
DETECTION OF EVENTS CAUSING PLUGGAGE OF A COAL-FIRED BOILER: A DATA MINING APPROACH

ANDREW KUSIAK*
ALEX BURNS
SHITAL SHAH
NICK NOVOTNY

Intelligent Systems Laboratory, Department of
Mechanical and Industrial Engineering, The University
of Iowa, Iowa City, Iowa, USA

This paper presents an approach to analyze events leading to pluggage of a boiler. The proposed approach involves statistics, data partitioning, parameter reduction, and data mining. Two independent data mining algorithms have been applied to detect both static and dynamic relationships among the process parameters. The multi-angle data mining approach increases the ability to locate rare events as well as the reliability of the results. The proposed approach has been tested on a 750 MW commercial coal-fired boiler affected with an ash fouling condition that leads to boiler pluggage, thus resulting in unscheduled shutdowns. The cause of the boiler pluggage is not known. The rare event detection method presented in the paper identifies several critical time-based data segments that are indicative of the boiler pluggage. The events define a set of general guidelines that when followed should reduce the likelihood of boiler pluggage. The knowledge extracted by the data mining algorithm is an important component of an intelligent alarm system.

Keywords: Data mining, fault detection, combustion process, temporal data mining, boiler pluggage, process control

Received 12 August 2004; accepted 15 May 2005.

*Address correspondence to andrew-kusiak@uiowa.edu

INTRODUCTION

Rare events are often characterized by normal parameter conditions sequenced in abnormal combinations that are hidden in large volumes of routinely collected temporal data (Narayanswamy et al., 1998). They may lead to a catastrophic system failure. The ability to predict and avoid these rare events in time series data is a challenge that could be addressed by data mining approaches. Difficulties arise from the fact that often a significant volume of data describes normal conditions and only a small amount of data may be available for rare events. The latter poses a challenge to the traditional data mining algorithms. This problem is further exacerbated by the fact that traditional data mining does not account for the time dependency of the data. The approach presented in this paper overcomes these concerns by defining time windows and utilizing two independent data mining algorithms.

In this paper a data-driven approach rather than the physics-based approach is used. One of the advantages of the data-mining approach is its ability to generate previously unknown models using essentially unlimited number of quantitative and qualitative parameters. The discovered models are then validated and tested without the need for full understanding of the process physics. This quick discovery of usable models involving qualitative and quantitative parameters with data-mining algorithms may offer an effective paradigm for combustion research.

The key to successful detection of rare events with a data mining approach is in parameter reduction, data preprocessing, statistical analysis, time segmentation, and event labeling. These steps lead to partitioning the problem into a series of smaller sub-problems. Solving each of these sub-problems augments the probability of the rare event detection. Some of the rare events detection methods published in the literature are briefly discussed next.

Chance discovery in medicine can be viewed as locating some critical actions/occasions that describe dangerous possibilities. That is, rare but risky events occur with low prevalence, as shown by Tsumoto (2003), who developed various approaches for the detection of rare events and applied them to clinical data for motor neuron disease. The temporal aspect of data was handled by employing an extended moving average approach. In Tsumoto's approach, initially the overall average for each variable was calculated followed by the maximum and minimum values of each variable. A vector for each data record was created with respect

to the various time window values. The categorical parameters were classified as constant, ranking, and variable. Next a customized extended moving average approach was applied. The parameters were standardized into seven categories and their qualitative trends were classified into ten classes. A rule induction algorithm extracted from this new data set non-temporal, short-term, as well as long-term knowledge.

Narayanswamy et al. (1998) discussed a cascade of data elimination strategies for rare event detection. Template matching, neural networks, integrated optical density and morphological processing were examined for data processing. They applied these techniques to automated detection of aberrant cells in cervical smear slides. In the second stage processing their strategy had the overall event detection probability of 0.87, a significant improvement over the initial probability of 0.01.

The vibration in power plant machinery such as pumps, turbines, and so on, is monitored using online diagnostic tracking system. Vibration spikes assist in detection of process changes as well as rare but detrimental events leading to system shutdown. Branagan and Wasserman (1992) developed a two-phase algorithm for detecting vibration spikes. The deviation of amplitude from an exponential forecasting technique was used for initial screening of abnormal spikes. This was followed by classification of events into no-spikes, positive-going spikes and negative-going spikes using a pre-trained Gaussian probability-density-function-based neural network. The approach resulted in a high rate of detecting vibration spikes when applied to the overall-vibration and thrust-position data set.

The approach presented in this paper uses two concepts. The first is a rule induction algorithm that captures the subtle parameter relationships that cause the rare events to occur (Quinlan, 1986). Static events in the data are captured without distinguishing the trends of the parameter values, rather the current value. The time segmentation and time windows provide some insights into the time dependency of parameter fluctuations. The second concept uses data transformation and clustering to identify events based on the trends and value changes. The two independent data mining approaches increase the ability to discover events of interest.

This study concentrates on an ash fouling condition that causes boiler shutdowns (several times a year) on a 750 MW tangential coal-fired boiler. The ash fouling leads to pluggage in the reheater section of the boiler. Once the build up becomes substantial the boiler performance

is negatively affected. This leads to the derating and the eventual shut-down of the boiler. The cleaning of the boiler during the shutdown requires 1 to 3 days. At present there is no method to determine the level of ash pluggage without shutting down the boiler to physically inspect the area. Furthermore, the data analysis has shown that all parameters were within specifications, so there was no obvious single parameter indicating the pluggage.

