CHAPTER 4

MINING OPTIMAL BOILER DESIGN VALUES USING CLUSTERING

4.1 DATA MINING AN INTRODUCTION

The power sectors accumulates huge amount of data due to rapid advancement in data collection through day to day transactions. Performing data analysis to extract the required information from this massive size of data storage is extremely a challenging task. The traditional data analysis methods cannot be applied due to the presence of non-traditional properties of the industrial data. The concept of data mining emerges with sophisticated algorithms for analysing huge amount of data with the objective of extracting essential inferences. Here, the data mining is defined as the process of discovering hidden information in the thermal data repository. The traditional query languages retrieve only the data not the patterns that always remains unknown. Data mining methods are widely used in all the business sectors for performing predictive task that enhances the performance of information retrieval systems.

Data mining algorithms have been carried out as a series of transformation steps to convert the raw data recorded in the repository to meaningful patterns. The purpose of data transformation steps includes merging data from multiple sources, removing noisy data, determines duplicate observations and selecting the features that are more suitable to perform data mining task. The primary goal of the data mining techniques are data sampling, data estimation, searching, modelling complex systems, machine learning, artificial intelligence and pattern recognition. The
secondary goal is to integrate the ideas from other disciplines like optimization, evolutionary computing, information retrieval, signal processing, visualization and information theory. The primary and secondary goals mainly focus on the motivating challenges such as scalability, dimensionality, heterogeneous data and data distribution from various business aspects. The patterns discovered through data mining techniques provide the guidelines to achieve stable operations and to have effective performance monitoring process. The Figure 4.1 shown below symbolizes the computing activities associated with knowledge discovery process.

**Figure 4.1 Sequential Activities in Knowledge Discovery Process**

The purpose of data mining algorithm is to perform predictive or descriptive task of any real time applications. The objective of the predictive task is to predict or estimate the value of a target attribute based on the values of a source attributes. The intention of the descriptive task is to derive patterns that are defined in terms of clusters, correlations and trends. These are exploratory in nature. Also, the algorithms are required to integrate with post processing techniques and to execute the evaluation task. The predictive or
The descriptive task is decided based on the issues that are specific to the particular applications. Some of the issues are timeliness, relevance, knowledge about the data and intended use expected by the user. Based on the behaviour, the algorithms are categorized into clustering, classification and association analysis and anomaly detections.

The synthetic thermal data derived from the predicted model is stored in thermal data set. After the basic data analysis task gets over, based on the impact of each operational parameter, the desired amount of inputs are selected and recorded as sample mining dataset. Some sought of pre-processing steps are carried out to remove incomplete instances. The pre-processed data are then moved to perform transformation process that converts it into the standard format essential to carry out further analysis practice. The K-Means clustering algorithm is applied over the transformed data that results in mined optimal boiler design values. Through proper interpretation of the visualized representation, the optimal designed values are mined that provides sufficient knowledge to proceed with parameter estimation.

4.2 CLUSTER ANALYSIS

The Cluster Analysis is the widely used mechanism that divides the raw data into similar groups (Andrew Kusaik et al. 2008). The resultant groups are denoted as clusters which should capture the natural structure of the data. The significance of the cluster analysis is to improve,

- Understandability : Domain Knowledge
- Utility : Representative Object for summarization, compression and finding nearest neighbour
The cluster analysis forms the group of data objects based on the information resided in the data property that defines the objects and their relationships. The clustering is mainly categorized as follows,

a) **Hierarchical Versus Partitioning**: The Hierarchical clustering follows tree-like organization of clusters, in that each node denotes as a cluster to act as a union of its children and the root node contains all the data objects. In Partitioning cluster or non-overlapping cluster, the data objects reside exactly in one cluster.

b) **Exclusive Versus Overlapping**: In Exclusive clustering, the data objects are exclusively assigned to a single cluster. In overlapping clustering, the same objects may be assigned to different clusters based on the defined criteria.

c) **Complete Versus Partial**: The complete clustering assigns every object to a cluster but in partial clustering, few data objects may not be assigned to any of the desired cluster.

The groups are not predefined like classification technique, instead the grouping is formed by determining the similarities that exist between the data objects based on its characteristics. In all the categories of clustering methods, the data points are grouped based on the similar or dissimilar features (Osama 2008). The similarity exist between the properties of two objects is a numerical measure of the degree to which those objects are alike. Similarities between the data points reside in the cluster are usually denoted by non-negative integer either 0 for non-similarity and 1 for complete similarity. The disimilarity associated between the characteristics of two data objects is a numerical measure of degree to which those objects are different.
The dissimilarities are denoted between the interval \([0,1]\) or the range from 0 to \(\infty\).

