Development of a Decision Support System for LEED for EB Credit Selection Based on Climate Factors

Jack C.P. Cheng[1] and Jun Ma[2]

[1]Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology; email: cejcheng@ust.hk
[2]Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology; email: jmaae@ust.hk

ABSTRACT

LEED is a credit-based rating system that provides a third-party verification of green buildings worldwide. A building can obtain a Platinum, Gold, Silver or Certified grade based on the number of LEED credit points achieved. Selection of target credits is important and challenging for LEED managers due to limited budget, tight project schedule, and limited resources in many green building projects. Local climate factors like temperature can affect the selection of green building technologies and hence the LEED credits adopted. However, no research has been done to suggest LEED target credits based on climate factors. This paper aims to develop a decision support system based on climate factors for LEED credit selection using data mining techniques. The LEED for Existing Buildings version 2009 was focused in this study. Information of 912 certified green building projects and their surrounding climate circumstances was collected and studied. Classification models for 48 LEED credits that use credit achievement as the class and climate factors as the variables were then constructed and optimized using three data mining algorithms - Random Forests, AdaBoost Stumps and Support Vector Machine (SVM). The results were incorporated in a web-based decision support system. A case study was then conducted to illustrate and evaluate the system. The results showed that our decision support system has a high accuracy.

INTRODUCTION

Leadership in Energy and Environmental Design (LEED) is one of the major rating systems for evaluating green building performance. LEED is a credit based rating system, in which each credit evaluates one perspective of the building's performance and grade relative points. After summing up all the LEED credit points, the system assigns a corresponding green building level.

Since retrofitting existing building has a high potential impact to the built environment (Ma & Cheng, 2013), increasing numbers of buildings have started to pursue the certification of LEED for Existing Buildings (LEED-EB). Therefore, LEED-EB was selected as the focus of this study. LEED for Existing Building version 2009 (LEED-EB v2009) is the latest rating system for existing buildings in

LEED. Prerequisites and credits in the LEED-EB v2009 address 7 topics: Sustainable Sites (SS), Water Efficiency (WE), Energy and Atmosphere (EA), Materials and Resources (MR), Indoor Environmental Quality (IEQ), Innovation in Operations (IO), and Regional Priority (RP). Buildings are certified with the LEED-EB v2009 according to the following scale: 40–49 points for Certified, 50–59 points for Silver, 60–79 points for Gold, and 80 points and above for Platinum (USGBC, 2008).

The LEED-EB v2009 system consists of totally 52 credits, each of which addresses different aspects of building performance. Due to limited budget, tight project schedule, and limited resources in many green building projects, selection of target credits is often an important and challenging problem for LEED project managers. Researchers have attempted to analyze the relationships between decision making factors and LEED credits. For example, (Madanayake & Ruwanpura, 2012) studied the cost, schedule, and environmental aspects of most LEED credits and how they affect the target credit selection. Besides these aspects, climate factors like sunshine percentage could also affect the selection of target credits. That is why projects in different locations have different preferences for target credits even with the same target level. However, LEED managers seldom refer to the local weather when planning for a green building project and there is no research conducted to study the suggestions on target credit selection based on climate factors.

This paper presents a decision support system that we developed based on the climate factors using data mining techniques. Fig.1 shows the framework of the decision support system, which provides users with a list of suggested target credits for their LEED projects according to the project location and target LEED grade. To develop this system, 912 certified LEED-EB v2009 projects were studied. The climate conditions of the 912 projects were then gathered from the U.S. National Climate Data Center. By using the climate factors as variables and credit achievement as the class, classification models were built for each LEED credit. The models were optimized using three classifiers – Random Forests, AdaBoost Decision Stumps, and Support Vector Machine. The classification results are evaluated and illustrated with a case study in this paper.

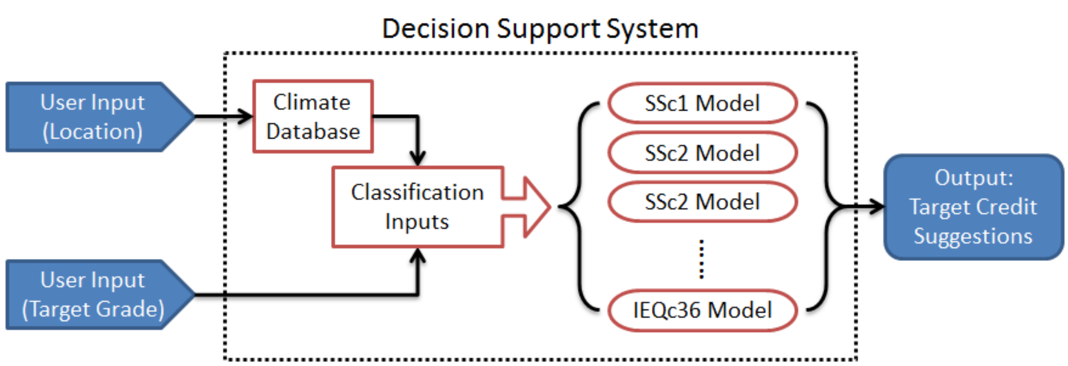

Fig.1. The framework of our decision support system.