Ash fouling is a significant issue in the energy industry. It has been reported that 37% of all pulverized coal-boilers are affected by frequent ash fouling problems (Valero and Cortes, 1996). Several approaches have been developed that continuously monitor coal properties utilizing electron scanning microscopes (ESM) (Erickson et al., 1995). This approach scans the coal prior to being fed into the boiler. The ESM locates coal chemical properties, such as high sulfur dioxide, that may lead to ash fouling. These types of approaches have demonstrated promising results, but their implementation is expensive and far from ideal. Coal treatment methods have also been investigated and have demonstrated some success. Vuthaluru (1999) experimented with wet mineral treatment and sodium compounds with mineral additives. The results demonstrated a reduction of ash fouling in various controlled experiments, but due to the constant variations of coal chemical composition and properties the implementation is difficult. The data mining approach overcomes these concerns by using existing control systems to alter the relationships between the boiler process parameters. The suggested parameter changes can be generated by an intelligent pluggage avoidance system.

To investigate the problem, data was collected for 173 different boiler parameters. This included flows, pressures, temperatures, controls, demands, and so on. The data was collected in one-minute intervals over the course of three months. The data collection began directly following a shutdown where the reheater section of the boiler had no pluggage on February 2002. The collection period ended approximately three months later on May 2002 when the boiler had to be shut down for pluggage removal. This data set contained over 168,000 observations. To test the accuracy of the proposed data mining approach, three confirmation data sets were considered. Two of the data sets exhibited the pluggage and one represented a best case scenario where no pluggage occurred. The two data mining approaches coupled with a statistical analysis and the comparison of time windows led to the discovery of relationships reducing the likelihood of pluggage. The

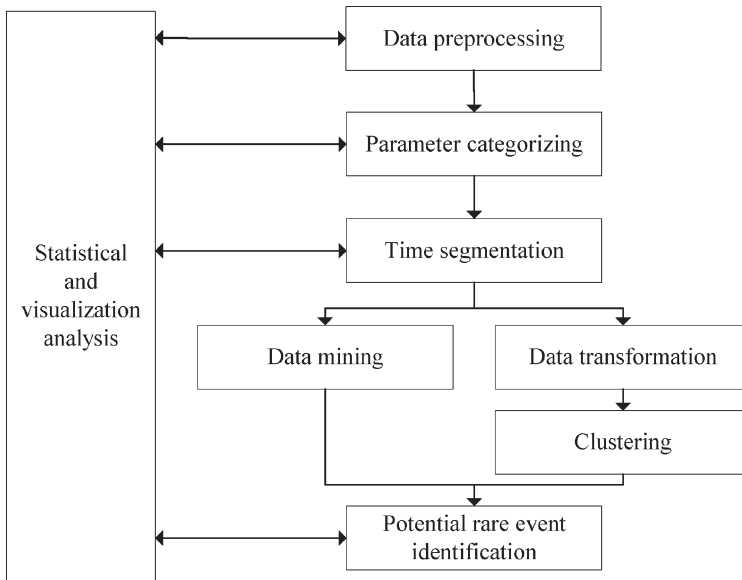


Figure 1. The procedure for detecting events causing boiler pluggage.

event detection method identified several specific events that may potentially cause the pluggage. The method for identifying and analyzing the rare events that lead to boiler pluggage are discussed in the next section.

THE EVENT DETECTION PROCEDURE

The rare events that contribute to the reheater pluggage can be detected by applying the five-step procedure discussed in this section. These steps include the categorization and reduction of the parameters, the time segmentation of the data, statistical and visual analysis to validate the results obtained in the two previous steps, the application of the rule inducing data mining algorithms, clustering, and the analysis of knowledge and validation (see Figure 1). This method fully integrates both the dynamic and static events to determine the cause of the pluggage.

Step 0: Data Preprocessing

The raw data was preprocessed for consistency, bad values, sensor errors, sensor calibrations, and so on. Parameter readings that were

above a defined threshold (e.g., above 10 standard deviations) were considered as unknown and were marked as “?”. Also, the data was transformed using a 20-minute moving average. This was necessary to smooth out the volatile oscillations occurring between data points.

Step 1: Parameter Categorization

Parameter categorization is the process of identifying all the relevant parameters as either impact or response parameters. Domain experts and statistical analysis were required to categorize the parameters. For example, computing correlation coefficients and standard deviations, and visually analyzing the data facilitated the categorization of the data. The categorization of parameters leads to the reduction of parameters, which decreases the computational complexity of the problem.

Response parameters are those that change values due to a rare event or a failure, e.g., an air leak in a pressurized chamber. Response parameters were ignored for the majority of the research; however, some are useful in the construction of time windows and a knowledge base. Thus, analyzing changes and shifts of response parameters may substantiate the designated time windows. Impact parameters are defined as parameters that are either directly or indirectly controllable and may cause the rare event. These are the parameters that are of greatest interest in determining rare events. Parameters that do not fall into a clear category, e.g., the parameters that are loosely related to the rare event and are not controllable are categorized based on the problem domain. An example is a “demand” parameter that is not inherently controllable, but may be relevant or impact the failure.

A response parameter that is highly correlated with the outcome could result in high prediction accuracy, but little or no knowledge gain to the users. Furthermore, the process of removing response parameters from analysis reduces the computational complexity and increases the generalization of the data mining algorithms.

All the 173 parameters, which included both response and impact parameters, were analyzed. The parameter list was reduced to include 26 impact parameters. This parameter categorization and reduction was accomplished with the assistance of domain experts, as well as statistical analysis such as correlation and multivariate analysis. The list of selected parameters is shown in Table 1.

