Machine learning for predicting energy efficiency of buildings: a small data approach

Izonin, Ivan; Tkachenko, Roman; Mitoulis, Stergios; Faramarzi, Asaad; Tsmots, Ivan; Mashtalir, Danylo

DOI:
10.1016/j.procs.2023.12.173

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Izonin, I, Tkachenko, R, Mitoulis, S, Faramarzi, A, Tsmots, I & Mashtalir, D 2024, 'Machine learning for predicting energy efficiency of buildings: a small data approach', *Procedia Computer Science*, vol. 231, pp. 72-77. https://doi.org/10.1016/j.procs.2023.12.173

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 231 (2024) 72–77

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2023)
November 7-9, 2023, Almaty, Kazakhstan

Machine learning for predicting energy efficiency of buildings: a small data approach

Ivan Izonin[a,c*], Roman Tkachenko[b], Stergios Aristoteles Mitoulis[c], Asaad Faramarzi[c], Ivan Tsmots[d], Danylo Mashtalir[a]

[a] *Department of Artificial Intelligence, Lviv Polytechnic National University, Kniazia Romana str., 5, Lviv, 79905, Ukraine*
[b] *Department of Publishing Information Technologies, Lviv Polytechnic National University, S. Bandera, str., 12, Lviv, 79013, Ukraine*
[c] *Department of Civil Engineering, School of Engineering, University of Birmingham, Birmingham B15 2TT, United Kingdom*
[d] *Department of Automated Control Systems, Lviv Polytechnic National University, Kniazia Romana str., 5, Lviv, 79905, Ukraine*

Abstract

This paper provides a method for predicting the energy efficiency of buildings using artificial intelligence tools. The scopes is twofold: prediction of the levels of the heating load and cooling load of buildings. A feature of this research is the performance of intellectual analysis in conditions of a limited amount of data when solving the stated tasks. An improved method of augmentation and prediction (input-doubling method) is proposed by processing data within each cluster of the studied dataset. The selection of the latter occurs due to the use of the fast and easy-to-implement k-means method. Next, a prediction is made using the input-doubling method within each separate cluster. The simulation of the method was performed on a real-world dataset of 768 observations. The proposed approach was found to have a high prediction accuracy in the absence of overfitting and high generalization properties of the improved method. Comparison with existing methods showed an increase in accuracy by 40-46% (MSE) compared to SVR with rbf kernel, which is the basis for the improved method, and by 5-12% (MSE) compared to the closest existing hierarchical predictor.

* Corresponding author. Tel.: +380988889687.
E-mail address: ivanizonin@gmail.com

1877-0509
10.1016/j.procs.2023.12.173

Keywords: artificial intelligence; civil engineering; energy efficiency; small data approach; prediction; clustering; input-doubling method

1. Introduction

Electric energy plays a key role in the modern development of post-industrial society. It is one of the most important forms of energy that ensures the efficient operation of various industries and sectors of the economy. The rapid development of industry requires the use of increasing amounts of electric energy. However, the constant growth of energy prices makes it necessary to find ways of efficient energy use to ensure the maximum result with minimum energy consumption. These include energy saving, optimization of production processes, use of renewable energy sources, and improvement of energy efficiency of technologies and buildings [1,2].

The energy efficiency of buildings is important for various purposes. Designing and constructing buildings with energy-efficient materials and technologies, using quality energy-efficient heating, ventilation, and air conditioning systems, using natural lighting, and other measures can reduce energy consumption for heating and cooling. The evaluation of the last two parameters will help to determine the necessary ventilation systems for a comfortable life for the building's residents while minimising the resources needed for heating and cooling the buildings. Prediction of these indicators based on the characteristics of the building will save time for the engineer and allow designing and construction of an energy-efficient building.

