International Journal of Smart Home
Vol. 10, No. 9 (2016), pp. 287-296
http://dx.doi.org/10.14257/ijsh.2016.10.9.27

Prediction Methods and Precise Electricity Energy Prediction of School Facility

Hanguk Ryu*[1] and Sebo Kim[2]

[1]Dept. of Architectural Engineering, Changwon National University
20 Changwondaehak-ro Uichang-gu Changwon-si, Gyeongsangnam-do 641-773, South Korea
[2]Dept. of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA
[1]hgryu@changwon.ac.kr, [2]sebkim@cise.ufl.edu

Abstract

There are many obvious evidences supporting a correlation between school facility and student behavior and performance. With the increasing awareness of sustainable school facility, incorporation of various operation cost impact into the consideration of school facility management is attracting a lot of attention. So the Green-School Project in South Korea aims to transform existing deteriorated elementary, middle, and high school facilities into eco-friendly energy saving schools through environmentally friendly materials and techniques and full-scale renovation and repair work. However, the total number of educational facilities in South Korea as of 2015 is 11,590 (5,978 elementary schools, 3,219 middle schools, and 2,393 high schools). Overall reconstruction of these deteriorated educational facilities is realistically difficult. Expenditure by school systems must stay within the limit of their available funding. So in order to plan exact operating cost, this paper presents a prediction improving method of the amount of electricity consumption of elementary school in South Korea by using two regressions, i.e., SVR (Support Vector Regression) and GPR (Gaussian Process Regression) and outlier detection methods, EE (Elliptic Envelope) and EM (Expectation and Maximization) algorithms. As a result, this study enables school facility managers to straightforwardly predict the electricity consumption of elementary school. This method can also extend to prediction of the amount of electricity usage for middle school and high school as well as elementary school.

Keywords*: Electricity Consumption, Elementary School, Operation Cost, Regression Algorithm, Support Vector Machine, Gaussian Process Regression, Elliptic Envelope, Expectation and Maximization*

1. Introduction

As schools are centralized locations where learning is encouraged, effective and efficient operation and maintenance of school facilities is a critical need for all students, teachers, related educational government department and every society. There are many obvious evidences supporting a correlation between school facility and student behavior and performance. School facility factors such as building age and condition, quality of maintenance, and temperature can affect student health, safety and psychological state [1]. Lawrence, B. (2003) pointed out that poorly maintained school facilities may have adverse health and safety impacts in causing asthma attacks, drowsiness, lethargy and a resulting inability to concentrate [2]. School facility maintenance participants now widely

*Corresponding Author

ISSN: 1975-4094 IJSH

recognize the needs to actively engage in the sustainable school facility. Because establishing an effective and efficient school facility management system is very important because a safe and comfortable study environment enhances educational performance [3].

Most of the elementary and middle schools in South Korea, built around the mid-to late 1980s, are also old and not energy efficient, and urgently require improvement [4, 5]. The Green-School Project in South Korea aims to transform existing deteriorated elementary, middle, and high school facilities into eco-friendly energy saving schools through environmentally friendly materials and techniques and full-scale renovation and repair work. The total number of educational facilities in South Korea as of 2015 is 11,590 (5,978 elementary schools, 3,219 middle schools, and 2,393 high schools) [5]. However, the overall reconstruction of these deteriorated educational facilities is realistically difficult [6]. In addition, expenditure by school systems must stay within the limit of their available funding. As such, there is a budget constraint. When there are needs deemed to be more important, those needs take priorities in funding. When tackled with a revenue deficit in the short term, most schools endeavor to defer maintenance on school facilities. The improperly maintained schools will be incongruously deteriorated in the long term.

School facility maintenance has been usually subject to deferral by school systems in favor of building new schools and making additions to existing schools. Throughout the lifecycle of the school facilities, portions of funding intended for maintenance have tended to be deferred or reassigned. This continuously accumulates the total amount of deferred maintenance thereby resulting in poorly maintained school facilities [7].

The energy consumption characteristics of 10 elementary schools in Daegu, located in the southern part of South Korea, were analyzed by year, unit area, and per capita, based on the data from January 2006 to December 2010, to present the specific values for the energy-saving goals of the elementary schools in South Korea [8]. Since the 2000s, South Korea has been distributing IT devices and cooling/heating facilities in schools to improve the educational environment, and the schools have been modernized [9].

