Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 110 (2017) 371 – 376

Energy Procedia

1st International Conference on Energy and Power, ICEP2016, 14-16 December 2016, RMIT University, Melbourne, Australia

Developing a hybrid model of prediction and classification algorithms for building energy consumption

Saeed Banihashemi[a]*, Grace Ding[a], Jack Wang[b]

[a]*School of the Built Environment, University of Technology Sydney (UTS), 15 Broadway, Ultimo 2007, Australia*
[b]*Faculty of Engineering and IT, University of Technology Sydney (UTS), 15 Broadway, Ultimo 2007, Australia*

Abstract

Artificial intelligence algorithms have been applied separately or integrally for prediction, classification or optimization of buildings energy consumption. However, there is a salient gap in the literature on the investigation of hybrid objective function development for energy optimization problems including qualitative and quantitative datasets in their constructs. To tackle with this challenge, this paper presents a hybrid objective function of machine learning algorithms in optimizing energy consumption of residential buildings through considering both continuous and discrete parameters of energy simultaneously. To do this, a comprehensive dataset including significant parameters of building envelop, building design layout and HVAC was established, Artificial Neural Network as a prediction and Decision Tree as a classification algorithm were employed via cross-training ensemble equation to create the hybrid function and the model was finally validated via the weighted average of the error decomposed for the performance. The developed model could effectively enhance the accuracy of the objective functions used in the building energy prediction and optimization problems. Furthermore, the results of this novel approach resolved the inclusion issue of both continuous and discrete parameters of energy in a unified objective function without threatening the integrity and consistency of the building energy datasets.

Peer-review under responsibility of the organizing committee of the 1st International Conference on Energy and Power.

Keywords: Energy optimisation; machine learning; prediction; classification; residential buildings.

1. Introduction

The potential to save energy by systematic building management and optimisation is known to be significant and

* Corresponding author. Tel.: +61-406-639380.
E-mail address: Seyed.S.BanihashemiNamini@student.uts.edu.au

Peer-review under responsibility of the organizing committee of the 1st International Conference on Energy and Power.
doi:10.1016/j.egypro.2017.03.155

it could be estimated from 5% to 50% [1]. With this respect, numerous machine learning methods have been applied in the recent decade for predicting, classifying and optimising the energy consumption of buildings focusing on the different important parameters but each optimisation algorithm requires specific objective function as its fundamental unit to model the parameters and minimise the values. In the energy optimisation context, developing the most appropriate objective function is a critical task as much as the reliable energy dataset generation particularly where both types of qualitative and quantitative data are involved with the optimisation solver. It can be inferred from literature that two types of machine learning algorithms; namely prediction [2] and classification [3] are run for continuous and discrete parameters of building energy consumption, respectively. However, there is a salient gap on the investigation of hybrid objective function development for energy optimization problems including qualitative and quantitative datasets in their constructs. Although in some cases, transformation techniques can be utilized for converting the continuous and discrete variables into each other [4] but some shortcomings such as losing integrity or randomness may be arisen [5]. Therefore, this study is to introduce a hybrid energy objective function covering both categorical and continuous data in the unified approach in which two well-known algorithms of Artificial Neural Network (ANN) and Decision Tree (DT) were developed integrally in order to find the best solution for energy optimisation functions and enhance the accuracy of data-driven energy modelling and prediction.

1.1. ANN & DT

ANN is a mathematical or computational model that tries to simulate the structure or functional aspects of biological neural networks. One of the applications of ANN in engineering field is to predict the outcome of non-linear statistical problems which is usually utilized to model complex relationships between inputs and outputs or to find patterns in datasets [6]. The thermal equations used to analyses and calculate energy loads are complex, making ANN a good platform to be used for this purpose. In this form, the network is presented with datasets obtained from simulations and the values of inputs are fed into each neuron or nod. The weights are then adjusted through learning algorithms iteratively until a suitable output is produced. A suitable output, in this case, suitable predicted annual energy load is the one which is as close as to the simulation results.