Many papers are focus on the energy performance of residential buildings. The ones that use the machine learning methodologies [3,4] are the most promising because of the accuracy and speed of prediction using such tools [5]. Despite this, existing ML-algorithms do not always provide a sufficient level of accuracy when analyzing complex non-linear relationships between data. The problem becomes more challenging in the case of the need to analyze a limited set of data. In this case, the generalization ability of the machine learning algorithm, which is trained under conditions of data scarcity, may be quite low. In addition, there are overfitting problems that make it impossible to use the chosen method in practice

When it is difficult to select data for training at a certain point in time, data augmentation methods can be used. However, classical approaches to solving this problem have many significant drawbacks that limit their application [6]. To avoid these shortcomings, a new augmentation procedure was developed in the cycle of investigations [6–8], which provides a significant increase in prediction accuracy using ML-based algorithms. However, the quadratic increase of the dataset significantly slows down the operation of machine learning methods and requires a lot of resources for their implementation [9,10].

The paper aims to improve the accuracy of the predicting energy efficiency of buildings in the case of analysis of short datasets due to the consistent use of clustering, augmentation, and ML-based regressors in each separate data cluster.

2. Materials and methods

Intelligent analysis of short datasets faces many challenges which are described in detail in [11–13]. To overcome such limitations, in the cycle of research [6–8] a new method of augmentation and prediction in the case of short sets of tabular data (input-doubling method) is proposed for the first time. According to the author's algorithm of data augmentation, which is based on the principles of axial symmetry of the response surface, it provides a quadratic increase in the number of vectors in the tabular dataset with a simultaneous doubling of the number of features of each vector.

The main step of the data augmentation procedure is the pairwise concatenation of all vectors of the available training dataset [6]. The output attribute in this case is formed as the difference of outputs of both vectors that are concatenated $output_i = (y_i - y_k)$. As a result of this step, the training data set increases quadratically. In this way, we get a dataset ready for training.

The Support Vector Regression (SVR) was chosen to implement the training procedures of the improved method. This choice is due to several reasons, in particular, [6,14,15]: high speed and ease of implementation, accurate results when analyzing short datasets, and the ability to work with significantly non-linear data due to the use of different kernels of the method.

In the testing mode of the method, one vector of the test sample concatenated with each of N-vectors initial training sample, prediction of temporary values $output_{k,i}^{predicted}$ for each k-th vector of such a temporary sample of concatenated vectors, and formation of the final result y_k using the following equation [6]:

$$y_k \cong \frac{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} output_{k,i}^{predicted}}{N} \quad (1)$$

The advantage of this approach is that it provides a significant increase in prediction accuracy using simple procedures within the available dataset. However, its limitation is its quadratic increase in the dataset, which in the case of analyzing samples of more than 100 observations is time-consuming and requires resources to implement the training algorithms.

To avoid this drawback, and increase the prediction accuracy, this paper proposes the use of a clustering procedure for data preprocessing. It consists in selecting the optimal number of clusters for the studied dataset. In the case of clustering for the small-sized or middle-sized dataset, individual data clusters may contain a very small number of observations, which will significantly limit or even make it impossible to use classical machine learning methods for solving the prediction task. This is also true for the hierarchical predictor [16]. Therefore, the analysis of each data cluster in this paper is proposed to be performed using the input-doubling method. This approach leads to a significant reduction in the duration of the method's training procedure since the analysis of each data cluster can be implemented in parallel. In addition, prediction in the case of close vectors, specific to each data cluster, can also improve the prediction accuracy of the method [15].

A fast, easy-to-implement and established k-means method was chosen for clustering in this paper. Clustering by the k-means method involves the user determining the required number of clusters, randomly selecting the centroids of these clusters, and assigning each observation to the corresponding cluster based on the determination of the smallest Euclidean distance to the centre of the corresponding cluster [17,18].

The main advantage of k-means, which is also described in the main steps of this method, is its simplicity [19]. However, performing the first step of clustering using K-means, namely the determination by the user of the number of clusters, is a significant drawback of this algorithm, particularly within the framework of the method developed in this paper. In particular, choosing the wrong number of clusters can lead to the misinterpretation of the data or failed clustering. To avoid this, an automated procedure for determining the optimal number of clusters during clustering using the K-means algorithm was used. It is based on the use of an objective metric to determine this indicator, the silhouette coefficient [20]. It measures how well each object fits into its cluster compared to other clusters. The higher the value of the silhouette coefficient, the better the clustering.

