J. Comp. Int. Sci. (2019) 10(2):-59-68
http://epacis.net/jcis/PDF_JCIS/JCIS-0160.pdf
jcis@epacis.net

Pan-American Association of Computational Interdisciplinary Sciences

Prediction of energy load of buildings using machine learning methods

Grasiele Regina Duarte[a1], Priscila Capriles[a] Afonso Lemonge[a] and Leonardo Goliatt[a]

[a]Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil

Received on Dez, 2018 / Accepted on Feb, 2019

Abstract

The climatic conditions within the residential buildings in warm climates can be modified with the use of air conditioners equipments. However, constant use of these equipments may result in high energy consumption. To reduce the use of such equipments and maintain the desired internal temperature is possible to design energy-efficient buildings. To measure the energy efficiency of a building is necessary to estimate its heating and cooling loads, considering some of its physical characteristics defined in the design. Machine Learning Methods can be applied to this problem estimating an answer from a set of inputs. These methods require a training phase, called supervised training, which considers a database drawn from selected variables in the problem domain. This work evaluates the performance of four Machine Learning Methods in the prediction of cooling and heating loads of residential buildings. The training database consists of eight input variables and two output variables, derived from building designs. Methods were selected according to an exhaustive search and adjusted by a strategy with cross-validation. To the evaluate were used four performance metrics. This strategy resulted in algorithms with optimized parameters and allowed to obtain competitive results with the literature.

Keywords: Energy efficiency, heating and cooling loads, machine learning

1 Introduction

The climatic conditions within the residential buildings can be determined with the use of technologies such as air conditioners and heaters.

[1]E-mail Corresponding Author: grasiele.duarte@engenharia.ufjf.br

However, the intense and numerous use of such equipments can generate a high energy consumption. An alternative to reduce the use of cooling and heating equipments, maintaining the desired indoor climate conditions, is to design energy-efficient buildings able to produce such conditions. To assess the energy efficiency of a building, it is necessary to estimate and analyze its heating and cooling loads based on physical characteristics defined during the design process. Moreover, informations as global location, purpose of building, occupation and activity level, should be considered.

Among the computational tools for this purpose are those that simulate scenarios which often produce accurate results. These tools may require advanced knowledge of the user due to the multidisciplinary aspect present in many of them. In addition, simulations may consume considerable computational time and cost and the results may vary depending on the software used.

Another computational resource that can be applied in this type of problem are Machine Learning Methods, identified in this work as MLM. These tools are trained (to learn) from a database elaborated based on a set of selected variables in the problem domain. This type of training is known as supervised training and consists basically in to reduce the error that can exist between the value produced by the MLM and the value contained in the database.

Some MLM have been used in the context of energy performance of buildings [1, 2, 3, 4]. Between many MLM available in the literature it is important to know which are better suited to the problem mentioned, that produce good results or even what should not be applied (do not have the profile that meets the conditions of the scenario). This work applies 4 MLM (DT, MLP, RF and SVR) with a procedure of model selection for the prediction of cooling and heating loads of residential buildings.

2 Database and Machine Learnig Methods

2.1 Database

The dataset used in this study was presented in [5]. The data were obtained by the simulation of a set of buildings using the software Ecotect. The set is composed by eight input variables and two output variables mentioned in Section 1. From now on, the input variables will be identified as x_1, x_2, x_3, x_4, x_5, x_6, x_7 and x_8 and the output variables as y_1 and y_2, as described in Table 1.

In [5] were simulated 12 buildings located in Athens – Greece, composed

Table 1: Representation of the input and output variables [5].

Representation	Variable[a]	Number of Possible Values
x_1	Relative Compactness (RC)	12
x_2	Surface Area	12
x_3	Wall Area	7
x_4	Roof Area	4
x_5	Overall Height	2
x_6	Orientation	4
x_7	Glazing Area	4
x_8	Glazing Area Distribution	6
y_1	Heating Load	586
y_2	Cooling Load	636

[a] Input (x_i) and output (y_i) variables.

by 18 blocks of dimension $3.5 \times 3.5 \times 3.5m$, leading to a volume equal to $771.75m^3$ for each simulated building. In [5] was considered that all buildings were constructed with the same material, all of them with the lowest U-value. The characteristics used (U-values in parentheses) were: walls (1.780), floor (0.860), roofs (0.500) and windows (2.260). In terms of surface area and design dimensions the buildings are different, as shown in Figure 1.

