A Review of using Data Mining Techniques in Power Plants

Waleed Hamed Ahmed Eisa [1], Naomie Bt Salim [2]

Faculty of Computer Sciences, Sudan University of Science and Technology1
Faculty of Computer Science and Information Systems, University Teknologi Malaysia[2]
mds.waleed@gmail.com

Received: 02/08/2016
Accepted: 1910/2016

ABSTRACT- Data mining techniques and their applications have developed rapidly during the last two decades. This paper reviews application of data mining techniques in power systems, specially in power plants, through a survey of literature between the year 2000 and 2015. Keyword indices, articles' abstracts and conclusions were used to classify more than 86 articles about application of data mining in power plants, from many academic journals and research centers. Because this paper concerns about application of data mining in power plants; the paper started by providing a brief introduction about data mining and power systems to give the reader better vision about these two different disciplines. This paper presents a comprehensive survey of the collected articles and classifies them according to three categories: the used techniques, the problem and the application area. From this review we found that data mining techniques (classification, regression, clustering and association rules) could be used to solve many types of problems in power plants, like predicting the amount of generated power, failure prediction, failure diagnosis, failure detection and many others. Also there is no standard technique that could be used for a specific problem. Application of data mining in power plants is a rich research area and still needs more exploration.

المستخلص- خلال العقدين السابقين تطورت تقنيات التنقيب في البيانات مما أدى لانتشار استعمالها في العديد من المجالات. في هذه الورقة تم تجميع و تلخيص استخدام تقنيات التنقيب في البيانات المستعملة في أنظمة الكهرباء و بالذات محطات التوليد الكهربائي, و ذلك من خلال مراجعة ما يزيد على ستة و ثمانين بحثاً من مختلف المجلات العلمية العالمية. و لأن الورقة تهدف لدراسة تطبيق تقنيات التنقيب في البيانات على محطات التوليد الكهربائي, فقد بُدأت الورقة بمقدمة عن التنقيب في البيانات ثم أساسيات أنظمة توليد الكهرباء. كذلك تم تصنيف هذه البحوث بناء على ثلاثة محاور وهي هدف البحث, و التقنية المستعملة لحل المشكلة, و مجال البحث. من خلال هذه الورقة نخلص الى أن تقنيات التنقيب في البيانات يمكن أن تستعمل لإيجاد الحلول المناسبة لكثير من مشاكل محطات التوليد الكهربائي و التي يصعب حلها بالطرق التقليدية. ومثال ذلك توقع كمية الكهرباء, و توقع أعطال المحطة, و تحليل تلك الأعطال. كما نلاحظ من خلال الورقة أنه ليس هناك اتفاق على استعمال تقنية محددة لحل مشكلة بعينها. إن تطبيق تقنيات التنقيب في البيانات في محطات التوليد الكهربائي هو مجال بحثي واسع يحتاج الى مجهود مكثف من الباحثين.

Keywords*: Power Plant , Thermal Power Plant , Feature Selection , Prediction, Regression, Data Mining*.

INTRODUCTION

Most of the electric power systems no a days are equipped with SCADA (Supervisory Control And Data Acquisition**)** systems, that ease the collection of real time data. This huge amount of data which is collected instantly encourages the application of data mining techniques in power systems. However, this area is new and still faces several difficulties to benefit from data mining [1]. The first difficulty is that: mining power systems data is an interdisciplinary task, that requires electromechanical engineers and data scientists to work as a team in order to achieve their goals. The Second one is the limitation in data storage capacities, which leads to automatic purging. Consequently, data is available for short period, less than what is required by a data mining tool. A third difficulty is the lack of standardized benchmarks, this is very clear from all researches presented here after, where researchers are using proprietary datasets, which makes it difficult to compare algorithms and reproduce results.

Because it is an interdisciplinary research, this paper started by giving electrical engineers an introduction about data mining and CRISP-DM model. On the other site an introduction of power systems and Rankine cycle is provided for data scientists and computer engineers. After that a comprehensive review about application of data mining in power systems is provided. In this part some interesting articles which are not related to power plants were discussed, like *Prediction of thermodynamic properties of refrigerants using data mining* – by E.U.Kucuksille et al [2], their

research focused in prediction of thermodynamics properties. The last part focused in the application of data mining in power plants, in this part many researches were discussed from different types of power plants. Prof. Andrew Kusiak from University of Iowa, is the principal investigator of research project that aims to propose a novel approach to modeling the performance of wind turbines. This project provide real innovations and clear vision of using data mining in power plants problems. Then the whole idea is wrapped up in the conclusion part.