3. Results and discussion

Modeling of the proposed method took place using a short set of data on the assessment of the energy efficiency of buildings [21]. It contains 768 observations, 6 independent attributes (Wall Area, Relative Compactness, Overall Height, Glazing Area, Orientation, Roof Area, Surface Area, Glazing Area Distribution), and two dependent attributes (Heating Load, Cooling Load). The authors of this dataset consider the problem of energy efficiency of buildings as the problem of assessing the heating and cooling load requirements of buildings. The prediction of these parameters can be built as an approximation of the function of building parameters. Considering the two dependent attributes, in this paper we considered the task of predicting both of these parameters as two separate tasks.

The simulation was carried out using the software developed by the authors. It was performed on a computer with the following : AMD Ryzen 5 1600 3.2GHz, 6 kernels, Nvidia Geforce 1050TI, 16 GB DDR4-2400GHz. The available dataset was randomly divided into two parts: training (75%) and test (25%) data samples. Performance evaluation was conducted using Mean Absolute Error (MAE) and Mean Square Error (MSE).

The method developed consists of two steps: The first step is to perform data clustering with the determination of the optimal number of clusters. The k-means method was used as a fast and easy-to-implement clustering method. The optimal number of clusters was chosen taking into account the silhouette coefficient. Fig. 1 shows the dynamics

of the values of the silhouette coefficient when the number of clusters into which the studied data set was divided is changed.

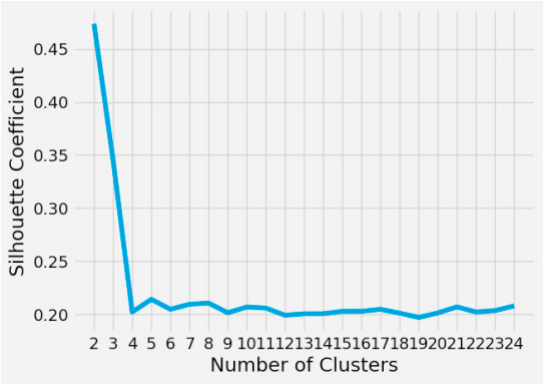

Fig. 1. The value of the silhouette coefficient for different numbers of clusters

As can be seen from Figure 1, the optimal number of clusters for the studied dataset is 2.

The second step of the method involves the application of the input-doubling method for prediction within each of the studied clusters.

The general method's results in both training and application modes based on MAE and MSE for both stated tasks are summarized in Table 1.

Table 1. Method's results.

Prediction outcomes	Mode	MAE	MSE
Heating load prediction	Train	1.407	4.098
	Application	1.410	4.363
Cooling load prediction	Train	1.582	5.552
	Application	1.726	6.438

As can be seen from Table 1, overfitting, which can happen when applying the data augmentation methods, is not observed. The training error is smaller than the application error. In addition, the method demonstrates high generalization properties especially when solving the heating load prediction task (the difference between training and application errors is small). This indicates the possibility of using the proposed method when solving real-world applied tasks.

To evaluate the effectiveness of the proposed method, its results were compared with the results of known methods, in particular:

- Classic SVR with rbf kernel with basic parameters from the library https://scikit-learn.org/;
- Hierarchical predictor [15], which used the same number of clusters as the developed method and the classical SVR with rbf kernel to perform the prediction procedure in each separate cluster.

The performance errors of all studied methods (MAE and MSE) when solving the building heating load and building cooling load tasks are shown in Fig. 2 and Fig. 3 respectively.

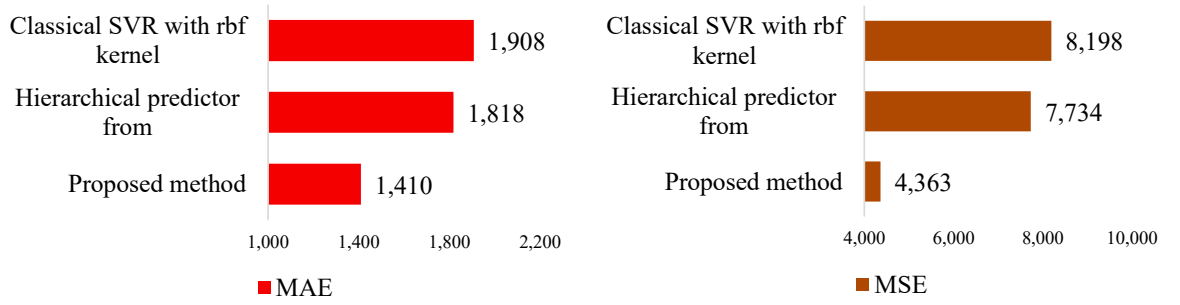

Fig. 2. Comparison of different ML-based methods results for buildings heating load prediction based on (a) MAE; (b) MSE.

