Using data mining in optimisation of building energy consumption and thermal comfort management

Conference Paper · July 2010

Source: IEEE Xplore

CITATIONS
27

READS
965

4 authors, including:

Yang Gao
University College Cork
30 PUBLICATIONS **586** CITATIONS
SEE PROFILE

Karsten Menzel
Technische Universität Dresden
104 PUBLICATIONS **636** CITATIONS
SEE PROFILE

Using Data Mining in Optimisation of Building Energy Consumption and Thermal Comfort Management

Yang Gao, Emmanuel Tumwesigye, Brian Cahill, Karsten Menzel
Department of Civil and Environmental Engineering
University College Cork
College Road, Cork, Ireland
y.gao@ucc.ie

Abstract—Performance monitoring using wireless sensors is now common practice in building operation and maintenance and generates a large amount of building specific data. However, it is difficult for occupants, owners and operators to explore such data and understand underlying patterns. This is especially true in buildings which involve complex interactions, such as ventilation, solar gains, internal gains and thermal mass. Performance monitoring requires collecting data concerning energy consumption and ambient environmental conditions to model and optimise buildings' energy consumption. This paper details the use of data mining techniques in understanding building energy performance of geothermal, solar and gas burning energy systems.

The paper is part of an outgoing research into optimisation of building performance under hybrid energy regimes. The objective of the research presented in this paper is to predict comfort levels based on the Heating, Ventilating, and Air Conditioning (HVAC) system performance and external environmental conditions.

A C4.5 classification methodology is used to analyse a combination of internal and external ambient conditions. The mining algorithms are used to determine comfort constraints and the influence of external conditions on a building's internal user comfort. To test the performance of classification and its use in prediction, different offices, one to the south and the other to the north of the building are used.

Classification rules being developed are analysed for their application to modify control algorithms and to apply results to generalise hybrid system performance. The results of this study can be generalised for an entire building, or a set of buildings, under a single energy network subject to the same constraints.

Index Items-sensors, HVAC, energy, performance, multi-dimension, classification, data mining

I. INTRODUCTION

Measuring and collecting large volumes of data relevant to environmental performance is becoming increasingly affordable with falling prices of sensing and computing technology and the increasing availability of Building Energy Management Systems (BEMS). To make the best use of this abundance of data, it is necessary to apply statistical analysis and data mining techniques to extract relevant information useful to the building services operator. Intelligence with respect to energy in a smart building consists of management of cooling and heating systems, lighting, and occupancy to achieve optimized user comfort [12]. Measuring and predicting energy requirements in these systems provides significant benefits for energy management. The models are trained and tested on data obtained from a subject building's south facing room. The data contains measurements relevant to external and internal environmental conditions. The data is then tested considering only internal environmental measurements. This process is validated with a north facing room.

Building performance analysis, based on Wireless sensor Networks (WSNs) data, has a wide spectrum of exciting applications. The aim of this research is to assess how data mining (DM), especially the Decision Trees (DTs) method, can enable facility operators to predict room temperature, and lead to potential optimization of building energy and equipment costs. This research uses Weka [18] for data mining, based on the Knowledge Discovery in Databases (KDD) process. The aim is to predict multiple room comfort temperature. This technique can also assist to:

- Make recommendations on the number and types of sensors required to support optimum building intelligence.
- Carry out comparative analysis to develop generic rules on heating requirements for rooms, zones and buildings.

Data mining classification models have been successfully used for different kinds of data in many areas, such as the support of building control systems. The C4.5 Weka classification model was used in this research to develop relevant classification rules. Results of this analysis demonstrate the value of data mining to analyse large data sets and prove a useful tool in developing general patterns.

A development of classification data mining models is reported to study building performance data and develop user

comfort concepts. Multi-dimensional data of internal and external weather conditions are used to train and evaluate models. Similar data are then used to make predictions and develop classification rules. These rules may be used as precursors to building control algorithms. Under the same constraints, different rooms' performance may be analysed without necessarily maintaining extensive sensing systems.

II. RELATED WORK

Data mining is a sophisticated data search capability that uses classification algorithms to discover patterns and correlations within a large volume of data. This paper presents the selection and application of data mining techniques on maintenance data of buildings. The results of applying such techniques and potential benefits of utilising their results to identify useful patterns of knowledge and correlations to support decision making that will improve the management of building life cycle processes are presented and discussed.