4.2.1 K-Means Algorithm

4.2.1.1 Introduction

The K-Means approach is a prototype based effective clustering algorithm to deal with numeric data. The K-Means is an iterative clustering mechanism in that the data items are moved among defined sets of cluster until it reaches the desired group. The effectiveness of this clustering algorithm purely depends on the way in which the resultant groups are interpreted and essential information is extracted. This approach helps the user to discover the knowledge from domain perspective based on the input attributes to attain stable performance.

The K-Means algorithm begins its execution based on the assigned synthetic data. It produces different results at each run due to the random selection of cluster centroid. After an iteration is over, the centroids of the cluster is re-computed based on the number of thermal data points assigned to it. It terminates its execution when centroids of two successive iterations are same.

Among different types of measure, euclidian distance is the common proximity measure that defines the notion of closeness for the current data object assigned to the closest centroid. Euclidean measure is the common measure incorporated with K-Means to compute the distance of the data points.
4.2.1.2  **K-Means Algorithm**

The K-Means algorithm follows the basic philosophy of receiving input from the user, perform the required computation and visualize the obtained results. The steps involved in this algorithm is stated as,

**Step:1** Initialize the input values.

Sample Data (\( D_s \)) = \{ r_1, r_2, \ldots, r_n \} - set of elements.

\[ N = \text{Desired number of clusters}. \]

**Step:2** Specify the target output.

\[ C = \text{Cluster Set}. \]

**Step:3** Executional task.

a) Random assignment of initial centroid (\( a_1, a_2, \ldots, a_n \)).

b) Repeat

i) Assign each data item (\( V_t \)) from \( D_s \) to the cluster that contains the nearest centroid value.

ii) Compute new centroid value for each cluster.

Until Convergence or termination criteria is met.

4.2.1.3  **Clustering using K-Means**

The massive data from thermal power plants for different loads with varying calorific values of coal is useful in understanding the plant behaviour. Further, some design inferences could possibly be inferred with data patterns emerging from proper queries. The predicted performance values normally supplied to owners of the plant by Original Equipment Manufacturer (OEM) is made use of for arriving certain valuable information.

The derived inferences form as a base for framing optimal design criteria. Thermal data analysis leads to excellent prediction of equipment
essential parameters based on their correlations as stated in (Andrew Kusaik et al. 2012).

4.2.1.4 Representation of optimal design criteria

The aim of partitioning the predicted thermal data set into $k$ number of clusters is visualized in the following figures. The visualization represents the resultant cluster that contains the heat absorption pattern of thermal components as the source. Each colour in the below pictorial representation indicates different clusters of data. The observations in each cluster clearly portrays the heat absorption pattern. The Figure 4.2 below (a) shows the cluster patterns over heat absorption in Re-heaters.

![Clustered Heat Absorption Pattern of Re-heating Process](image)

a) Clustered Heat Absorption Pattern of Re-heating Process

Based on the given $k$ value, the number of instances belongs to the clusters vary. The selected data points are grouped into various clusters based on the characteristics of the data. The cluster represents the heat absorption pattern of each thermal component from 250MW, 300MW, 400MW, 500MW respectively as shown in the associated figures. Each run makes the difference in the output due to the variation exist in the $k$ value. But the interpretation of the obtained cluster results in same level of information extraction.
The Figure 4.2 below (b) shows the cluster patterns over the Net Heat input.

b) Clustered Net Heat Input Pattern for Combustion Process

The Figure 4.2 below (c) shows the cluster patterns over the heat absorption in Super-heater 1.

c) Clustered Heat Absorption Pattern of Super-Heater 1
The Figure 4.2 below (d) shows the cluster patterns over the heat absorption in Super-heater 2&3.

d) Clustered Heat Absorption Pattern of Super-Heater 2&3

The Figure 4.2 below (e) shows the cluster patterns over the heat absorption in Economizer.

e) Clustered Heat Absorption Pattern of Economizer
The Figure 4.2 below (f) shows the cluster patterns over the heat absorption in Water walls system.

f) Clustered Heat Absorption Pattern of Water Wall System

The Figure 4.2 below (g) shows the cluster patterns over the heat absorption in Primary Air Heater.

g) Clustered Heat Absorption Pattern of Primary Air Heater
The Figure 4.2 below (h) shows the cluster patterns over the heat absorption in Secondary Air Heater.

h) Clustered Heat Absorption Pattern of Secondary Air Heater

The Figure 4.2 below (i) shows the cluster patterns over the predicted values of Burner Tilt.

i) Clustered Pattern of Burner Tilt
The Figure 4.2 below (j) shows the cluster patterns over the predicted values of the influencing parameter coal flow.