METHODOLOGY

Classification. In data mining, classification is an important technique to identify to which set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known (Han, Kamber, & Pei, 2011). In this study, the achievement of a single credit was set as the target while the surrounding climate factors of the project were set as the variables so as to build classification models for each of the selected 48 LEED credits. In this case, each classification model can predict (or suggest) whether a new project will achieve the relative credit. The summary of the predictions (or suggestions) to all the credits is the outcome of the decision support system. To build and optimize the classification models for each credit, this study used three commonly used classifiers – Random Forests, AdaBoost Decision Stumps, and Support Vector Machine (SVM), which will be described below.

Random Forests. The Random Forests is based on decision trees, and is an ensemble method that use “voting” technique to ensemble the prediction results of all the trees. Each tree is constructed by randomly selecting a number of cases with replacement for training and the rest for testing. In this case, each case could be used as a training case as well as a testing case. The splitting algorithm for the decision tree is based on the Gini index, which is calculated as follow.

$$\mathrm{Gini}(D) = 1 - \sum_{i=1}^{m} P_i^2$$

where P_i is the probability of class i of the attribute D, and m is the number of classes of this attribute. There are a number of advantages in using this classifier. For example, there is no need for cross-validation and the testing error is proven to be unbiased (Breiman, 2001). In addition, Random Forests reveals good performance in many classification problems (Breiman, 2001).

AdaBoost Decision Stumps. This classifier is also developed from the decision tree. However, instead of using the whole tree, only the stump is used for AdaBoot Decision Stumps. That means each tree grows only one level. The trick lies in the boosting method. In fact, this classifier is also an ensemble method. Based on this method, the decision tree is adapted over time based on the performance. The trees are built one by one, and examples that are mislabeled by a previous tree are chosen more often than those that are correctly classified. The output prediction H(x) of this algorithm can be described in the following equation.

$$\mathrm{H(x)} = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$$

where α_t is the updated weight of round t, and $h_t(x)$ is the prediction of round t. The number of iterations T is defined by the user and usually it is the parameter that needs to be optimized for this algorithm (Han et al., 2011).

Support Vector Machine (SVM). SVM is a well-known classifier due to its capacity in finding the global optimal value while classifying the data. It tries to find the largest margin between two classes by optimizing the margin value using the mathematical kernel function. Based on the difference in the kernel functions, SVM is capable of classifying complicated non-linear separable classes (Han et al., 2011). In this study, the well-known Gaussian kernel was used, as shown in the following equation. Here σ is the deviation of the kernel.

$$\mathrm{k}(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/(2\sigma^2))$$

DATA COLLECTION

LEED Project Data. Data on certified LEED-EB v2009 projects were collected from the United States Green Building Council (USGBC) website using a C++ program that we developed. In total, 1012 cases were collected (USGBC, 2013). Since different countries have different policies and markets in retrofitting existing buildings, this study only focuses on the 912 projects that are located in the United States. A summary of the collected cases is shown in Table 1.

Table 1. Summary of the LEED-EB projects in this study

Category	Platinum	Gold	Silver	Certified	Total
V2009	52	419	297	144	912

Climate Data. The climate data in this study were collected from two datasets from the U.S. National Climate Data Center. One is the NOAA's 1981-2010 Climate Normals (NCDC, 2011), and the other is the 2011 Comparative Climatic Data (NCDC, 2013). The first one contains comprehensive data directly extracted from the national weather stations, while the second one only contains the data on the big cities in each state. A summary of the collected 15 climate factors is shown in Table 2.

Table 2. Climate Datasets Summary

Collected Annual Average Climate Data (Unit)	Number of Stations or Cities
Mean Temperature (°F)	7501(S)
Diurnal Temperature Range (°F)	7501(S)
Heating Degree Day (°F)	7501(S)
Cooling Degree Day (°F)	7501(S)
Precipitation (inch)	9307(S)
Snowfall (inch)	6377(S)
Number of Days with Precipitation more than 0.001 inches (day)	7484(S)
Number of Days with Snowfall more than 0.01 inches (day)	6377(S)
Number of Cloudy Days (day)	261(C)
Number of Partly Cloudy Days (day)	261(C)
Number of Clear Days (day)	261(C)
Percentage of Humidity in the Morning (%)	264(C)
Percentage of Humidity in the Afternoon (%)	264(C)
Percentage of Sunshine (%)	165(C)
Wind Speed (miles/per hour)	265(C)

*Note: The (S) and (C) refer to whether the data is from a station or a city.

