Residential buildings conceptual cost estimates with the use of support vector regression

Michał Juszczyk^{1,*}

¹Cracow University of Technology, Faculty of Civil Engineering, Warszawska 24, 31-155 Kraków

Abstract. Cost analyses, and the conceptual cost estimates among them, are of the key importance for the construction projects successes. Implementation of neural networks or machine learning methods provides broad possibilities for this specific type of cost. The aim of the paper is to present some results of the studies on the use of support vector regression as a machine learning tool for conceptual cost estimates of residential buildings. Results for three models based on support vector regression and radial basis kernel functions are introduced.

1 Introduction

Cost estimation is one of the key processes in the course of a construction project. Cost estimates serve for parties involved in a project as a basis for various decisions. Conceptual cost estimates, prepared in the early design stage are specific due to the two facts. Firstly the estimates have to rely on imprecise, incomplete information about the building. Secondly the ability to influence building's characteristics and cost are the greater in the early design stage, than later in the course of the project. Both of the aforementioned stimulate the research on development models, that are based on various approaches and methods, capable of supporting conceptual cost estimates.

The aim of this paper is to present the results of investigations on applicability of support vector regression (SVR) method for the purposes of residential buildings conceptual cost estimates. The paper is a contribution to the author's broader research on the development of cost estimation models based on artificial intelligence and machine learning methods, (some exemplary earlier works can be mentioned here [1-3]).

Support vector regression (SVR) originates in support vector machines (SVM) as a machine learning method. The theory and fundamentals of SVM can be found in [4-6]. Applications of SVM are reported for variety of problems, also those that belong to the field of construction management, such as: development of evolutionary support inference system for construction management [7], automating document classification [8], supporting legal decisions [9], risk prediction [10], estimates of time and cost at completion for construction projects [11], predicting project success [12-13], cost estimates of school buildings [14].

* Corresponding author: mjuszczyk@izwbit.pk.edu.pl

The paper content includes: assumptions made for the application of SVR for the conceptual cost estimates for residential buildings as well as some theoretical background, presentation of results, brief discussion of results and conclusions in the summary section.

2 Assumptions for implementing support vector regression in conceptual cost estimates of residential buildings

The aim of the study was to investigate applicability of SVR in the problem of conceptual estimating of residential and residential-commercial buildings' cost. The model based on SVR was developed under the assumptions that follow.

The dependent variable of the model, that is cost of a building, was represented by y . The cost was considered in respect of construction works including: site preparation, shell and core, internal and external finishing, installations and services as well as all the primary equipment. (The cost of land improvement and landscaping were not taken into account.)

Independent variables included 13 features (cost descriptors), related to the information about buildings available in the early stage of design, adequately for the conceptual cost estimates. The independent variables formed a vector, later referred to as x . (Analysis and selection of the independent variables is presented in previous work [17]).

In general the problem can be defined as regression analysis which aim is to find mapping $x \rightarrow y$. Table 1 presents the variables taken into account for the purpose of model development.

Table 1. Font styles for a reference.

Symbol	Description	Type and value
y	the cost of a building	numerical (net cost – excluding value added tax, discounted for a base year) [mln PLN]
x_1	footprint	numerical (building's footprint area - m^2)
x_2	cubature	numerical (building's cubic capacity - m^3)
x_3	number of storeys	numerical (number of building's storeys)
x_4	type of foundation	descriptive (RC continuous footing, RC foundation slab, special foundation)
x_5	building structure	descriptive (traditional, RC skeletal structure, RC column-wall structure, RC solid-wall structure, RC monolith structure)
x_6	roof structure	descriptive (RC - flat roof, wooden structure - steep roof)
x_7	number of structural segments of a building	numerical (number of segments)
x_8	number of elevator shafts	numerical (number of shafts)
x_9	ground conditions	descriptive (simple, complex, complicated) [†]
x_{10}	usable area of dwellings	numerical (total usable area of dwellings - m^2)
x_{11}	usable area of commercial premises	numerical (total usable area of commercial premises - m^2)
x_{12}	usable area of underground garages	numerical (total usable area of underground garages - m^2)
x_{13}	building finishing standard	descriptive (developer's standard, turn-key standard)

[†] values adopted according to law regulations in Poland

According to the scope of the research the mapping $\mathbf{x} \rightarrow y$ was supposed to be approximated by a regression function f :

$$y = f(\mathbf{x}) \quad (1)$$

Function f was supposed to be found with the use of SVR method. (The theoretical background of the SVR method, given in the forthcoming part of the paper, is compiled on the basis of [4-6], [15-16]). The SVR machine learning was applied to approximate the regression function as a linear hyperplane, after mapping input examples into a higher dimensional feature space:

$$\Phi: X \rightarrow H, \Phi(\mathbf{x}) \in H \quad (2)$$

The mapping $\Phi(\mathbf{x})$ is supposed to increase the expressive power of the representation and allows to compute the approximation function in the mapped space H . Under the aforementioned assumptions the sought-for linear hyperplane where $\mathbf{x}$ is replaced by $\Phi(\mathbf{x})$ can be given as:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + w_0 \quad (3)$$