![Figure 4.2: Clustering using K-Means](image)

**j) Clustered Coal Flow Pattern of Combustion Process**

**Figure 4.2 Clustering using K-Means**

The extracted inference from each cluster’s defines the optimal design value for 250MW, 300MW, 400MW, 500MW and 550MW respectively.

**4.2.2 Self-Organizing Map Clustering (SOM)**

**4.2.2.1 Introduction**

The Kohonen SOM is a neural network based effective clustering and visualization technique (Kohonen 2013). The objective of SOM is to determine the collection of desired centroid set to assign objects that provides best approximation. At a time, a single data object is processed and the closest centroid to it gets updated. The process repeats and terminates when convergence criteria is met. The SOM imposes the topographic structure on the centroids which are termed as neurons. The centroids in SOM have a predefined topographic ordering relationship. During the training process,
centroids are updated in topographical order by using the data points which resides in it.

The fundamental difference between SOM and K-Means clustering algorithm is that SOM produces ordered set of centroids for the selected samples. The clusters that are neighbors are more related to each other than the clusters that are not. The user must select the values for parameter setting, neighborhood function, SOM structure and the number of centroids.

4.2.2.2 SOM algorithm

The SOM algorithm follows the fundamental idea of getting input from the user, execute the mandatory computation and visualize the obtained outcome. The steps involved in this algorithm is stated as,

**Step:1** Initialize the desired number of centroids in terms of dimensions.

Data Vector (X) = \{ x_1, x_2, \ldots, x_n \} - set of elements.

Weight Vector W =\{w_{1j}, w_{2j}, \ldots, w_{nj}\}

**Step:2** Specify the target output.

C = Cluster Set.

**Step:3** Executable task.

a) Repeat 3

i) Select the current data object.

ii) Compute the similarity exist between X and i as
\[ \text{sim}(X, i) = \sum_{j=1}^{n} w_{ij}x_j \]

iii) Assign the data object \( (x_i) \) from \( X \) to the cluster that contains the nearest cluster mean value. Then,

Update the weight vector value for each cluster.

Until

Convergence or termination criteria is met.

4.2.2.3 Clustering using SOM

The SOM algorithm follows competitive unsupervised learning, in which the nodes are permitted to compete and the winner takes all (Kohonen et al. 1996). The number of input node is based on the number of selected attributes. During network creation, the connection between the nodes are established and viewed as a grid like structure. Initially weights are randomly assigned between the range of 0 to 1 and denoted in weight vector. The weights are updated at each iteration between input, intermediate and output node. During the learning process, the weights are adjusted to produce enhanced results. The arc associated between the nodes is to represent appropriate weight value. The basic idea of SOM learning process is that after each data point in the training set, winning node and its neighbours mean value changes to be closer to that of the data point. The output node generates the outcome that carries information or patterns based on the input data.

4.2.2.4 Representation of the SOM design

SOM representations are well suited for converting high dimensional thermal process data into low dimensional data that are used for analysis with the objective of deriving hidden information.
The Figure 4.3 (a) and (b) visualizes the basic notations of designed neural network model as shown below.

![SOM Topology](image1)

**a) Network Topology**

![SOM Neighbor Connections](image2)

**b) Cluster Neighbor Connections**

The Figure 4.3 (c) and (d) visualizes the distance existing between the network nodes and includes trained hits as shown below.

![SOM Neighbor Weight Distances](image3)

**c) Neighbor Weight Distance**

![Hits](image4)

**d) Instance Hits**
The Figure 4.3 (e) visualizes the weight values among network layer as shown below.

\[ \text{Weights from Input 1} \]
\[ \text{Weights from Input 2} \]

**e) Input Weights between Network Layer**

The weights of the neurons are updated at each training phase based on their proximity towards the winning neuron.

The Figure 4.3 (f) and (g) visualizes the optimal design values of NHI and SH as shown below.

\[ \text{SOM Weight Positions} \]

**f) Clustered Pattern of NHI**

**g) Clustered Pattern of SH1**
The Figure 4.3 (h) and (i) visualizes the optimal design values for SH2 and RH as shown below.

h) Clustered Pattern of SH2  
i) Clustered Pattern of RH

The SOM clustering needs the weight vector to be initialized randomly based on the initial vector to have successful groupings. The insufficient data in weight vector will increase the randomness in groupings. The enhanced ability is needed to select the correct data of weight vector, that decides the accuracy of SOM. The Figure 4.3 (j) and (k) visualizes the optimal design values of Economizer and Water Walls as shown below.

j) Clustered Pattern of ECO  
k) Clustered Pattern of WW
The Figure 4.3 (l) and (m) visualizes the optimal design values of PAH and SAH as shown below.

