An integrated approach for operational knowledge acquisition of refuse incinerators

Yang-Chi Chang *, Po-Chuan Lai, Meng-Tsung Lee

Department of Marine Environment and Engineering, National Sun Yat-sen University, 70 Lien-Hae Road, Kaohsiung 804, Taiwan, ROC

Abstract

Refuse incinerator operation poses a tremendous challenge for efficient supervision due to the highly complexity of physical and chemical mechanisms inside the systems. It is difficult to comprehend operational knowledge without thorough study and long-term on site experiments. Fortunately, many sensors are installed in incineration plants and tremendous amounts of raw data about daily practices and system states are recorded to assist operations. Without proper analysis, however, these data are not beneficial to operators. An integrated approach is adopted in the current study using feature selection and data mining techniques. Feature selection was initially applied to cope with the heavy computation burden due to the huge data set. Data dimension can be reduced by discarding redundant information and leaving only relevant features for further analysis. Data mining analysis is then utilized to build two decision tree models based on steam production and NO\textsubscript{x} emission target attributes. Implicit incinerator system relations, represented by production rules and predicting accuracies, can be acquired from the decision tree models. Such rule-based knowledge is expected to facilitate on-site operations and enhance refuse incinerator efficiency.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Rule-based knowledge; Refuse incinerator; Data mining; Feature selection; Decision tree classification

1. Introduction

Although trash incineration is an efficient and primary method for dealing with daily trash in Taiwan, some critical problems still exist. For example, most incinerators are equipped with heat recovery systems consisting of a boiler unit and a cogeneration process, and current cogeneration benefits have become the most important income for refuse incinerator systems (Chung & Poon, 1996; Sonesson, Bjorklund, Carlsson, & Daleno, 2000). Managers must therefore search for a solution of how to keep highly efficient power generation within air pollution control. Although some physical and chemical mechanisms within the incinerator were studied, the research results may do little help to the incineration manager due to lack of long term on-site experiments. Otherwise, even though standard controls can handle ordinary operations, some system parameter adjustments are still needed in a variety of situations. However, those knowledge patterns or experiential rules are not easily discovered without proper analyses.

Many sensors are installed in a refuse incineration plant to support a highly efficient operation through system monitoring. The rationale is to embed an implicit system characteristic about plant operation in the collected raw data. However, it is infeasible for a manager to identify the desired information by either data observation or simple calculations. Therefore, without proper analyses, system monitoring and data collection are useless.

Fortunately, the difficulty of retrieving useful information through huge data sets can be overcome by advanced computer power and information techniques. An integrated approach was adopted by the current study using feature selection and data mining techniques. Feature
selection was initially applied to solve the problem of heavy computation burden due to numerous data attributes in the raw data set. Data can be reduced by discarding redundant information and leaving only relevant attributes or features for further analysis. The data mining (DM) technique was then utilized to build two decision tree models based on steam production and NO\textsubscript{x} emission target attributes. The implicit relations between each incinerator sub-system, represented by production rules and predicting accuracies, can be acquired from the decision tree models. Such rule-based knowledge is expected to facilitate on-site operations and enhance refuse incinerator efficiency.

2. Methodologies

2.1. Feature selection techniques

Real world data are often large and more complex in structure than experimental data. Thus, it would take a tremendous amount of computational effort to analyze real world data. Computation time is one of the prime factors affecting real world data application. With rapid growth of data collection by automatic sensors or recorders, the needed computational effort has exceeded current computer capability for conducting such data analysis. Some software approaches have been developed to solve real world problems, such as principal component analysis (PCA), discrete wavelet transform (DWT), feature selection, etc. Among those approaches, feature selection is considered to perform better in real world applications (Liu & Setiono, 1998).

Feature selection is also referred to as “feature subset selection” in some literatures. It is important to pattern classification or data mining in the machine learning domain due to its speedy learning process capability. In addition to finding relevant features, feature selection can also greatly improve classification performance, and remove redundant or irrelevant information (Blum & Langley, 1997; Pudil, Novovičová, & Šomol, 2002). The feature selection algorithm is valuable for improving prediction accuracy while reducing attribute numbers (John, Kohavi, & Pileger, 1994; Kohavi, 1995).

The Sequential Forward Floating Search (SFFS) algorithm proposed by Pudil, Novovičová, and Kittler (1994) is a revised algorithm based on a “sub-optimal” feature subset selection. It is proven to be a better feature selection algorithm than other feature selection algorithms (Anil & Douglas, 1997). Three major steps are identified in the SFFS algorithm: the first is “Inclusion”, the second is “Test”, and the last is “Exclusion”. SFFS begins with the inclusion process to select a feature with best performance. The searching process is followed by conducting a test on every feature selected in the same iteration to specify features that will decay the whole performance. If such a feature exists, SFFS would commence the exclusion process to ignore such a feature. The algorithm will continue looking for other better features until all features are examined.

