УДК 519.23+004.8
DOI: 10.25206/2310-4597-2023-1-42-45

ПРЕДВАРИТЕЛЬНАЯ ПОДГОТОВКА ДАННЫХ ДЛЯ ПОСЛЕДУЮЩЕГО КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ СОЛНЕЧНОЙ ЭЛЕКТРОЭНЕРГИИ

PRELIMINARY PREPARATION OF DATA FOR SUBSEQUENT SHORT-TERM FORECASTING OF SOLAR ELECTRICITY

Д. И. Васина, А. Ю. Горшенин
Омский государственный технический университет, г. Омск, Россия

D. I. Vasina, A. Y. Gorshenin
Omsk state technical university, Omsk, Russia

Аннотация. В статье обсуждается важность точного прогнозирования выработки солнечной энергии при управлении солнечными электростанциями и интеграции солнечной энергии в электрическую сеть. В статье представлен алгоритм программы, предназначенной для оптимизации форматов данных с целью сокращения использования памяти в больших наборах данных, особенно для данных, собранных с двух солнечных электростанций. Программа выполняет разведочный анализ данных (EDA) для оптимизации формата данных и сравнения эффективности двух установок. Результаты показывают, что оптимизированные форматы данных могут сократить использование памяти и повысить производительность моделей машинного обучения.

***Ключевые слова*:** солнечная энергия, прогнозирование, машинное обучение, оптимизация данных, анализ данных, эффективность памяти, солнечные электростанции, электросети, разведочный анализ данных, производство солнечной энергии

Abstract. The article discusses the importance of accurate forecasting of solar energy generation in the management of solar power plants and the integration of solar energy into the electric grid. The article presents an algorithm of a program designed to optimize data formats in order to reduce memory usage in large datasets, especially for data collected from two solar power plants. The program performs exploratory data analysis (EDA) to optimize the data format and compare the efficiency of two installations. The results show that optimized data formats can reduce memory usage and improve the performance of machine learning models.

***Keyword*:** solar energy, forecasting, machine learning, data optimization, data analysis, memory efficiency, solar power plants, power grids, exploratory data analysis, solar energy production

I. Введение

Прогнозирование выработки солнечной энергии является важной областью исследований в области возобновляемых источников энергии [1]. С растущим внедрением солнечной энергии в качестве жизнеспособного источника энергии точное прогнозирование выработки солнечной энергии стало решающим для эффективного управления солнечными электростанциями и интеграции солнечной энергии в электрическую сеть [2].

Прогнозирование солнечной энергии предполагает использование математических моделей и алгоритмов для оценки количества солнечного излучения, которое будет получено солнечными панелями за данный период времени. Обычно это делается с использованием данных прогнозов погоды, спутниковых снимков и исторических данных о производстве солнечной энергии [3].

Точное прогнозирование производства солнечной энергии имеет ряд преимуществ. Во-первых, это позволяет операторам солнечных электростанций лучше управлять производством энергии, оптимизируя свою деятельность и снижая затраты [4]. Кроме того, точное прогнозирование солнечной энергии может помочь операторам электросетей управлять колебаниями выработки солнечной энергии, обеспечивая лучшую интеграцию солнечной энергии в электрическую сеть [5].

Существует несколько проблем, связанных с прогнозированием солнечной энергии, включая изменчивость солнечной радиации из-за облачного покрова и других погодных условий, а также потребность в точных данных [6]. Однако достижения в области машинного обучения и методах анализа данных показывают многообещающие результаты в повышении точности прогнозов солнечной энергии [7, 8].

II. Теория и результаты экспериментов

Основной проблемой в прогнозировании выработки солнечной энергии определяется – неточные данные, или вовсе их отсутствие в свободном доступе для проведения исследований и дальнейшего изучения. Для последующего прогнозирования выработки необходимо проанализировать входные данные, а также провести их оптимизацию.

Оптимизация данных – важный процесс в анализе данных и машинном обучении, поскольку он может помочь уменьшить размер данных и сделать их более управляемыми, что особенно важно при работе с большими наборами данных. Оптимизация данных может повысить эффективность обработки данных, снизить требования к хранилищу и повысить производительность моделей машинного обучения. Кроме того, оптимизированные данные легче визуализировать и анализировать, и они могут быть более надежными, поскольку уменьшают вероятность ошибок и несоответствий в данных.

Поскольку размер наборов данных растет экспоненциально, объем памяти становится решающей проблемой, которую необходимо учитывать при выполнении анализа данных. Неэффективность памяти может не только замедлить время обработки, но и сделать невозможным выполнение анализа на определенных машинах или системах. В статье будет описан алгоритм программы, которая оптимизирует форматы данных для уменьшения объема памяти в больших наборах данных.

Программа специально разработана для анализа данных о выработке солнечной энергии двумя установками.

Эти данные были собраны на двух солнечных электростанциях в Индии в течение 34 дней. Он содержит две пары файлов – каждая пара содержит один набор данных о выработке электроэнергии и один набор данных о показаниях датчиков. Наборы данных о выработке электроэнергии собираются на уровне инвертора.