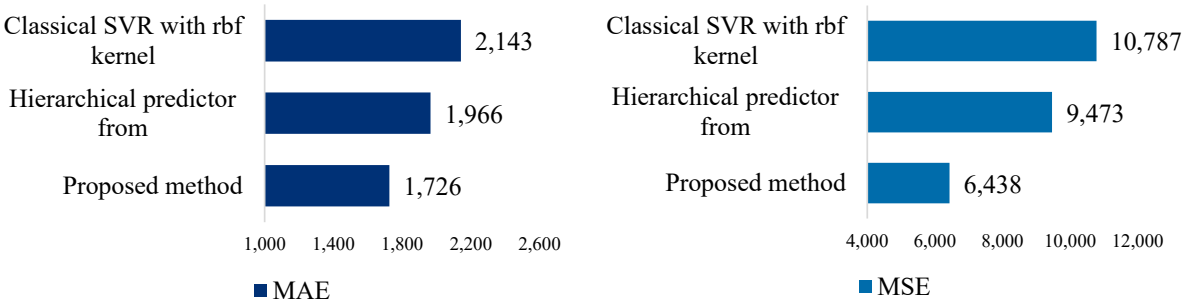

Fig. 3. Comparison of different ML-based methods results for buildings cooling load prediction based on (a) MAE; (b) MSE.

As can be seen from both Figures, the use of the prediction strategy, within each cluster into which the given dataset is divided, fully justified itself. In particular, the method developed in this paper demonstrates better results than its analog, a hierarchical predictor, which also works within each data cluster, but does not use augmentation, which is very important in the case of short datasets and is significantly higher than using the classic SVR with RBF kernel on the whole dataset. According to MSE, the developed method demonstrated for the first task a 5.5% better result than the known one and a 46% better result than using classical SVR. For the second task, such improvements were 12% and 40%, respectively.

4. Conclusions

This paper considers the task of predicting the energy efficiency of buildings using intelligent data analysis. The authors predict the level of the heating load and cooling load of buildings based on their parameters. For this purpose, the paper uses the procedures of data augmentation and prediction based on the input-doubling method. We proposed the improvement of this method due to the use of preliminary data clustering, which ensured an increase in the accuracy of the obtained results.

The modeling of the proposed method was performed on a real-world dataset of 768 observations. To evaluate the effectiveness of the proposed method, its results were compared with the results of known methods. The results showed that, according to MSE, the developed method for the first task is 5.5% better than the known one and 46% better than using classical SVR when solving the heating load prediction task. For the second task, the cooling load prediction task, such improvements were 12% and 40%, respectively. This is strong evidence that the proposed method can be used in practical applications.

Acknowledgments

This research is supported by the British Academy's Researchers at Risk Fellowships Programme.

References

[1] Medykovskvi M, Pavliuk O, Sydorenko R. Use of Machine Learning Technologys for the Electric Consumption Forecast. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, 2018, p. 432–5. https://doi.org/10.1109/STC-CSIT.2018.8526617.

[2] Pavliuk O, Steclik T, Biernacki P. The forecast of the AGV battery discharging via the machine learning methods. 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan: IEEE; 2022, p. 6315–24. https://doi.org/10.1109/BigData55660.2022.10020968.

[3] Tsanas A, Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings 2012;49:560–7. https://doi.org/10.1016/j.enbuild.2012.03.003.

[4] Ibrahim DM, Almhafdy A, Al-Shargabi AA, Alghieth M, Elragi A, Chiclana F. The use of statistical and machine learning tools to accurately quantify the energy performance of residential buildings. PeerJ Computer Science 2022;8:e856. https://doi.org/10.7717/peerj-cs.856.