The simplified SFFS flow chart is shown in Fig. 1 (Pudil et al., 1994; Somol, Pudil, Novovičová, & Paclík, 1999).

2.2. Classification analysis in DM

DM, the emerging information technique, effectively deals with the discovery of hidden knowledge, unexpected patterns, and implicit database rules (Adriaans & Zantinge, 1998; Han & Fu, 1999; Tung, Huang, Chen, & Shih, 2005). Raw data is seldom of direct benefit to people. The true value of data, contributing useful information for decision support and interpreting the phenomena represented by data sources, can be shown by the DM technique. Classification analysis is a DM technique, often used to establish relationships between classes and attribute value via a set of training instances. Given a set of finite and pre-determined class values, the classes in a classification problem are the ultimate objectives to be determined. Meanwhile, attribute values of each instance represent the instance’s characteristics, which affect potential classification outcome. Among the available classification methods, decision tree analyses have been used extensively in terms of predictive accuracy, computational speed, robustness, scalability, and interpretability (Elmasri & Navathe, 2000; Giordana & Neri, 1995). Therefore, the current study has also applied the decision tree analysis for acquiring operational knowledge in the designated incinerator plant.

A decision tree is a flow-chart-like tree structure, where each internal node represents an attribute test, and each branch represents a test outcome, and leaf nodes represent classes. The basic decision tree induction concept is based on a greedy algorithm that constructs decision tree in a top–down, recursive, and divide-and-conquer manner.
(Greene & Smith, 1993). Many decision tree induction algorithms have been developed, including ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), CART (Breiman, Friedman, Olshen, & Stone, 1984), CHAID (Kass, 1980) and OC1 (Murthy, Kasif, & Salzberg, 1994). Among the techniques, C4.5 is the most widely used in classification applications. C4.5 algorithm is a descendant algorithm of ID3 that employs an entropy-based measure as a heuristic for selecting the attribute that will best separate the samples into individual classes. The entropy is also referred to as the information gain in the C4.5 algorithm. The attribute with the highest information gain is chosen as the target attribute for each node. If the attribute $A_i$ is chosen as the branching attribute for training instances $T$, the concept of information gain of $A_i$ is written as the following equation:

$$\text{gain}(A_i) = E(T) - E(A_i),$$

where

$$E(T) = -\sum_j p_j \log_2(p_j) \quad \text{and} \quad E(A_i) = \sum_k (n_k/n)E(T_k)$$

$E(T)$ is the entropy of the set of instances $T$, and $p_j$ is the probability that an instance belongs to class $j$. $E(A_i)$ is the resultant entropy of $T$ if $T$ is partitioned according to $A_i$, in which $T_k$ is a disjoint subset of $T$ based on $A_i$’s value, $n$ is the total number of instances in $T$, and $n_k$ is the number of instances in $T_k$. With this algorithm, the information gain for each of the variables defining the samples in $T$ can be obtained. The variable with the highest information gain is considered the most discriminating variable of the given set.

3. Data collection and cleaning process

3.1. Data collection

The data set was acquired from the designated incineration plant in south Taiwan. There are 10755 records in the data, and each record has 509 attributes, respectively. According to the incineration plant functionalities, all attributes are mainly grouped into three sub-systems, including the refuse incineration sub-system, the cogeneration sub-system, and the air pollution control sub-system. The simplified flow chart for the incineration plant refuse process is shown in Fig. 2. The interaction mechanisms between these sub-systems are highly complex and non-linear. By observing the data or performing simple statistic analysis, it is impossible for the managers and the operators to grasp such relations.

The objective of this study is to discover the inherent mechanisms embedded in the original data set. The focus is on the interactions between the refuse incineration sub-system and two main system output variables, the steam production and $NO_x$ emission. The rationale is that the steam and the waste gas ($NO_x$), originating from the refuse incineration sub-system, are critical elements to incinerator management. Therefore, only refuse incineration sub-system attributes are taken into consideration as important parameters for current knowledge extraction modeling.

Data pre-processes are required before the data-driven analysis. The pre-process steps are often referred to as “data cleaning”, or removing outliers and inconsistent records, as well as pruning redundancies and abnormalities existing in the data set. Data cleaning avoids the “garbage in, garbage out” problem, meaning that if invalid data enter analysis processes, the resulting output will also be invalid and useless. There are many abnormal values within the original raw data, acquired from the incineration plant that may fail the data analysis process. Therefore, a two-stage cleaning process, including vertical cleaning and lateral cleaning, were used in this study to ensure the data integrity for subsequent analyses.

