ЭЛЕКТРОТЕХНИЧЕСКИЕ КОМПЛЕКСЫ И СИСТЕМЫ / ELECTRICAL COMPLEXES AND SYSTEMS

DOI: https://doi.org/10.23670/IRJ.2024.141.37

АНАЛИЗ МЕТОДОВ КЛАСТЕРИЗАЦИИ ВРЕМЕННЫХ РЯДОВ ЭЛЕКТРОПОТРЕБЛЕНИЯ

Научная статья

Каретников М.С.[1, *]
[1] ORCID : 0009-0007-6239-0656;
[1] Самарский государственный технический университет, Самара, Российская Федерация

* Корреспондирующий автор (maksim.karietnikov[at]mail.ru)

Аннотация

В статье рассматривается применение методов кластеризации временных рядов K-means и Soft DTW BaryCenter для анализа поведенческого потребления электроэнергии. Основной акцент делается на важности предварительной обработки и подготовки данных перед кластеризацией. В статье также рассматриваются различные подходы к подготовке данных.

Проведен сравнительный анализ методов кластеризации K-means и Soft DTW BaryCenter, указаны их преимущества и недостатки. Приводятся методы оценки кластеризации с использованием метрик, которые отражают компактность и разделенность кластеров. Сделано заключение, что выбор метода кластеризации зависит от задачи исследования, а также от качества и объема исходных данных.

Ключевые слова: анализ данных, электропотребление, кластеризация, временные ряды.

AN ANALYSIS OF METHODS FOR CLUSTERING TIME SERIES OF ELECTRICITY CONSUMPTION

Research article

Karetnikov M.S.[1, *]
[1] ORCID : 0009-0007-6239-0656;
[1] Samara State Technical University, Samara, Russian Federation

* Corresponding author (maksim.karietnikov[at]mail.ru)

Abstract

The article examines the application of K-means and Soft DTW BaryCenter time series clustering methods to the analysis of behavioural electricity consumption. The main focus is on the importance of data pre-processing and preparation before clustering. The work also discusses various approaches to data conditioning.

A comparative analysis of K-means and Soft DTW BaryCenter clustering methods is carried out, their pros and cons are pointed out. Methods of clustering evaluation using metrics that reflect the compactness and separateness of clusters are presented. It is concluded that the choice of clustering method depends on the research task, as well as on the quality and volume of the original data.

Keywords: data analysis, electricity consumption, clustering, time series.

Введение

Проблема кластеризации данных временных рядов в контексте потребления электроэнергии является весьма актуальной. С развитием интеллектуальных сетей возросли требования к краткосрочному прогнозированию спроса на электроэнергию. Точное прогнозирование потребления электроэнергии позволяет поставщикам энергии планировать свои производственные мощности и оптимизировать закупки электроэнергии на оптовом рынке электроэнергии и энергоснабжения.

Более глубокое понимание того, как потребители используют электроэнергию, может быть получено с помощью кластеризации временных рядов потребления электроэнергии. Это может помочь поставщикам электроэнергии лучше понимать потребности своих клиентов и повысить точность их прогнозирования. Неточное прогнозирование потребления электроэнергии может привести к недостаточному или чрезмерному количеству закупаемой электроэнергии и финансовым рискам для поставщика.

Метод K-means является популярным методом кластеризации из-за его простоты и эффективности [1], [2]. Однако он чувствителен к выбросам в данных [3]. С другой стороны, метод Soft DTW BaryCenter обеспечивает более плавный эффект и может уменьшить влияние выбросов. Однако это может быть сложным с точки зрения вычислений, особенно при работе с большими наборами данных временных рядов [4].

В литературе также обсуждается важность предварительной обработки данных и их подготовки перед кластеризацией. Это включает в себя удаление выбросов, стандартизацию и нормализацию данных, преобразование данных в числовые значения, выбор наиболее информативных функций и удаление ненужных данных [5], [6], [7], [8], [9].

Наконец, в литературе обсуждается оценка качества кластеризации. Упоминается несколько методов, включая метод локтя, метод силуэта, индекс Данна и индекс Дэвиса-Булдина [10], [11]. Авторы отмечают, что одной метрики часто недостаточно для адекватной оценки кластеризации, и также может потребоваться визуальный контроль [12].

Цель исследования – проанализировать методы подготовки данных в контексте потребления электроэнергии для проведения качественной кластеризации и определить, какой метод кластеризации, K-means или Soft DTW-BaryCenter, является более подходящим в зависимости от качества и объема изучаемых данных.