Table 1. The reduced parameter list

Parameter	Parameter
NO3_O2_PROBE_M	GENERATOR_GROSS_OUTPUT_C_M
NO6_O2_PROBE_M	GENERATOR_NET_OUTPUT_CHA_M
PER_FUEL_FLOW_CHART_RECORDE_M	MAIN_STEAM_FLOW_M
COAL_FEEDER_101_FLOW_M	SUPHTR_DESUP_SPRAY_FW_F_2_M
COAL_FEEDER_102_FLOW_M	HOT_REHEAT_102_STEAM_T_M
COAL_FEEDER_103_FLOW_M	COLD_SECONDARY_AIR_P_M
COAL_FEEDER_104_FLOW_M	FD_FAN_101_AIR_FLOW_M
COAL_FEEDER_105_FLOW_M	FD_FAN_102_AIR_FLOW_M
COAL_FEEDER_106_FLOW_M	SECONDARY_AIR_FLOW_M
COAL_FEEDER_107_FLOW_M	W_TILT_M
COAL_FEEDER_TOT_FLOW_M	STACK_FLOW_M
E_TILT_M	SUPERHEAT_SPRAY_FLOW_CHA_M
GENERATOR_DEMAND_M	SODA_ASH

Step 2: Time Segmentation

Time segmentation deals with partitioning and labeling the data into time windows (TWs). A time window is defined as a set of observations in chronological order that describe a specified number of continuous observations. This step allows the data mining algorithms to account for the temporal nature of the data. The most effective method to segment the data is by determining/estimating the approximate date of failure and marking it as the last observation of the final time window. In this application the failure event was defined by the date when the boiler was derated due to the pluggage. The cause of the shutdown was confirmed through visual inspection of the affected region. This date was then set as the last day of the final time window (TW6).

The windows were set to be approximately 1 week long. A week was chosen for several reasons. First, the boiler was inspected approximately 1 month prior to its derating. During the inspection the reheater section of the boiler was completely free of pluggage. This information indicated that the pluggage required less than 1 month manifesting itself to the point of shutdown. It was hypothesized that the pluggage develops over several days. Based on this information, 1 week was deemed to be an adequate time window. One week also provided a sufficient number of observations (over 10,000 per window) for the data mining algorithms.

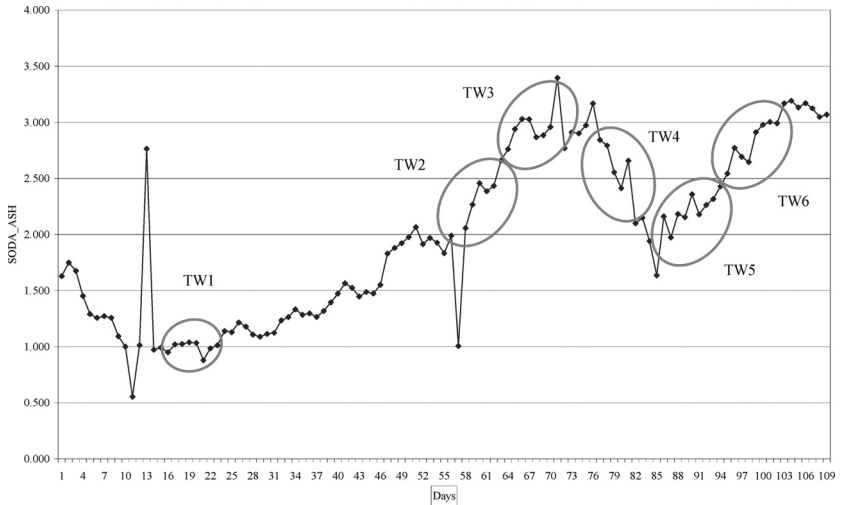


Figure 2. Time windows for ash fouling application.

Using the derate date and a 1-week-long time window, the data was divided into 6 time windows shown in Figure 2. Time window 1 (TW1) ensured that there was adequate data to describe the normal operating conditions.

The fact that the reheater section was visually inspected and determined to be free of any pluggage between TW3 and TW4 has significant impact on the study. This allows for the assumption that the events that led to the ash pluggage most likely occurred in either TW4 or TW5.

Step 3: Data Mining

Rule inducing data mining algorithms discover relationships among parameters and an outcome in the form of IF...THEN rules and other constructs (e.g., decision tables) (Pawlak, 1991; Quinlan, 1986). Data mining is a natural extension of more traditional tools such as neural networks, multivariable algorithms, or traditional statistics. In the detection of rare events, the rule induction algorithms are explored for two reasons. First, the algorithms generate explicit knowledge that is understandable to a user. The user is able to comprehend the extracted knowledge, assess its usefulness, and learn new and interesting concepts. Secondly, the data mining algorithms have been shown to produce highly

accurate knowledge in many domains. An example of a decision rule produced by the decision-rule algorithm is presented next:

*Rule 1: IF BOILER_MASTER \leq -0.53 AND AIR_MASTER
> -1.5 AND AIR_FUEL_RATIO > 0.13 AND AVG_MID_TEMP
> -0.42 THEN Time Window = 2[1120]*

This rule includes a premise and a conclusion, and is assigned metrics describing its quality, e.g., strength (the number in the square brackets). The relationships among parameters represented by the rules allow for the analysis of complex processes based on the data. The intuitive rules enhance understanding of the cause of rare events.

In this research, data mining algorithms were applied to generate a knowledge base and to predict the previously defined time windows (outcomes). The rules derived from the data mining algorithms define the parameter relationships and interactions that cause the failure events as well as normal operating conditions.

