



**HAL**  
open science

## Feature selection for support vector regression in the application of building energy prediction

H. X. Zhao, F. Magoules

► **To cite this version:**

H. X. Zhao, F. Magoules. Feature selection for support vector regression in the application of building energy prediction. 9th IEEE International Symposium on Applied Machine Intelligence and Informatics (SAMI 2011), Jan 2011, Smolenice, Slovakia. hal-00617937

**HAL Id: hal-00617937**

**<https://hal.science/hal-00617937v1>**

Submitted on 31 Aug 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Feature Selection for Support Vector Regression in the Application of Building Energy Prediction

Hai-xiang ZHAO, Frédéric MAGOULÈS

Applied Mathematics and Systems Laboratory, Ecole Centrale Paris  
Châtenay Malabry, France

Email: haixiang.zhao@ecp.fr, frederic.magoules@hotmail.com

**Abstract**—When using support vector regression to predict building energy consumption, since the energy influence factors are quite abundant and complex, the features associated with the statistical model could be in large quantity. This paper focuses in feature selection for the purpose of reducing model complexity without sacrificing performance. The optimal features are selected by their feasibility of obtaining and the evaluation of two filter methods. We test the selected subset on three datasets and train support vector regression with two different kernels: radial basis function and polynomial function. Extensive experiments show that the proposed method can select valid feature subset which guarantees the model accuracy and reduces the computational time.

**Keywords**—support vector regression; feature selection; building; energy consumption

## I. INTRODUCTION

Support vector regression (SVR) has shown remarkable generalization abilities in the prediction of energy load [1–3]. In our application, the statistic model is trained from the historical consumption data and then the derived model is used to predict the energy behaviors of one building in the future or for some new buildings. The energy behavior of buildings is influenced by many factors, such as ambient weather conditions, building’s structures, occupants’ behaviors, inner facilities, etc, therefore, the possible features of the samples could be in large quantity. However, not all of the features that are important to the statistic model. How to reasonably choose the useful features to be used in model learning is one of the key issues for machine learning methods on such applications.

To the best of our knowledge, there is little work concerning the feature selection for machine learning methods in the application of predicting building energy consumptions. Most of the existing works derive their own model using previously established sets of features. Azadeh et al. [4] and Maia et al. [5] forecast electrical energy consumption through analyzing the varying inner targets without any contributory variable involved. Neto et al. [6] built their neural network based on the input of daily average values of dry bulb temperature, relative humidity, global solar radiation and diffuse solar radiation. Tso et al. [7] used more than 15 features in their assessment of traditional regression analysis, decision tree and neural networks. Yokoyama et

al. [8] considered only two features, air temperature and relative humidity in their neural network model. Other similar approaches can be found in [1, 3, 9, 10].

This paper proposes an heuristic approach for selecting subset of features, and systematically analyzes how it will influence the model performance. The models are trained by support vector regression (SVR) with two different kernel methods based on three datasets. The feature selection method is evaluated by comparing the models’ performances before and after feature selection is performed. Next, section II briefly introduces the principles of SVR model and feature selection algorithms, section III will introduce the extensive experiments and evaluate the proposed feature selection method, and finally, section IV concludes this paper.

## II. SVR AND FEATURE SELECTION

### A. The Principles of SVR

The basic idea of SVR is to find a model function  $f(x)$  to represent the relationship between the features and the target [11]. We represent the whole training data as  $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$ , where vector  $x_i$  is the  $i$ th sample of the features,  $y_i$  is the corresponding target,  $l$  is the number of samples. If there are  $n$  features in the training data, then the dimensionality of  $x_i$  is  $n$ . The model function has the form  $f(x) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K(x_i \cdot x) + b$ , where  $K$  is kernel function which implicitly maps the problem from lower dimensional space to higher feature space,  $\alpha_i^*$  and  $\alpha_i$  are variables which are optimized by maximizing the following quadratic function:

$$W(\alpha_i^*, \alpha_i) = -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i \cdot x_j) + \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) - \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i)$$