According to Korea Energy Economics Institute, the energy consumption of the elementary schools in South Korea is increasing annually due to the improved building facilities, the distribution of educational apparatuses and IT systems for the curriculum, and the increase in the school meal systems [9].

Accordingly, the energy consumption of schools is continuously increasing. Of the diverse energy sources, the consumption of electric energy is especially rapidly increasing as a main energy source for schools [10]. For the energy consumption by energy type, electric energy is used the most (71.4%), followed by gas (22.3%) and oil (6.3%), it seems that the electric power consumption continues to increase because of the replacement of cooling/heating systems by electric systems, the installation of electric IT equipment, and the changes in the outdoor temperature; and the energy consumption per unit study area of the elementary schools in South Korea in 2010 was 1040 MJ/m2y for electricity, 325 MJ/m2y for gas, and 92 MJ/m2y for oil [8].

With the increasing awareness of sustainable school facility, incorporation of various operation cost impact into the consideration of school facility management is attracting a lot of attention. This study enables school facility managers to straightforwardly predict the electricity consumption. This paper presents prediction improvement of the amount of electricity consumption of elementary school in South Korea by using two regressions, i.e., SVR (Support Vector Regression) and GPR (Gaussian Process Regression) and outlier detection methods, EE (Elliptic Envelope) and EM (Expectation and Maximization) algorithms.

2. Prediction Performance of the Amount of Electricity Consumption of Elementary School using Regression Algorithms

2.1. Data Samples to Predict the Amount of Electricity Consumption of Elementary School

Our data consists of 4,674 number of samples. Each sample represents individual elementary school information. For all samples, there are six independent variables (features); region information, establishment year, the number of student, the number of class, the number of teaching staff, and the architectural area. All variables have ordinal values except region information feature which has categorical values. Therefore, we should preprocess this region information feature for repressor to be able to learn the data properly.

We exploit one-hot encoding to preprocess the region information. One-hot refers to a group of bits among which the legal combinations of values are only those with a single high (1) bit and all the others low (0). Through this way, we make the dimension of region information feature from one to 16 (because region information has 16 different main city values in South Korea). The weather independent variables are considered for 24 hours monthly average data for one year except for summer (August) and winter (January) vacation in South Korea. With these independent variables, every sample has a dependent variable that is the amount of electricity consumption for 2014 year [4].

2.2. Prediction Performance using SVR (Support Vector Machine) and GPR (Gaussian Process Regression)

For our experiment, we mainly use two regression algorithms; SVR and GPR. The one of special characteristics of GPR is that it can produce pointwise confidence interval when predicting test samples. This function of GPR offers the criteria for how strong we can believe a particular prediction. However, the critical problem of GPR is that time and space complexity grow rapidly with the increase of dimension of data. In our experiment environment (Intel Core i7-4790 with 12GB RAM), ~20 independent variables are allowed. On the other hand, SVR, a few thousand independent variables are easily dealt with, does not have this dimension limitation in learning samples even if it cannot produce pointwise confidence interval.

We use 5-fold CV (cross validation) and R^2 score for checking how well the combination of the elementary school data and two regression models can predict test samples. 5-fold CV randomly partitions the original dataset into five subsets. We use four of these subsets for training, and the remaining one to test the R^2 score of the regression method. We repeat the CV process five times, each with a different test dataset as the validation data. We then report the average of the resulting five R^2 scores. R^2 score is defined as equation (1), where $\hat{y}_i$ is the predicted value of the i-th sample, y_i is the corresponding true value, and equation (2). If the regression method creates perfect predicted values for all samples, then the score is 1.0. The lower bound of the score is minus infinity because the model can be arbitrarily worse.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2} \tag{1}$$

$$\bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i \tag{2}$$

Figure 1 presents R^2 score bar-graph of the five-fold CV results. Our results suggest that prediction performance of SVR is almost similar with that of GPR. We call establishment year, the number of student, the number of class, the number of teaching staff, and the architectural area as basic 5 features. There are additional weather information features. Red bar in the figure indicates the result when using 5 basic features and blue bar indicates the result when using 5 basic features and all weather information available. Including additional weather information definitely improves the prediction performance of both regression methods [4].