- Stage 1, Vertical Cleaning: Vertical cleaning removes abnormal attributes, with a missing data rate larger than 10% in the vertically aligned records. The results also imply that the sensors associated with abnormal data attributes should be checked and repaired if necessary.
- Stage 2, Lateral Cleaning: Lateral cleaning focuses on eliminating redundancies in the temporal scale. To correct such errors, some records reduplicated horizontally in the data set should be removed. In addition, some sensor malfunctions have been detected in the lateral cleaning process, and those records with many missing values are also removed.

After the data cleaning process, 3427 records remain for the feature selection and decision tree analyses, as demonstrated in the following section.

4. Feature selection and decision tree model

In many cases, previous experience, when applied for decisions, is transformed into lexical representation in a classification application (Helen, Alexander, & David, 2002). Classification can predict a categorical value of a specific data item, usually represented by a small number of ordinal levels, for example, high, medium, etc. The
advantageous partitioning of continuous values into ordinal values is stressed below (Helen et al., 2002):

- Ordinal values in classification may improve analysis results in classification applications.
- Ordinal information is rather simple compared to continuous information in the traditional decision tree model. It leads to a reduced number of rules and nodes with simultaneous error rate reduction.
- Also, an Ordinal problem can lead to a much richer set of conclusions and provide managers or operators a quick understanding of the phenomena.

All attributes are partitioned into three categories, high, medium and low, for a better classification in this study, and each level covers a specific numerical data range.

The refined attributes in the refuse incineration system are applied for feature selection after the data cleaning and categorizing process. The SFFS algorithm implemented by Robert P.W. Duin in his toolbox “PRTools,” was used for conducting the feature selection. The PRTools is a Matlab® based toolbox for pattern recognition, and is freely used for academic research. There are several feature selection algorithms supported by PRTools, and SFFS has proven to perform well in terms of computational efficiency among them by using 1-nearest-neighbor classification as the “criterion function” (Zongker and Jain, 1996). Results generated by the SFFS algorithm on the target attributes are shown in Table 1. Among numerous attributes in the refuse incineration sub-system, the SFFS algorithm selected five attributes most related to steam production and six attributes for NO\textsubscript{x} emission.

After data cleaning and the feature selection process, the selected attributes are then analyzed using the decision tree induction technique. The software Weka was applied to perform decision tree analysis in this study. Weka is a collection of machine learning algorithms for data mining implemented in Java and contains various tools for data preprocessing, classification (decision tree), regression, clustering, association rules, and visualization (Witten & Frank, 2005). It also provides many widely adapted decision tree induction algorithms for building a classification model, such as C4.5 and ID3. The current study preferred using the C4.5 algorithm for conducting the decision tree analysis due to its better computational efficiency.

Before commencing the decision tree analysis, some parameters must be assigned for the C4.5 algorithm. The “Confidence factor” parameter determines the confidence value to be used when pruning the tree, that is, removing branches that provide little or no statistical accuracy gain of the model. It has been proven that the default value of 0.25 works reasonably well in most cases, and therefore this parameter was left unchanged in this study (Murthy et al., 1994; Winston, 1992).

Another parameter needed for decision tree building is the minimum number of instances that must be presented for a new leaf to be expanded in the decision tree. A more generalized or specialized tree could be expected by adjusting this parameter. In general, a higher number of such parameters will create a more generalized tree and a lower number will create a more specialized tree. It was normally suggested that the number of instances be set to twenty per class for better performance (Kirchner, Tölle, & Krieter, 2004). The decision tree model for the steam production uses the parameter of 20 instances per class, as shown in Fig. 3.

A ten-fold cross-validation algorithm was applied for accuracy rate estimations of the decision tree models. Cross-validation is a method for estimating generalization accuracy based on “resampling” (Weiss & Kulikowski, 1991). When conducting k-fold cross-validation, the entire data set needs to be randomly divided into k subsets of equal size. The training processes will then be performed k times, where each time one of the k-subsets will be reserved for testing the model accuracy rate, acquired from training the remaining k-1 subsets. Prediction accuracy results are often used to evaluate decision tree performance.

Another method for accuracy rate estimation is the “split-sample” or “hold-out” method. In the split-sample method, only one validation subset is used to estimate generalization accuracy, while k different subsets are used recursively to estimate generalization accuracy in a k-fold cross-validation method. Therefore, the cross-validation method would provide more unbiased information in terms of the models’ accuracy for data mining applications (Elsner & Schmertmann, 1994; Rivals & Personnaz, 1999; Schaffer, 1993; Shao & Tu, 1995). Using 10-fold cross-validation, the current study found 86% predicting accuracy for the steam production model and 80% predicting accuracy for the air pollution (NO\textsubscript{x}) model, all within the acceptable range.