PREPROCESSING

Datasets Connection. The building's surrounding climate data is represented by the data of its nearest station or city. Thus, the latitude and longitude data of both stations and cities were obtained through Google Map. Furthermore, the geographical distance was then calculated in order to find the nearest station or city. The largest distance between the project location and the nearest station or city was 197km.

Transformation of LEED Credits. Credits in LEED-EB v2009 can be divided into two categories. One is the binary credits, which use 0/1 to represent the credit achievement. Another one is the multi-point credits, in which higher points mean a higher green building performance. In order to maintain the consistency, the multi-point credits were transformed into two variables. One variable represents whether the project has successfully achieved this credit by labeling "yes" or "no" (_YN). The other variable represents whether this project achieved "high" points or "low" in that credit compared with the non-zero average score (_HL), which is the average point of all the "yes" projects. Table 3 shows the multi-point credits. The credits in the categories "Regional Priority" and "Innovation in Operation" were excluded because many work tasks performed to achieve these credits are unique and cannot be generalized for data mining.

DATA MINING

By setting the climate factors as the variables and the credit achievement as the target class, 54 classification models (48 credits) were developed using the three classifiers. All the classifiers and models were coded with R (3.0.2). The performance of the classifiers was evaluated using 10-fold cross-validation and the criterion was based on prediction accuracy, which is a major consideration of the decision support system. Therefore, the classifier which gives the highest prediction accuracy was selected as the classifier for that credit in the decision support system. Parts of the mining results are shown in Table 4.

Table 3. Transformed multi-point credits.

Original Credit	Original Points	Transformed Credits	Non-zero Average Score
SSc4	15	SSc4_YN, SSc4_HL	9.1259
WEc1	2	WEc1_YN, WEc1_HL	1.8588
WEc2	5	WEc2_YN, WEc2_HL	4.0802
WEc3	5	WEc3_YN, WEc3_HL	3.6548
EAc1	18	EAc1_YN, EAc1_HL	12.4018
EAc3.2	2	EAc3.2_YN, EAc3.2_HL	1.3855
EAc4	6	EAc4_YN, EAc4_HL	4.6764

Table 4. Examples of the optimized classification models

Credit Code	Credit	Best Classifier	Accuracy
SSc1	LEED Certified Design and Construction	SVM	0.9474
SSc2	Building Exterior and Hardscape Management Plan	Ada	0.8564
SSc3	Integrated Pest Management, Erosion Control, and Landscape Management Plan	SVM	0.7551
SSc4_YN	Alternative Commuting Transportation	Ada	0.8167
SSc4_HL	Alternative Commuting Transportation	Ada	0.8756
SSc5	Site Development—Protect or Restore Open Habitat	RF	0.6959
SSc6	Stormwater Quantity Control	Ada	0.8321
SSc71	Heat Island Reduction—Non-Roof	RF	0.7966
SSc72	Heat Island Reduction—Roof	RF	0.7640
SSc8	Light Pollution Reduction	Ada	0.7231
WEc1_YN	Water Performance Measurement	RF	0.8612
WEc1_HL	Water Performance Measurement	RF	0.7323
WEc2_YN	Additional Indoor Plumbing Fixture and Fitting Efficiency	SVM	0.9436
WEc2_HL	Additional Indoor Plumbing Fixture and Fitting Efficiency	RF	0.6298
WEc3_YN	Water Efficient Landscaping	RF	0.7765
WEc3_HL	Water Efficient Landscaping	Ada	0.8474
WEc4	Cooling Tower Water Management	RF	0.8446

*Note: Ada refers to AdaBoost Decision Stumps, and RF refers to Random Forests

IMPLIMENTATION AND CASE STUDY

After building and optimizing the classification models, the decision support system was developed. The framework of the system is shown in Fig.1. The system works in three steps. First is the input interpretation, which links the climate data to the project and finds the surrounding climate data based on the latitude and longitude provided by the user. Together with the target grade of the project, these will be the inputs of the classification models. The second step is classification, in which each of the 54 models generates a relative prediction (or suggestion) of the achievement of each credit class based on the inputs. Lastly, the system outputs all the suggestions on the target credits selection to the user. This framework was then developed into a web-based tool, the interface of which is shown in Fig. 2.

