DOI: 10.20535/2523-4455.2019.24.6.186996

Електронні системи та сигнали

УДК 621.311.1

Машинне навчання для прогнозування споживання та генерації електроенергії

Заруба[f] Д. С., ORCID 0000-0003-3918-6300
Швець[f] М. Ю., ORCID 0000-0002-6996-6650
Хохлов[s] Ю. В., к.т.н. доц., ORCID 0000-0002-2034-6979
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського» kpi.ua
Київ, Україна

***Анотація*—Стаття присвячена підготовці і аналізу даних для покращення прогнозування кількості використаної та згенерованої електроенергії методами машинного навчання, а також оцінка важливості та впливу на прогнозування періоду доби, місяця, року, температури, вологості повітря, атмосферного тиску та інших ознак. Набір даних, що використовувався в даній статті, містить відомості про використання та генерацію електроенергії, а також погодні показники за 11 місяців з періодом фіксації даних 1 хвилина. Оброблення даних ґрунтувалось на статистичних методах обробки інформації, визначенні кількості пропущених даних, лінійних залежностях між ознаками, сумісності типів даних. Для оцінки точності прогнозування було використано коефіцієнт детермінації.**

Ключові слова — машинне навчання; коефіцієнт кореляції Пірсона; коефіцієнт детермінації; модель «Випадковий ліс».

I. Вступ

На сьогодні швидкими темпами відбувається розвиток та розповсюдження технології MicroGrid, основна задача якої забезпечення енергоефективності з використанням альтернативних джерел електроенергії як основних елементів мережі електроживлення. Тому необхідно забезпечити взаємовигідні умови для споживання і виробітку відновлювальної електроенергії.

Прогнозування грає ключову роль при формуванні балансу електроенергії в енергосистемі, впливаючи на вибір режимних параметрів і розрахункових електричних навантажень. Баланс генерації і споживання електроенергії - це основа технологічної стійкості енергосистеми, його порушення позначається на якості електроенергії (відбувається деградація частоти і напруги в мережі), що знижує ефективність роботи обладнання. Крім того, правильний прогноз дозволяє забезпечити оптимальний розподіл навантаження між об'єктами енергосистеми. Короткострокове прогнозування навантаження (КПН) в основному націлене на прогнозування навантаження системи з випередженням часу від однієї години до семи днів, що необхідно для адекватного планування і роботи енергосистем. КПН традиційно є важливим компонентом систем управління енергоспоживанням (СУЕ) [1]–[3], оскільки воно надає вхідні дані для аналізу потоку навантаження і аналізу непередбачених обставин. Прогнозування навантаження також стало важливим компонентом енергетичних брокерських систем [4]. Це дає можливість управляти вартістю покупки електроенергії шляхом регулювання завантаження устаткування, переводячи, наприклад, основні обсяги генерації електроенергії в години і зони оптового ринку енергії з найменшою ціною. У цьому новому контексті висока точність і швидкість прогнозування потрібні не тільки для надійної роботи системи, але і для адекватної роботи на ринку, так як недооцінки, так і переоцінки можуть привести до збільшення експлуатаційних витрат і втрати доходів [5].

Метою статті є порівняльний аналіз та вибір методів машинного навчання для розв'язання задачі прогнозування обсягів генерації та споживання електричної енергії у MicroGrid на базі аналізу великої кількості різнотипних параметрів, а також дослідження можливості підвищення точності прогнозування за рахунок попередньої обробки даних.

II. Аналіз та підготовка даних

В якості даних було обрано датасет розумного будинку, обладнаного сонячними батареями для генерації власної електроенергії, яка частково покриває потреби будинку. В датасеті присутні наступні ознаки: час («time»), використана електроенергія («use [kW]»), згенерована електроенергія («gen [kW]»), температура («temperature»), вологість («humidity»), видимість («visibility»), тиск («pressure»), швидкість вітру («windSpeed»), хмарний

DOI: 10.20535/2523-4455.2019.24.6.186996

покрив («cloudCover»), напрям вітру («windBearing»), температура, яка відчувається людиною («apparentTemperature»), інтенсивність опадів («precipIntensity»), точка роси («dewPoint»), імовірність опадів («precipProbability»). Дані фіксувались протягом 11 місяців, часовий інтервал між записами – 1 хвилина.