under the constraints  $\sum_{i=1}^l \alpha_i^* = \sum_{i=1}^l \alpha_i$  and  $0 \leq \alpha_i^*, \alpha_i \leq C$ ,  $i = 1, 2, \dots, l$ . Then the problem can be solved by many off-the-shelf quadratic problem solvers. Usually, only certain parts of the samples can satisfy the property:  $\alpha_i^* - \alpha_i \neq 0$ , they are called support vectors (SVs) since only them have contributions to the model. The popular choices of the kernel

are linear function, sigmoid function, radial basis function (RBF) and polynomial function. They represent different mappings from lower space to higher feature space. In our data training process, we choose the last two kernels for the purpose of solving non-linear problems.

### B. Algorithms of Feature Selection

The aim of feature selection is to select the most useful feature set in order to establish a good predictor for the concerned learning algorithm. The irrelevant and unimportant features are discarded in order to reduce the dimensionality. Several advantages will be achieved if we wisely perform the selection. The first one is to simplify the calculation while keeping the dimensionality minimized, which could contribute to avoid the curse of dimensionality. The second one is the possible improvement of accuracy of the developed model. The third one is the improved interpretability of the models. The last one is the feasibility of obtaining accurate feature samples, especially for some time series problems in practice.

In our practice, two feature selection methods are used to rank each feature, then the features are filtered according to their ranks. The first one gives each feature a score depending on the correlation coefficient between this feature and the target. We use CC to stand for this method. It is defined as:

$$CC(f) = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}}$$

The other one is called Regression, Gradient guided feature selection (RGS), which is firstly proposed by A. Navot et al. when they use regression to analyze brain neural activities [12]. The basic idea is to assign a weight to each feature and then optimize the weight vector of all the features simultaneously by gradient ascent. The non-linear function K-Nearest-Neighbor is applied in the procedure of optimization. The estimated target of sample  $x$  is defined as:

$$\hat{f}_w(x) = \frac{1}{Z} \sum_{x' \in N(x)} f(x') e^{-d(x,x')/\beta}$$

where  $N(x)$  is the set of  $K$  nearest neighbors of sample  $x$ .  $d(x, x') = \sum_{i=1}^n (x'_i - x_i)^2 w_i^2$  is the distance between sample  $x$  and one of its nearest neighbors  $x'$ ,  $n$  is the number of features,  $w_i$  is the specific weight assigned to  $i$ th feature and  $w$  is the weight vector.  $Z = \sum_{x' \in N(x)} e^{-d(x',x)/\beta}$  is a normalization factor where  $\beta$  is a Gaussian decay factor. Then the optimal  $w$  can be found by maximize the following evaluation function:

$$e(w) = -\frac{1}{2} \sum_{x \in S} (f(x) - \hat{f}_w(x))^2$$

where  $S$  is the samples for model training. Since  $e(w)$  is smooth almost everywhere in a continuous domain, one can solve the extremum seeking problem by gradient ascent.

## III. NUMERICAL EXPERIMENTS

### A. Preparing Datasets of Building Energy Consumption

Hourly electricity consumption together with hourly values of several influencing factors is recorded as the building energy consumption data. This data is generated by *EnergyPlus* which is a widely applied, state-of-the-art building energy simulation tool [13]. The test buildings are single-story mass-built buildings for office use located in five major cities of France, which are Paris-Orly, Marseilles, Strasbourg, Bordeaux and Lyon. They represent the office energy requirements under five typical weather conditions in France. All of the buildings are located in urban area. Each building has similar structures, e.g., one rectangle room with attic roof and four windows in four directions without shading. Electrical equipments including lighting system, fans, water heaters, are scheduled as common office use. The number of occupants depends on the housing space, with the average of 0.2 people per zone floor area. Since the buildings are designed for office use, the energy consumption in holidays (including weekends) is quite small compared to weekdays. In order to simplify the model in practice, we only record the consumption data in weekdays.

