Semi-Automatic Data Labelling of Smart Meter Data for Electricity Theft Detection

Kwok Tai Chui[1], Lap-Kei Lee[1], Ryan Wen Liu[2], Mingbo Zhao[3] and Miltiadis D. Lytras[4,5]

[1]School of Science and Technology, Hong Kong Metropolitan University, Ho Man Tin, Kowloon, Hong Kong SAR, China
[2]Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan 430063, China
[3]School of Information Science and Technology, Donghua University, Shanghai 200051, China
[4]School of Business & Economics, Deree College—The American College of Greece, Gravias 6, 153 42 Aghia Paraskevi, Greece
[5]King Abdulaziz University, Jeddah P.O. Box 34689, Saudi Arabia

Abstract

The global penetration rate of smart meter installation is ever-growing. The smart meters provide 24/7 continuous recordings of electricity data which is comprised of useful information related to electricity usage and consumers' behaviors. Anomalies such as electricity theft can be detected using machine learning algorithms. To build a robust and accurate detection model, sufficient labelled data is important to ensure good generalization that adapts properly to unseen data. Nevertheless, manual electricity data labelling is costly and unrealistic in a large-scale population. In this paper, a semi-automatic data labelling algorithm is proposed based on deep convolutional neural network to label the electricity data, with limited amount of labelled data. Results reveal that the algorithm could serve as tradeoff between costly manual labelling and performance of the detection model.

Keywords

Advanced metering infrastructure, data labelling, data science, electricity theft, machine learning, non-intrusive load monitoring, smart meters, smart grid

1. Introduction

Reduction of carbon emission from electricity consumption has been the leading global vision to tackle with global warming which can wreak havoc on human beings. Environmental experts have emphasized that global warming leads to severe ice and permafrost melting which releases a large amount of methane, about 30 times powerful greenhouse effect compared to carbon dioxide[1][2]. It may become irreversible positive feedback for melting if the increase in sea-level temperature reaches the threshold.

In many countries, traditional electric grid has been migrated to smart grid to address the challenge [3][4]. The deployment of smart meters has played a crucial role in smart grid which supports continuous collection of electricity data in apartments and buildings[5]. Current works estimated that the number of smart meters has reached 200 million, 96 million, 70 million and 2.9 million in Europe, China, U.S., and U.K. respectively, with a market penetration of over 50%[6][7]. This has built up a solid foundation for further analysis of massive amount of electricity data. In the light of the introduction of electricity load disaggregation (ELD) algorithm (it is also called non-intrusive load monitoring, NILM), electricity data is disaggregated into electricity consumption of individual

International Conference on Smart Systems and Advanced Computing (Syscom-2021), December 25–26, 2021

EMAIL: jktchui@hkmu.edu.hk (A. 1); lklee@hkmu.edu.hk (A. 2); wenliu@whut.edu.cn (A. 3); mzhao4@dhu.edu.cn (A. 4); mlytras@acg.edu (A. 5)
ORCID: 0000-0001-7992-9901 (A. 1); 0000-0001-9619-6041 (A. 2); 0000-0002-1591-5583 (A. 3); 0000-0003-0381-4360 (A. 4); 0000-0002-7281-5458 (A. 5)

electric appliance, that brings valuable insight and impact to the general public, electric company, and government [8][9]. Electricity data also provides insight to detect electricity theft, which is accounted for the lost of 25%, 16%, and 6% of electricity supply in India, Brazil, and China, respectively [10]. Besides, the consequences of electricity theft include the lack of safety of utilities and users, revenue loss, and increase in electricity demand [11].

Data labeling is preferable approach to provide the ground truth of electricity data. In traditional data collection process, data is costly to be labeled for supervised machine learning algorithm[12][13]. Possible labeling methods include manual labeling [14] and adding a control system to record the on and off statues of appliance [15]. Nevertheless, these are costly and limited by large scale deployment of smart meters. Therefore, a computational approach via automatic data labeling algorithm is desired.

2. Literature review of automatic data labelling algorithms

The discussion of automatic data labeling techniques is algorithmic-based approach instead of costly manual-based or control-based approaches [16][17]. In general, existing works adopted the concept of semi-supervised learning. To summarize, the following works [18][19][20][21] shared the concept of single label assignment. Started with small amount of labeled electricity data and large amount of unlabeled electricity data, 1-Nearest Neighbor was adopted for the assignment of unlabeled data to the class of the nearest neighbor [18]. A two stages graph-based approach was proposed in [19]. Both labeled and unlabeled data are firstly transformed into graph representation with the adjacency graph setting of G=(V,E) where V and E are nodes and edges respectively. The closer the connection on the graph, the higher the chance of sharing the same label. In [20], the presented research work considered the introduction of unlabeled data to the consistency learning loss function and formed composite loss with the labeled data. Another approach combined the decision tree and nearest-neighbor techniques for data labeling that the former was acted as eager learner and the latter was served as lazy learner [21].