На адекватність вимірів фізичних величин можуть чинити негативний вплив внаслідок збоїв в роботі датчиків, відсутність з'єднання або живлення, а також інші зовнішні збурення. Тому перед аналізом даних і подальшим навчанням моделей машинного навчання необхідно провести обробку отриманих даних.

В першу чергу було досліджено кількість пропущених та нульових значень ознак в отриманих даних, а також їх долю у відсотках від загальної кількості даних. Результати для ознак, що мають пропущені та нульові значення, наведено у табл. 1. Значення останньої колонки табл.1 округлюються до десятих.

Як видно з таблиці 1, в даних присутні показники з кількістю нульових і пропущених даних більше 50% - інтенсивність опадів («precipIntensity») та імовірність опадів («precipProbability»). Але проаналізувавши ці два параметри, робимо висновок, що велика кількість нульових значень в них спричинена реальною відсутністю опадів, тому видаляти їх немає необхідності.

Наступний крок – позбутися викидів, тобто значень, які знаходяться на «аномальній» відстані від інших значень у випадковій вибірці. Вони можуть бути пов'язані з помилками вимірювань, помилками в одиницях виміру або бути коректними, але екстремальними значеннями.

Для визначення викидів було прораховано нижній квартиль Q_1 та верхній квартиль Q_3 наявних даних для графіку розподілу значень використаної електроенергії. Для розрахунку квартилів треба поділити варіаційний ряд медіаною на дві рівні частини, а потім в кожній з них знайти свою медіану. Також було прораховано міжквартильний розмах IQ, що визначається як різниця між верхнім і нижнім квартилем.

Далі відбувається процедура очищення – з початкового набору видаляються всі дані, для яких справедливий вираз:

$$Q_1 - 3 \cdot IQ > W_{вик} > Q_3 + 3 \cdot IQ . \qquad (1)$$

Результати очищення наведено на рис. 1.

Для ефективного тренування моделі в алгоритмі машинного навчання необхідно обрати ознаки, що є найбільш суттєвими та підходять для навчання. Багато з наявних ознак розглянутого датасету є надлишковими, тому що для деяких з них наявний високий ступінь кореляції. Наприклад, залежність «temperature» від «apparentTemperature» (рис. 2) має коефіцієнт кореляції 0.993, що буде негативно впливати на модель при її навчанні у випадку, якщо одночасно при тренуванні будуть розглядатися обидві ознаки.

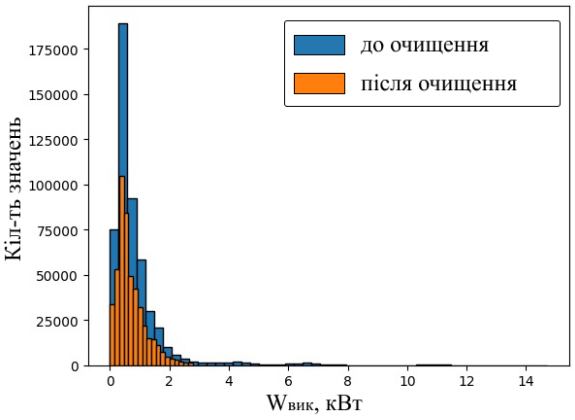


Рис. 1. Розподіл значень використаної електроенергії до та після очищення

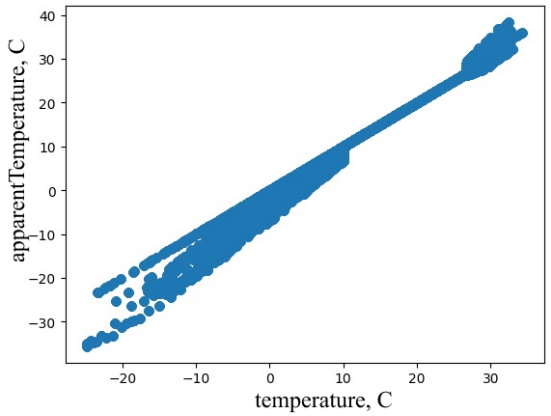


Рис. 2. Графік залежності між реальною температурою і значенням, яке відчувається людиною