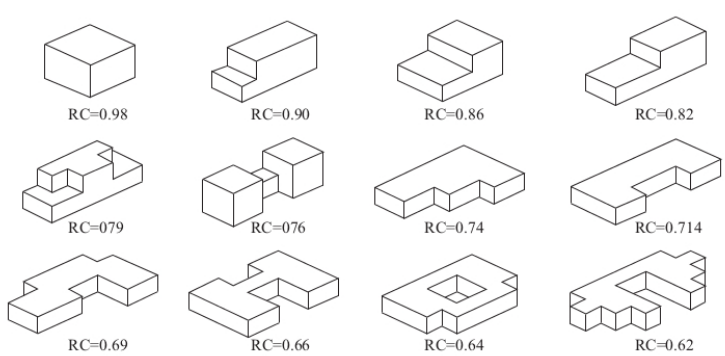


Figure 1: Relative Compactness coefficient variation – RC (x_1) according to [4].

In [5] was considered that each block was occupied by seven persons with sedentary life. The internal design of the blocks was defined as: clothing: 0.6 *clo*, humidity: 60%, air speed: $0.30m/s$, lighting level: 300 Lux. The sensitivity for internal gain was 5 and latency $2W/m^2$. The air infiltration

rate was 0.5 and the air change rate with wind sensibility was 0.25. The buildings were simulated in [5] with temperatures between $[19, 24]^o$C and was considered that the activities in the buildings occurred between 15 and 20 h on weekdays and 10 to 20 h on weekends.

For each of the 12 simulated buildings in [5] was considered three types of glazing areas defined as percentages of the floor area: 10%, 25%, and 40%. For each of them were applied five different distributions: (1) Uniform: 25% glazing on each side of the building; (2) North: 55% of glazing on the north side and 15% on the other sides; (3) South: 55% of glazing on the south side and 15% on the other sides; (4) East: 55% of glazing on the east side and 15% on the other sides; (5) West: 55% of glazing on the west side and 15% on the other sides. Besides, buildings were simulated without glazing areas. Finally, all the buildings were rotated to face the four cardinal points. Based in these information it is concluded that the database used in this work is composed by $12 \times 3 \times 5 \times 4 + 12 \times 4 = 768$ samples of buildings.

2.2 Machine Learning Methods

In this paper the following MLM's were used to approximate the Heating and Cooling Loads from data set: Decision Trees (DT) [6] (Binary Trees.), Random Forests (RF) [7], Support Vector Machines (SVR - Support Vector Regression) [8] and Artificial Neural Networks (MLP - Multilayer Perceptron) [9]. Detailed description of the methods employed here and their parameters can be found in [10] and [11].

2.3 Grid Search

Grid Search is a strategy for automatic and optimized adjustment of MLM parameters. This technique builds a mesh from sets of predefined values for each parameter. For each possible combination of values the method is trained with a portion of the data, generating a set of outputs. The best parameters values are those that produced the best set of outputs. The number of configurations for the method is given by $S = \prod_{k=1}^{K} N_k$ where K is the number of parameters and N_k is the number of values chosen for the k-th parameter [12].

2.4 Cross Validation – k-Fold

After the training step it is necessary to test the MLM's ability to adapt to a new data with similar characteristics. The strategy known as k-Fold CV was adopted, which shares the data set into k sets. The model is adjusted

with $k-1$ sets and validated with the remaining part. Training and testing steps are repeated k times alternating the sets of training and testing. Figure 2 illustrates the application of k-Fold CV. In this study $k = 10$ was set.

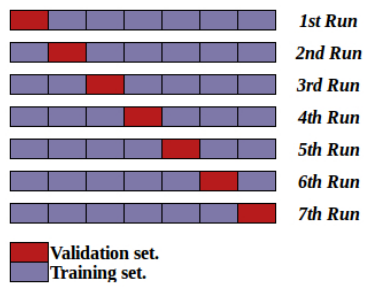


Figure 2: Training and testing of the MLM applying k-Fold CV with $k = 7$.

2.5 Performance evaluation

Given a data set composed by N observations, the following performance metrics are written as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}_i|, \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}_i|^2}. \tag{2}$$

$$MRE = 100 \times \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \bar{y}_i|}{y_i}. \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{4}$$

where y_i is the expected value for the output variable y with the input $\mathbf{x}_i$, $\bar{y}_i$ is the predicted value to y for the same input $\mathbf{x}_i$ and $\hat{y}_i$ is the average predicted value to the output variable y.

3 Computational Experiments

To achieve the main objective of this work the computational experiment can be described by:

1. Train and evaluate MLM listed in Section 2.2 to problem described in Section 1:
   - Identify the best values for the method parameters by applying the Grid Search.
   - Train and evaluate the method by applying the technique k-Fold CV.
2. Evaluate, compare and identify the best method based on performance metrics.
   - MAE, RMSE, MRE, R^2.