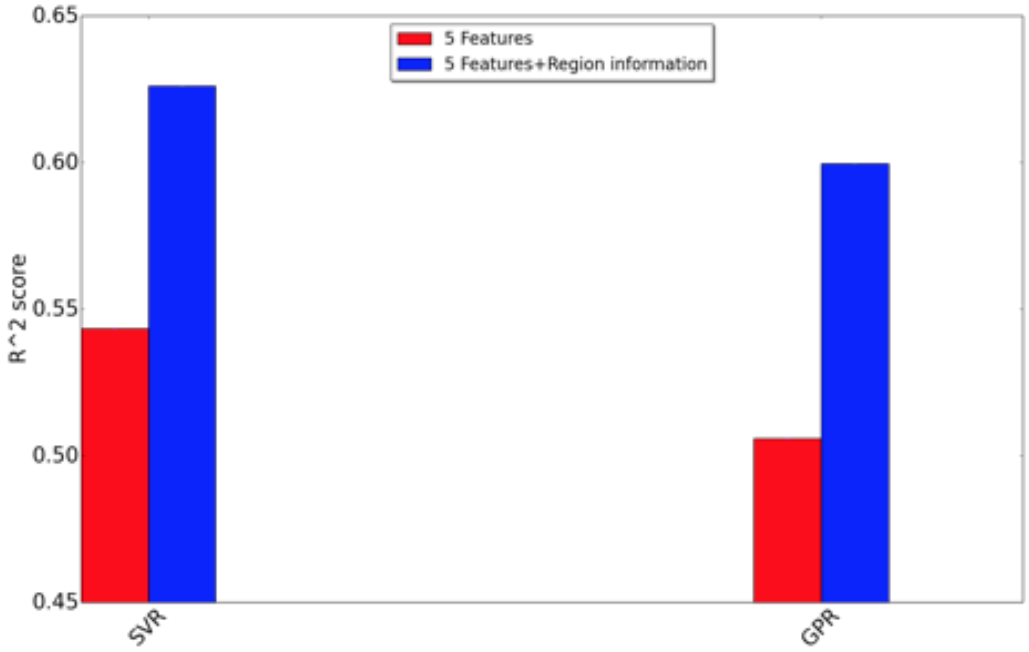

Figure 1. Performance Comparison between SVR and GPR with Using Five Features

3. Elementary School Energy Prediction Performance Improvement by Applying Outlier Detection Method

3.1. Outlier Detection Method

We assume that if we exclude outlier samples in our dataset we can boost the performance of the regression method. Moreover, analyzing these outlier samples can give insight of which schools have to be inspected for school remodeling or renovation in the perspective of energy efficiency more thoroughly.

Assume there are bunch of samples whose feature values are more or less similar each other. Their electricity consumptions are likely to be similar each other because those samples share input feature values. Roughly speaking, outlier samples are the samples whose electricity consumption abnormally deviates from the electricity consumption of major samples.

In this section, we consider how we define outlier samples in our dataset. Assume that we have two groups ‘A’ and ‘B’. We have ‘n1 number of’ samples that belong to ‘A’ and ‘n2’ number of samples that belong to ‘B’. We call input features as independent variables in the context when dealing with the regression. Each sample contains values of independent variables. For example, sample ‘s1’ may contain [1.0, 1.1, 1.05], sample ‘s2’ may contain [2.0, 2.5, 3.0], sample ‘s3’ may contain [1.01, 1.11, 1.049], and sample ‘s4’

may contain [2.05, 2.53, 3.04]. We can calculate the distance between two samples using any distance metric such as Euclidean. If we group those 4 samples into two groups, samples 's1' and 's3' will be in the group 'alpha', and samples 's2' and 's4' will be in the group 'beta' because that grouping minimizes the sum of distances between any two samples in the group 'alpha' and the group 'beta' respectively. Developing this kind of clustering algorithm is not an easy task when there are many numbers of samples and high dimensional input features for each sample. Though finding perfect solution, i.e., finding global minima that minimizes as much as possible, the sum of distances between any two samples in each group respectively is not a computationally solvable problem. Finding almost perfect solution, i.e., finding local minima, is possible with EM (Expectation and Maximization) algorithm. The input of EM algorithm is values of independent variables of all samples and the input parameter of EM is the number of components, i.e., how many clusters or groups of algorithm partition into.