ТАБЛИЦЯ 1. ОЗНАКИ З НАЙБІЛЬШОЮ КІЛЬКІСТЮ НУЛЬОВИХ І ПРОПУЩЕНИХ ЗНАЧЕНЬ

Назва	Нульові значення	Пропущені значення	Кіл-ть нульових і пропущених значень	% нульових і пропущених значень
precipProbability	416607	1	416608	82,7
precipIntensity	416607	1	416608	82,7
cloudCover	68236	59	68295	13,6
windBearing	1787	1	1788	0,4
windSpeed	230	1	231	0
gen [kW]	64	1	65	0
use [kW]	1	1	2	0

DOI: 10.20535/2523-4455.2019.24.6.186996

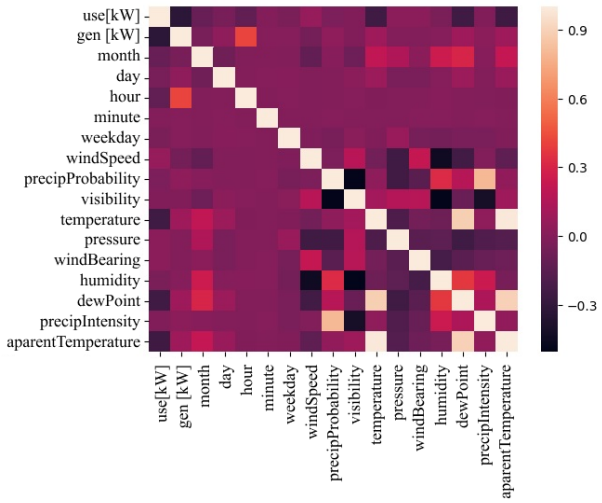


Рис. 3. Графічне зображення матриці кореляції Пірсона

Для того, щоб кількісно оцінити ступінь і позитивність лінійних зв'язків між парами ознак, а також оцінити ступінь впливу ознак на значення спожитої та згенерованої електроенергії, використано коефіцієнт кореляції Пірсона [6,7].

На рис. 3 у графічному вигляді наведено результати попарного обчислення кореляції між ознаками розглянутого датасету, причому більш світлий колір відповідає більшій кореляції (білі комірки по головній діагоналі). Для ознак, що попарно характеризуються високим ступенем кореляції, з метою підвищення ступеню адекватності, узагальнення аналізу та інтерпретації результатів необхідно і достатньо залишити лише одну з них [8]. Виключенням є випадок, коли ознаки корелюють з цільовою ознакою, а не між собою. Тому за результатами кореляційного аналізу було видалено три параметри: «apparentTemperature», «dewPoint» і «precipProbability».

III. Прогнозування використаної та спожитої електроенергії методами машинного навчання

Для навчання і тестування початкову базу даних було розділено наступним чином: 70% значень розглядалися як навчальна вибірка, 30% - як тестова. Це дозволяє оцінити залежність прогнозованих значень генерації та споживання як від попередніх значень тих самих параметрів, так і від інших незалежних параметрів, присутніх у аналізованому датасеті.

Для порівняння було обрано три моделі машинного навчання з бібліотеки scikit-learn мови програмування Python: «Лінійна», «Випадковий ліс», «k найближчих сусідів» [11]. В якості метрики для оцінки точності було використано коефіцієнт детермінації R^2 - показник, що використовується в статистичних моделях як міра залежності варіації залежної змінної від варіації незалежних змінних [9]. На рис.4 наведено діаграми числових значень коефіцієнта R^2 для ознак генерації та споживання електроенергії з використанням різних методів машинного навчання. Як видно з рис.4, в даному випадку найбільш придатною виявилася модель машинного навчання «Випадковий ліс».