DATA MINING & DATA MINING PROCESSES

This section is divided into three parts, the first one gives an introduction about data mining. The second is about CRISP-DM which is a complete blueprint for conducting a data mining project. The third section presents some regression algorithms, because this survey focuses on the application of regression techniques in power plants.

A) Data Mining

Data mining is defined as the process of discovering patterns in data [3]. To discover this pattern; first we must explore the past then we can predict the future. Exploring the past is done by describing the data by means of statistical and visualization techniques. While future prediction is done by developing prediction models. Prediction models could be classified to the following categories [4]:

1. Classification: if the model outcome is categorical. There are four main groups of classification algorithms:

- **i.** Frequency Table: which contains ZeroR, OneR, Naive Bayesian and Decision Tree algorithms.
- **ii.** Covariance Matrix: which contains Linear Discriminant Analysis, and Logistic Regression.
- **iii.** Similarity Functions: which contains K Nearest Neighbors
- **iv.** Others: which contains Artificial Neural Network, and Support Vector Machine

2. Regression: if the model outcome is numerical. There are four main groups of regression algorithms:

- **i.** Frequency Table: which contains Decision Tree
- **ii.** Covariance Matrix: which contains Multiple Linear Regression
- **iii.** Similarity Functions: which contains K Nearest Neighbors
- **iv.** Others: which contains Artificial Neural Network, and Support Vector Machine

3. Clustering or descriptive modeling: is the assignment of observations into clusters so that observations in the same cluster are similar. There are two main groups of clustering algorithms:

- **i.** Hierarchical which contains Agglomerative, and Divisive.
- **ii.** Partitive: which contains K Means, and Self-Organizing Map.

4. Association rules: can find interesting associations amongst observations. One of the most famous algorithms in this group is Apriori algorithm which was proposed by Agrawal and Srikant in 1994 [5]. The algorithm generates the association rules by finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support). Then the algorithm uses the larger itemsets to generate the association rules that have confidence greater than or equal to a user specified minimum confidence [6].

B) The CRISP-DM Reference Model [7]

CRISP-DM is a comprehensive process model and data mining methodology that provides anyone with a complete blueprint for conducting a data mining project. As shown in figure (1) CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The arrows in the figure indicate the most important and frequent dependencies between the phases, while the outer circle shows the cyclic nature and continual improvement of data mining itself , i.e. lessons learned during the data mining process and from the deployed solution can trigger new, business questions.

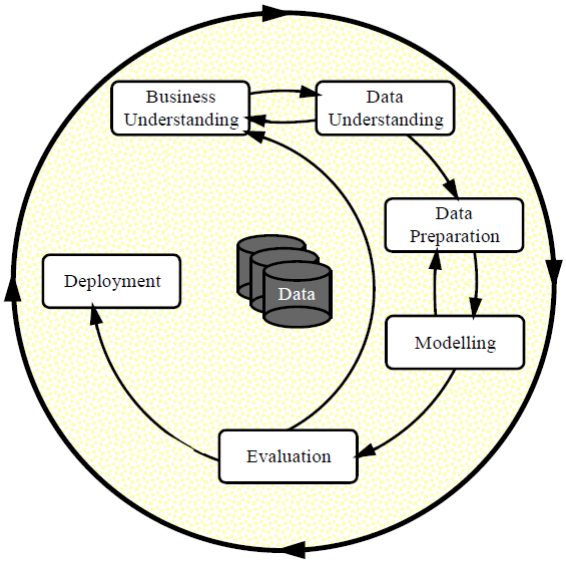

Figure 1: Phases of the CRISP-DM [7]

1. Business Understanding : This is the most important phase, it is the initial business understanding phase which focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how. This phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan.

2. Data Understanding : This phase starts with an initial data collection, then the analyst start to increase familiarity with the data, to identify data quality problems, discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. This phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

3. Data Preparation : This phase covers all activities to construct the final data set that will be fed into the modeling tool(s) from the initial raw data. This phase consists of five steps : the selection, cleansing, construction, integration, and data formatting.

4. Modeling : In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Several techniques exist for the same data mining problem type. Some techniques have specific data requirements, therefore, stepping back to data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models.

5. Evaluation : In this phase the model is evaluated and its construction is reviewed to be certain it properly achieves the business objectives, and consider all important business issues. At the end of this phase, the project leader should decide how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps.

6. Deployment : The knowledge gained must be organized and presented in a way that the customer can use it, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up front what actions must be taken in order to actually make use of the created models. The key steps here are plan deployment, plan monitoring and maintenance, the production of the final report, and review of the project.