Our dataset has not only independent variables but also a dependent variable. Remind that the purpose of regression methods is to predict the value of a dependent variable when we are given values of independent variables. We call these as predicted values. To predict the value, regression methods need to fit or learn the data before it actually predict something. We call values of a dependent variable that are used in learning phase as true values. To explain more about how we define outlier samples, remind that we assume we have 'n1' number of samples that belong to group 'A', and 'n2' number of samples that belong to group 'B'. Samples in each group are close to each other in terms of values of independent variables. Thus, the values of a dependent variable in the same group are supposed to be similar together. However, this rarely happens because data usually contains outlier samples. We say that outlier samples are the samples whose values of a dependent variable are far away from values of a dependent variable of major samples in one group. We summarize that major samples are the samples where all values of both independent variables and a dependent variable are grouped together while outlier samples are the samples where only values of independent variables are grouped together.

We use EE (Elliptic Envelope) algorithm to filter out outlier samples. The input of the EE is values of a dependent variable of all samples in one group and the input parameter of EE is the proportion of outliers. For instance, if the total number of samples in one group is 100 and we set this input parameter as 0.1, then EE algorithm considers 10 samples as outlier samples. Remind that the input of EM algorithm is values of independent variables of all samples and the input parameter of EM is the number of components, i.e., how many clusters or groups of algorithm partition into as mentioned. Overall, the main procedure of outlier detection method has two steps: 1) EM algorithm clusters data samples into 'm' number of groups and 2) For each group, find the most 'epsilon' percent outlier samples using EE.

3.2. Outlier Detection using EM Algorithm

We tried to apply this outlier detection method to improve the performance of prediction. In clustering groups, we used two different strategies. First strategy was that we used all independent variables to create 'm' number of groups using EM algorithm. Because our hypothesis is that it is the most important factor that affects the amount of electricity consumption. Second strategy was that we used only one independent variable that is the number of students to create 'm1' number of groups, and used EM algorithm again to create 'm2' number of groups for each clustered groups. Therefore, total created groups in the second strategy is ('m1' * 'm2') number of groups. Table 1 shows the prediction performance applying outlier detection method of the first strategy, using EM algorithm. 'm' in Table 1 is the number of groups created, and 'alpha' is the proportion of outliers in the data set.

Table 1. First prediction performance of outlier detection using EM algorithm

	Alpha=0.1	Alpha = 0.2	Alpha = 0.3
m=30	0.6906	0.68	0.6944
m=60	0.6975	0.7023	0.7104
m=90	0.6952	0.7038	0.7016
m=120	0.6984	0.693	0.6955

The problem of the first strategy grouping is that very different samples are grouped into the same cluster. Both the fact that the small numbers of samples are available and the number of independent variables is relatively large may cause this phenomenon. The second strategy is helpful to mitigate this issue. Table 2 and Figure 2 show the result of second prediction performance of outlier detection using EE algorithm.

Table 2. Second Prediction Performance of Outlier Detection using EE and EM Algorithm

	Alpha=0.1	Alpha = 0.2	Alpha = 0.3
m1=10, m2=2	0.7858	0.8076	0.83
m1=10, m2=4	0.7753	0.7927	0.8111
m1=10, m2=6	0.7292	0.7485	0.7553
m1=10, m2=8	0.7175	0.7248	0.7275