Three datasets are generated for model training and testing. The first dataset is one year's consumption of one building which is located in Pairs-Orly. The hourly electricity demands together with hourly behaviors of 23 variables are recorded through one year. The consumption series from January 1st to October 31st is used for training and the samples in remaining two months, from November 1st to December 31st, is used for model testing. In other two datasets, we record the consumption for multiple buildings in winter season which means from November 1st to March 31st. We design 20 and 50 buildings to be the training data for the second and the third dataset respectively. For model testing purpose, we simulate one more building. The differences among these multiple buildings mainly rely on weather conditions, building structures and number of occupants. The buildings have diverse characteristics with randomly generated parameters: length, width, height and window/wall area ratio. The features of one building are listed in Table I.

### B. Model Selection

To fully investigate how feature selection on these three datasets influence SVR models, two kernels are involved. The first one is RBF kernel:  $\exp(-\gamma \|x_i - x_j\|^2)$ . Three parameters of the learning algorithm,  $C$ ,  $p$  and  $\gamma$  are chosen by stepwise 5-fold cross validation as described in the technical document of Libsvm [14]. The initial searching spaces are  $\{2^{-3}, 2^{-2}, \dots, 2^8\}$ ,  $\{2^{-10}, 2^{-9}, \dots, 2^{-5}\}$  and  $\{2^{-10}, 2^{-9}, \dots, 2^2\}$  for  $C$ ,  $p$  and  $\gamma$  respectively. In order to avoid numerical problems in the calculations, before training the model, we scale the training set into the range  $[0, 1]$  and then apply the scaling function to scale the testing set.

We also test the feasibility of polynomial kernel ( $\gamma * x'_i * x_j + coef$ )<sup>d</sup> which is also applicable on non-linear problems. The kernel parameters *coef* is set to zero,  $\gamma$  and *d* are estimated by 5-fold cross validation in searching spaces  $\{2^{-10}, 2^{-9}, \dots, 2^2\}$  and  $\{2, 3, \dots, 7\}$  respectively. In the evaluation step, Mean Squared Error (MSE) and Squared Correlation Coefficient (SCC) are adopted to evaluate the performance of the predictor.

The efficient, widely applied open source tool Libsvm [14] is used to practice our SVR training.

### C. Selection of Features

In practice, our feature selection is based on two criteria. The first one is that the selected features can be easily obtained in reality, i.e., they could be collected from measurements, surveys, related documents like building plans and so on. The second one is that they should potentially be the most important ones to the predictor. The combination of RGS and CC algorithms is used to evaluate the usefulness of features for SVR model. The scores for each feature are listed in Table I columns 2 and 3. As our extensive experimental results have shown, the features with the highest scores under RGS are generally more useful than those with the highest ranks according to CC. This indicates that RGS method is more applicable to SVR than CC method. However, since the feature subsets with low scores are still possibly useful for the learning algorithms [15], we take both RGS and CC into consideration while choosing the features.

The weather condition is normally considered to be the most important factor which influence the building energy consumption. In our selection, we keep two weather features that have the highest scores under RGS, which are dry bulb temperature and outdoor air density while other weather features like relative humidity are discarded no matter how their variations could contribute to energy requirement as we thought naturally. The water main temperature, electrical equipment heat gain such as lights heat gain are determined by their power and occupants' schedule, they could probably be measured or assessed in actual buildings. Zone mean air temperature and zone infiltration volume could also be measured or estimated in a normally operated building. All of the above selected features have scores not less than 1. A special case we have to consider is the number of occupants. This feature takes a middle place under RGS, but since it can be easily counted in real life and has a very high score under the evaluation of CC, we choose to keep it in the final subset. All other features will be discarded since they get low scores or are hard to be collected in reality. For example, Zone Total Internal Total Heat Gain is difficult to be obtained directly and District Heating Outlet Temp is useless according to CC. The selected features are indicated with stars in column Case1 in Table I.

In order to evaluate if the selected feature set is optimal, we also design three other subsets for comparison, they are

Table II  
COMPARISON OF MODEL PERFORMANCE ON DIFFERENT FEATURE SETS.  
(NF: NUMBER OF FEATURES, MSE: MEAN SQUARED ERROR, SCC: SQUARED CORRELATION COEFFICIENT.)