DT is also considered as the most applied type of machine learning algorithms in classification problems thanks to its wide use in practice [7]. The reputation of this algorithm is largely hinged to its interpretability and accuracy in delivering predictive models with understandable structure which generates useful information on the corresponding domain. In addition, DT is capable of processing both numerical and categorical parameters. However, this method is more appropriate and accurate in handling the categorical parameters rather than numerical data [3]. There are different types of DT algorithms including Simple Tree, Medium Tree, Complex Tree and Bagged Trees which follow the similar fundamental principles but different degrees of complexity in combining Trees. It applies a flowchart like tree structure to separate the dataset into different predetermined categories for presenting the interpretation, categorisation and generalisation on data [7]. With reference to the mentioned characteristics of ANN and DT, a comprehensive dataset should be collected for feeding the algorithms with inputs and outputs.

2. Dataset development

For data collection purposes, a four-story building consisting of four units on each floor was selected for simulation representing the conventional type of low-rise residential apartments. Each level area is 400 m^2 summing up to the total area of 1600 m^2. The building was modelled in Rhino 5, parameterised in Grasshopper software and simulated by EnergyPlus for annual energy estimation. For energy simulation, building calculation program was set to low-rise apartment and kitchen, bedroom, bathroom and dining room in each unit were defined as a zone which each zone had its own thermal properties. This approach enables the thermal engine to precisely quantify adjacencies and inter-zonal connections [8]. The thermostat was set between 18-26 ^oC to provide thermal comfort for occupants and activate HVAC devices below or above this range. Four cities of Sydney, Moscow, Kuala Lumpur and Phoenix were chosen as representatives of Temperate, Cold, Tropical and Hot-arid Climates, respectively. The procedures taken for the data collection were resulted in generating 4435 of datasets including 13 inputs (the variables) leading to the output (annual energy consumption) [9] consisting of 1053, 1138, 1114 and 1130 data for the cities of Sydney, Phoenix, Kuala

Lumpur and Moscow, respectively. To obtain an overview about the generated dataset, descriptive statistics were performed using SPSS on the measures of central tendency and dispersion. Different statistical measurements of minimum, maximum, range, mean, median and standard deviation were computed to specify the probability distribution and the dispersion of the data. Table 1 indicates that considering both continuous and categorical data, the generated output covers a wide range of distribution which is of advantage as to enabling an accurate energy objective function development.

Table 1. Descriptive statistics of the developed dataset (*categorical parameters)

Parameters	Range	Minimum	Maximum	Mean	Std. Deviation	Median
Wall*	5	NA	NA	NA	1.513	NA
Insulation*	6	NA	NA	NA	1.860	NA
Roofing material*	1	NA	NA	NA	0.468	NA
Windows glazing*	1	NA	NA	NA	0.480	NA
Floor ground system*	1	NA	NA	NA	0.464	NA
Type of main space heating*	2	NA	NA	NA	0.62	NA
Type of main space cooling*	1	NA	NA	NA	0.495	NA
Building orientation	315	0	315	102.98	87.023	90.00
Window to wall ratio	0.2	0.2	0.4	0.25	0.088	0.20
Ceiling height (m)	1	3	4	3.14	0.225	3.00
Meter square of rooms heated	5	10	15	11.77	2.391	10.00
Meter square of rooms cooled	5	10	15	12.39	2.498	10.00
Lighting (Lux)	40	0	40	2.39	2.581	2.00
Energy load	10535.58	309.88	10845.45	2725.16	2015.30	2208.66

3. Hybrid objective function development

The procedure of hybrid objective function development was started with running DT and ANN on the discrete and continuous inputs along with the energy consumption output and then followed with averaging the output of these two separately trained models in a hybrid function [10].

$$f(\vec{x}) = \sum_k w_k f_k \vec{x} \tag{1}$$

Where k is the index for each algorithm and w denotes the weight of $f(x)$ for each objective function. Based on the weighted average output of the hybrid function (Equation 1), the final result of each function should be homogenous. On one hand, results from DT algorithm which leads to the classified level of energy consumption of buildings such as low, medium, high and excessive could not be arithmetically averaged with the numerical and continuous results of ANN model. On the other hand, using transformation techniques may lead to the randomness or losing integrity. Additionally, with respect to the energy optimization purposes, optimisation algorithms generally work better with numerical outputs and it is of high priority to obtain the actual-numerical level of energy consumption from the hybrid objective function. Therefore, C4.5 algorithm, according to the concept of entropy, was used to construct the Bagged DT including the seven discrete variables of wall, roof and insulation materials, floor ground system, glazing type and the types of main space heating and cooling systems as the inputs and annual energy consumption as the output. Taking advantage of the flexibility of C4.5 in handling numerical outputs, it was intended to apply standard deviation reduction instead of Information-Gain in calculating the homogeneity of the splits of DT [11]. This approach allows for obtaining the outputs applicable for being averaged with the arithmetic results of ANN in the hybrid function, in the meanwhile of maintaining the original form of DT algorithm. In this method, standard deviation is zero if the splits of DTs are completely homogenous. In fact, the decrease in standard deviation, after a dataset is split on an attribute, shapes the process of DT construction (Equation 2).