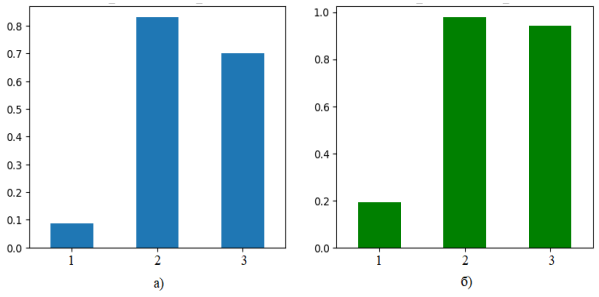


Рис. 4. Значення коефіцієнта детермінації для використаної (а) та згенерованої (б) електроенергії з використанням моделей машинного навчання: «Лінійна» (1), «Випадковий ліс» (2), «k найближчих сусідів» (3)

Таблиця 2 Вклад ознак в модель машинного навчання

Ознака	Вклад у $W_{вик}$, %	Вклад у $W_{зген}$, %
Місяць ("month")	7,47	4,8
День ("day")	4,35	4,9
Години ("hour")	16,6	55,6
Хвилини ("minute")	33,89	5,2
День тижня ("weekday")	3,33	2,6
Швидкість вітру («windSpeed»)	5,01	3,8
Інтенсивність опадів («precipIntensity")	1,03	1
Видимість («visibility»)	2,36	1,8
Температура («temperature»)	9,09	6
Тиск («pressure»)	7,06	5,7
Напрям вітру («windBearing»)	5,18	4
Вологість («humidity»)	4,64	4

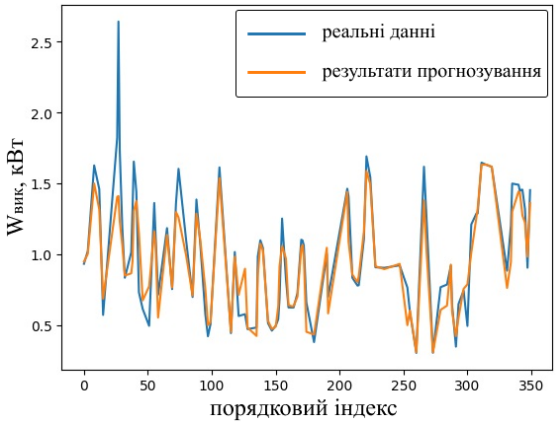
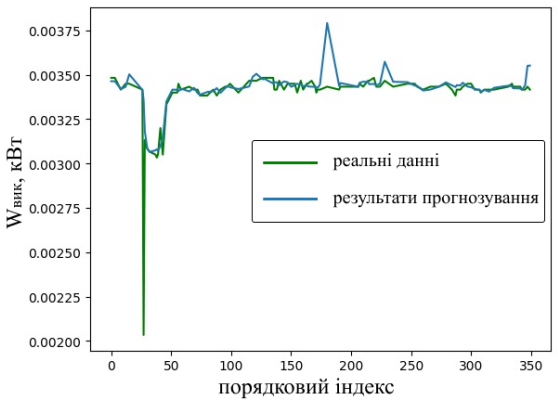


Рис. 5. Результати прогнозування використаної електроенергії



DOI: 10.20535/2523-4455.2019.24.6.186996

Рис. 6. Результати прогнозування згенерованої електроенергії

В таблиці 2 наведено вклад ознак в модель машинного навчання «Випадковий ліс», які були визначені за допомогою функцій бібліотеки scikit-learn мови програмування Python, а результати прогнозування для кількості використаної $W_{вик}$ та згенерованої $W_{зген}$ електроенергії наведено на рис. 5 та 6. По горизонтальній вісі на рис.5 та рис.6 відкладено порядковий індекс – номер кроку обчислення.

Серед протестованих моделей машинного навчання найкращий результат точності був відзначений у моделі «Випадковий ліс» (84% для використаної електроенергії, та 95% - для згенерованої).

Використання описаних методів попередньої обробки даних дозволяє підвищити точність прогнозування (на 25% для використаної електроенергії, та на 2% для згенерованої).