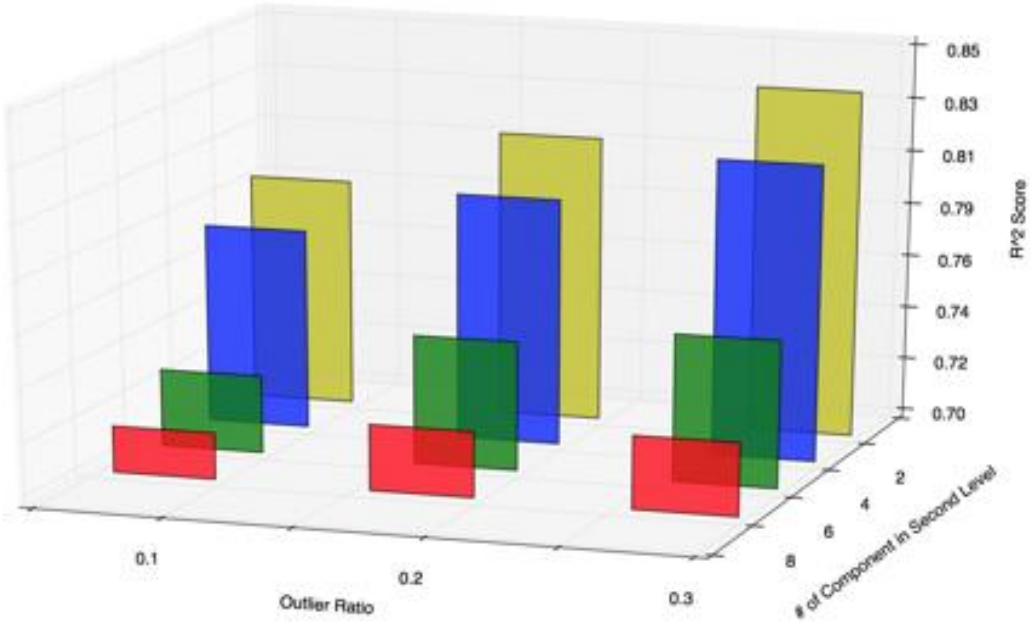

Figure 2. Second Prediction Performance Graph of Outlier Detection using EE and EM Algorithm

3.3. Scatter Plots of Predicted Value and True Value

Figure 3, 4, and 5 present scatter plots of predicted value and true value. We divide samples into train and test samples in random way. After training, we can get true values and predicted values of test samples. Figure 3 shows each test sample of true value and predicted value by SVR using only five features, and figure 4 shows each test sample of true value and predicted value by SVR using five features with additional weather

information features, and figure 5 shows each test sample of true value and predicted value that is survived after filtering out outliers by SVR using five features with additional weather information features.

Ideal scatter plot would be a direct line plot (y=x) if the prediction algorithm can predict exact true value. As we see, adding additional weather information increases the performance of the regression method. Samples tend to be located nearby the direct line as shown from figure 3 to figure 4. Filtering out outlier samples further increases the performance. After filtering out outliers, the samples show stronger tendency that they are located nearby the direct line as shown from figure 4 to figure 5.

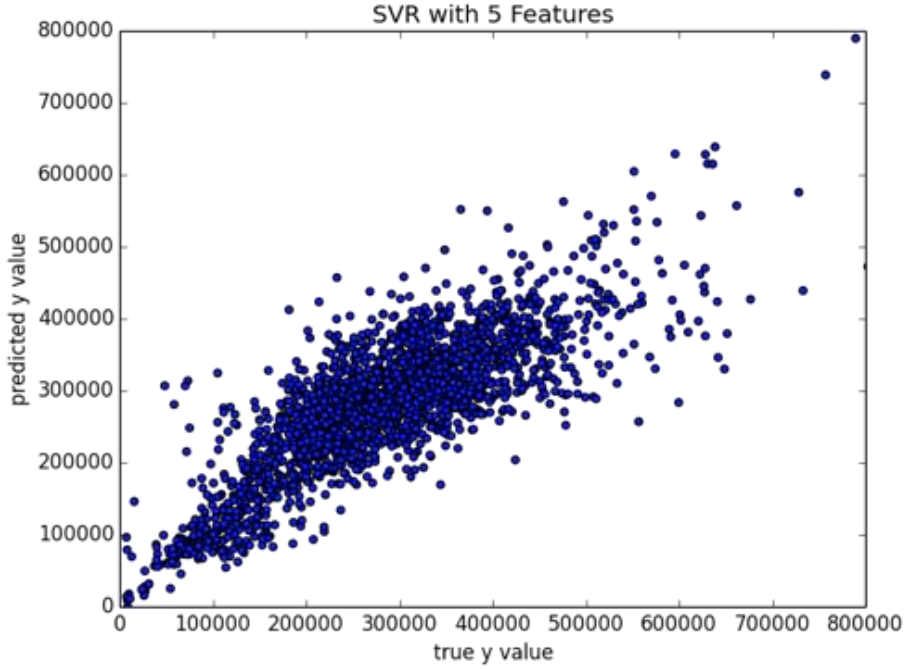

Figure 3. Test Samples of True Value and Predicted Value by SVR using only Five Features