In engineering, data mining is particularly useful in situations where the physics is either not well understood, mathematically elaborated or where physical models are not available or accessible. Often data mining is used to quickly obtain comparable results to physical models. Knowledge Discovery in Databases (KDD) is one data mining approach that is appropriate for large databases.

KDD refers to the broad process of finding knowledge in data, and emphasises the "high-level" application of particular data mining methods [6]. It has attracted a great deal of attention in the information industry, due to the need for turning large amounts of data into useful information and knowledge. Using results from simulation engines, KDD has been used in fault detection and energy prediction by testing different machine learning algorithms [2]. Genetic algorithms have been used for optimisation and multi-criteria analysis of building designs and performance [11], [17], [19].

The process uses a variety of data analysis tools, or algorithms, to discover patterns and relationships in data that may be used to make valid predictions. The tools, or algorithms, predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. It can answer business questions that traditionally needed a long time to resolve. These algorithms' are used to search through databases for hidden patterns, finding predictive information that experts may overlook because it lies outside their expectations and vision.

Data Mining divides the KDD process into three parts – pre-processing, data mining and post-processing [8], [17]. Fig. 1 presents the research steps based on the KDD pre-processing part, and includes data Selection, Sampling, Cleansing and Transformation. The pre-processing part may involve attribute importance analysis, evaluation of statistical independence and cross correlation. The Data Mining part analyses data and generates outputs. The post-processing part includes Visualisation and Evaluation of mining results.

Data Mining and the KDD processes research in building performance analysis is rare and only beginning to be developed in the built environment data analysis domain.

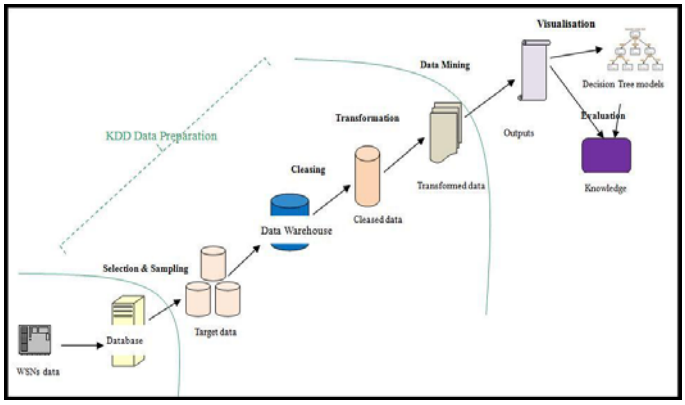

Figure 1. The standard KDD process outline the main activities undertaken

Morbitzer et al. [14] applied data mining to obtain information from simulation results. Their approach uses clustering as a particular useful analysis technique and illustrates its potential in enhancing the analysis of building simulation performance predictions.

On the other hand Reffat et al. [16] investigated the potential of applying data mining techniques on maintenance data of buildings to identify impediments to facilitate better performance of building assets. Reffat et al demonstrated what knowledge can be found in maintenance records. Decision rules are identified that may be used to improve building performance, satisfy occupants and minimise the operational cost of maintenance.

III. CASE STUDY FOR BUILDING PERFORMANCE ANALYSIS

This research was carried out using data compiled from sensor meters which are installed in the building of the Environmental Research Institute (ERI), University College Cork, Ireland [10]. It is a full scale test bed that looks not only at how sustainable buildings can perform but also at the methods used to assess a building's performance.

Data is streamed from sensors into a central data warehouse system. The Building was inaugurated in 2006 and has offices and laboratories for over one hundred researchers. Individual rooms have wired and wireless sensors to measure temperature, carbon dioxide, humidity and lux levels. Since the summer of 2007, over 20 million records have been collected.

A. Dataset Introduction

Two of the ERI rooms, the *Analytical Chemistry Lab (ACL)* and *Open Plan Office (OPO)*, which are located on opposite sides of the ERI building, were selected as test cases for this paper. These rooms presented this research with a comparison of analysis outputs from north and south facing rooms.

