twins of a researched object.

The purpose of the study as a whole is to create a digital twin of the researched object based on a set of temporal multimodal data with previously unknown relationships, which will allow predictions with greater accuracy than a single linear approximation. The input data set is considered as complete and synchronized. This paper focuses on the use of clustering to analyse the sets of temporal multimodal data that characterize the researched object.

The paper presents a method for dividing the data space into intervals, where linear approximations will be more accurate, by clustering based on the values of data points and their statistical characteristics for independent variables that show a nonlinear relationship with the dependent variable. As a result, the accuracy in models that use a linear approximation for a given value has increased (the value of the mean square error used as an accuracy metric has decreased by 11%). At the same time, linear models have much

better accuracy due to algorithms for calculating parameters that are less prone to overfitting and are more numerically stable. However, the proposed method is more computationally expensive due to the need to perform clustering, calculate intermediary approximations and store more models that describe the system. If there is more data, modalities and variations in the behaviour of the system, their number can be much larger and can lead to some reduction in productivity and accuracy.

Keywords: digital twins, temporal multimodal data, data analysis.

Сулема Євгенія Станіславівна - д.т.н., зав. кафедри ПЗКС Київського політехнічного інституту ім. Ігоря Сікорського.
Пеня Олександр Романович - аспірант кафедри ПЗКС Київського політехнічного інституту ім. Ігоря Сікорського.

Sulema Yevgenia - D. Eng., assistant professor, Igor Sikorsky Kyiv Polytechnic Institute.
Penia Oleksandr - graduate student, Igor Sikorsky Kyiv Polytechnic Institute.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)