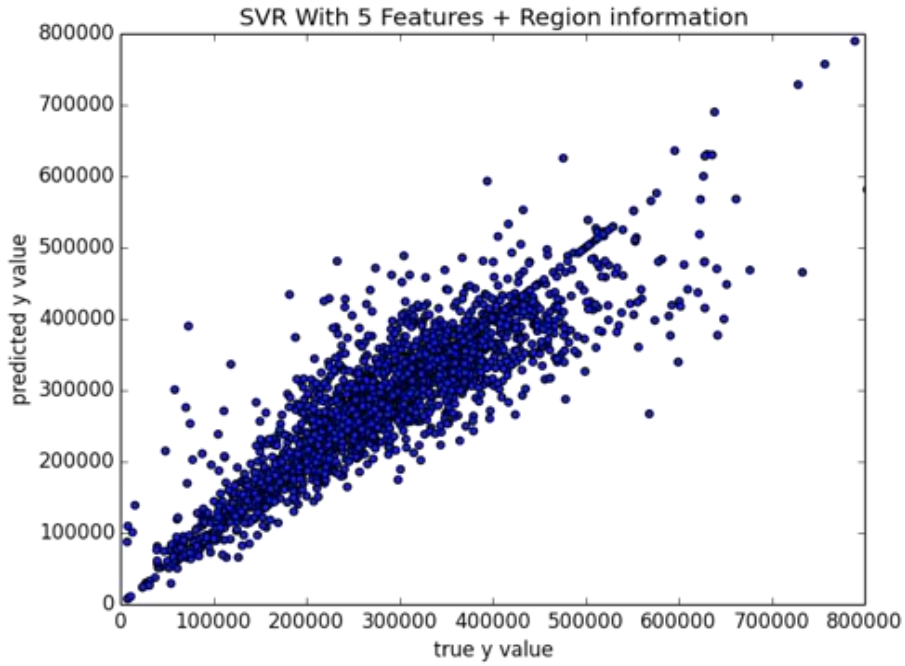

Figure 4. Test Samples of True Value and Predicted Value by SVR using Five Features with Additional Weather Information Features

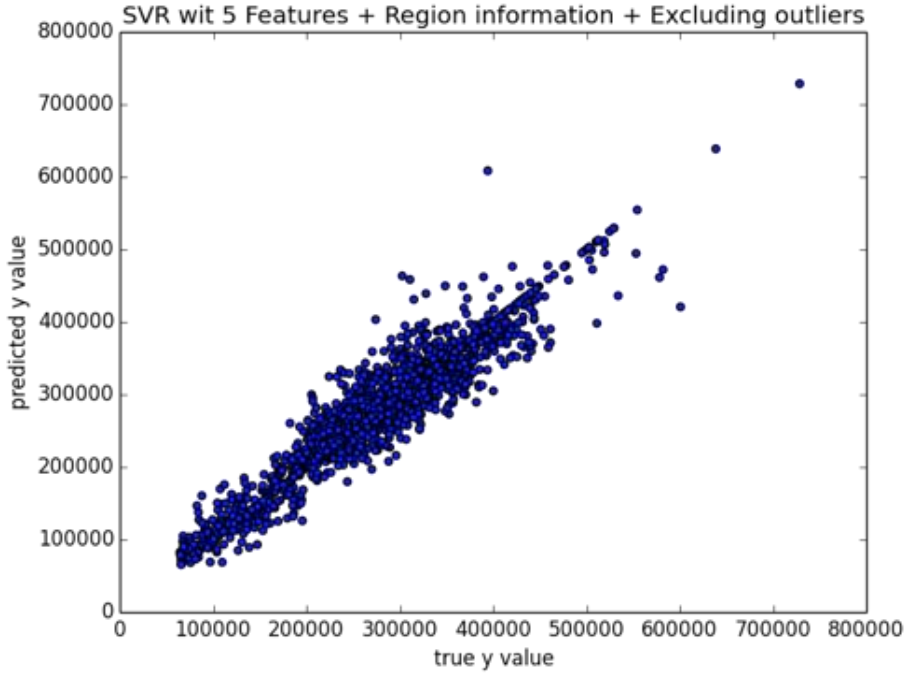

Figure 5. Test Samples of True Value and Predicted Value that is Survived after Filtering out Outlier Samples by SVR using Five Features with Additional Weather Information Features