Data relevant to the laboratory room in the ERI includes under floor inflow and outflow heating water measurements, radiant and room temperature, carbon dioxide and humidity records. External environmental conditions are recorded by the ERI weather station through its BEMS. All data records from sensors are taken every 15 minutes. However, as in any other

TABLE I. DESCRIPTION OF MODEL INPUT ATTRIBUTES

Attributes	Explanation	Sample Values
READINGTIMESTAMP	Data time stamp.	29-AUG-08 08:30:00.000 000000
V_ACL_CO2, V_OPO_CO2	CO2 level for room ACL and OPO, CO2 (ppm)	404.62
V_ACL_HUMIDITY, V_OPO_HUMIDITY	Humidity for room ACL and OPO, Humidity (%)	57.8
V_ACL_ROOM_TEMP, V_OPO_ROOM_TEMP	Room temperature for room ACL and OPO, degrees Celsius, categorized to four comfort levels as a nominal data type.	23.42, (cold/ moderate/ comfortable/ hot)
V_OUTSIDE_AIR_HUMIDITY	The outside air humidity, Humidity (%)	70.05
V_OUTSIDE_AIR_TEMPR	The outside air temperature, Temperature (degrees Celsius)	14.42
V_OUTSIDE_WIND_SPEED	The outside wind speed, Speeed (m/s)	0.2
DM_UF3	The under floor input flow temperature, Temperature (degrees Celsius)	29.61
DM_UF4	The under floor output flow temperature, Temperature (degrees Celsius)	28.87

database, there are gaps where data are not collected for one reason or another. The input characteristics are as follows:

- A four-season (one year) period was selected.
- About 68,000 data points were used.
- Seven input attributes were used as shown in Table I.

B. *Data Pre-procession*

A sensor network's data streams almost inevitably present complex issues related to data quality. Data is often missing, and when available, may be subject to potentially significant noise and calibration effects. In addition, because sensing relies on some form of physical coupling, the potential for faulty data is a concern. It is necessary to ensure that the data integrity is maintained and that samples remain representative.

Classification algorithms do not appropriately treat NULL values and they must therefore be removed [3]. Data gaps exceeding two hours were deleted from the set while other missing data were estimated using the cubic spline interpolation method [4]. The usual measures of central tendency and spread were tested to ensure that data represented the case in point. The cleansed data set has 29,887 records for each room with the same timestamps.

The target measure is room temperature which is strongly correlated with room air temperature, thus the room air temperature was removed as an attribute. The Chartered Institution of Building Services Engineers (CIBSE) environmental design guide recommends comfort temperature in offices as 21-23ºC [1]. Room temperatures were modified into nominal values of cold, moderate, comfortable and hot, based on this standard.

C. *Data Mining*

This research used decision trees which is one of the classification methods implemented by Weka. The process of classification and data mining is shown in Fig. 2.

From the data mining process different parameters and methods within the J48 were implemented to achieve an optimised pruned tree. J48 algorithm is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, e.g., humidity or temperature.

Following a path from the root to the leaves of the tree, a sequence of such tests is performed resulting in a decision about the appropriate class of comfort. The decision trees are constructed in a top-down fashion by choosing the most appropriate attribute each time. An information-theoretic measure is used to evaluate features, which provides an indication of the "classification power" of each feature. Once a feature is chosen, the training data are divided into subsets, corresponding to different values of the selected feature, and the process is repeated for each subset, until a large proportion of the instances in each subset belong to a single class. Decision tree induction is an algorithm that normally learns a high accuracy set of rules.

In this research, other algorithms such as the M5 tree and the Naïve Bayes methods were used. What is reported here is only the best method, which for our case was the J48 classification algorithm.

The data classification was carried using training and test data sets. Weka J48 model provides a mechanism to split data into training, testing and cross-validation sets as shown in Fig. 3. Adjustments were carried out to determine the most optimal division between test and training data sets.

