Journal of Natural Sciences and Engineering, Vol. 1, (2019)
DOI number: 10.14706/JONSAE2019114

Quantitative estimation of cooling load capabilities of residential buildings using machine learning

Nedret Bećirović, Ismail Bejtović, Jasmin Kevrić
International Burch University, Sarajevo, Bosnia and Herzegovina
nedret.becirovic@stu.ibu.edu.ba
ismailbejtovic@hotmail.com
jasmin.kevric@ibu.edu.ba

***Abstract* – Based on previous research on energy efficiency of the buildings, particularly their cooling load capabilities we will develop a collection of machine learning methods for detecting buildings with best cooling load capabilities. This collection will study the influence of 8 input variables (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution) on one output parameter, that is cooling load of buildings. The results of this study support the practicability of using machine-learning software to estimate building parameters as a convenient and accurate approach, as long as the methods chosen are well suited for the type of data in question.**

Keywords – cooling load, energy efficiency, machine learning, neural network.

1. Introduction

Considering growing electrical energy consumption in the residential sector [1] and Global Warming it is noticeable that energy consumption for cooling will surpass energy consumption for heating in the foreseeable future. Heating and cooling load are two very important parameters in the efficient building design. These two parameters are closely related to the materials that the building is made of, so construction decisions made early on have a great impact on the final result. There has been a considerable body of research [2] on this field and on this dataset but with no focus on the cooling load itself. Various software for simulation of energy consumption has been used over the years often in conjunction with architectural design. Accuracy of the simulation varies often across from one software package to another [3]. Therefore this work is envisaged as an addition to the existing software solutions.

It is often the case that building parameters are compared separately with cooling and heating load, and simple correlation has been sought [4]. Multiple regression analysis was very popular for prediction of energy consumption until it was proven that a simple Neural Network is much better than Multiple Linear Regression Analysis with a large database [5].

For architects it is very important to single out and rank parameters that have the strongest impact since normality assumptions do not hold for very complicated problems. For example, glazing areas will have minimal impact on the cooling load. Surface area and overall height are parameters with strongest impact.

This work is done in hope it will help future architects, energy advisors for building smart buildings and generally in the field of energy efficiency. Further studies could help with choosing suitable materials for the construction.

2. Data

This study is based on UCI database made, non-gaussian dataset made by a CAD software Ecotect. Dataset represents 12 different building forms, where each form is composed of 18 building blocks of the same volume (3.5 x 3.5 x 3.5), and houses have also the same volume, which is 771.75 m^3, but different height and surface area. Materials used in these 18 blocks are all contemporary and with best U-values which are well defined for walls, floors etc with variations in glazing area and orientation [2].

With twelve building forms and three glazing area variations with five glazing area distributions each, and for four orientations, (12x3x5x4) 720 building samples. 12 building types are considered without glazing but with four sides of orientation (4x12). In all it gives 768 different building types. [2]

Since parameters are identified which have the strongest impact a new dataset can be constructed where some parameters can be locked in value and others can be varied.

Data-mining is the identification of the parameter which has the greatest influence of the result. Statistical tools will be used tools but also inputs from builders, architects, masons etc. will give great value to the study. They can also provide knowledge of feasibility of building parameters. How much a particular building feature costs in the real world.

This is a well understood, relatively large dataset with 786 buildings each having 8 parameters. This is not a skewed dataset, so this dataset is not treated as such, meaning that data were not sifted through. Some light pruning, or trimming of data is an essential part of the random and best first search methods.

Data are though skewed in another way. Dataset is non-gaussian, and it is of great importance to find any bias that may have influenced the dataset using classical statistical analysis which visually gives an outlying parameter. There were not any parameters which should be given more or less weight in the neural network model. Finding a dataset of real buildings or extracting data from buildings with a great cooling load was

also a goal for this work. Glazing area did not have much importance in this data set for finding cooling load. New modern types of materials are changing the paradigm of the builders' philosophy and focus of this work changed back on the study of the virtual buildings i.e. our dataset. It would be best to actively follow the research on the field, particularly if there has been a report on a construction of the buildings based on research using this or a similar dataset. Dataset has been normed, quantified and classified in a very understandable and logical way by Xifara-Tsanas, (see Table 1).