C) Some Regression Algorithms

1. Linear regression (LR)

Linear regression is a well known method of mathematical modeling of the relationship between a dependent variable and one or more independent variables. Regression uses existing (or known) values to forecast the required parameters. In the simplest case, regression employs standard statistical techniques such as linear regression. Unfortunately, many real world problems are not simply linear projections of previous values. So, more complex techniques (e.g., logistic regression, decision trees or neural networks) may be necessary to forecast future values [8].

2. Pace regression (PR)

Pace regression improves the classical ordinary least squares regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their

contribution to the overall regressions. Under regularity conditions, pace
regression is provably optimal when the number of coefficients tends to infinity. It consists of a group of estimators that are either overall optimal or optimal under certain conditions [3].

3. Multiple Linear Regression (MLR)

Multiple Linear Regression (**MLR**) is a method used to model the linear relationship between a dependent variable (target or class) and one or more independent variables (predictors or attributes). In equation (1) below: y is the observed target, x_1, x_2 .. x_p are the predictor. y' is the predicted target, b_0 is constant, b_1,b_2 ..b_p are attributes' weights, ϵ is The model error.

$$observed\ data \rightarrow y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \epsilon$$
$$predicted\ data \rightarrow \hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \epsilon$$
$$Error \rightarrow \epsilon = y - \hat{y} \quad \textbf{(1)}$$

MLR is based on ordinary least squares (**OLS**), the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized. The MLR model is based on several assumptions (e.g., errors are normally distributed with zero mean and constant variance). Provided the assumptions are satisfied, the regression estimators are optimal in the sense that they are *unbiased*, *efficient*, and *consistent*. Unbiased means that the expected value of the estimator is equal to the true value of the parameter. Efficient means that the estimator has a smaller variance than any other estimator. Consistent means that the bias and variance of the estimator approach zero as the sample size approaches infinity [4].

4. SMOReg

SMO (Self-Organizing Maps (SMO) algorithm for regression) implements a non-linear method for sequential minimal optimization to train a support vector regression using polynomial or radial basis function (RBF) kernels. Multi-class problems are solved using pair wise classification. To obtain the proper probability estimates, we use the option that fits logistic regression models to the outputs of the support vector machine [9].

5. SVMReg(Support Vector Machine Regression)

The SVR model maps data nonlinearly into a higher-dimensional feature space, in which it undertakes linear regression. Rather than obtaining empirical errors, SVR aims to minimize the upper limit of the generalization error [4].

6. KStar

Is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. The underlying assumption of instance-based classifiers is such as KStar [4].

7. M5 Model Tree Algorithm

An algorithm for generating M5 model trees. M5 builds a tree to predict numeric values for a given instance. The algorithm requires the output attribute to be numeric while the input attributes can be either discrete or continuous. For a given instance the tree is traversed from top to bottom until a leaf node is reached. At each node in the tree a decision is made to follow a particular branch based on a test condition on the attribute associated with that node. Each leaf has a linear regression model associated with it. As the leaf nodes contain a linear regression
model to obtain the predicted output, the tree is called a model tree .
To build a model tree, using the M5 algorithm, we start with a set of training instances. The tree is built using a divide-and-conquer method. At a node, starting with the root node, the instance set that reaches it is either associated with a leaf or a test condition is chosen that splits the instances into subsets based on the test outcome. A test is based on an attributes value, which is used to decide which branch to follow. [3]

8. RepTree

Quinlan first introduced Reduced Error Pruning (REP) as a method to prune decision trees. REP is a simple pruning method though it is sometimes considered to overprune the tree. A separate pruning dataset is required, which is considered a downfall of this method because data is normally scarce. However, REP can be extremely powerful when it is used with either a large number of examples or in combination with boosting. The pruning method that is used is the replacement of a subtree by a leaf representing the majority of all examples reaching it in the pruning set. This

replacement is done if this modification reduces the error, i.e. if the new tree would give an equal or fewer numbers of misclassifications [3].

9. Decision table (DT)

Decision table summarizes the dataset with a 'decision table', a decision table contains the same number of attributes as the original dataset, and a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. This implementation employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller, more condensed decision table [4].

POWER SYSTEMS

The availability of real time data in the electric power industry encourages the adoption of data mining techniques. Data mining is defined as the process of discovering patterns in data [3]. However, there is some obstacles that faces researchers and engineers to benefit from data mining in this area. The first one is the interdisciplinary nature of such a research, because it requires deep knowledge in both IT and electromechanical engineering. Another obstacle is the lack of standard analysis methods and benchmarks, this leads to usage of different methods and datasets [1].