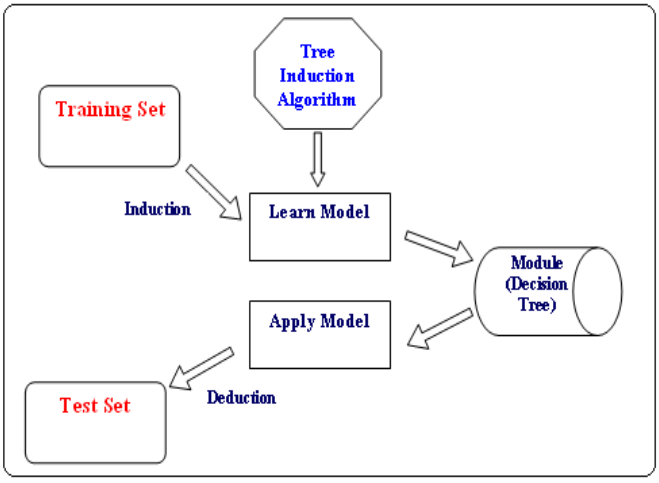

Figure 2. Decision Tree working process

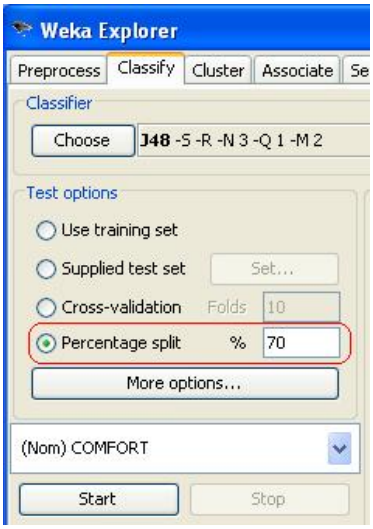

Figure 3. Decision Tree split percentage sample

IV. RESULTS AND ANALYSIS

The evaluation on the test split part gives the error levels when applying the classifier to the training data from which it is constructed. Table II presents the evaluation summary of room ACL test dataset with both external and internal attributes. The most significant data within this table are the numbers of correctly and incorrectly classified instances, and the mean absolute error. The incorrectly classified instances percentage and the mean absolute error rate were expected to be as small as possible.

TABLE II. SUMMARY STATISTICS OF THE DECISION TREE MODEL EVALUATION RESULTS OF ROOM ACL WITH EXTERNAL AND INTERNAL ATTRIBUTES

=== ACL Evaluation Summary ===		
Name	**Number of instances**	**Percentages/rate**
Correctly Classified Instances	6440	79.8116 %
Incorrectly Classified Instances	1629	20.1884 %
Mean absolute error		0.1307
Root mean squared error		0.284

Table III presents the evaluation summary of room ACL test dataset with only internal attributes.

TABLE III. SUMMARY STATISTICS OF THE CLASSIFICATION MODEL EVALUATION RESULTS OF ROOM ACL WITH INTERNAL ATTRIBUTES

=== ACL Evaluation Summary ===		
Name	**Number of instances**	**Percentages/rate**
Correctly Classified Instances	5734	71.0621 %
Incorrectly Classified Instances	2335	28.9379 %
Mean absolute error		0.182
Root mean squared error		0.327

TABLE IV. SUMMARY STATISTICS OF THE DECISION TREE MODEL EVALUATION RESULTS OF ROOM OPO WITH EXTENTAL AND INTERNAL ATTRIBUTES

=== OPO Evaluation Summary ===		
Name	**Number of instances**	**Percentages/rate**
Correctly Classified Instances	6507	80.642%
Incorrectly Classified Instances	1562	19.358 %
Mean absolute error		0.1211
Root mean squared error		0.2819

TABLE V. SUMMARY STATISTICS OF THE CLASSIFICATION MODEL EVALUATION RESULTS OF ROOM OPO WITH INTERNAL ATTRIBUTES

=== OPO Evaluation Summary ===		
Name	**Number of instances**	**Percentages/rate**
Correctly Classified Instances	5038	70.2356 %
Incorrectly Classified Instances	2135	29.7644 %
Mean absolute error		0.1865
Root mean squared error		0.3343

Using the same test procedure, the room OPO's Incorrectly Classified Instance rate with external and internal attributes is 19.358%, and the Incorrectly Classified Instance rate with only internal attributes is 29.7644%. The Mean Absolute Error with external and internal attributes is 0.1211, and the Mean Absolute Error with only internal attributes is 0.1865. See Table IV and V. From the output data above, compared with the difference between the Incorrectly Classified Instance rate and Mean Absolute Error for each room, a significant increase is observed while the external attributes were removed from the analysis cases. This translates that external environmental parameters of a building do influence the prediction of room temperature.

The Confusion Matrix part illustrates, for each class, how instances from a relevant class received various classifications. The output with external and internal attributes of ACL is considered as an example. See Table VI.