Методы подготовки данных для кластеризации временных рядов электропотребления

Перед кластеризацией данных необходимо выполнить подготовку данных. В общем виде процесс подготовки данных состоит из нескольких шагов:

2.1. Удаление выбросов в наборе данных

Выбросом называют такой объект некоторого класса, значения признаков которого значительно отличаются от значений признаков другого класса [5]. Выбросы во временных рядах электропотребления могут быть вызваны различными факторами, такими как временные сбои в работе счетчика, нестабильность в электросети, повреждение счетчика или быть результатом ошибке в сборе данных. Выбросами в контексте электропотребления могут быть недели, в которых есть дни с необычно высоким или низким объемом электропотребления для этого дня недели. Такие выбросы можно определить с помощью инструментов визуализации данных или с помощью математических функций Z-score и IQR.

2.2. Обработка пропущенных значений

Целесообразным является сочетание метода Zet-алгоритма для восстановления групп пропущенных значений и метода сплайн-интерполяции для восстановления одиночных пропущенных значений [6], [7], [13].

Для оценки точности восстановления пропущенных значений часто используется мера корня из среднеквадратичной ошибки (RMSE).

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}\left(t_i - t_i'\right)^2}\ ,$$

где h – количество пропущенных значений, ti и $t'i$ — фактическое и восстановленное значения временного ряда соответственно [14].

2.3. Нормализация данных

Нормализация данных – это процесс приведения данных к определенному диапазону значений, обычно от 0 до 1. Формула нормализации данных:

$$x_{\text{норм}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}\ ,$$

где $x_{норм}$ – нормализованное значение, x_{min} – минимальное значение, x_{max} – максимальное значение, x – фактическое значение.

Нормализация данных полезна, когда данные имеют разные масштабы или когда требуется привести данные к определенному диапазону для работы с определенными алгоритмами.

Учитывая высокую дисперсию в данных, нормализация данных является неотъемлемо частью кластеризации и может сыграть важную роль в интерпретации результатов [7].

Методы кластеризации временных рядов

Метод K-means является одним из наиболее популярных методов кластеризации из-за своей простоты реализации и эффективности [1], [4], [12]. Он используется для разделения набора данных на несколько кластеров, где каждый кластер представляет собой группу точек схожих между собой. Алгоритм начинается с инициализации k центроидов, затем точки данных присваиваются ближайшему центроиду, после чего центроиды пересчитываются на основе средних значений точек в каждом кластере. Этот процесс повторяется до сходимости.

Однако поскольку метод K-means относится к методам четкой кластеризации (каждая точка данных принадлежит только одному кластеру) выбросы в данных могут существенно повлиять на результаты кластеризации [3].

Метод Soft DTW BaryCenter позволяет учитывать сходство между временными рядами, используя взвешенное среднее значение. Он позволяет временным рядам принадлежать к нескольким кластерам, поэтому имеет сглаживающий эффект и может уменьшить влияние выбросов в данных. Однако в отличие от метода K-means, метод Soft DTW BaryCenter может быть вычислительно сложным, особенно при работе с большими наборами временных рядов.

В работе [1] были использованы методы кластеризации K-means и Soft DTW BaryCenter, где после исследования K-means для улучшения результата был применен метод Soft DTW BaryCenter который поспособствовал наиболее точному описанию профилей электропотребления потребителей.

K-means может использовать различные методы инициализации центроидов, такие как случайная инициализация или K-means++. Авторы статьи [2] использовали метод K-means, где центроиды были инициализированы с использованием алгоритма K-means++. В K-means++ начальные центроиды выбираются с учетом расстояний между точками данных, чтобы они были равномерно распределены по всему набору данных. Это позволяет избежать проблемы с попаданием в локальные минимумы, которая может возникнуть при случайной инициализации центроидов в K-means.

Soft DTW BaryCenter не требует явной инициализации центроидов, так как он вычисляет взвешенное среднее значение на основе имеющихся временных рядов.

Методы оценки качества кластеризации

При оценке качества кластеризации мы также стараемся определить оптимальное число кластеров. Это важно для того, чтобы понять, насколько хорошо данные были разделены на группы и какое количество кластеров наиболее подходит для данного набора данных. Нужно учитывать, что при всем разнообразии потребителей меньшее количество кластеров облегчает интерпретацию, а потому предпочтительнее, чем большое количество кластеров.

Существует несколько подходов к определению оптимального числа кластеров, включая метод локтя, метод силуэта, метод индекса Данна и метод индекса Дэвиса-Болдина [10]. Эти методы стремятся к тому, чтобы кластеры были компактными и хорошо разделенными, с минимальной внутрикластерной дисперсией и максимальным межкластерным разделением и таким образом позволяют оценить, какое количество кластеров наилучшим образом представляет структуру данных. Согласно выводам авторов [12], наиболее часто используемым подходом является

индекс Дэвиса-Болдина и метод силуэта. Также авторы работы [12] подчеркивают, что одной метрики недостаточно для адекватной оценки кластеризации. Иногда значения метрик противоречат друг другу, и тогда приходится полагаться на оценку с помощью визуального осмотра, но и она может быть предвзятой из-за интерпретации визуального представления [11], [12].