Точність прогнозування обсягів використаної електроенергії найбільше залежить від даних про час доби і температуру. Точність прогнозування обсягів згенерованої електроенергії найбільше залежить від часу доби, температури та тиску.

Додаткове підвищення точності може бути досягнуто з використанням більшої кількості навчальних вибірок та аналізованих ознак, додаткових методів попередньої обробки даних, а також за рахунок збільшення проміжку спостережень.

Висновки

Таким чином, в результаті порівняльного аналізу було обрано метод машинного навчання для розв'язання задачі прогнозування обсягів генерації та споживання електричної енергії у MicroGrid на базі аналізу великої кількості різнотипних параметрів. Використання попередньої обробки даних дозволяє підвищити точність прогнозування на величину від 2% до 25% для розглянутого датасету розумного будинку.

Перелік посилань

[1] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Trans. Power Syst.*, vol. 5, no. 4, pp. 1535–1547, 1990, **DOI:** 10.1109/59.99410.

[2] E. A.-J. Al-Shareef, A.J., E.A. Muhammad, "One Hour Ahead Load Forecasting Using Artificial Neural Network for the Western Area of Saudi Arabia," *Int. J. Electr. Syst. Sci. Eng.*, vol. 37, p. 7, 2008, **URL:** https://www.researchgate.net/publication/238738216_One_Hour_Ahead_Load_Forecasting_Using_Artificial_Neural_Network_for_the_Western_Area_of_Saudi_Arabia.

[3] E. A. Feinberg and D. Genethliou, "APPLIED MATHEMATICS FOR POWER SYSTEMS, Chapter 12; LOAD FORECASTING," *Short-Term Load Forecast. Proc. IEEE*, vol. 75, pp. 269–285, 1987, **URL:** http://www.ams.sunysb.edu/~feinberg/public/lf.pdf

[4] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1484–1491, 1989, **DOI:** 10.1109/59.41700.

[5] G. M. Rao, I. Narasimhaswamy, and B. S. Kumar, "Deregulated power system load forecasting using artificial intelligence," *2010 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2010*, pp. 136–140, 2010, **DOI:** 10.1109/ICCIC.2010.5705745.

[6] S. Dowdy, S., Wearden, "Statistics for research, by Shirley Dowdy and Stanley Wearden. New York: Wiley, 1983, 537 pp.," *J. Policy Anal. Manag.*, vol. 3, no. 4, pp. 637–637, 2007, **DOI:** 10.1002/pam.4050030448.

[7] D. P. Francis, A. J. S. Coats, and D. G. Gibson, "How high can a correlation coefficient be? Effects of limited reproducibility of common cardiological measures," *Int. J. Cardiol.*, vol. 69, no. 2, pp. 185–189, 1999, **DOI:** 10.1016/S0167-5273(99)00028-5.

[8] N. R. Draper and H. Smith, "Applied Regression Analysis: Third Edition," *Wiley Ser. Probab. Stat.*, vol. 47, no. 3, p. 706, 1998, **DOI:** 10.1198/tech.2005.s303.

[9] Bakhrushin V. Ye., "Programmnaya Realizatsiya Metodov Analiza Nelineynykh Statisticheskikh Svyazey V Sisteme [Software Implementation of Non-Linear Statistical Relations Analysis Methods in a System]," *Software Systems And Computing Methods*, vol. 2, no. 2, pp. 228–238, 2014, **DOI:** 10.7256/2305-6061.2014.2.11477.

[10] S. A. Glantz and B. K. Slinker, "Primer of Applied Regression and Analysis of Variance," *McGraw-Hill*, p. 777, 1990, **URL:** https://books.google.com/books?id=UcR6QgAACAAJ&pgis=1.

[11] A library of scikit-learn Python programming languages. **URL:** https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

Надійшла до редакції 22 вересня 2019 р.