Table 1. Mathematical representation of the input and output variables to facilitate the presentation of the subsequent analysis and results.

Mathematical representation	**Name**	**Number of possible values**
x1	Relative compactnes	12
x2	Surface area	12
x3	Wall area	7
x4	Roof area	4
x5	Overall height	2
x6	Orientation	4
x7	Glazing area	4
x8	Glazing area distribution	6
y2	Cooling load	636

3. Methods

Classical statistical tools like histograms and scatter plots are firstly applied to dataset. Seeing the data on the graph is a great help in understanding the data. It gives the idea in which direction study has to go. Improving a model can take two different directions: make the model simpler or add complexity. Making a simpler model involves feature reduction, pruning branches and removing learners from an ensemble. Adding complexity means fine-tuning involving model-combination or adding more data sources [6].

Out of many software tools, WEKA is chosen because it is easy to use and it is easily accessible. Searching for the best computer intelligence method that is suitable for artificial dataset was the first step. Which algorithm to use is to be based on dataset form and trial and error method. Getting a good result from the start with a random forest method gave indication in which direction to go.

For the analysis of the available data set, five different regression algorithms were used:

- Linear Regression
- Random Forest
- REPTree
- SMOreg
- Multilayer Perceptron

These algorithms are recommended for these types of datasets [7]. Regression analysis was helpful to model the relationship between dependent variables (cooling load) and independent variables (8 attributes in our dataset), and because a class from a data set (cooling load) has a large number of different instances. Cross validation was used with ten folds, to get insight of how the model will behave to an unknown dataset.

All of the above algorithms are regression algorithms, with the same goal, but working in different ways. **Linear regression models** are linear predictor functions whose model parameters are estimated from the data. Linear regression models are often fitted using the least square approach, but they may be fitted in many other ways [8].

Random forest is an ensemble method, which creates multitude of decision trees, and gives as output mean prediction of individual trees. This algorithm applies bootstrap aggregating, or bagging, to its tree learners. Compared to decision tree random forest tends to provide more accurate classification of a feature, because of the decreased bias and variance. The more decision trees are chosen the more computational power is required [9].

Reduced Error Pruning Tree (**REPTree**) is a fast decision tree learner, which creates multiple trees in different iterations and selects the best one from all created trees. REPTree builds regression tree information gain and prunes it using reduced-error pruning. For numeric attributes it sorts values only once [10].
SMOreg uses a support vector machine for regression. RegSMOImproved for SMOreg are used to learn parameters, but many other algorithms can be used, like Platt's SMO [11].

Multilayer perceptron is a class of feedforward artificial neural networks. It consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. It is by far the most popular architecture because of its structural flexibility, good representational capabilities, and the availability of a large number of training algorithms [12].

Feature selection is a key part of the applied machine learning process, just as model selection is. Feature selection should be considered as a part of the model selection process. If not, bias can inadvertently be introduced into models and it results in overfitting.

DOI number: 10.14706/JONSAE2019114

Feature selection must be included within the inner-loop when using accuracy estimation methods such as cross-validation. This means that feature selection is performed on the prepared fold right before the model is trained [7].

Dataset used in this work is small both in number of features and samples and it does not suffer from the "curse of dimensionality" [13] p.4. Feature selection and feature extraction methods are not recommended for this type of datasets with a small number of features [13] but extracting the information about which variables are most important, is important in this type of study. Choosing this particular approach is a type of rudimentary data mining.

Four attribute evaluators and two search methods combinations are used:

- CfsSubsetEval and BestFirst
- ClassifierAttributeEval and Ranker
- ClassifierSubsetEval and BestFirst
- CorrelationAttributeEval and Ranker

CfsSubsetEval creates subsets of attributes, where predictive ability of each feature and level of redundancy is considered. Features need to be highly correlated with class and low intercorrelation. Best first search method is used with CfsSubsetEval.

ClassifierAttributeEval evaluates the worth of an attribute by using a user-specified classifier. For example if we use linear regression on our dataset, linear regression needs to be chosen for the classifier attribute evaluator. Ranker search method is used with classifier attribute evaluators.