$$SP(X) = \sum_{D \in X} A(D) S(D) \tag{2}$$

Where $SP(X)$ indicates the split of data based on the attribute (A) of standard deviation of D. Likewise, the attribute

with the largest standard deviation reduction is selected as the split attribute for each tree node. Considering the default configuration of DTs [11], the algorithm was set to train 100 complex trees and the test errors and cross-validation errors were computed. It was observed that the Bagged Tree performs well in the cross-validation state as compared to the testing condition. However, the lowest error was recorded at 1.6 from the tenth complex tree, without a significant fluctuation toward the end of the training trees. Applying the standard deviation in the structure of the Bagged Tree opens a new window of opportunity in using ensemble regularization technique which is a process of removing weak learners from the DT structure and improves the performance in a way that fewer number of trees are required to train the algorithm. Therefore, this feature, as a significant achievement in the hybrid function, was run to decrease the time of the training and increase the speed of objective function. The regularization procedure specifies a well-trained learner weights that could minimize the errors in the below Equation:

$$\mathrm{Re}\,g(DT)=\sum_{n=1}^{N}W_n g\left(\left(\sum_{t=1}^{T}\alpha_t h_t(x_n)\right),y_n\right)+\lambda\sum_{t=1}^{T}|\alpha_t| \tag{3}$$

Where α_t is the optimal set of learner weights, $\lambda \geq 0$ is the lasso parameter, h_t is a weak learner in the ensemble trained on N observations with predictors of x_n, responses of y_n, and weights of w_n and g (f,y) = $(f - y)^2$ is the squared error. The minimised α_t could be achieved with minimising Mean Squared Error (MSE) of the above equation. Usually, an optimal range could be found in which the accuracy of the regularized ensemble is better or comparable to that of the full ensemble without regularization. In this process, if a learner weight; α_t is calculated to be 0, this learner is excluded from the regularized ensemble. In the end, an ensemble with improved accuracy and fewer learners is obtained. As a result of this procedure, the reduced Bagged Tree contained 15 complex trees in its structure along with generating approximately 0.8 of cross-validated MSE (Figure 1). This reduced ensemble gives low loss while using many fewer trees.