In Table VI the columns represent the predictions, and the rows represent the actual class. E.g. for class "a", it shows that 3644 instances were correctly predicted as "comfortable" level. The table also shows that 80 instances were correctly predicted as "cold" level. Correct predictions always lie diagonally on the table. Conversely, on the second row, it shows that 520 instances were predicted as a "comfortable" level when they were, in fact, a "moderate" level. It can be seen here that the incorrect instance numbers are always much less than the correct numbers. This shows this decision tree method has desirable accurate results. Hence the model is good at predicting the ACL room temperature comfort level.

In decision tree classifier, rules are a way of representing information. A rule-based classifier uses a set of IF-THEN rules. An IF-THEN rule is an expression of the form "IF condition THEN conclusion" [9]. In the tree model, each leaf is equivalent to a classification rule, each branch represents a rule.

TABLE VI. RESULTING CONFUSION MATRIX WITH THE COUNT FOR THE CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES OF ROOM ACL

=== Confusion Matrix ===					
a	***b***	***c***	***d***	←	***classified as***
3644	417	260	4	|	***a*** = comfortable
520	**1508**	18	9	|	***b*** = moderate
362	5	**1208**	0	|	***c*** = hot
4	28	2	**80**	|	d = cold

TABLE VII. RULES EXTRACTED FROM THE DECISION TREE MODEL TO SUPPORT CONTROL WHERE THERE ARE NO TEMPERATURE SENSORS

Attributes	Under Floor water outflow temperature (UF3), Under Floor water inflow temperature (UF4), ACL room CO_2 (AC), ACL room Humidity (AH), Outside Air Temp (OAT), Outside Air Humidity (OAH), Outside Wind Speed (OWS)
COMFORT STATE	**RESULTING RULES**
Hot	IF AH>51.18, AND AWS<=3.87, AND OAH<=80.87, AND AC>417.59, AND OAT<=12.21
Comfortable	IF AH>43.32, AND OAH>76.26, AND UF3<=23.11, AND OAT>12.21
Moderate	IF 43.35=>AH>41.13, AND 417.59=>AC>413.27, AND UF3<=27.62, AND OWS >2.04, AND OAT<=12.21
Cold	IF 23.54=>UF3>20.99, AND 94.65=>OAH>75.32, AND 41.13=>AH>34.66, AND OAT<7.88, AND 430.69=>AC>419.39, AND OWS>1.74

Using the ACL dataset which includes external and internal attributes as an example, Table VII lists some of the rules which have a 100% accurate result from the decision tree as the samples for each comfort level to express how the rules are represented. For instance, the second rule in the table below can be described as: IF the ACL room Humidity is greater than 43.32%, AND the Outside Air Humidity is greater than 76.26%, AND the Under Floor water outflow temperature is less than and equal to 23.11 degrees Celsius, AND the Outside Air Temperature is greater than 12.21 degrees Celsius, THEN the ACL room comfort level is Comfortable.

V. CONCLUSION

Building performance analysis provides mechanisms to efficiently operate and maintain buildings. The necessity of performance analysis becomes more apparent with a drive to efficiently utilise energy sources and reduce energy consumption. To achieve objectives for energy efficiency and building energy costs, it is necessary to accurately measure and evaluate information contained in sensed building data.

Data mining has proved to be the best approach to analyse huge quantities of data especially where patterns are not obvious. In this research, data mining classification is applied to wireless and wired sensor measurements.

Amongst data mining tools, classification models were chosen as the most appropriate tool to predict comfort under different environmental conditions. External and internal conditions were analysed to obtain trends and gain information from these relevant data sets. From the results it was established external conditions strongly affect a building's internal environment. This was particularly valuable since it is in agreement with the underlying building physics. External heat gains from sunlight, wind impacts and external temperatures influence internal comfort conditions.

The objective of this analysis supports evaluating energy consumption patterns, analysis of comfort requirements, as they relate to heating and cooling systems, and predicting sensor network requirements in similar buildings. By relating external to internal environmental conditions, it was possible to evaluate energy requirements under different external conditions.

As is often the case, heat generated within buildings is wasted or disposed of unless there are mechanisms for predicting its generation, demand and usage behaviour. By carrying out data mining analysis it was possible to generate recommendations for control strategies as feedback mechanisms and modifications to control processes. The intelligent control of an actuator system would positively affect the energy consumption of buildings.