Step 4: Clustering

Clustering provides an alternative approach to locating and identifying the events that lead to pluggage. The clustering algorithm captures the dynamic events that may contribute to adverse events. Clustering observations required several steps listed in Figure 3. Heuristic algorithms were developed to reduce the complexity and expedite the search for rare events. The *k*-means algorithm was used for clustering the data (Aha, 1992).

Step 4.1. Data Transformation. Before the *k*-means clustering algorithm was applied, a data transformation step was performed to label the data (I = increasing, D = decreasing, SI = steadily increasing, SD = steadily decreasing, P = peak, V = valley, N = no change). The categorical values represent the dynamic nature of the parameters across the entire data set (i.e., how the parameter slopes are fluctuating from observation to observation) and comprise the dynamic event list.

The preprocessed data (Step 0) set was used to derive the dynamic event list. The data transformation is a multi-step procedure consisting of data normalization and 20-minute slope calculations. This procedure defines primary slope events and delta slope events. These slope events

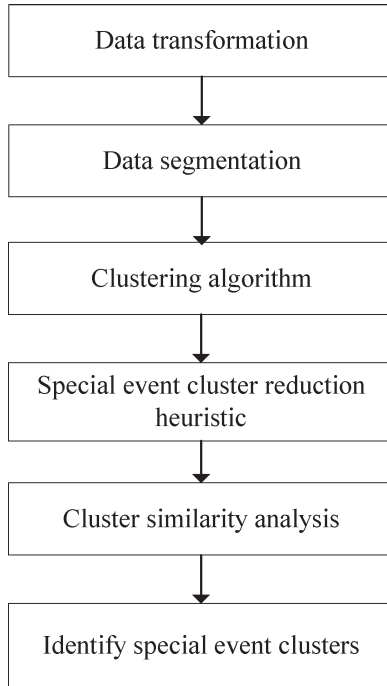


Figure 3. Clustering dynamic events method.

are used with a template-matching scheme to label each observation as a member of the dynamic event list.

The initial raw data for each parameter was normalized using the z-transform (Table 2). The normalization of the data improves the generalization of the transformation procedure. The slope of each parameter was generated for the previous 20 minutes. If the value of the slope was larger than the defined threshold (e.g., ± 0.001) then the primary slope event was assigned a value “1”. Conversely, if the observed negative change in slope was less than the negative threshold value, then the primary slope event was assigned a value “-1”. If the slope was between the thresholds the observation was assigned the value “0”. These values represent the primary slope events.

The delta slope events were generated by totaling the slopes for the previous 10 observations. These values were then compared to the threshold total slope (e.g., ± 0.15). This was performed to capture the trends that were potentially missed by the primary slope events.

Table 2. Example of data transformation

Time	P1 raw	P1 normalized	20 min. slope	Primary slope event	10 min. slope delta	Delta slope event	Final event list	Label pattern
70	65	-0.08	0	0	0	0	0	N
71	63	-0.08	0	1	0	1	1	SI
72	62	-0.08	0.01	1	0	1	1	SI
73	60	-0.08	0.01	1	0	1	1	SI
74	59	-0.08	0.01	1	0	1	1	SI
75	58	-0.08	0.01	1	0	1	1	SI
76	56	-0.08	0.01	1	0	1	1	I
77	54	-0.08	0.01	0	0	0	0	N
78	53	-0.08	0.01	0	0	0	0	N
79	50	-0.06	0.01	0	0	0	0	N
80	49	-0.06	0	0	0	0	0	N
81	45	-0.06	0	0	0	0	0	N
82	30	-0.15	0	0	0	0	0	N
83	20	-0.25	0	-1	0	-1	-1	SD
84	21	-0.25	-0.01	-1	0	-1	-1	SD
85	20	-0.25	-0.01	-1	0	-1	-1	SD
86	19.8	-0.25	-0.01	-1	0	-1	-1	SD
87	17	-0.25	-0.01	-1	0	-1	-1	SD

The delta slopes used the same assignment scheme as the primary slope events. The values of this column represent the delta slope event column in Table 2.

The two events, i.e., primary slope event and delta slope events were combined to generate the final event list. Here, if both slope events agree then the event list is assigned the corresponding event value (i.e., 1, 0, and -1). If one event has a positive or negative assignment and the other is a zero, then either 1 or -1 is assigned. In the case of conflict, i.e., one slope event determines positive trend and the other says negative trend, then a conservative policy of assigning no change (i.e., 0) is used. Thus the final event list is comprised of 1, 0, and -1 for each time period.

The final event list was scanned with the templates seen in Table 3. The event templates were used to transform the final event list from numeric values to more specific temporal events such as peaks, valleys, steadily increasing, steadily decreasing, increasing, decreasing, and normal. A peak is defined as a pattern where there is an increase in the slope and it remains steady for certain time period (1 to 8 units),

Table 3. Event templates

Event	Pattern	Event label
Peak	{1, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1}, {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1}, ..., {1, 0, -1}	P
Valley	{-1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1}, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1}, ..., {-1, 0, 1}	V
Steadily increasing	{1, 1, 1, 1, 1}	SI
Steadily decreasing	{-1, -1, -1, -1, -1}	SD
Increasing	{1}	I
Decreasing	{-1}	D
Normal	{0}	N

followed by a decrease in the slope. All the time units representing these patterns are labeled “P”. Steadily decreasing (i.e., “SD”) labels are assigned to the time periods when five consecutive time periods have negative slope events. If there are no event slopes defined then the time unit is assigned a normal label (“N”), i.e., no change. The patterns are scanned sequentially, by starting with the peaks, followed by valleys, steadily increasing, steadily decreasing, increasing, and decreasing patterns. If none of the above patterns are matched, then that time unit is assigned a label “N.”