	Case1	Case2	Case3	Case4
NF	8	8	14	6
MSE	6.2e-4	7.5e-4	2.1e-3	9.2e-4
SCC	0.97	0.96	0.90	0.96

indicated by columns Case2, Case3 and Case4 in Table I. Compared with case 1, in case 2, we change three of the selected features to other three unselected ones. In case 3, all of the selected features are substituted with other unselected ones except Zone Total Internal Total Heat Gain which is not easily obtained in practice. In case 4, two features which gain lowest scores are removed from selected subset.

### D. Numerical Results

Based on the above considerations, we generate four sub-datasets of the first dataset for model training and testing. We show the results of all four cases in Table II. Two conclusions can be reached accordingly, the first one is that the designed feature selection method is valid due to model performance in case 1 outperforms other three cases. The other one is that SVR model with RBF kernel has stable performance since high prediction accuracy is always achieved on all of the four subsets.

We put the model performance before and after feature selection on three datasets in Table III. After feature selection, the accuracy of the prediction on 50 buildings' consumptions improves significantly. With regard to 20 building's consumptions, MSE increases to a certain extent, indicating certain decrease in prediction accuracy. However from the standpoint of SCC, the performance of the model with RBF kernel is quite close to the situation without feature selection performed. With regard to polynomial kernel, when training on the original datasets, the prediction ability of the model is just as good as RBF kernel, indicating that polynomial kernel is also applicable on such problem. After adopting feature selection, the performance of the model becomes better for the case of 50 buildings than for 20 buildings. Same trend can also be found for RBF kernel. These phenomena indicate that proposed feature selection approach could give better performance to the models when relatively more training samples are involved.

Another advantage of feature selection we can find in the results is the reduction of training time. For the case of 1 building, 20 buildings and 50 buildings, when we train the model with RBF kernel, before feature selection, the training time is 19.9s, 8208.6s, 40816.6s respectively, while after feature selection, the time decreases to 7s, 2018.4s and 29523.8s respectively. This reduction is remarkable for all of the three cases. We note that for each dataset, we have to assign different model parameters for the learning algorithm,

Table I  
THE SCORES OF FEATURES EVALUATED BY RGS AND CC SELECTION METHODS. THE STARS INDICATE SELECTED FEATURES IN THAT CASE.

Features	RGS	CC	Case1	Case2	Case3	Case4
Outdoor Dry Bulb	1.61	0.29	*	*		*
Outdoor Relative Humidity	0.62	0.26		*		
Wind Speed	0.52	0.01		*	*	
Direct Solar	0.54	0.47			*	
Ground Temperature	0.99	0.07			*	
Outdoor Air Density	1.26	0.20	*			*
Water Mains Temperature	1.30	0.07	*			*
Zone Total Internal Total Heat Gain	1.01	0.67				
People Number Of Occupants	0.93	0.68	*	*		
People Total Heat Gain	0.93	0.68			*	
Lights Total Heat Gain	1.13	0.05	*	*		*
Electric Equipment Total Heat Gain	1.06	0.69	*	*		*
Window Heat Gain for each wall	1.03	0.62			*	
Window Heat Loss for each wall	0.93	0.50			*	
Window Heat Gain for each wall	0.82	0.35			*	
Window Heat Loss for each wall	0.82	0.48			*	
Window Heat Gain for each wall	0.73	0.56			*	
Window Heat Loss for each wall	0.82	0.48			*	
Window Heat Gain for each wall	0.89	0.56			*	
Window Heat Loss for each wall	0.95	0.50			*	
Zone Mean Air Temperature	1.14	0.22	*			*
Zone Infiltration Volume	1.00	0.34	*	*		
District Heating Outlet Temp	0.95	7.35e-4		*	*	

Table III  
PREDICTION RESULTS OF SVR WITH TWO KERNEL METHODS ON THREE DATASETS. (BD: BUILDING, BF: BEFORE FEATURE SELECTION, AF: AFTER FEATURE SELECTION.)