[5] Argyroudis SA, Mitoulis SA, Chatzi E, Baker JW, Brilakis I, Gkoumas K, et al. Digital technologies can enhance climate resilience of critical infrastructure. Climate Risk Management 2022;35:100387. https://doi.org/10.1016/j.crm.2021.100387.

[6] Izonin I, Tkachenko R, Gregus M, Zub K, Lotoshynska N. Input Doubling Method based on SVR with RBF kernel in Clinical Practice: Focus on Small Data. Procedia Computer Science 2021;184:606–13. https://doi.org/10.1016/j.procs.2021.03.075.

[7] Izonin I, Tkachenko R, Fedushko S, Koziy D, Zub K, Vovk O. RBF-based Input Doubling Method for Small Medical Data Processing. Advances in Intelligent Systems and Computing 2021;ICAILE2021: The First International Conference on Artificial Intelligence and Logistics Engineering:(in press).

[8] Izonin I, Tkachenko R, Gregus ml. M, Zub K, Tkachenko P. A GRNN-based Approach towards Prediction from Small Datasets in Medical Application. Procedia Computer Science 2021:(in press).

[9] Kotsovsky V, Batyuk A. Feed-forward Neural Network Classifiers with Bithreshold-like Activations. 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine: IEEE; 2022, p. 9–12. https://doi.org/10.1109/CSIT56902.2022.10000739.

[10] Kotsovsky V, Batyuk A, Voityshyn V. On the Size of Weights for Bithreshold Neurons and Networks. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), LVIV, Ukraine: IEEE; 2021, p. 13–6. https://doi.org/10.1109/CSIT52700.2021.9648833.

[11] Bisikalo O, Kharchenko V, Kovtun V, Krak I, Pavlov S. Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis. Entropy 2023;25:184. https://doi.org/10.3390/e25020184.

[12] Krak I, Kuznetsov V, Kondratiuk S, Azarova L, Barmak O, Padiuk P. Analysis of Deep Learning Methods in Adaptation to the Small Data Problem Solving. In: Babichev S, Lytvynenko V, editors. Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making, vol. 149, Cham: Springer International Publishing; 2023, p. 333–52. https://doi.org/10.1007/978-3-031-16203-9_20.

[13] Bodyanskiy Y, Chala O, Kasatkina N, Pliss I. Modified generalized neo-fuzzy system with combined online fast learning in medical diagnostic task for situations of information deficit. MBE 2022;19:8003–18. https://doi.org/10.3934/mbe.2022374.

[14] Chumachenko D, Piletskiy P, Sukhorukova M, Chumachenko T. Predictive Model of Lyme Disease Epidemic Process Using Machine Learning Approach. Applied Sciences 2022;12:4282. https://doi.org/10.3390/app12094282.

[15] Building sharp regression models with K-Means Clustering + SVR. Paperspace Blog 2021. https://blog.paperspace.com/svr-kmeans-clustering-for-regression/ (accessed July 15, 2023).

[16] Shakhovska N, Izonin I, Melnykova N. The Hierarchical Classifier for COVID-19 Resistance Evaluation. Data 2021;6:6. https://doi.org/10.3390/data6010006.

[17] Bodyanskiy YV, Tyshchenko OK, Kopaliani DS. An evolving connectionist system for data stream fuzzy clustering and its online learning. Neurocomputing 2017;262:41–56. https://doi.org/10.1016/j.neucom.2017.03.081.

[18] Babichev S, Škvor J. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. Diagnostics 2020;10:584. https://doi.org/10.3390/diagnostics10080584.

[19] Zomchak L, Melnychuk V. Creditworthiness of Individual Borrowers Forecasting with Machine Learning Methods. In: Hu Z, Ye Z, He M, editors. Advances in Artificial Systems for Medicine and Education VI, vol. 159, Cham: Springer Nature Switzerland; 2023, p. 553–61. https://doi.org/10.1007/978-3-031-24468-1_50.

[20] Berezsky O, Pitsun O, Dubchak L, Berezka K, Dolynyuk T, Derish B. Cytological Images Clustering of Breast Pathologies. 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine: IEEE; 2020, p. 62–5. https://doi.org/10.1109/CSIT49958.2020.9321867.

[21] Tsanas A, Xifara A. Energy efficiency 2012. https://doi.org/10.24432/C51307.