The use of ε -insensitive loss-function is assumed to measure the errors in the course of the training process:

$$l(f(\mathbf{x}), y) = |y - f(\mathbf{x})|_\varepsilon \quad (4)$$

where:

$$|y - f(\mathbf{x})|_\varepsilon = 0 \quad \text{for} \quad |y - f(\mathbf{x})| \leq \varepsilon \quad (5)$$

$$|y - f(\mathbf{x})|_\varepsilon = |y - f(\mathbf{x})| - \varepsilon \quad \text{for} \quad |y - f(\mathbf{x})| > \varepsilon \quad (6)$$

In the given loss-function ε defines a tube of insensitiveness around the true values y , which results in toleration of deviations smaller than ε and allows to trade-off function complexity. The optimization problem to be solved:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\zeta_i - \zeta_i^*) \rightarrow \min \quad (7)$$

subject to the constraints for the both sides of ε -tube:

$$\mathbf{w}^T \Phi(x_i) + w_0 - y_i \leq \varepsilon + \zeta_i \quad (8)$$

$$y_i - (\mathbf{w}^T \Phi(x_i) + w_0) \leq \varepsilon + \zeta_i^* \quad (9)$$

$$\zeta_i, \zeta_i^* \geq 0 \quad (10)$$

where C represents complexity of a model, ζ_i and ζ_i^* are slack variables that penalize predictions out of the ε -tube.

The optimization problem can be solved by introducing Lagrange multipliers, which leads to the approximation function in the form:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + w_0 = \sum_i (\alpha_i - \alpha_i^*) \Phi(x_i)^T \Phi(x_j) + w_0 \quad (11)$$

where α_i, α_i^* are the Langrange multipliers for the optimal solution. Explicit calculation of the dot products $\Phi(x_i)^T \Phi(x_j)$ as well as the choice of an appropriate mapping are difficult and computationally complexed. Instead, the dot products are calculated with the use of kernel functions:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{12}$$

In turn the approximation function can be written finally as:

$$f(\mathbf{x}) = \sum_i (a_i - a_i^*) K(x_i, x_j) + w_0 \tag{13}$$

where $K(x_i, x_j)$ is the chosen kernel function.

Before the start of machine learning all the values of models' variables were scaled to the range $<0 ; 1>$. (In case of numerical values linear scaling was applied, whereas the descriptive values were transformed with use of pseudo-fuzzy scaling.) For the purposes of machine learning the whole set of training data, including 151 samples, was divided into the learning subset and testing subset in that took part in relation 70% to 30% respectively.

As the kernel function $K(x_i, x_j)$, radial basis function was chosen:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|) \tag{14}$$

Parameters of the model were selected in respect of 10-fold cross validation. The mesh of the parameters C and ε that were investigated was as follows: C varied between 5 and 30, ε varied between 0,1 and 0,3. The cross validation allowed to select optimal nod of the mesh. The process of SVR based machine learning was repeated several times for different variants of learning and testing subsets sampling. All of the computations were made with the use of STATISTICA® software.

3 Results of the study

Table 2 presents characteristics of three models selected from the set of investigated models in terms of the number of support vectors and bounded vectors, as well as parameters C and ε . Additionally table 1 includes correlation coefficients between the real life values and predicted values of the dependant variable y .

Table 3 includes measures of cost prediction errors computed for the models A, B and C. The measures include mean squared error (MSE) and mean absolute percentage errors ($MAPE$).

(To be distinguished, the three models are given letters A, B and C respectively. In case of the values present in table 2 and table 3 the subscripts indicate that the values were calculated for learning subset - subscript L, testing subset - subscript T or learning and testing subsets together - subscript L&T.)

Table 2. Characteristics of selected models.

Model	Number of support vectors	Number of bounded support vectors	C	ε	R_L	R_T	$R_{L\&T}$
A	22	3	28	0,1	0,992	0,992	0,992
B	23	1	30	0,1	0,992	0,991	0,992
C	20	2	27	0,1	0,993	0,992	0,993

Both number of support vectors and bounded support vectors do not differ significantly between models A, B and C. The complexity of models also varied very little – C ranges between 27 and 30, whereas the value of ε is the same the three models. Correlation coefficients R are almost the same in case of the three selected models.

Table 3. Cost prediction errors computed for the selected models.