A) Power System

A typical power system which is also known as the grid is shown in figure (2), it is divided into three components: the generator which produce the power, the transmission system that carries the power from the generators to the load centers and the distribution which delivers power to the end users.

There are many types of generators (also known as power plant) normally these power plants contain one or more generators which is a rotating machine that converts mechanical power into electrical power. Then the motion between a magnetic field and a conductor creates an electrical current. Most power plants in the world burn fossil fuels such as coal, oil, and natural gas to generate electricity. Others use nuclear power, but there is an increasing use of cleaner renewable sources such as solar, wind, wave and hydroelectric [1].

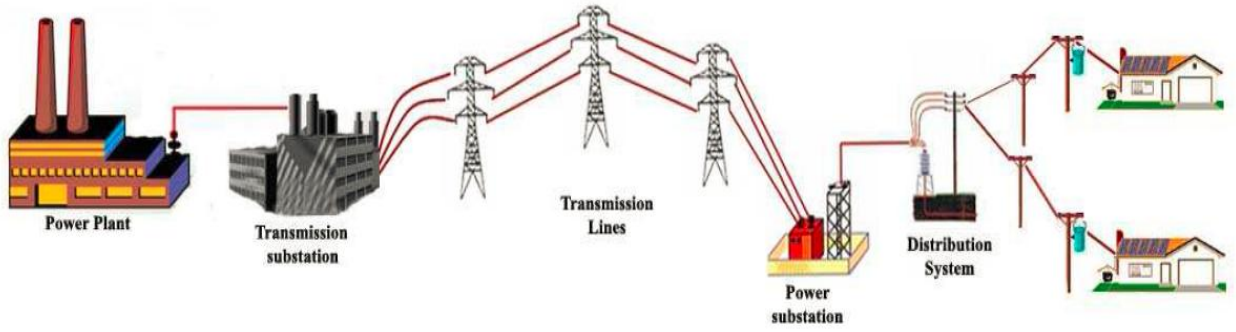

Figure 2: Components of power system [1]

B) Rankine Cycle

This research focus on thermal power plants that uses oil as energy source, these types of power plants uses Rankine Cycle to generate power. More theoretical investigation about Rankine Cycle could be found in [10]. Rankine Cycle is a closed system consists of four main components, that are interconnected together to build one system as shown in figure (3) below. These components are [11] :

1. **Steam Turbine which** uses the superheated steam that is coming from the boiler to rotate the turbine blades.
2. **Condenser**: uses external cooling water to condense the steam which is exhausted from turbine to liquid water.
3. **Feed water Pump**: to pump the liquid to a high pressure and bush it again to boiler.
4. **Boiler** which is externally heated to boil the water to superheated steam.

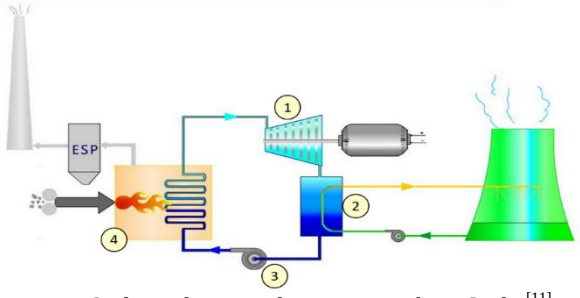

Figure 3:Thermal Power Plant using Rankine Cycle [11]

C) The use of data mining in power plants.

Recently the use of data mining application in electricity power systems have been increased. Many papers were found in the literature, each is focusing in one area of the power system. Some are focusing in the distribution system like Ramos 2008 [12] who used decision tree to classify the consumers. Saibal, 2008 [13] used WN (Wavelet Networking is an extension of perceptron networks) for the classification of transients. Figueiredo, 2005 [14] used decision tree for the classification of electric energy consumer. Dola, 2005 [15] used decision tree and neural network for faults classification in distribution system. Mori, 2002 [16] used regression tree and neural network for load forecasting. Other researcher focused in the transmission line problems like: Hagh 2007 [17] , Silva, 2006 [18] , Costa, 2006 [19] , Vasilic, 2002 [20] all of them used neural network to study faults detection, classification and locations in transmission lines. Dash, 2007 [21] used support vector machine for the classification and identification of series compensated. Vasilic 2005 [22] and Huisheng 1998 [23] used Fuzzy/ neural network for faults classification. Some other researchers focused in **Power Generation** part (power plants) like Andrew Kusiak et. al. [24] who used a multi objective optimization model to optimize wind turbine performance. Others like Ecir Ug˘ur Kucuksille et. al. [2] used data mining to predict thermodynamic properties, they used many algorithms to predict enthalpy, entropy and specific volume for specific types of refrigerants. Other researchers focused on work process optimization and performance monitoring, Water and Power Plant Fujairah (FWPP) in the United Arabic Emirates is a true success story of data mining usage, where more than 4% of of the total consumption have been achieved [25]. Softstat showed the superiority of data mining tools to traditional approaches like DOE (Design Of Experiments), CFD (Computational Fluid Dynamics). In their research they started by feature selection then apply DM algorithms to get better performance of Flame temperature. Finally recommendations from the model were deployed [26].