Step 4.2. Clustering. The k -means clustering algorithm was applied to the event labeled data in each time window. The clustering algorithm was executed three times to produce 45 clusters for each time window. The original data set consisting of 165,000 observations was reduced to 270 clusters by the k -means clustering algorithm.

To identify clusters of data that potentially cause pluggage 45 different clusters per time window were analyzed. The TW4 clusters are especially important since the boiler pluggage was initiated during this time window.

The clusters of data potentially causing the blockage are referred to as adverse event clusters. It was assumed that the adverse event cluster(s) are represented by dynamically changing parameters exhibiting features (SI, SD, P, V). Also, it was assumed that the adverse event cluster(s) that caused the pluggage on the boiler tubes occurred over a fairly short period of time. In order to narrow the search among the 45 clusters, a heuristic approach was used accounting for the two assumptions. Nine

Table 4. Nine potential adverse event clusters from TW4

Cluster	Parameter 1	Parameter 2	Parameter 3	Parameter 22	Parameter 23	Parameter 24
1	N	SD	P	P	P	P
2	N	SI	SD	SD	SD	SD
3	N	SD	SI	SI	SI	N
4	N	SD	SD	SD	SD	N
5	N	SI	SD	SD	SD	SD
6	N	SD	SI	SI	SI	SI
7	N	SD	SI	SI	SI	SI
8	N	SI	SI	SI	SI	SI
9	N	SI	SD	SD	SD	SD

clusters conformed to the two conditions of the heuristic and were labeled as potential adverse event clusters (see Table 4).

Step 5: Analysis of Knowledge and Validation

After the completion of steps 1 through 4 several potential adverse event clusters were identified and rules and knowledge have been extracted from the original data set. An additional three data sets were analyzed to further validate and isolate the specific events that lead to the boiler pluggage (Table 5). The validation approach is discussed in the next section.

COMPUTATIONAL RESULTS

Several methods were utilized to validate the knowledge, including statistical and visual analysis, cross-testing, and correlation analysis. The previously discussed approaches, i.e., preprocessing, rule induction, and clustering were applied to the three confirmation data sets.

Table 5. Details of confirmation data sets

Confirmation data	Dates	Blockage
Data set 1	1/26/02–4/26/02	Yes
Data set 2	10/6/03–1/2/04	No
Data set 3	1/12/04–3/27/04	Yes
Original data	2/12/03–5/9/03	Yes

Statistical and Visual Analysis

All the parameters from the original data set were analyzed and graphed against the data from the confirmation data sets. Several interesting observations were made from these graphs. The graph of NO3_O2_PROBE_M is shown in Figure 4. There was no data for this parameter from confirmation data set 1, so it was ignored in this analysis. It is evident that confirmation data set 2 (no pluggage) is always lower than the two data sets that exhibited the pluggage.

The NO6_O2_PROBE_M graph shown in Figure 5 exhibits a trend similar to that of the NO3_O2_PROBE_M graph. The average value of NO6_O2_PROBE_M (excluding TW6) is 3.11 in data set 2, and the average value for the original data set and confirmation data set 3 is 3.14. The analysis of these trends leads to the general conclusion that lower O2 readings are advantageous.

The east and west tilts were also examined. The analysis of the E_TILT_M is shown in Figure 6. Some data in the confirmation data set 1 was missing. The graph demonstrates that the values of the E_TILT_M are lower in the period where the pluggage occurred. This trend is also present in the west tilt, but like the NO6_O2_PROBE_M, it is not as visibly obvious.

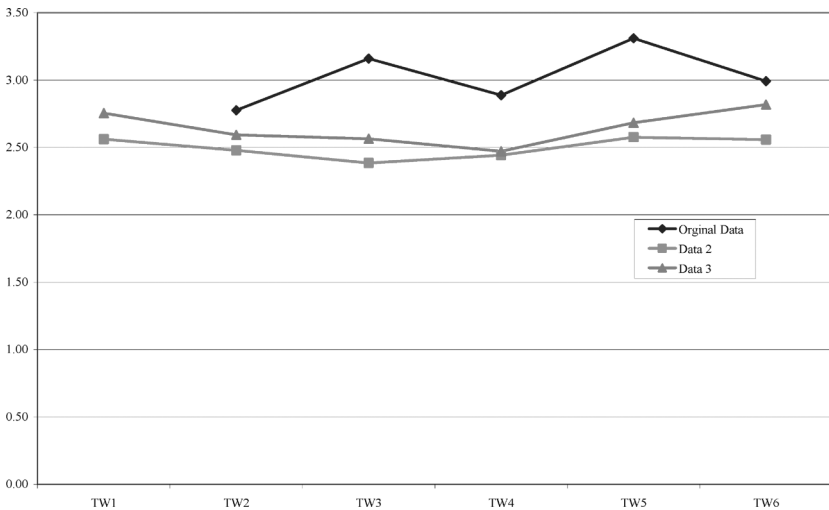


Figure 4. Comparison of NO3_O2_PROBE_M.