5. Rules generation

The primary advantage of a decision tree model for incinerator operation, compared with other black box algorithms, is its natural knowledge representation method. A decision tree model can be easily transferred to a production rule base using the form “IF antecedence \{x_1,x_2,x_3, \ldots, x_n\}, THEN consequence \{y\}”. In the previ-
Fig. 3. Decision Tree for the target attribute of the steam production.

CCT: Combustion Chamber Temperature
PAV: Preheating Air Temperature
QPA: Quantity of Primary Air
QPA: Quantity of Secondary Air
QPA: Quantity of Primary Air of Regulator
QSA: Quality of Steam Product Ion
QSP: Quality of Steam Production
RHV: Refuse Heating Value
ous IF-THEN rule, the antecedence set contains the branching attributes \( x_1, x_2, x_3, \ldots, x_n \) and the consequence \( y \) is the target attribute in the original decision tree. Each branch in the decision tree model is equivalent to an IF-THEN rule with a corresponding accuracy rate as shown in Fig. 4. There are 22 and 41 rules generated from the steam production and air pollution model decision trees, accordingly.

Since the IF-THEN rule bases are intuitive to humans, they powerfully facilitate on-site decision making. An incineration plant operator can use the heuristic rules to assess target attribute outcomes under various operational scenarios represented by branching attribute combinations in the antecedence set. Some of these attributes may not be manually adjusted in the central control room, since they indicate on-site initial status. For example, refuse heating value is prescribed depending on the incoming refuse material characteristic. Nevertheless, operators can manipulate controllable attributes to achieve desired outcomes. The IF-THEN rule presented in Fig. 3 is given below and its associated knowledge in terms of incinerator operation is explained in the following paragraph.

\[
\text{IF} \ (\text{CCT} \geq 1020.10) \ \text{AND} \ (6970.00 \leq \text{QSA} \leq 9462.00) \ \text{AND} \ (\text{QPA6} \geq 2530.00) \ \text{AND} \ (\text{RHV} \geq 2187.00) \\
\text{THEN} \ (\text{QSP} \geq 36.40)
\]

The rule demonstrated above is based on the steam production target attribute, interpreted as Fig. 5. Two initial conditions, the uncontrollable attributes of combustion chamber temperature (CCT) and refuse heating value (RHV), have been detected either by the sensors or human observation in advance. If CCT is larger than or equal to 1020.1°C and RHV is larger than or equal to 2187 cal/kg, the operators may need to maintain two specific attributes within a certain range for achieving quantity of steam production (QSP) larger than 36.4 tons/h. Here the two controllable attributes are quantity of secondary air (QSA) supply and quantity of primary air supply for roller #6 (QPA6). The operators in the central control room must adjust the associated controllers to maintain the QSA value between 6970 and 9462 cubic meters per minute (cmm) and the value of QPA6 larger than 2530 cmm. When applying this rule, the operators must also be aware of the embedded uncertainty in the rule, implying that this rule may not be correct all the time since the predictive accuracy is about 80%. In summary, the rule-based knowledge acquired from the decision tree models should be of great help for operators. Since only a few crucial attributes rather than numerous attributes deserve the operators’ attention, explicit results can be expected for better refuse incinerator operation.

6. Conclusions

This paper proposes an integrated approach using two data-driven analyses for operational knowledge acquisition for the refuse incinerator. Tremendous amounts of data were generated and recorded via digital revolution, through real-time incineration plant, monitoring, providing a sound basis for advanced analysis. Hidden information in huge data sets cannot be easily identified as knowledge to support operations or decision making unless proper and trustworthy methods are applied. In light of the need discussed above, the feature selection technique and the decision tree analysis were utilized to accommodate practicable computational effort and explicit knowledge representation issues. The feature selection technique is capable of detecting most relevant features and thus a more concise and precision data set can be acquired for better subsequent analysis. Decision tree induction analysis, one of the data mining techniques, is well-known for its outstanding rule-extraction and prediction ability.
The current study successfully created two sets of rule-based knowledge, focusing on steam production and NO\textsubscript{x} emission. Following the process of data cleaning and feature selection, two decision tree models were formulated based on the huge historical data. Decision tree performance was acceptable as evaluated by the accuracy rate index, using the 10-fold cross-validation algorithm. Finally, the more intuitive know-how domain expression, presented as a production rule, was derived from the decision tree models. The rules can greatly help incinerator plant on-site operators so that more efficient refuse incineration can be expected.

Future issues beyond the one described here include: (1) a finer division of continuous values into ordinal values, other than high, medium, and low; (2) other related target attributes besides steam production and NO\textsubscript{x} emission; and (3) a broader model formulation to cover all the cogeneration sub-system and air pollution control sub-system attributes if the computational capacity problem is resolved.

References