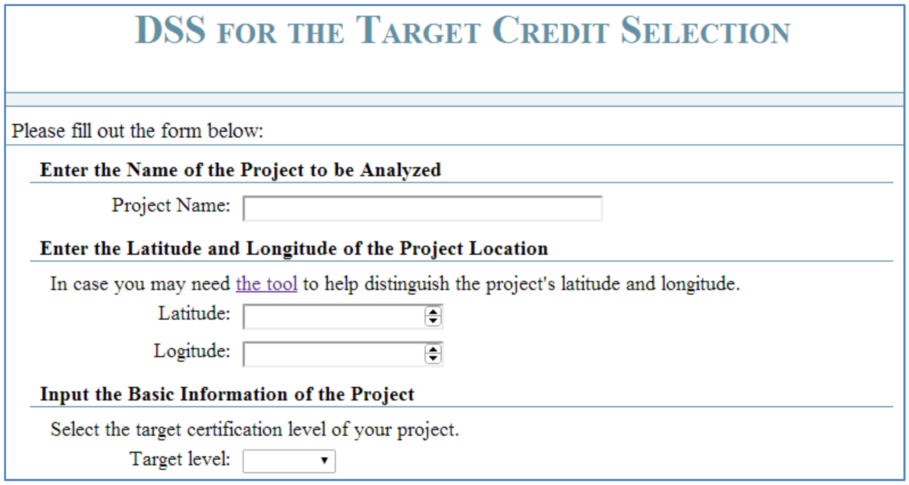

Fig.2. The screenshot of the interface of the decision support system.

The performance of the system was then tested using a project which was neither included in the previous training nor in the testing data. Basic information about the project is shown in Table 5. Table 6 shows a comparison of the actual credit achievements of the project and the "suggestions" generated by our decision support system. It can be seen that the correct suggestion rate of our system for this Gold targeted project is 85.19%.

Table 5. The basic information of the tested project

Attribute	Information
Building Name	Two Twenty Two Berkeley by EOP
Location	Boston, MA 02116, United States
Certification	LEED O+M: EB v2009 **GOLD**
Total Score	65

CONCLUSIONS

In this study, data mining techniques were used to predict the credit achievements of LEED-EB v2009 projects. Three different classification methods were implemented to improve the performance of the prediction. The study reveals promising implementation of data mining techniques in green building related studies. The result of the experiment of the Gold project shows that around 85.19% of the credits can be accurately suggested by our system.

Due to limited information, our decision support system did not considered other building features like building type, project budget and schedule. We believe the system will work better if that information could be included. This is considered as part of our future work. Besides, we will also try to extend the system to suggest target credits for new construction projects.

Table 6. Comparison between the actual credit achievement and our "suggestions"

Credit Category	Number of Models	Number of Correctly Predicted Models
Sustainable Site	10	9

Water Efficiency	7	7
Energy & Atmosphere	12	10
Materials & Resources	10	8
Indoor Environmental Quality	15	12
Total	46/54	
Correct Suggestion Rate	85.19%	

ACKNOWLEDGEMENT

The authors would like to acknowledge the support by the Hong Kong Research Grants Council, Grant No. 611513. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the Hong Kong Research Grants Council.

REFERENCES

Breiman, Leo. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Han, Jiawei, Kamber, Micheline, & Pei, Jian. (2011). *Data Mining: Concepts and Techniques 3rd Edition* (3 ed.): Morgan kaufmann.

Ma, Lucky J., & Cheng, Jack C.P. (2013). *A Study on the Influence of Climate Factors on LEED Green Building Credit Selection Using Data Mining Techniques*. Paper presented at the ICCBEI 2013, Tokyo, Japan.

Madanayake, HL Sulochana P, & Ruwanpura, Janaka Y. (2012). *Multi-Criteria Decision Making to Improve Performance in Construction Projects with LEED Certification.* Paper presented at the Construction Research Congress 2012@ sConstruction Challenges in a Flat World.

NCDC. (2011). NOAA's 1981-2010 Climate Normals. Retrieved April 16, 2013 http://www.ncdc.noaa.gov/oa/climate/normals/usnormals.html

NCDC. (2013). Comparative Climatic Data for the United States through 2011. Retrieved April 16, 2013 http://www1.ncdc.noaa.gov/pub/data/ccd-data/

USGBC. (2008). LEED 2009 for Existing Buildings: Operations & Maintenance Rating System: Washington, DC: US GBC.

USGBC. (2013). LEED for Existing Building Projects Directory. Retrieved July 9, 2013, from http://www.usgbc.org/projects/existing-buildings