Model	MSE_L	MSE_T	$MSE_{L\&T}$	$MAPE_L$	$MAPE_T$	$MAPE_{L\&T}$
A	3,434	3,452	3,439	10,01%	8,19%	9,49%
B	3,169	3,863	3,385	9,50%	7,00%	9,00%
C	3,042	2,889	2,996	9,41%	7,89%	8,87%

The overall measures of models errors, as presented in the table 3, do not vary between the models A, B and C considerably. However one can see that for model A MSE values computed for learning and testing were almost the same. In case of model B MSE was smaller for learning subset, in case of model C MSE was smaller for testing subset. For all models $MAPE$ values were smaller for testing subsets.

Figure 1 depicts scatter plot of the real life values of the construction cost of residential buildings (introduced as y) and the costs predicted by the models (introduced as $\hat{y}$) for the samples belonging to the testing subset. Figure 2 depicts scatter plot of the real life values of the construction cost of residential buildings (introduced as y) and corresponding absolute percentage errors (introduced as APE) calculated for the samples belonging to the testing subset.

Fig. 1. Scatter plot of the real life values of cost and cost predictions made by the models A, B ad C for the testing subset.

In the figure 1 one can see that the points are distributed along the line $y = \hat{y}$ that indicates perfect fit. Points are distributed evenly on both sides of the line, so the cost predictions made with the use of the model for the testing subset belong either to overestimates or underestimates.

In the figure 2 the distribution of APE shows that the predictions made with the use of models are more reliable for residential buildings for which the real life values y pertain to the range 30 – 55 million PLN (APE values in the range 0 to 10% for the testing subset). Less reliable predictions can be expected for the buildings that real life construction cost range between 20 – 30 million PLN (APE values in the range 0 to 15% for the testing subset). Finally the least reliable predictions can be expected in case of buildings for which real life construction cost belong to the range 5 – 20 million PLN (most of APE values in the range 0 to 25%, some in the range 25 to 35%, for the testing subset).

Fig. 2. Scatter plot of the real life values of cost and absolute percentage errors of cost predictions made by the models A, B and C for the testing subset.

4 Summary

The research investigated the use of support vector regression based machine learning for the purposes of conceptual cost estimates of residential buildings. The paper presents results for the three models selected from a number of examined ones. In case of all of the selected models radial basis kernel function was applied, parameters C and ε have been found with the use of cross validation.

Parameters of the selected models were similar. Correlation coefficients between the real life values and models' predictions were almost the same for all models. The overall errors' measures (MSE as well as $MAPE$) do not differ significantly. Analysis of the APE errors allowed to specify the reliability of models' predictions for the clusters of buildings, distinguished in respect of the construction cost.

In the light of the results' analyses quality of the developed models can be considered satisfactory. Obtained results justify the statement that application of support vector machine learning approach can bring benefits for the conceptual cost estimates.

References

1. M. Juszczuk, *Procedia Engineering*, **196**, 415-422, (2017)
2. A. Leśniak, M. Juszczuk, *Archives of Civil and Mechanical Engineering*, **18(3)**, 973-982, (2018)
3. M. Juszczuk, A. Leśniak, K. Zima, *Complexity*, **2018**, Article ID 7952434, (2018)
4. V. Vapnik, *Statistical Learning Theory*, (John Wiley & Sons, New York, 1998)
5. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 2013)
6. A. J. Smola, B. Schölkopf, *Statistics and computing*, **14.3**, pp 199-222, (2004)
7. M.Y. Cheng, Y.W. Wu, *Automation in Construction*, **18(5)**, pp 597-604, (2009)
8. C.H. Caldas, L. Soibelman, *Automation in Construction*, **12(4)**, pp 395-406 (2004)

9. T. Mahfouz, A. Kandil, Construction Research Congress 2010: Innovation for Reshaping Construction Practice, pp. 879-888 (2010)
10. J.H. Chen, J.Z. Lin, Automation in construction, **19(6)**, pp 702-708 (2010)
11. M.Y. Cheng, H.S. Peng, Y.W. Wu, T.L. Chen, Automation in Construction, **19(5)**, pp 619-629, (2010)
12. M.Y. Cheng, Y.W. Wu, C.F. Wu, Automation in Construction, **19(3)**, pp 302-307, (2010)
13. Y.R. Wang, Y.C. Yu, H. H. Chan, International Journal of Project Management, **30(4)**, pp 470-478, (2002)
14. G.-H. Kim, J.-M. Shin, S. Kim, Y. Shin, Journal of Building Construction and Planning Research, **2013.1**, pp. 1-7 (2013)
15. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-based Learning Methods)*, (Cambridge University Press, Cambridge, 2000)
16. S.R. Gunn, *Support Vector Machines for Classification and Regression* (Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997)
17. M. Juszczak, A. Leśniak, K. Zima, Selected Scientific Papers-Journal of Civil Engineering, **8(1)**, pp 25-34, (2013)