The implementation of all methods were based on package scikit-learn [13] version 0.15. Each MLM (using the best configuration defined with the Grid Search) was trained and validated in 50 independent runs. Table 2 lists the parameters of each MLM and the values selected for application of Grid Search strategy. Other parameters involved in the methods, not defined for this step, were kept with the default values set in the implementations of the methods in scikit-learn package.

Figure 3 illustrates the values of MAE, RMSE, MRE and R^2 calculated from the average of the predicted values to the output variables y_1 and y_2 among 50 runs of each MLM. One can note observing Figure 3(d) that is possible to predict with quality the heating (y_1) and cooling (y_2) loads applying methods evaluated in this work. It is clear from the Figures 3(a), 3(b) and 3(c) that the cooling load is more difficult to be defined, probably due to the high number of possible values for this variable when compared with the number of samples as presented in Table 1. Thus, the results for this variable are very important to evaluate the methods in this problem. Considering only the variable y_1, Figure 3 shows that the performance difference among the methods is relatively small. Also, it is possible to note in the Figures 3(a), 3(b) and 3(c) that the SVR method presents the less attractive values for two of these three metrics whereas for MRE the difference is relatively high. Observing carefully the variable y_2, the graphics of Figure 3 make clear that DT and RF methods have the worst performance to predict this variable. Between these two, the method DT presents the biggest errors, probably due to the limitation of the number of nodes, which can result in some conditions involving possible values for this variable no longer included in the model. Considering the two variables, note that in all graphs of Figure 3 the MLP method showed the best performance. The best values produced for all metrics are related to this method.

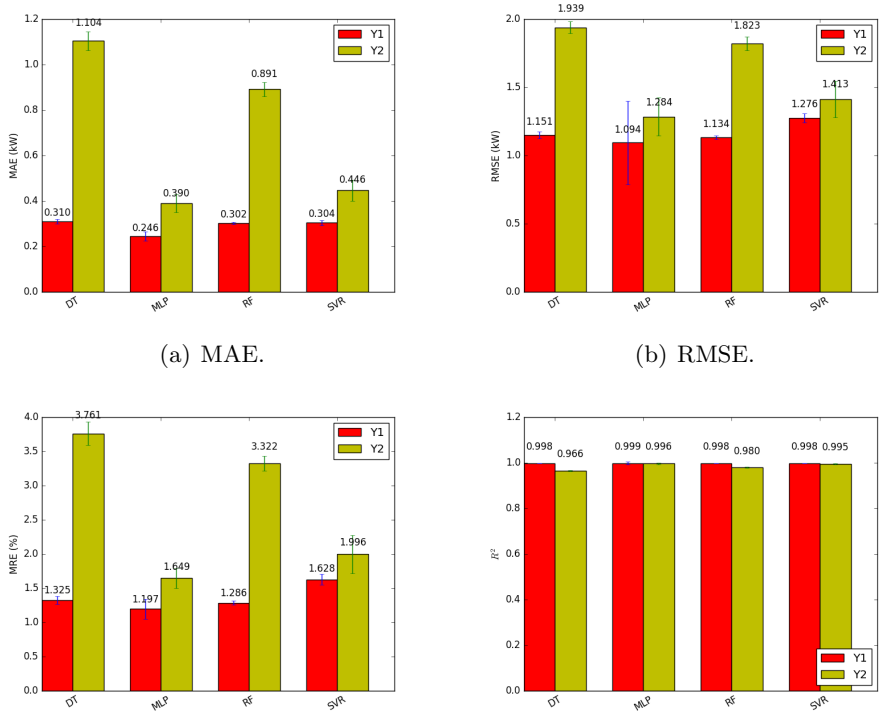
Table 2: Parameters and their values for application of Grid Search. A detailed explanation of all parameters can be found on Reference [13].