The grid like clustering representation makes the process easy to interpret to understand to observe the similarities exist between data points (Muhammad et al. 2010). The Figure 4.3 (n) and (o) visualizes the optimal design values of Coal flow and Burner Tilt as shown below.

Figure 4.3 Clustering using SOM
The mapping of multidimensional thermal data into lower dimensional subspaces are clearly depicted in the above representations. The relationships exist between the points indicate their similarity. The dimensionality reduction supports the user to interpret the data based on the requirement. The SOM clustering provides too little information that may not be nearing the accuracy or become informative as compared to the K-Means clustering.

4.2.3 Optimal Design Values

The stable combustion process requires optimal design criteria to avoid inefficient and unsafe conditions. The average optimal designs values of thermal components are derived from two categories of clustering techniques with respect to different grades of coal are tabulated below. These optimal values are useful to maintain the boiler combustion process in a stable manner. The values are suited only to the associated loads specified in the unit. The Table 4.1 shown below illustrates the derived optimal values of different grades of coal for 250MW.

Table 4.1 Optimal Design Value 250 MW

<table>
<thead>
<tr>
<th>Component/Grade</th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHI</td>
<td>183982</td>
<td>186535</td>
<td>187966</td>
</tr>
<tr>
<td>ECO</td>
<td>18004</td>
<td>18887</td>
<td>19507</td>
</tr>
<tr>
<td>WW</td>
<td>55551</td>
<td>53808</td>
<td>53375</td>
</tr>
<tr>
<td>SH1</td>
<td>16120</td>
<td>17420</td>
<td>17796</td>
</tr>
<tr>
<td>SH23</td>
<td>38102</td>
<td>38210</td>
<td>37266</td>
</tr>
<tr>
<td>RH</td>
<td>22490</td>
<td>22553</td>
<td>22678</td>
</tr>
<tr>
<td>PAH</td>
<td>4631</td>
<td>5400</td>
<td>5904</td>
</tr>
<tr>
<td>SAH</td>
<td>7595</td>
<td>7337</td>
<td>7137</td>
</tr>
</tbody>
</table>

The optimal values from each cluster is derived based on the synthetic thermal data and converted that into real thermal data by applying standard combustion process. It is represented in the Table 4.1.
The Table 4.2 shown below illustrates the derived optimal values of different grades of coal for 300MW.

### Table 4.2 Optimal Design Value 300 MW

<table>
<thead>
<tr>
<th>Component / Grade</th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHI</td>
<td>214759</td>
<td>217104</td>
<td>219980</td>
</tr>
<tr>
<td>ECO</td>
<td>21628</td>
<td>22413</td>
<td>23424</td>
</tr>
<tr>
<td>WW</td>
<td>62465</td>
<td>62295</td>
<td>61489</td>
</tr>
<tr>
<td>SH1</td>
<td>19790</td>
<td>20216</td>
<td>22950</td>
</tr>
<tr>
<td>SH23</td>
<td>41817</td>
<td>42311</td>
<td>42736</td>
</tr>
<tr>
<td>RH</td>
<td>27063</td>
<td>27067</td>
<td>27344</td>
</tr>
<tr>
<td>PAH</td>
<td>5268</td>
<td>6094</td>
<td>7695</td>
</tr>
<tr>
<td>SAH</td>
<td>10589</td>
<td>10323</td>
<td>9060</td>
</tr>
</tbody>
</table>

The tabulated values guide to learn the sequence of heat absorption pattern for Grade-1, Grade-2 and Grade-3 of coal. The heat absorption trends for the entire load with different grades of coal fall in the same sequential line. The tabulated values are clearly indicating this aspect. The heat absorption between thermal components is either increase or decrease based on the heating value of the coal.

The Table 4.3 illustrates the derived optimal values of different grades of coal for 400MW.

### Table 4.3 Optimal Design Value 400 MW

<table>
<thead>
<tr>
<th>Component / Grade</th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHI</td>
<td>278318</td>
<td>282393</td>
<td>285084</td>
</tr>
<tr>
<td>ECO</td>
<td>29066</td>
<td>30780</td>
<td>31402</td>
</tr>
<tr>
<td>WW</td>
<td>80155</td>
<td>79581</td>
<td>79615</td>
</tr>
<tr>
<td>SH1</td>
<td>26896</td>
<td>27619</td>
<td>30459</td>
</tr>
<tr>
<td>SH23</td>
<td>51267</td>
<td>50446</td>
<td>50178</td>
</tr>
<tr>
<td>RH</td>
<td>34906</td>
<td>35334</td>
<td>35150</td>
</tr>
<tr>
<td>PAH</td>
<td>6780</td>
<td>8462</td>
<td>10352</td>
</tr>
<tr>
<td>SAH</td>
<td>16085</td>
<td>15804</td>
<td>14114</td>
</tr>
</tbody>
</table>
The Table 4.4 shown below illustrates the derived optimal values of different grades of coal for 500 MW.