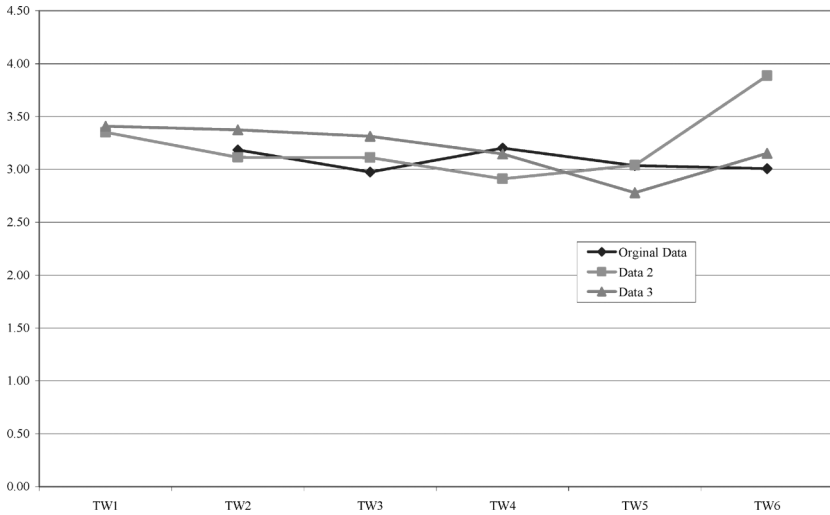


Figure 5. Comparison of NO6_O2_PROBE_M.

The analysis of the STACK_FLOW_M demonstrated an interesting trend (see Figure 7). The graph depicts the general relationship that lower values of stack flow are present in the data set without boiler pluggage.

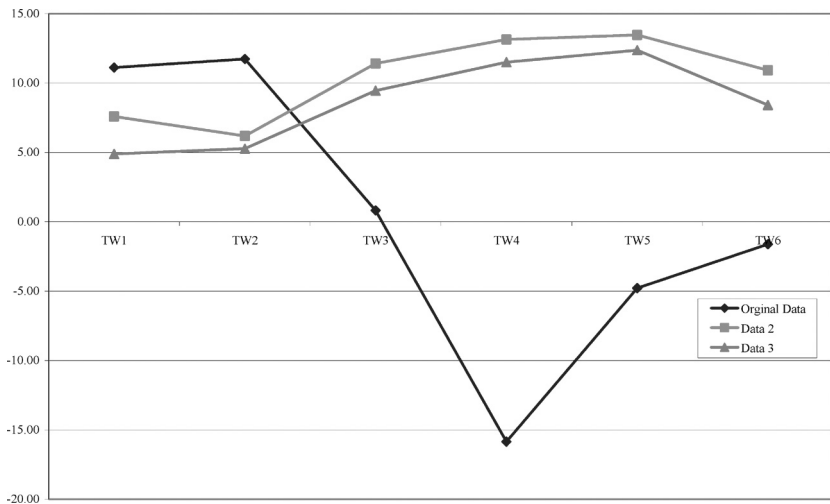


Figure 6. Comparison of E_TILT_M.

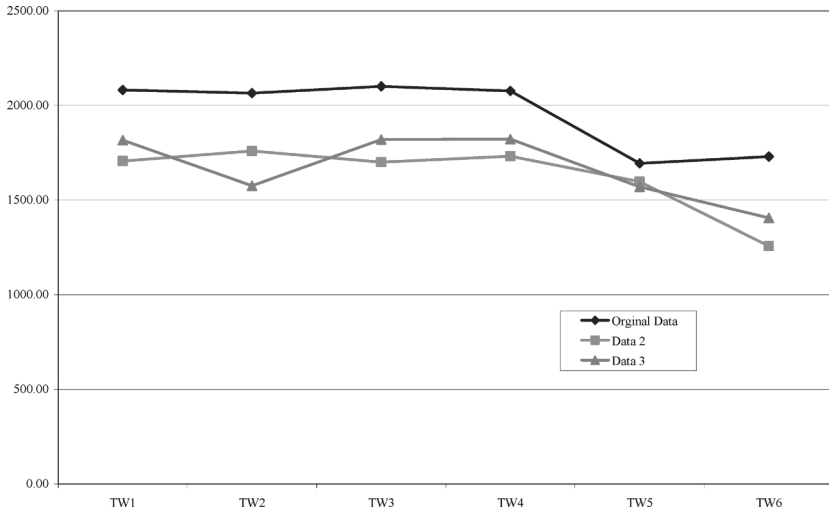


Figure 7. Comparison of STACK_FLOW_M.

Statistical and visual analyses have detected another relationship that relates to the coal feeders. In each of the three data sets affected by the pluggage at least one of the seven coal feeders was turned off for at least one time window (approximately one week). For data set 2 all feeders were on throughout the six time windows. This relationship may indicate that the pluggage could be due to changes in boiler output or sudden demand shifts.

The average value of each parameter of interest was computed as shown in Table 6. These values can be further examined to provide the basis for general guidelines and target values to avoid pluggage.

Rule Validation

The validation of the rules extracted from the original data set has two significant results. First, it allows an analysis of the quality of the rules

Table 6. Average values of parameters with and without pluggage

Boiler condition	NO6_O2_ PROBE_M	NO3_O2_ PROBE_M	E_TILT_M	W_TILT_M	Stack_ Flow_M
Pluggage	3.14	2.82	4.44	-3.79	1813
No pluggage	3.11	2.50	10.45	3.63	1625

Table 7. Cross-testing results for confirmation data set 1

		Predicted value					
		TW1	TW2	TW3	TW4	TW5	TW6
Actual value	TW1	4341	3697	391	212	0	0
	TW2	7256	1127	66	190	0	0
	TW3	7992	676	1401	7	4	0
	TW4	7140	1573	620	32	0	7
	TW5	8984	405	576	114	0	0
	TW6	5411	2166	1068	0	0	0

tested with the confirmation data sets. Second, it allows for the extraction and identification of specific events that may cause the pluggage.