Первым шагом является загрузка данных из CSV-файлов и выполнение разведочный анализа данных (EDA) для этих данных. EDA выполняется с использованием библиотеки pandas на Python. Программа загружает данные для установки 1 и выполняет различные операции для оптимизации формата данных.

Первое наблюдение, сделанное в EDA, заключается в том, что данные очень чистые, без нулевых значений, отрицательных или бесконечных. Второе наблюдение заключается в том, что названия столбцов написаны в верхнем регистре и будут изменены на строчные. Третье замечание заключается в том, что столбец ДАТА_ВРЕМЯ находится в текстовом формате и будет преобразован в временную метку. Четвертое заключается в том, что СИЛА_ПОСТОЯННОГО_ТОКА и ПЕРЕМЕННАЯ_МОЩНОСТЬ, по-видимому, имеют проблему с масштабированием, поскольку СИЛА_ПОСТОЯННОГО_ТОКА должен быть схожим на ПЕРЕМЕННАЯ_МОЩНОСТЬ, но вместо этого кажется в 10 раз больше. Наконец, столбец ПАНЕЛЬ_ID содержит единственное значение во всем наборе данных, и этот столбец будет удален, а значение сохранено во внешней переменной, чтобы уменьшить объем памяти фрейма данных.

Результирующая гистограмма на рис. 1 обеспечивает визуальное представление распределения ненулевых значений в каждом столбце отфильтрованного фрейма данных, что может помочь выявить закономерности и тенденции в данных.

Следующий шаг – оптимизация форматов. Функция optimize_formats(df) предназначена для оптимизации форматов набора данных. Функция принимает фрейм данных pandas в качестве входных данных и возвращает новый фрейм данных с оптимизированными форматами и словарем, содержащим дополнительные данные. Функция проверяет, написаны ли названия столбцов в верхнем регистре, и если да, то приступает к процессу оптимизации. Затем функция создает словарь для хранения данных, которые должны быть удалены из фрейма данных. Функция изменяет имена столбцов на строчные, кодирует "ИДЕНТИФИКАТОР" в целые числа, сохраняет исходный "ИДЕНТИФИКАТОР" в отдельной переменной, удаляет столбец "ПАНЕЛЬ_ID" и сохраняет его значение во внешней переменной. Функция также изменяет столбец 'ДАТА_ВРЕМЯ' со строки на pd. Отметка времени. Наконец, функция выводит начальный и конечный размер фрейма данных и процентное сокращение объема памяти.

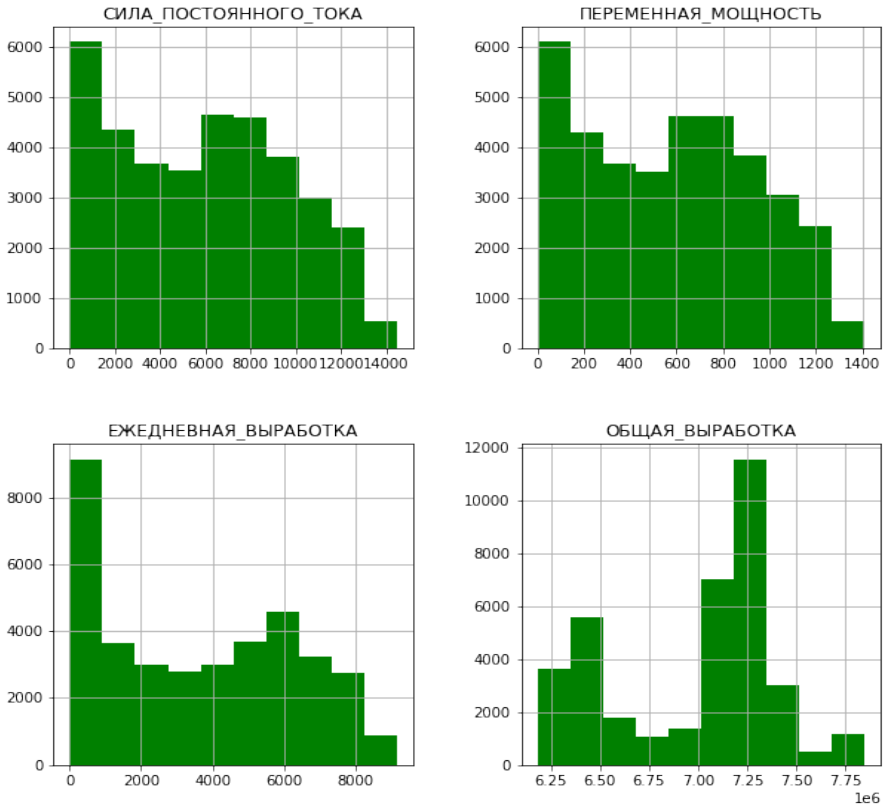

Рис. 1. Распределение ненулевых значений

Мы отфильтровываем строки, в которых значения ЕЖЕДНЕВНАЯ_ВЫРАБОТКА и ПЕРЕМЕННАЯ_МОЩНОСТЬ равны нулю, поскольку они представляют периоды, когда солнечная установка не вырабатывает никакой энергии, и, таким образом, они не предоставляют никакой полезной информации для анализа. Затем мы строим гистограммы для всех оставшихся столбцов в sample_df, чтобы визуализировать распределение значений в каждом столбце. Это позволяет нам получить представление о данных и выявить любые закономерности или аномалии, которые могут присутствовать. Устанавливая цвет гистограмм на зеленый, мы просто меняем внешний вид графиков.