Заключение

Новизна исследования заключается в сравнительном анализе методов кластеризации временных рядов K-means и Soft DTW BaryCenter в контексте данных о потреблении электроэнергии. Результаты исследования подчеркивают важность предварительной обработки и подготовки исходных данных при проведении кластеризации. В статье сравниваются методы кластеризации K-means и Soft DTW BaryCenter, рассматриваются преимущества и недостатки каждого из них. В случае частых и небольших выбросов в данных применение метода Soft DTW BaryCenter более целесообразно, чем применение метода K-means из-за более высокой чувствительности к выбросам метода K-means. Также установлено, что Soft DTW BaryCenter является более вычислительно сложным при работе с большими наборами временных рядов. Установлено, что метод кластеризации по временным рядам показывает более высокую эффективность, чем метод, основанный на общем объеме электроэнергии. Таким образом следует заключить, что выбор одного из этих методов для кластеризации временных рядов определяется качеством и объемом исходных данных. Выявлено, что наиболее часто используемыми метриками оценки кластеризации являются индекс Дэвиса-Болдина и метод силуэта по причине простоты их интерпретации. Результаты этого исследования согласуются с результатами других исследований в этой области. Например, работа [4] также подчеркивает эффективность кластеризации временных рядов для понимания спроса на электроэнергию. Аналогичным образом, исследования [2] и [10] также подчеркивают важность предварительной обработки и подготовки данных при кластеризации.

Однако в этом исследовании представлено более подробное сравнение методов K-means и Soft DTW BaryCenter с обсуждением их преимуществ и недостатков в контексте данных о потреблении электроэнергии.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Demirer H. A Clustering and Benchmarking Based Monthly Electricity Consumption Analysis for Creating Energy Efficiency Insights to the Utility End-Users / H. Demirer, E. Yıldıztepe, R.N. Kalem et al. // 4th South East European Council of CIGRE (SEERC) Conference. — 2023. — p. 587-594.

2. Laurinec P. Adaptive Time Series Forecasting of Energy Consumption Using Optimized Cluster Analysis / P. Laurinec, M. Lóderer, P. Vrablecová et al. // IEEE. — 2016. — 16. — p. 398-405. — DOI: 10.1109/ICDMW.2016.0063.

3. Егоров А. В. Особенности методов кластеризации данных / А. В. Егоров, Н. И. Куприянова // Известия ЮФУ. Технические науки. — 2011. — 11.

4. Hyojeoung K. Time-series Clustering and Forecasting Household Electricity Demand Using Smart Meter Data / K. Hyojeoung, P. Sujin, K. Sahm // Energy Reports. — 2023. — 9. — p. 4111-4121. — DOI: 10.1016/j.egyr.2023.03.042..

5. Волченко Е.В. Метод удаления выбросов в данных на основе взвешенных обучающих выборок w-объектов / Е.В. Волченко // Восточно-Европейский журнал передовых технологий. — 2014. — 4. — с. 31-36.

6. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко — Новосибирск: Издательство Института математики, 1999. — 270 с.

7. Абраменкова И. В. Методы восстановления пропусков в массивах данных / И. В. Абраменкова, В. В. Круглов // Программные продукты и системы. — 2005. — 2.

8.

9. Пономарев Д. С. Иерархическая кластеризация на языке R для производственно-экономических показателей пенитенциарной системы / Д. С. Пономарев // Экономика. Информатика. — 2023. — 3. — с. 655-668.

10. Okereke G.E. K-means Clustering of Electricity Consumers Using Time-domain Features from Smart Meter Data / G.E. Okereke, M.C. Bali, C.N. Okwueze et al. // Journal of Electrical Systems and Information Technology. — 2023. — 10. — DOI: 10.1186/s43067-023-00068-3.

11. Gogolou A. Comparing Similarity Perception in Time Series Visualizations / A. Gogolou, T. Tsandilas, T. Palpanas et al. // IEEE. — 2019. — 1. — DOI: 10.1109/TVCG.2018.2865077.

12. Toussaint W. Clustering Residential Electricity Consumption Data to Create Archetypes that Capture Household Behaviour in South Africa / W. Toussaint, D. Moodley // South African Computer Journal. — 2020. — 32. — p. 1-34. — DOI: 10.18489/sacj.v32i2.845.