Again, by comparing these conditions in north facing and south facing rooms, it was possible to evaluate the impacts of external conditions on room temperature. In situations where physical models are not available or expensive, data mining may be used to achieve a similar result.

Sensor network placement optimisation is also another factor of importance in building performance monitoring. If there is no added value to additional placement of sensors or if existing sensors can be replaced, savings can be made. This research demonstrates that where conditions are similar, individual room measurements can be generalised for a building or a set of buildings.

The results given in relation to temperature, humidity, CO_2 etc. from sensors that determine user comfort, and based on decision tree analysis, will allow managers to determine the optimum usage of energy. This will allow operators fine tune energy equipment to the extent that not only can the information be used for one building but also be used for buildings of a similar construction and energy provision. This means that based on parameters determined by one particular building, designers and facility managers can apply a model based on an existing building to similar building stock.

REFERENCES

[1] CIBSE, Guide A, "Environmental design", London: Chartered Institution of Building Services Engineers, 1999.

[2] D. Holcomb, W. Li, and S. A. Seshia, "Algorithms for Green Buildings: Learning-Based Techniques for Energy Prediction and Fault Diagnosis", Google Scholar, UCB/EECS-2009-138, 2009.

[3] E. Frank, and I. H. Witten, "Data Mining: Practical machine learning tools and techniques with Java implementations", Morgan Kaufmann, 2005.

[4] F. N. Fritsch, and R. E. Carlson, "Monotone piecewise cubic interpolation", SIAM Journal on Numerical Analysis 17, no. 2, pp.238-246, 1980.

[5] G. Piatetsky-Shapiro, C. J. Matheus, P. Smyth, and R. Uthurusamy, "Kdd-93: Progress and challenges in knowledge discovery in databases", AI magazine 15, no. 3, pp.77-82, 1994.

[6] G. Piatetsky-Shapiro, W. J. Frawley, S. Brin, R. Motwani, J. D. Ullman, C. C. Aggarwal, P. S. Yu, B. Liu, and W. Hu, "Discovery, Analysis, and

Presentation of Strong Rules", In Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA, 16, pp.191-200, ENST, 2005.

[7] H. Doukas, C. Nychtis, J. Psarras, "Assessing energy-saving measures in buildings through an intelligent decision support model", Building and environment, 2008.

[8] J. Han, and M. Kamber, "Data mining: concepts and techniques", Morgan Kaufmann, 2006.

[9] J. R. Quinlan, "C4. 5: programs for machine learning", Morgan Kaufmann, 2003.

[10] K. Menzel, D. Pesch, B. O'Flynn, M. Keane, and C. O'Mathuna, "Towards a Wireless Sensor Platform for Energy Efficient Building Operation", Tsinghua Science & Technology, pp.381-386, 2008.

[11] L. G. Caldas, and L. K. Norford, "Genetic algorithms for optimization of building envelopes and the design and control of HVAC systems", Journal of Solar Energy Engineering, 2003.

[12] M. Himanen, "The Intelligence of Intelligent Buildings: The Feasibility of the Intelligent Building Concept in Office Buildings", VTT PUBLICATIONS, 2003.

[13] M. V. Assen, G. V. D. Berg, P. Pietersma, "Key Management Models – The 60+ models every manager needs to know", Prentice Hall, 2009.

[14] C. Morbitzer, P. Strachan and C.Simpson, "Application of Data Mining Techniques for Building Simulation Performance Prediction Analysis". Proc. of Eighth IBPSA. Eindhoven, Nehterlands, 2003.

[15] P. N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining", Pearson Addison Wesley Boston, 2005.

[16] R. Reffat, J. Gero and W. Peng, "Using data mining on building maintenance during the building life cycle", In Proceedings of the 38th Australian & New Zealand Architectural Science Association (ANZASCA) Conference, pp. 91–97, 2004.

[17] T. T. Chow, G. Q. Zhang, Z. Lin, and C. L. Song, "Global optimization of absorption chiller system by genetic algorithm and neural network", Energy & Buildings 34, no. 1, pp.103-109, 2002.

[18] University of WAIKATO, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java", Weka, Available online: http://www.cs.waikato.ac.nz/mll/weka/, 2009.

[19] W. Huang, and H. N. Lam, "Using genetic algorithms to optimize controller parameters for HVAC systems", Energy & Buildings 26, no. 3, pp.277-282, 1997.