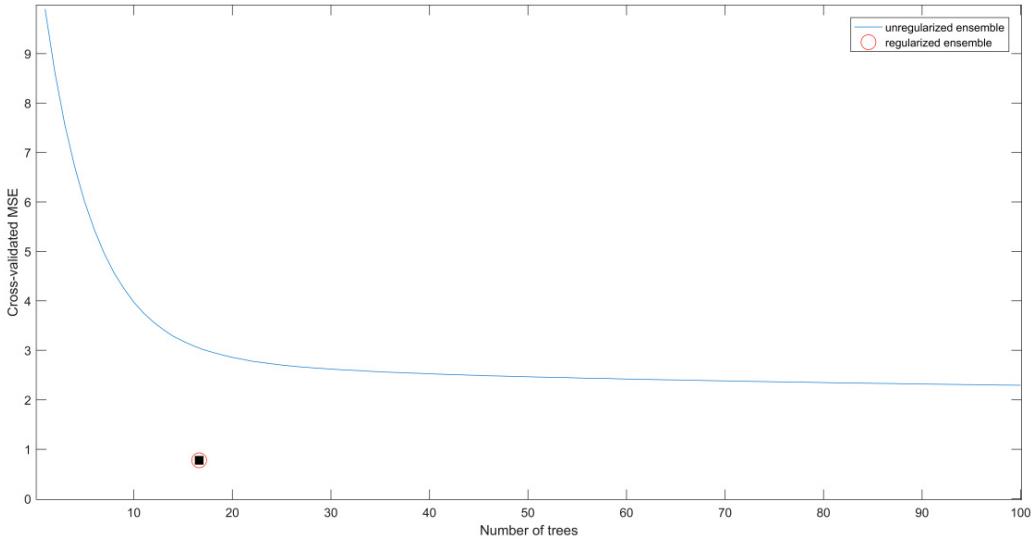

Fig. 1. Regularized vs. un-regularized ensemble in the hybrid model

Similar to DT model development for qualitative data (categorical parameters), the ANN model was also developed upon quantitative data; six continuous parameters of building orientation, window to wall ratio, ceiling height, meter square of rooms heated and cooled and lighting based on the optimum configurations of ANN, identified from the trial and error process. This ANN was tested with different combinations of the number of neurons in the input, hidden and output layers, as recommended by Shahidepour et al. [12] and the model with 6, 7 and 1 neurons in the input, hidden and output layers and comprising of 70%, 15% and 15% of data for training, testing and validating, was respectively structured. Hyperbolic tangent function was the activation function chosen for the input layer, sigmoid

transfer function was applied between the hidden layer and the output layer and the Levenberg-Marquardt Back propagation algorithm was set as the learning algorithm. The model was fixed to 1,000 iterations and the best validation performance was recorded at the 109th iteration with MSE of 0.40974 (Figure 2). As a result, this ANN model was trained up to 142 epochs and stopped on the rule of 6 consecutive runs without any decrease in the performance error.

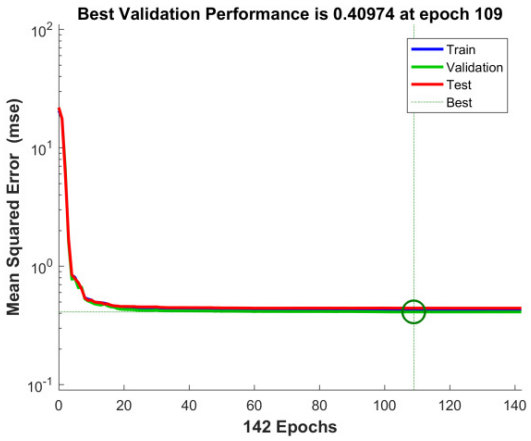

Fig. 2. Validation performance of ANN in the hybrid model

Ultimately, referring to the Equation (1), the hybrid model was composed from the ensemble of the Bagged DT and ANN model covering both continuous and discrete parameters in one objective function (Figure 3). Since throughout algorithms generation, MSE was considered as the main accuracy driver, this criterion was employed in demonstrating the improved accuracy of the hybrid model too. The MSE of the hybrid model was computed approximately at 0.6 which is very low in the error rate. In order for validating the improved performance of the hybrid model against single objective models, first, the normalized values of the predictions of single ANN were figured out upon the whole dataset. Second, the performance error of single DT on the whole dataset was plotted and finally, the associated results along with the performance of hybrid models were illustrated vis-à-vis the normalized actual outputs of energy consumption indicating the comparative performance of each model. As shown in Figure 4, the approximate linear trend-line of the normalized values predicted by the hybrid model is more match with the equality state in comparison with that of single ANN and DT models. This observation confirms the superior performance of the hybrid model in generating the predictive data as close as to the baseline data and provides more robust objective function.