УДК 621.311.1

Машинное обучение для прогнозирования потребления и генерации электроэнергии

Заруба[f] Д. С., ORCID 0000-0003-3918-6300
Швец[f] М. Ю., ORCID 0000-0002-6996-6650
Хохлов[s] Ю. В., к.т.н. доц., ORCID 0000-0002-2034-6979

DOI: 10.20535/2523-4455.2019.24.6.186996

Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского» kpi.ua
Киев, Украина

***Аннотация*—Статья посвящена подготовке и анализу данных для улучшения предсказаний количества использованной и генерируемой электроэнергии методами машинного обучения, а также определению степени важности и влияния на прогнозирование таких параметров, как время суток, месяц, год, температура, влажность воздуха, атмосферного давление и других факторов. Набор данных, используемый в данной статье, содержит сведения о потреблении и генерации электроэнергии, а также погодные показатели за 11 месяцев с периодом фиксации данных 1 минута. Обработка данных основывалась на статистических методах обработки информации, определении количества пропущенных данных, линейных зависимостях между признаками, совместимости типов данных. Для оценки точности предсказаний был использован коэффициент детерминации.**

Ключевые слова - машинное обучение; коэффициент корреляции Пирсона; коэффициент детерминации; модель «Случайный лес».

UDC 621.311.1

Machine Learning for a Power Consumption and Generation Prediction

D. S. Zaruba[f], ORCID 0000-0003-3918-6300
M. Yu. Shvets[f], ORCID 0000-0002-6996-6650
Yu. V. Khokhlov[s], PhD Assoc.Prof., ORCID 0000-0002-2034-6979
National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" kpi.ua
Kyiv, Ukraine

***Abstract*—The paper is devoted to the preparation and analysis of data sets in order to improve the prediction of the amount of consumed and generated electrical energy volumes using machine learning methods. The importance level and influence on predicting the time of day, month, year, temperature, humidity, atmospheric pressure, and other factors were determined. The dataset used in this article contains the data of smart house equipped by photovoltaic cells for the own generation of electrical energy that covers the part of house's demand. There are following values in dataset: «time», consumed electrical energy («use [kW]»), generated electrical energy («gen [kW]»), «temperature», «humidity», «visibility», «pressure», «windSpeed», «cloudCover», «windBearing», the temperature as it felt by human «apparentTemperature», precipitation intensity «precipIntensity», «dewPoint», precipitation probability «precipProbability». The data was collected during 11 months with a data fixing period of 1 minute.**

Before the data analysis and further learning it's necessary to execute preliminary processing. At first stage, it was investigated how large is the part of missed and zero values in dataset. The second stage includes elimination of outliers that are situated at anomaly distance from other values in random sample. These outliers could be caused by measurement errors, wrong measuring units use. Also, it could be correct but extremum values. The purification procedure includes defining the lower and the upper quartiles of existing data for the distribution of used energy.

For effective learning of the model it is necessary to choose the values that are most important and suitable for training. Pearson's correlation coefficient was used to estimate numerically the level and positivity of linear connections between the pairs of values as well as to estimate their influence to the used and generated energy. Among the values with the high level of correlation only one was chosen that helped increasing adequacy, generalization and results interpretation. As a result of correlation analysis three parameters were selected for the training - «apparentTemperature», «dewPoint» and «precipProbability». Use of proposed preprocessing methods allows increasing the predictions exactness by 25% for the used energy and by 2% for the generated energy.

The initial dataset was divided as follows: 70% of values were considered as the training samples and 30% - as testing ones. To compare the training methods three models of machine learning from the library Scikit-learn in programming language Python were considered: «Linear», «Random forest», «k nearest neighbors». The determination coefficient R^2 was used as a metrics to estimate the exactness. The diagrams of numerical values of R^2 coefficient for the parameters of generation and consumption of electrical energy and for three considered models of machine learning were built. Among the tested model the best result was demonstrated for the "Random forest" model (84% for the used energy and by 95% for the generated energy).

Additional exactness increasing could be reached by use of more amount of testing samples and parameters during the analysis and more time intervals of observation as well as additional methods of data preprocessing.

Keywords — machine learning; Pearson correlation coefficient; determination coefficient; random forest model