Method	**Parameters Description**	**Range of values**
DT	Max Depth	$[None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50]$
MLP	Number of hidden layers and neurons	$[[5], [10], [20], [50], [100], [5, 5], [10, 10], [20, 20], [5, 5, 5], [10, 10, 10]]$
	Training algorithm	[tnc, l-bfgs, sgd, rprop]
	Bias	True
	Connectivity	[mlgraph, tmlgraph]
RF	Number of trees	$[10, 20, 30]$
	Boot Strap	[True, False]
	Max Depth	$[None, 1, 2, 4, 8, 16, 32]$
	Max Features	$[auto, 1.0, 0.3, 0.1]$
	Minimum Samples Leaf	$[1, 3, 5, 9, 17]$
SVR	Max iterations	100000
	C	[1, 10, 1e2, 1e3, 1e4, 1e5, 1e6]
	σ	[1e0, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6]
	Base function	[rbf]
	ε	[1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6]

Tables 3 and 4 present the best values produced in this work for each performance metric. These values are compared to the results obtained in other studies. For the cooling load prediction, the method identified in this work as the best (MLP) produced the best values for 3 of the 4 metrics applied with a significant difference. What can be concluded is that apply MLP adopting Grid Search for adjusting its parameters from discrete sets of values is a viable and competitive alternative.

4 Conclusion

This work evaluated the application of four MLM (DT, MLP, RF and SVR) in the prediction of the energy efficiency of residential buildings based on a data set of 768 simulated buildings. After measuring the accuracy of



(a) MAE. (b) RMSE.

(c) MRE. (d) R^2.

Figure 3: Values of MAE, RMSE, MRE and R^2 considering the average of the prediction value to the output variables in 50 runs of each MLM.

Table 3: Heating load – Comparison of results with other references (best results are highlighted in boldface).

Reference	MAE (kW)	RMSE (kW)	MRE (%)	R^2
[5]	0.510	–	2.180	–
[4]	0.340	0.460	–	**1.000**
[2]	**0.236**	**0.346**	–	0.999
[1]	0.380	–	**0.430**	–
This paper	0.246	1.094	1.197	0.999

the MLMs with four statistical metrics (MAE, RMSE, MRE and R^2) one can conclude that is possible to predict with reasobly quality the heating load and cooling load of residential buildings with the analyzed methods.

Table 4: Cooling load – Comparison of results with other references (best results are highlighted in boldface).

Reference	MAE (kW)	RMSE (kW)	MRE (%)	R^2
[5]	1.420	–	4.620	–
[4]	0.680	**0.970**	–	0.990
[2]	0.890	1.566	–	0.986
[1]	0.970	–	3.40	–
This paper	**0.390**	1.284	**1.649**	**0.996**

It is observed that the four methods present similar performance to predict heating load. The performance differences between methods appeared in the cooling load prediction. The MLP was what produced the best results for both output variables. The DT method was the one that produced the less attractive results. The SVR method tends to behave similarly to the MLP method, but showed lower performance.

Based on the results obtained, among the MLMs evaluated here, the MLP is considered the best choice for predicting heating load and cooling load of residential buildings when the eight input variables considered in this study are applied. As future work, with the expectation to improve the results presented here, others MLMs will be discussed. Also, the Grid Search strategy will be replaced by an optimization evolutionary algorithm for setting the parameters of the MLMs.

Acknowledgments. The authors thank the Graduate Program in Computational Modeling (UFJF) and Brazilian Agencies CNPq (grant 305099/2014-0) and FAPEMIG (grants TEC PPM 528/11, TEC PPM 388/14 and TEC APQ 01606/15) for the financial support.

References

[1] Castelli M., Trujillo L., Vanneschi L. and Popovi A. Prediction of energy performance of residential buildings: A genetic programming approach, *Energy and Buildings*, 102:67–74, 2015.

[2] Chou J. and Bui D. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design, *Energy and Buildings*, 82:437–446, 2014.

[3] Crawley D. B., Hand J. W., Kummert M. and Griffith B. T. Contrasting the capabilities of building energy performance simulation programs. *Building and Environment*, 43(4):661–673, 2008. Part Special: Building Performance Simulation.

[4] Cheng M. and Cao M. Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines, *Applied Soft Computing*, 22:178–188, 2014.

[5] Tsanas A. and Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[6] Hastie T., Tibshirani R. and Friedman J. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, Verlag, New York, 2 edition, 2009.

[7] Breiman L. Random forests, *Machine Learning*, 45(1):5–32, 2001.

[8] Gunn S. R. Support vector machines for classification and regression. Technical report, University of Southampton, 1998.

[9] Haykin S. O. *Neural Networks and Learning Machines*, Prentice Hall, New Jersey - USA, 3 edition, 2008.

[10] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R. and Dubourg V. Scikit-learn: Machine learning in python, *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Wojciechowski M. Solving differential equations by means of feedforward artificial neural networks. In *Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing-Volume Part I*, ICAISC'12, pages 187–195, Berlin, Heidelberg, 2012. Springer-Verlag.

[12] Bergstra J. and Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, February 2012.

[13] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.