4. Conclusions

Many participants including school administration, principal, school facility manager, constructor and others are now also increasingly accountable for their impacts on the safety, construction cost and maintenance cost or sustainability of school facility. Cost analyst, owner, engineer, contractor and facility manager have a difficulty to obtain and facilitate basic analyzed data required to plan and execute correctly the operation cost, especially the amount of consumption of electricity usage for a specific type of elementary school type. For example, it is shown that physical and location differences of elementary school may cause especially the amount of consumption of electricity usage

This paper presents prediction improvement of the amount of electricity consumption of elementary school in South Korea by using two regressions, i.e., SVR, GPR, and outlier detection methods using EE and EM algorithms which can supply more accurate prediction performance. After training samples in random way, we can get true values and predicted values of stronger tendency to ideal direct line plot by SVR using five features with additional weather information features as shown in scatter plots of predicted value and true value.

As a result, this study enables school facility managers to straightforwardly predict the electricity consumption of elementary school. This method can also extend to prediction of the amount of electricity usage for middle school and high school as well as elementary school.

Acknowledgments

This research was supported by the Basic Science Research Program of the National Research Foundation of Korea, which is funded by the Ministry of Education (No. NRF-2014R1A1A2055797).

References

[1] H. Ryu and S. Kim, "Prediction Improvement of the Amount of Electricity Consumption of Elementary School in South Korea", Asia-pacific Proceedings of Applied Science and Engineering for Better Human Life, Jeju, South Korea, (2016), August 16-19.
[2] Korean Educational Development Institute, 2015 Brief Statistics on Korean Education, (2015).
[3] T. Hong, H. Kim and T. Kwak, "Energy-Saving Techniques for Reducing CO2 Emissions in Elementary Schools", Journal of Management in Engineering, vol. 28, no. 1, (2011), pp.39-50.
[4] M. Bello and B. Loftness B, "Addressing Inadequate Investment in School Facility Maintenance", School of Architecture paper 50, Carnegie Mellon University, (2010).
[5] E. Young, H. Green, L. Roehrich-Patrick, L. Joseph and T. Gibson, "Do K-12 School Facilities Affect Education Outcomes?", Tennessee Advisory Commission on Intergovernmental Relations, (2003).
[6] B.K. Lawrence, "Save a Penny, Lose a School: The Real Cost of Deferred Maintenance", Policy Brief Series on Rural Education. Rural School and Community Trust, Washington DC, (2003).
[7] J. Choi and H. Ryu, "Multidimensional Representation of Educational Facility LCC (Life Cycle Cost) Data", Indian Journal of Science and Technology, vol. 8, no. 25, (2015), pp.1-8.
[8] T.W. Kim, K.G. Lee and W.H. Hong, "Energy Consumption Characteristics of the Elementary Schools in South Korea", Energy and Buildings, vol.54, (2012), pp.480–489.
[9] Korean Educational Development Institution (KEDI), Research for Development Zero Energy-Eco School Model (II), KEDI, Seoul, South Korea, (2009).
[10] Korea Ministry of Education, Science and Technology (KMEST), 2009 Statistical Yearbook of Education, KMEST, Seoul, South Korea, (2009).

Authors

Hanguk Ryu, He is an associate professor of Department of Architectural Engineering of Changwon National University in South Korea. His major is construction management covering construction time, cost, human resource, policy, claim, and school facilities maintenance. His interestes are in the field of construction management application with information technology.

Sebo Kim, He is a Phd student of Department of Computer and Information Science and Engineering in University of Florida, Gainesville, FL, USA. His interestes are in the field of machine learning, and bioinformatics.