**Table 4.4 Optimal Design Value 500 MW**

<table>
<thead>
<tr>
<th>Component /Grade</th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHI</td>
<td>341173</td>
<td>346185</td>
<td>348612</td>
</tr>
<tr>
<td>ECO</td>
<td>38361</td>
<td>39164</td>
<td>40589</td>
</tr>
<tr>
<td>WW</td>
<td>97490</td>
<td>97296</td>
<td>96368</td>
</tr>
<tr>
<td>SH1</td>
<td>34640</td>
<td>35392</td>
<td>36108</td>
</tr>
<tr>
<td>SH23</td>
<td>58618</td>
<td>57250</td>
<td>55968</td>
</tr>
<tr>
<td>RH</td>
<td>43187</td>
<td>43187</td>
<td>42652</td>
</tr>
<tr>
<td>PAH</td>
<td>8470</td>
<td>11212</td>
<td>13337</td>
</tr>
<tr>
<td>SAH</td>
<td>21878</td>
<td>21210</td>
<td>19578</td>
</tr>
</tbody>
</table>

The Table 4.5 shown below illustrates the derived optimal values of different grades of coal for 550 MW.

**Table 4.5 Optimal Designs Value 550 MW**

<table>
<thead>
<tr>
<th>Component/Grade</th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHI</td>
<td>377612</td>
<td>383940</td>
<td>390830</td>
</tr>
<tr>
<td>ECO</td>
<td>42372</td>
<td>44702</td>
<td>47546</td>
</tr>
<tr>
<td>WW</td>
<td>109174</td>
<td>106306</td>
<td>104090</td>
</tr>
<tr>
<td>SH1</td>
<td>44372</td>
<td>40369</td>
<td>41764</td>
</tr>
<tr>
<td>SH23</td>
<td>58314</td>
<td>64248</td>
<td>62318</td>
</tr>
<tr>
<td>RH</td>
<td>45207</td>
<td>45207</td>
<td>45207</td>
</tr>
<tr>
<td>PAH</td>
<td>9328</td>
<td>12678</td>
<td>15028</td>
</tr>
<tr>
<td>SAH</td>
<td>24504</td>
<td>24157</td>
<td>22438</td>
</tr>
</tbody>
</table>

4.3 **SUMMARY**

The discovered patterns are represented in the form of feature subsets. The extracted patterns serve as a base to define optimal standards to achieve stable and consistent operations. Through the experimental study, it is clearly understood that the performance of any clustering technique is not stable for all the applications. The efficiency of the clustering techniques
depend on the chosen input data set and the perspective of interpretation. Here, the clusters are obtained individually for all the thermal components from the generated synthetic data. The performance of any technique is measured in terms of cluster cohesiveness.

The cluster analysis visualizes the observations that belongs to the same cluster are similar to each other than the observations that belongs to the other clusters. The central point of each cluster is considered as optimal value for the specific load measured in mega watts. The observation clearly indicates the linear heat absorption pattern of all the boiler thermal components and observations are visualized individually.

The discovered knowledge is visualized and it shows the descriptive nature of statistical philosophy of boiler operations. Both the clustering methods K-means and SOM provide the cluster information that helps to derive optimal design patterns needed to perform the following,

- To support planning stage queries.
- To recognize inter relationships exist among internal and external factors
- Provide support to determine root cause for performance deviations.
- To achieve accurate diagnosis that leads to effective optimization.
- Indicating trends through well-organized data acquisition method.
• Measure component efficiency and generate guidelines to establish test procedures.

• Aided to validate data to assure that the outcomes are reliable with standards based on performance characteristics.

The optimal points acquired from two clustering techniques are compared to ensure its accuracy. It is found that the deployment methodologies followed in both the approaches are different but the resultant patterns are more or less unique.

Optimal set points attained from K-Means are more informative than SOM that confirms through evaluating resultant optimal criteria by comparing with industrial predicted performance data. It identifies that the error rate that exists is negligible and it won’t deviate the performance. An optimal design criterion for different grades of coal with various loads helps in learning the characteristics of fuel and trends of operational parameters.