Existing works [18][19][20][21] typically adopted single label assignment. Practically, there are numerous and ever-growing types and brands of electric appliances. Electric appliance may share similar characteristics with multiple types of electric appliances. Single label assignment has shortcoming in the aspects of convergence and diversity.

3. Methodology

Since the focus of our research work is semi-automatic data labelling instead of electricity theft detection, a highly cited existing work [22] using deep convolutional neural networks is adopted as electricity theft detection algorithm. The following steps are performed for semi-automatic data labelling of electricity data:

Step 1: Implement the electricity theft detection algorithm [22] with training dataset of period X week. The dataset is initially labelled with ground truth data;

Step 2: The trained model in Step 1 will be used to predict the outputs of data in next X week;

Step 3: The outputs in Step 2 will be served as new labelled data.

Step 4: Train the electricity theft detection model again.

Step 5: Compare the performance between the models in Step 1 and Step 4. Repeat Steps 2-4 until the deterioration of the performance exceeds the threshold γ.

Step 6: When condition in Step 5 fulfills, manually label new training dataset of period X week. Repeat Steps 2-5.

It is expected the semi-automatic data labelling algorithm can reduce the number of manually labelled data. Period X will be varied and analyzed in next Section.

4. Results

As preliminary study, the period X is varied from 1 to 4 (week). Accordingly, the performance of the first ten retrained models (RMs) are summarized in Table 1. Several observations have been summarized.

Observation 1: Increasing the length of period X can yield a better performance of the model because of the chacracteristic of machine learning. More data is available to provide better understanding of the nature of the problem.

Observation 2: The accuracy of the retrained model is deteriorating with the increase in the number of retraining processes.

Table 1

Accuracy of the first ten RMs with varying period X.

Period X	**Retrained model (accuracy in percentage)**									
(in week)	**RM1**	**RM2**	**RM3**	**RM4**	**RM5**	**RM6**	**RM7**	**RM8**	**RM9**	**RM10**
1	80.5	79.8	78.3	77.5	76.2	75.1	73.7	72.2	71.1	70.6
2	81.8	81.2	80.3	79.5	78.7	77.8	77.1	76.3	75.4	74.5
3	82.3	81.6	81.0	80.1	79.3	78.9	78.5	77.9	77.2	76.4
4	83.2	82.4	81.7	81.2	80.4	79.6	79.3	78.5	77.9	77.2

5. Conclusion

The preliminary study of the semi-automatic data labelling algorithm provides a tradeoff between the reduction of manual labelling and performance of the detection model. Future research directions include (i) synthesizing extra training data with generative adversarial network [23] which can increase the accuracy of the model in the initial model training phase.; (ii) adopting incremental learning algorithm to keep updating the model without retraining; and (iii) enhancing the performance of detection model with boosting algorithms.

Funding

The work described in this paper was fully supported by the Hong Kong Metropolitan University, research grant number R6394 [ROU/E10/20].

6. References

[1] D.Yumashev, C. Hope, K. Schaefer, K. Riemann-Campe, F. Iglesias-Suarez, E. Jafarov, E. J. Burke, P. J. Young, Y. Elshorbany, G. Whiteman, "Climate policy implications of nonlinear decline of Arctic land permafrost and other cryosphere elements", Nature communications 10.1 (2019) 1900.

[2] A. K. Liljedahl, J. Boike, R. P. Daanen, A. N. Fedorov, G. V. Frost, G. Grosse, L. D. Hinzman, Y. Iijma, J. C. Jorgenson, N. Matveyeva, M. Necsoiu, M. K. Raynolds, V. E. Romanovsky, J. Schulla, K. D. Tape, D. A. Walker, C. J. Wilson, H. Yabuki, D. Zona, "Pan-Arctic ice-wedge degradation in warming permafrost and its influence on tundra hydrology", Nature Geoscience 9.4 (2016) 312.

[3] Gupta, S., & Gupta, B. B. (2015, May). PHP-sensor: a prototype method to discover workflow violation and XSS vulnerabilities in PHP web applications. In Proceedings of the 12th ACM International Conference on Computing Frontiers (pp. 1-8).