The rules extracted from the original data set were tested independently with the three confirmation data sets. The rules were utilized to make predictions for the confirmation data sets.

In this case, the overall classification accuracy is not as important as the type and quantity of predictions. Since it is assumed that the events leading to the pluggage in the original data set occur in either window TW4 or TW5, the predictions for these time windows are the most significant in validation. It is hypothesized that there should be more predictions in TW4 and TW5 for the confirmation data sets 1 and 3 than the confirmation data set 2. The cross-testing results for each of the three confirmation data sets can be seen in Table 7 through Table 9.

The confusion matrices indicate that there are 559 predictions for TW4 and TW5 in the first confirmation data set and 284 predictions for TW4 and TW5 in confirmation data set 3. These results are even more significant given the fact that there were zero predictions for these two time windows in the second confirmation data set. These results indicate that the rules captured the unique relationships and events that lead to the boiler pluggage. These observations were extracted and analyzed to provide even more details of the events. These events will be compared with the events identified using the clustering technique in the next section.

Cluster Analysis

To validate the adverse event clusters derived from the clustering algorithm, the adverse event clusters were compared to clusters generated from confirmation data sets 1 and 2. Recall that boiler tube pluggage

Table 8. Cross-testing results for confirmation data set 2

		Predicted value					
		TW1	TW2	TW3	TW4	TW5	TW6
Actual value	TW1	514	863	2954	0	0	4309
	TW2	1447	2115	1036	0	0	5481
	TW3	1199	1083	6641	0	0	1179
	TW4	740	342	8899	0	0	99
	TW5	102	9	9949	0	0	20
	TW6	923	1729	5469	0	0	1959

was present throughout the entire confirmation data set 1 and that boiler tube pluggage was initiated during TW4 of the original data set. Given this, it is hypothesized that the TW4 potential adverse event clusters from the original data set should be highly similar to all time window clusters of confirmation data set 1.

The hypothesis was tested by comparing the TW4 potential adverse event clusters from the original data set to the clusters from confirmation data sets 1 and 2 and proved to be accurate. It was found that the TW4 clusters of the original data set matched 37 of 132 (28%) of the confirmation data 1 clusters. It was also found that the TW4 clusters of the original data set matched 21 out of 262 (8%) of the confirmation data set 2 clusters.

Overall, the cluster analysis involved comparing the nine TW4 potential adverse event clusters from the original data set to the remaining original data set clusters, the confirmation data set 1 clusters, and the confirmation data set 2 clusters (See Table 10). In each of the three indi-

Table 9. Cross-testing results for confirmation data set 3

		Predicted value					
		TW1	TW2	TW3	TW4	TW5	TW6
Actual value	TW1	0	0	3952	21	51	5977
	TW2	575	0	4791	0	86	4547
	TW3	0	0	4449	16	84	5451
	TW4	178	136	2717	26	0	6943
	TW5	651	0	3014	0	0	6335
	TW6	103	0	3121	0	0	6776

Table 10. Potential adverse event clusters from Table 3

Potential adverse event clusters	Cluster no.					
Original data set analysis	1	3	4	5	8	9
Confirmation data set 1 analysis	1	x	4	5	x	9
Confirmation data set 2 analysis	2	7	4	5	8	9

vidual analyses, three smaller subsets of TW4 potential adverse event clusters were identified as possible, genuine adverse event clusters. The TW4 potential adverse event clusters 4, 5, and 9 from Table 4 appeared in all three of the smaller subsets as seen in Figure 5. This indicates that clusters 4, 5, and 9 are the genuine adverse event clusters that most likely caused the boiler tube blockage in TW4 of the original data set.

Correlation

The cluster-based events were then compared and analyzed with the events discovered by the knowledge extraction approach. The knowledge extraction events (KEE) were identified by locating all observations from data sets 1 and 3 that were predicted in TW4 or TW5 (Table 7 and Table 9). The observations from each data set were extracted and sorted in chronological order. Observations that were located in similar time groups were labeled as knowledge extraction events. The process located five knowledge extraction events from confirmation data set 1 and eight from confirmation data set 3. These events along with cluster based events were analyzed using correlations. The events from each method that were highly correlated can be seen in Table 11.

This comparison has a significant impact. The clustering approach is a different method and the fact that there are similarities between the knowledge extraction events and the cluster events is extremely important. This increases the confidence in the cluster events and more importantly in the knowledge (i.e., rule sets) that was derived from the original data.

CONCLUSIONS

A rule-based data mining, statistical, and clustering approach was utilized to identify events that lead to ash pluggage in the reheater section of an industrial coal-fired boiler. This approach increased the understanding and the potential causes of the pluggage and provided a knowledge base