Classifiersubseteval evaluates attribute subsets on training data or a separate hold out testing set. Same as classifier attribute evaluator it uses classifier to estimate how good are subsets. Bestfirst search method is used with ClassifierSubsetEval.

CorrelationAttributeEval evaluates the worth of an attribute by measuring the correlation between it and the class. Each value of an attribute is treated as an indicator. Ranker method is used with CorrelationAttributeEval.

Best-first search method searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. Bestfirst may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions.

Ranker search method ranks attributes by their individual evaluations, where it is used with attribute evaluators.

Journal of Natural Sciences and Engineering, Vol. 1, (2019)
DOI number: 10.14706/JONSAE2019114

4. Results and Discussion

Classical statistical tools like probability distribution were used firstly in order to get the sense of the data. Table 2 represents the attribute subset evaluator used on random forests. Random forests with Classifier Subset Evaluator and Best First search method gave the best results for all the combinations. Best First search method is a heuristic or informed search; it evaluates the second step before taking the first. Then it chooses which way to go. For this combination of methods only attribute nr.2 (Surface Area), is not considered. Since the volume of the buildings is fixed it is logical that surface area has a little variation and therefore a little impact on the result.

Table 2. Results for combination of random forest and search methods

Random Forest						
Attribute Evaluator and Search Method	Correlation Coefficient	Mean Absolute Error	Root Mean Squared	Relative Absolute Error	Root Relative Squared Error	Selected Attributes
CfsSubSetEval and BestFirst	0.9711	1.4319	2.2692	16.6687 %	23.8241%	3, 5, 6, 7
Classifier AttributeEval and Ranker	0.9582	1.0079	1.6320	11.7324 %	17.1345%	1, 2, 3, 4, 5, 6, 7, 8
Classifier AttributeEval and Ranker	0.959	2.0323	2.6933	23.3581 %	28.2775%	1, 2, 4, 5
ClassifierSubsetEval and BestFirst	0.9854	0.9967	1.6196	11.6030 %	17.0046%	1, 3, 4, 5, 6, 7, 8
CorrelationAttributeEval and Ranker	0.9852	1.0079	1.6320	11.7324 %	17.1345%	1, 2, 3, 4, 5, 6, 7, 8
CorrelationAttributeEval and Ranker	0.9841	1.0859	1.6904	12.6408 %	17.7479%	5, 1, 3, 7

Relationship between the volume of a built form and the surface area of its enclosure is called compactness. Roundness is a similar feature.

R. Buckminster Fuller, engineer and an architect claimed that round houses have best energy efficiency, and an attempt to extract this feature has been made, but with no results.

Surface area, attribute nr.2, directly shows compactness of the building and by similarity, roundness. Classifier attribute evaluators removed this feature and gave the best correlation coefficient meaning that compactness has no impact on cooling load.

Usage of geometric compactness for such evaluative purposes is criticized on multiple grounds. It does not capture the specific morphology of the building shape, disregards transparent blocks of the structure and does not correlate with orientation att. nr. 6 [14].

High correlation coefficient with all attributes included, except for surface area finally pointed that compactness does not affect thermal load. Our model gave similar results using the same dataset as Tsanas and Xifara [2] with slightly better correlation coefficient which is shown in Table 3 for classifier attribute evaluator and ranker, in Table 4 for correlation attribute evaluator and ranker.

Table 3. Ranking of attributes according to attribute evaluator and ranker

ClassifierAttributeEval and Ranker		
Mathematical representation	Name	Ranked
x1	Relative compactnes	6.8134
x2	Surface area	6.8134
x4	Roof area	5.5105
x5	Overall height	5.2827
x3	Wall area	2.3935
x7	Glazing area	0.1718
x8	Glazing area distribution	0.0306

Table 4. Ranking attributes according to correlation attribute evaluator and ranker

CorrelationAttributeEval and Ranker		
Mathematical representation	Name	Ranked
x5	Overall height	0.8958
x1	Relative compactnes	0.6343
x3	Wall area	0.4271
x7	Glazing area	0.2075

x8	Glazing area distribution	0.0505
x6	Orientation	0.0143
x2	Surface area	-0.673
x4	Roof area	-0.8625

Further study is to be done with different variations of cross folds with above-mentioned algorithms. Results would be standing stronger if another dataset to test our algorithm was available. “K-nearest neighbor” algorithm gave poor results. It is a “data sensitive” algorithm, vulnerable when faced with large amounts of data. Different datasets would be a great boost to this work to test methods against them.