[4] AlZu’bi, S., Hawashin, B., Mujahed, M., Jararweh, Y.,et al. (2019). An efficient employment of internet of multimedia things in smart and future agriculture. Multimedia Tools and Applications, 78(20), 29581-29605.

[5] Gupta, B. B., & Akhtar, T. (2017). A survey on smart power grid: frameworks, tools, security issues, and solutions. Annals of Telecommunications, 72(9), 517-549.

[6] Y. Wang, Q. Chen, T. Hong, C. Kang, “Review of smart meter data analytics: Applications, methodologies, and challenges”, IEEE Transactions on Smart Grid 10.3 (2019) 3125-3148.

[7] M. Papadimitrakis, N. Giamarelos, M. Stogiannos, E. N. Zois, N. I. Livanos, A. Alexandridis, “Metaheuristic search in smart grid: A review with emphasis on planning, scheduling and power flow optimization applications”, Renewable and Sustainable Energy Reviews 145 (2021) 111072.

[8] K. T. Chui, B. B. Gupta, R. W. Liu, P. Vasant, “Handling Data Heterogeneity in Electricity Load Disaggregation via Optimized Complete Ensemble Empirical Mode Decomposition and Wavelet Packet Transform”, Sensors 21.9 (2021) 3133.

[9] M. D. Lytras, K. T. Chui, “The recent development of artificial intelligence for smart and sustainable energy systems and applications”, Energies 12.16 (2019) 3108.

[10] X. Kong, X. Zhao, C. Liu, Q. Li, D. Dong, Y. Li, “Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM”, International Journal of Electrical Power & Energy Systems”, 125 (2021) 106544.

[11] A. Aldegheishem, M. Anwar, N. Javaid, N. Alrajeh, M. Shafiq, H. Ahmed, “Towards sustainable energy efficiency with intelligent electricity theft detection in smart grids emphasising enhanced neural networks”, IEEE Access 9 (2021) 25036-25061.

[12] A. Khan, F. G. Penalvo (2021), Blockchain Technology and Associated Challenges in Smart Healthcare Systems, Insights2Techinfo, pp.1

[13] A. Dahiya, J. Wu (2021), 5G Communication for IIoT Era and smart manufacturing, Insights2Techinfo, pp. 1

[14] K. T. Chui, K. F. Tsang, S. H. Chung, L. F. Yeung, Appliance signature identification solution using K-means clustering, in IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2013, pp. 8420-8425. doi:10.1109/IECON.2013.6700545.

[15] I. Abubakar, S. N. Khalid, M. W. Mustafa, H. Shareef, M. Mustapha, “Application of load monitoring in appliances’ energy management–A review”, Renewable and Sustainable Energy Reviews 67 (2017) 235-245.

[16] A. Zoha, A. Gluhak, M. Imran, S. Rajasegarar, “Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey”, Sensors 12.12 (2012) 16838-16866.

[17] I. Abubakar, S. N. Khalid, M. W. Mustafa, H. Shareef, M. Mustapha, “Application of load monitoring in appliances’ energy management–A review”, Renewable and Sustainable Energy Reviews 67 (2017) 235-245.

[18] A. Iwayemi, C. Zhou, “SARAA: Semi-supervised learning for automated residential appliance annotation”, IEEE Transactions on Smart Grid 8.2 (2017) 779-786.

[19] D. Li, S. Dick, “Residential household non-intrusive load monitoring via graph-based multi-label semi-supervised learning”, IEEE Transactions on Smart Grid 10.4 (2019) 4615-4627.

[20] Y. Yang, J. Zhong, W. Li, T. Gulliver, S. Li, S. “Semisupervised multilabel deep learning based nonintrusive load monitoring in smart grids”, IEEE Transactions on Industrial Informatics 16.11 (2020) 6892-6902.

[21] J. M. Gillis, W. G. Morsi, “Non-intrusive load monitoring using semi-supervised machine learning and wavelet design”, IEEE Transactions on Smart Grid 8.6 (2017) 2648-2655.

[22] Z. Zheng, Y. Yang, X. Niu, H. N. Dai, Y. Zhou, Y. “Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids”, IEEE Transactions on Industrial Informatics: 14.4 (2018) 1606-1615.

[23] K. T. Chui, M. D. Lytras, P. Vasant, “Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset”, Applied Sciences 10.13 (2020) 4571.