Table 11. Events with high correlation

Parameters	Cluster-based event			Confirmation set 1 KEE		Confirmation set 2 KEE			
	5	6	13	1	5	1	2	4	7
Per_FUEL_FLOW_CHART _RECORDE_M	88.9	90.4	70.2	49.7	69.1	99.3	101.2	100.6	100.5
COAL_FEEDER_101_FLOW_M	0	0	0	31.1	49.3	0	0	0	0
COAL_FEEDER_102_FLOW_M	64.9	65.8	55	53.8	26.8	64.4	75.3	76.7	76.4
COAL_FEEDER_104_FLOW_M	69	72.1	54.1	53.7	41.7	62.8	74.4	74.9	74.8
COAL_FEEDER_105_FLOW_M	62.6	63.2	49.4	52.4	40.2	57.3	0	67.8	67.9
COAL_FEEDER_106_FLOW_M	65.4	69.2	52.7	31.3	41.9	62.7	73.3	74.3	74.3
COAL_FEEDER_107_FLOW_M	65	71.1	44.6	52.4	41.5	64.1	75.1	0	0
COLD_SECONDARY_AIR_P_M	12.6	13.6	9.5	4.7	4.4	14.5	14.4	14.7	14.7
FD_FAN_101_AIR_FLOW_M	62.4	63.3	45	21.1	30.2	72.4	72.7	73.2	73
FD_FAN_102_AIR_FLOW_M	55.8	57.7	41.2	45.7	27	65.3	63	64.8	64.5
SECONDARY_AIR_FLOW_M	59.5	61.1	43.3	33.7	28.6	68.9	67.8	69	68.8
SUPERHEAT_SPRAY_FLOW _CHA_M	180.5	165.7	113.1	80.1	148.3	0	23.1	29.9	28.3

as well as guidelines that can be utilized to decrease the likelihood of the pluggage. The rules, knowledge, and events were verified using three confirmation data sets. The rules demonstrated the ability to detect and identify potential events in data sets that exhibited the pluggage. Three implementation strategies can utilize this information and analysis to further reduce the ash pluggage. The implementation ranges from immediate and basic rules in the form of guidelines to a fully automated intelligent pluggage avoidance system.

The study indicated that higher east and west tilt values (i.e., greater than 0) may reduce pluggage. The analysis also demonstrated that oxygen readings less than 3 are desirable for both the number 3 and 6 probes. Stack flows that are less than 1800 also appear to reduce the pluggage. Demand fluctuations that are characterized by shutdowns of one or more coal feeders may contribute to the ash pluggage. The significant rules along with the values that were derived by the statistical and visual analysis provide basic guidelines that can be immediately implemented to reduce the likelihood of pluggage.

An intermediate implementation approach can be achieved by automatically avoiding the events defined by the clustering and rule induction approaches. Avoiding these settings (Table 11) and investigating them

Table 12. Example control signature

Parameter	Event
Per_FUEL_FLOW_CHART_RECORDE_M	69.1
COAL_FEEDER_101_FLOW_M	49.3
COAL_FEEDER_102_FLOW_M	26.8
COAL_FEEDER_104_FLOW_M	41.7
COAL_FEEDER_105_FLOW_M	40.2
COAL_FEEDER_106_FLOW_M	41.9
COAL_FEEDER_107_FLOW_M	41.5
COLD_SECONDARY_AIR_P_M	4.4
FD_FAN_101_AIR_FLOW_M	30.2
FD_FAN_102_AIR_FLOW_M	27
SECONDARY_AIR_FLOW_M	28.6
SUPERHEAT_SPRAY_FLOW_CHA_M	148.3

further should additionally reduce the possibility of the pluggage. Furthermore, the settings could be incorporated into control signatures. The control signature for the confirmation data set 1 event 5 can be seen in Table 12. The control signatures can be integrated into the existing control system. If the current boiler settings are equal to the values in the control signatures, alarm signals could warn the operators.

The final and complete implementation strategy would come from the design of a real-time intelligent pluggage avoidance system. This system could be designed with a rule-based decision making approach that utilizes the knowledge from the rule induction algorithms. The system would read in the current boiler status and controller values as inputs from a data historian system. Using the current boiler operating conditions the systems would make predictions regarding the likelihood of the boiler plugging in the near future. An output would then be sent through the existing control system. The output could consist of three levels of pluggage predictions: stable, warning, and, immediate pluggage likely. This would be equivalent to a green, yellow, red warning light system.

The multi-angle data mining event detection method presented in this paper could be generalized to reduce the likelihood of ash fouling in a variety of boilers. This method is especially robust due to the fact it uses the existing control system to modify boiler conditions. This method also avoids using any data on coal properties, therefore eliminating the need for expensive scanning and pretreatment equipment.

REFERENCES

- Aha, D. (1992) Tolerating, noisy, irrelevant, and novel attributes in instance based learning algorithms. *Inter. J. Man-Machine Studies*, **36**(2), 267–287.
- Branagan, L.A. and Wasserman, P.D. (1992) Introductory use of probabilistic neural networks for spike detection from an on-line vibration diagnostic system. *Intelligent Engineering Systems Through Artificial Neural Networks*, **2**, 719–724.
- Erickson, T.A., Allan, S.E., McCollor, D.P., Hurley, J.P., Srinivasachar, S., Kang, S.G., Baker, J.E., Morgan, M.E., Johnson, S.A., and Borio, R. (1995) Modeling of fouling and slagging in coal-fired utilities boilers. *Fuel Proc. Technol.*, **44**, 155–171.
- Narayanswamy, R., Metz, J.L., and Johnson, K.M. (1998) Intelligent data elimination for a rare event application. *Proce. SPIE—Inter. Soc. Optical Eng.*, **3460**, 906–917.
- Pawlak, Z. (1991) *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston.
- Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, **1**(1), 81–106.
- Tsumoto, S. (2003) Chance discovery in medicine—Detection of rare risky events in chronic diseases. *New Generation Computing*, **21**(2), 135–147.
- Valero, A. and Cortes, C. (1996) Ash fouling in coal-fired utility boiler. Monitoring and optimization of on-load cleaning. *Prog. Energy Combust.*, **22**, 189–200.
- Vuthaluru, H.B. (1999) Remediation of ash problems in pulverized coal-fired boilers. *Fuel*, **78**, 1789–1803.