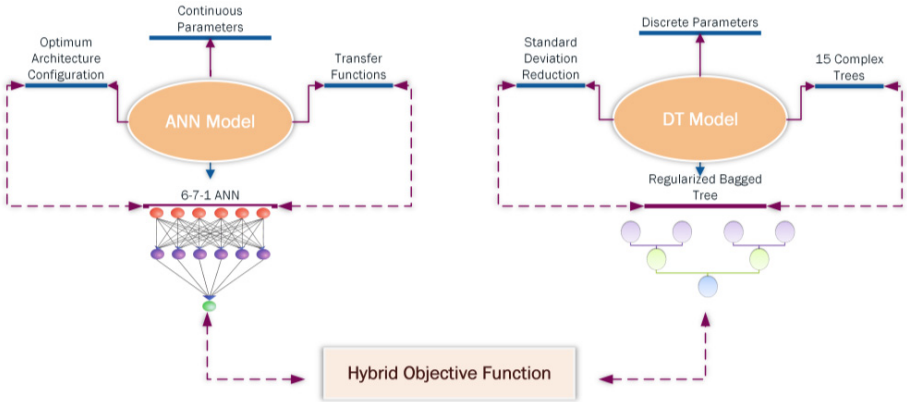

Fig. 3. The hybrid model structure

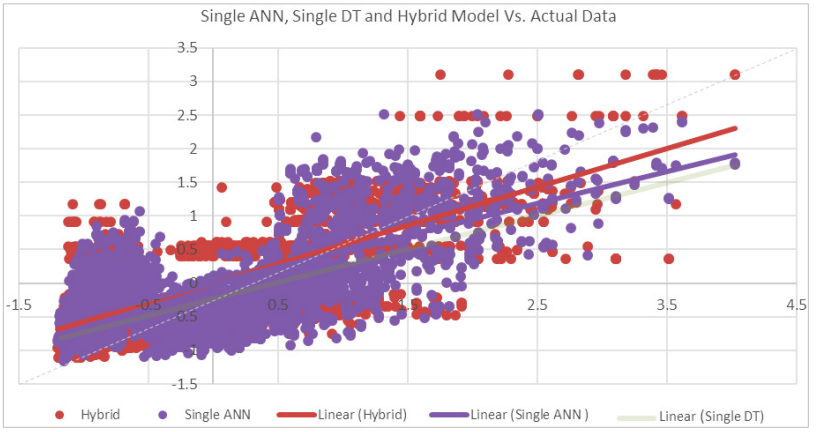

Fig. 4. Normalized predictive performance of single ANN, DT and hybrid model vs. normalized actual energy data

4. Conclusion

Driven by the gap in the body of the knowledge with regard to including discrete and continuous parameters in a homogenous objective function model, this study contributes to the field in different ways. It targets the hybrid objective function development for generating the predictive energy consumption data with the least error and the highest accuracy which paves the way of presenting a powerful engine for building energy optimisation. The outcome is an integrated platform containing both qualitative and quantitative variables of building energy consumption without affecting data consistency or requiring any data transformation procedures. The study also goes beyond the existing literature by revealing how a DT algorithm could be modified by replacing the information-gain concept with the standard deviation reduction in its structure and made ensemble with ANN algorithm. This achievement develops more error-free bagged DT by enabling the regularisation technique through removing weak learners and increasing the speed of hybrid objective function. However, the study findings should be considered with caution due to a number of limitations in conducting the present study. That is, the findings may not be directly applicable to other types of machine learning algorithms in prediction and classification as the different hybrid model may bring different attributes to attention. Moreover, the data collection was conducted considering 13 parameters as the input and four climates of temperate, tropical, cold and hot-arid. This calls for further investigation by validating the model in other contexts and using larger samples covering various building energy parameters and climates.

References

[1] M.M. Tahmasebi, S. Banihashemi, M.S. Hassanabadi, Assessment of the variation impacts of window on energy consumption and carbon footprint. Proc Eng 2011; 21:820-828.
[2] M. Shakouri Hassanabadi, S. Banihashemi, Developing an empirical predictive energy-rating model for windows by using Artificial Neural Network. Int J Green Energy 2012.
[3] Z. Yu, F. Haghighat, B.C. Fung, H. Yoshino, A decision tree method for building energy demand modeling. Energy&Bldgs 2010; 42:1637-1646.
[4] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Ren&Sustainable Energy Rev 2013; 23:272-288.
[5] C. Blum, A. Roli, Metaheuristics in combinatorial optimization: Overview and conceptual comparison, ACM Cmpt Surveys 2003; 35:268-308.
[6] J.A. FLORES. Focus on artificial neural network. New York: Nova Science Publishers; 2011.
[7] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: A review of classification techniques. 2007.
[8] S. Banihashemi, H. Golizadeh, M.R. Hosseini, M. Shakouri, Climatic, parametric and non-parametric analysis of energy performance of double-glazed windows in different climates, Int J Sustainable Built Env 2015; 4:307-322.
[9] S. Banihashemi, G. Ding, J. Wang, Identification of BIM-compatible variables for energy optimization of buildings-a delphi study, 40th AUBEA 2016, Cairns, Queensland, Australia. 2016: 281-291.
[10] C. Merkwirth, J. Wichard, M. Ogorzałek. ENTOOL-A Matlab Toolbox for Regression, Classification and Active Learning. 2007.
[11] J.R. Quinlan. C4. 5: programs for machine learning. Elsevier; 2014.
[12] M. Shahidehpour, H. Yamin, Z. li. Market Operation in Electric Power Systems. New York: Wiley-IEEE Press; 2002.