			One bd	20 bds	50 bds
RBF kernel	BF	MSE	4.8e-4	4.3e-4	4.4e-4
		SCC	0.97	0.97	0.97
	AF	MSE	6.2e-4	2.1e-3	3.7e-4
		SCC	0.97	0.96	0.97
Polynomial kernel	BF	MSE	8.0e-4	5.8e-4	5.9e-4
		SCC	0.96	0.96	0.96
	AF	MSE	2.1e-3	0.19	4.7e-4
		SCC	0.91	0.85	0.98

as explained in section III-B, which always have a great influence on the training speed.

#### IV. CONCLUSION

This paper proposes a feature selection method for support vector regression when applied to predict building's energy consumption. The features are selected according to their feasibility in practice and the evaluation under two selection algorithms which can give each feature a score according to its usefulness to the predictor: the gradient guided feature selection and the correlation coefficients. Experimental results show that the selected subset is valid and can provide acceptable predictors. Performance improvement is achieved in some cases, e.g., accuracy enhanced remarkably for the models with either radial basis function or polynomial kernel on fifty buildings' data, the time for model learning decreases to a certain extent. Besides radial basis function kernel, we proved that polynomial kernel is also applicable to this application. However it seems not as stable as

radial basis function kernel. Furthermore, it requires more complicated pre-processing work since compared to radial basis function, there are more kernel parameters need to be estimated.

This work serves as the first guide for selecting an optimal subset of features when applying machine learning methods to the prediction of building energy consumption. Future work will be carried out on developing appropriate feature selection algorithms for other modeling methods.

#### REFERENCES

- [1] B. Dong, C. Cao, and S. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Energy and Buildings*, vol. 37, no. 5, pp. 545–553, 2005.
- [2] M. Espinoza, J. Suykens, and B. Moor, "Fixed-size least squares support vector machines: a large scale application in electrical load forecasting," *Computational Management Science*, vol. 3, no. 2, pp. 113–129, April 2006.
- [3] H. X. Zhao and F. Magoulès, "Parallel support vector machines applied to the prediction of multiple buildings energy consumption," *Journal of Algorithms & Computational Technology*, vol. 4, no. 2, pp. 231–249, 2010.
- [4] A. Azadeh and S. Tarverdian, "Integration of genetic algorithm, computer simulation and design of experiments for forecasting electrical energy consumption," *Energy Policy*, vol. 35, no. 10, pp. 5229–5241, 2007.
- [5] C. Maia and M. Gonçalves, "A methodology for short-term electric load forecasting based on specialized

- recursive digital filters,” *Computers and Industrial Engineering*, vol. 57, no. 3, pp. 724–731, 2009.
- [6] A. Neto and F. Fiorelli, “Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption,” *Energy and Buildings*, vol. 40, no. 12, pp. 2169–2176, 2008.
- [7] G. Tso and K. Yau, “Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks,” *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.
- [8] R. Yokoyama, T. Wakui, and R. Satake, “Prediction of energy demands using neural network with model identification by global optimization,” *Energy Conversion and Management*, vol. 50, no. 2, pp. 319–327, 2009.
- [9] S. Wong, K. K. Wan, and T. N. Lam, “Artificial neural networks for energy analysis of office buildings with daylighting,” *Applied Energy*, vol. 87, no. 2, pp. 551–557, 2010.
- [10] F. Lai, F. Magoulès, and F. Lherminier, “Vapnik’s learning theory applied to energy consumption forecasts in residential buildings,” *International Journal of Computer Mathematics*, vol. 85, no. 10, pp. 1563–1588, 2008.
- [11] V. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
- [12] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia, “Nearest neighbor based feature selection for regression and its application to neural activity,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Scholkopf, and J. Platt, Eds. MIT Press, 2006, pp. 995–1002.
- [13] EnergyPlus, 2009, available online at: <http://www.EnergyPlus.gov>.
- [14] C. Chang and C. Lin, *LIBSVM: a library for support vector machines*, 2001, available online at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (Accessed October 2010).
- [15] G. Isabelle and E. André, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.