После оптимизации форматов данных программа выполняет дополнительный анализ для сравнения эффективности двух установок. Программа визуализирует шкалы мощности переменного и постоянного тока без нулей и сравнивает соотношение мощности постоянного и переменного тока (КПД) между двумя установками на рис. 2 и 3. Программа использует блок-схемы для сравнения масштабов мощности переменного и постоянного тока для обеих установок. На диаграмме показано, что существует значительная разница в масштабах мощности переменного и постоянного тока для обеих установок. Затем программа вычисляет эффективность обеих установок, используя максимальные значения мощности переменного и постоянного тока. Эффективность установки 1 сравнивается с эффективностью установки 2, и рассчитывается соотношение.

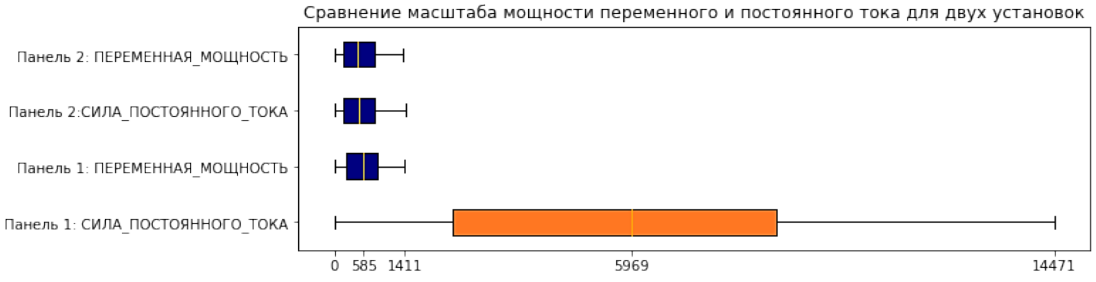

Рис. 2. Шкала мощности для двух установок

Рис. 3. Скорректированная шкала мощности для двух установок

III. Заключение

Оптимизация форматов данных имеет решающее значение для сокращения объема памяти в больших наборах данных. Эта программа предоставляет пример того, как оптимизировать форматы данных набора данных о производстве солнечной энергии, чтобы уменьшить объем памяти. Программа выполняет предварительный анализ данных, оптимизирует форматы данных и выполняет дополнительный анализ для сравнения эффективности двух установок. Программа может быть использована в качестве шаблона для оптимизации форматов данных других больших наборов данных.

Список литературы

1. Gorshenin A., Vasina D. Study of Methods for Forecasting Wind Power Generation Based on the Processing of Meteorological Data // 2022 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, Russian Federation. 2022. P. 1–5. DOI: 10.1109/Dynamics56256.2022.10014925.

2. Тюньков Д. А., Грицай А. С., Сапилова А. А. [и др.]. Нейросетевая модель для краткосрочного прогнозирования выработки электрической энергии солнечными электростанциями // Научный вестник Новосибирского государственного технического университета. 2020. № 4(80). С. 145–158. DOI 10.17212/1814-1196-2020-4-145-158.

3. Тюньков Д. А., Сапилова А. А., Грицай А. С. [и др.]. Методы краткосрочного прогнозирования выработки электрической энергии солнечными электростанциями и их классификация // Электротехнические системы и комплексы. 2020. № 3(48). С. 4–10. DOI 10.18503/2311-8318-2020-3(48)-4-10.

4. Pérez-Romero J., Lujano-Rojas J. M., Bremen L. Solar power forecasting: A review // Renewable and Sustainable Energy Reviews. 2017. Vol. 75. Pp. 10–28.

5. Zhang J., Li X., Yan W., Yang Y. A novel hybrid model for solar power forecasting using deep learning and auto-regressive integrated moving average // Renewable Energy. 2020. Vol. 146. Pp. 677–692.

6. Azaza O. S., Khiari M. Prediction of photovoltaic system performance using machine learning algorithms: A review // Energy Conversion and Management. 2021. Vol. 227. P. 113737.

7. Hong T., He W., Zeng Q. Solar power prediction based on machine learning and its application in energy internet // Energy. 2018. Vol. 153. Pp. 990–999.

8. Shuai Y., Shu L., Xu C., Li J. Deep Learning-Based Solar Power Forecasting Models: A Survey // IEEE Access. 2020. Vol. 8. P. 13955–13967.