13. Загоруйко Н.Г. Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм Zet) / Н.Г. Загоруйко, В.Н. Елкина, В.С. Тимеркаев // Вычислительные системы. — Новосибирск: Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, 1975. — Вып. 61. — с. 3-27.

14. Minor B.D. Learning Activity Predictors from Sensor Data: Algorithms, Evaluation, and Applications / B.D. Minor, J.R. Doppa, D.J. Cook // IEEE Transactions on Knowledge and Data Engineering. — 2017. — 12. — p. 2744-2757. — DOI: 10.1109/TKDE.2017.2750669

Список литературы на английском языке / References in English

1. Demirer H. A Clustering and Benchmarking Based Monthly Electricity Consumption Analysis for Creating Energy Efficiency Insights to the Utility End-Users / H. Demirer, E. Yıldıztepe, R.N. Kalem et al. // 4th South East European Council of CIGRE (SEERC) Conference. — 2023. — p. 587-594.

2. Laurinec P. Adaptive Time Series Forecasting of Energy Consumption Using Optimized Cluster Analysis / P. Laurinec, M. Lóderer, P. Vrablecová et al. // IEEE. — 2016. — 16. — p. 398-405. — DOI: 10.1109/ICDMW.2016.0063.

3. Egorov A. V. Osobennosti metodov klasterizatsii dannyh [Features of Data Clustering Methods] / A. V. Egorov, N. I. Kuprijanova // Proceedings of SFedU. Engineering Sciences. — 2011. — 11. [in Russian]

4. Hyojeoung K. Time-series Clustering and Forecasting Household Electricity Demand Using Smart Meter Data / K. Hyojeoung, P. Sujin, K. Sahm // Energy Reports. — 2023. — 9. — p. 4111-4121. — DOI: 10.1016/j.egyr.2023.03.042..

5. Volchenko E.V. Metod udalenija vybrosov v dannyh na osnove vzveshennyh obuchajuschih vyborok w-ob'ektov [Method for Removing Outliers in Data Based on Weighted Training Samples of W-subjects] / E.V. Volchenko // Eastern-European Journal of Enterprise Technologies. — 2014. — 4. — p. 31-36. [in Russian]

6. Zagorujko N.G. Prikladnye metody analiza dannyh i znanij [Applied Methods of Data and Knowledge Analysis] / N.G. Zagorujko — Novosibirsk: Institute of Mathematics Publishing House, 1999. — 270 p. [in Russian]

7. Abramenkova I. V. Metody vosstanovlenija propuskov v massivah dannyh [Methods of Repairing Gaps in Data Sets] / I. V. Abramenkova, V. V. Kruglov // Software Products and Systems. — 2005. — 2. [in Russian]

8.

9. Ponomarev D. S. Ierarhicheskaja klasterizatsija na jazyke R dlja proizvodstvenno-ekonomicheskih pokazatelej penitentsiarnoj sistemy [Hierarchical Clustering in R for Production and Economic Indicators of the Penitentiary System] / D. S. Ponomarev // Economics. Information Technologies. — 2023. — 3. — p. 655-668. [in Russian]

10. Okereke G.E. K-means Clustering of Electricity Consumers Using Time-domain Features from Smart Meter Data / G.E. Okereke, M.C. Bali, C.N. Okwueze et al. // Journal of Electrical Systems and Information Technology. — 2023. — 10. — DOI: 10.1186/s43067-023-00068-3.

11. Gogolou A. Comparing Similarity Perception in Time Series Visualizations / A. Gogolou, T. Tsandilas, T. Palpanas et al. // IEEE. — 2019. — 1. — DOI: 10.1109/TVCG.2018.2865077.

12. Toussaint W. Clustering Residential Electricity Consumption Data to Create Archetypes that Capture Household Behaviour in South Africa / W. Toussaint, D. Moodley // South African Computer Journal. — 2020. — 32. — p. 1-34. — DOI: 10.18489/sacj.v32i2.845.

13. Zagorujko N.G. Algoritm zapolnenija propuskov v empiricheskih tablitsah (algoritm Zet) [Algorithm for Filling Gaps in Empirical Tables (Zet algorithm)] / N.G. Zagorujko, V.N. Elkina, V.S. Timerkaev // Vychislitel'nye sistemy [Computing Systems]. — Novosibirsk: Institution of Science S.L. Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, 1975. — Iss. 61. — p. 3-27. [in Russian]

14. Minor B.D. Learning Activity Predictors from Sensor Data: Algorithms, Evaluation, and Applications / B.D. Minor, J.R. Doppa, D.J. Cook // IEEE Transactions on Knowledge and Data Engineering. — 2017. — 12. — p. 2744-2757. — DOI: 10.1109/TKDE.2017.2750669