Parameter tuning is an iterative process, and Weka makes it easy to use it, without need to understand how parameters work. Especially, when dealing with feature selection, bias can be inadvertently introduced into models as it can give unforeseen consequences, mostly overfitting [7] [15].

Numerical values calculated by software simulations, lies very closely to previous results. Close values as compared to similar studies on the same dataset is a characteristic of the machine learning scientific field and using different methods and coming to the same results is an achievement [16].

6. Conclusion

Results of the previous study were repeated [17], and further work was done with examining cooling load resulting in slightly better correlation coefficient than in article with high scientific impact [2].

Trial and error are at the core of machine learning. Choosing right algorithms is a trade-off between speed, accuracy, and complexity. Starting with simple combinations and then adding complexity is the core of dealing with machine learning while constantly having in mind what type of data is dealt with.

Empirical study gives answers to what algorithm to use or what parameters to choose. Knowing beforehand what method will work best is almost impossible. Constantly iterating different combinations of similar methods with systematic workflow and using Weka is a way forward. New and easy accessible software packages makes it easier to spot and exploit new research areas, which previously were inaccessible due to low computing capability.

REFERENCES

[1] Y-T. Chen, “The Factors Affecting Electricity Consumption and Sector – A Case of Taiwan”, 2017.

[2] A. Tsanas, A. Xifara, “Accurate quantitative estimation of energy performance of residential building using
statistical machine learning tools”, Science Direct, 2012, p 9.

[3] A. Yezioro, “An applied artificial intelligence approach towards assessing building performance simulation tools”,
Energy and Buildings, 2007, p 40.

[4] T. Catalina, J. Virgone, “Cooling energy demand evaluation by means of regression models”. Proceedings of the Eleventh International Conference Enhanced Building Operations, New York City 2011, pp 6.

[5] D. Datta, S. A. Tassou, D. Marriot, “Application of Neural Networks for the Prediction of the Energy Consumption”, 1997.

[6] Mathworks, “Mastering Machine Learning: A Steb-by-Step Guide with MATLAB.” Available at: https://www.mathworks.com/campaigns/offers/mastering-machine-learning-with-matlab.confirmation.html?ab_test=b_version.

[7] J. Brownlee, “Machine Learning Mastery With Weka”, Wellington: Jason Brownlee 2019.

[8] X. Yan, X. Su, “Linear Regression Analysis: Theory and Computing”, World Scientific, 2009.

[9] D. Natingga, “Data Science Algorithms in a Week”, 2017.

[10] S. Kalmegh, “Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News.”, IJISET- International Journal of Innovative Science, Engineering and Technology, 2015, Vol. 2 Issue 2.

[11] S. K. Shevade, “Improvements to the SMO Algorithm for SVM Regression”, IEEE Transactions on Neural Networks, 2000, vol. 11, no. 5-6.

[12] P. Thomas, M. C. Suhner, “A new Multilayer Perceptron Pruning Algorithm for Classification and Regression Applications”, Neural Processing Letters, Springer Verlag, 2015, p 31.

[13] M. S. Raza, U. Qamar, “Understanding and Using Rough Set Based Feature Selection – Concepts, Techniques and Applications”, Springer, 2017.

[14] W. Pessenlehner, A. Mahdavi, “Building Morphology, Transparence and Energy Performance”, Eight International IBPSA Conference, Netherlands, Eindhoven, 2003.

[15] M. Kosinski, Y. Wang, “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”, Journal of Personality and Social Psychology, 2018.

[16] J. Christian, “Statistician: Machine Learning Is Causing A Crisis in Science”, Available: https://futurism.com/machine-learning-crisis-science.

[17] A. Bajek, A. Hasandić, “Energy Efficiency of the buildings.” Sarajevo: International Burch University 2017.