DATA MINING FOR POWER SYSTEMS

A) Applications of data mining methods in Power Systems

Various data mining methods has been used by many researchers for different types of problems in power systems like: energy efficient building design, HVAC systems, energy demand modeling, electricity price forecast, prediction of properties of refrigerants, object tracking, optimization of wind turbine, cluster of load profiles, modeling of absorption heat transformer and analysis of fluidized-bed boiler. This section gives an overview about some of these researches.

Tso and Yau [27] have used regression analysis, decision tree and neural networks models to predict electricity consumption. Model with least squared errors were selected. In an electricity energy consumption study, the decision tree and neural network models outperformed the stepwise regression model in understanding energy consumption patterns and predicting energy consumption levels. Using data mining approach for predictive modeling, different types of models can be built in a unified platform: to assess, select and implement the most appropriate model.

Şencan [28] applied DM process to determine specific volume values of methanol/LiBr and methanol/LiCl used in absorption heat pump systems. Six algorithms were used: Linear regression (LR), pace regression (PR), sequential minimal optimization (SMO), M5 model tree, M5'Rules and back propagation neural network (BPNN).

Figueiredo et al. [29] presented an electricity consumer characterization framework based on a knowledge discovery in databases, supported by data mining techniques. This framework consists of two modules: The first one is the load profiling module which creates a set of consumer classes using a clustering and the representative load profiles for each class. The second is the classification module which uses this knowledge

to build a classification model to assign different consumers to the existing classes.

Kusiak et al. [30] applied data mining approach to analyze relationships between parameters of a circulating fluidized-bed boiler. The model can predict efficiency to the same degree of accuracy with and without the data describing the fuel composition or boiler demand levels. It is proved that data mining is applicable to different types of burners and fuel types.

Küçüksille et al. [2] has applied data mining approach for the modeling of thermodynamic properties of alternative refrigerants. In addition, mathematical equations in order to calculate enthalpy, entropy and specific volume values of each refrigerant were presented. The values calculated from obtained formulations were found to be good compared to actual values. The results showed that data mining is suitable for predicting thermodynamic properties of refrigerants for every temperature and pressure. More details about this article will be presented later in this section.

Lu et al [31] carried out electricity price forecast framework to predict the normal price and the price spikes. The model is based on a mining database including market clearing price, trading hour, electricity demand, electricity supply and reserve. The model can generate forecasted price spike, level of spike and associated forecast confidence level.

Hou et al. [32] combined rough set and an artificial neural network to developed a model that detects and diagnose sensor faults based on the past running performance data in heating, ventilating and air conditioning (HVAC) systems. The reduced information is used to develop classification rules and train the neural network to infer appropriate parameters. Results from a real HVAC system showed that only the temperature and humidity measurements can work very well as the to distinguish simultaneous temperature sensor faults of the supply chilled water (SCW) and return chilled water (RCW).

Yu et al. [33] built energy demand model using decision tree. This model applied to Japanese residential buildings for predicting and classifying building levels. The results have demonstrated that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data).

Kusiak et al. [34] presented data-driven approach to minimize the energy of air condition. Eight algorithms were applied to model the nonlinear relationship among controllable parameters (supply air temperature and supply air static pressure), and uncontrollable parameters. The multiple-linear perceptron (MLP) outperforms other models, so it is selected to model a chiller, a pump, a fan, and a reheat device. These four models are integrated into an energy optimization model with two decision variables. The optimization results have demonstrated the total energy consumed by the heating, ventilation, and air-conditioning system is reduced by over 7%.

Haiming Zhou1 et al. [35] introduced an operating analysis and data mining system for power grid dispatching. The system digs out the potential rule of the power grid operating by applying online analytical processing and data mining techniques to the operational data collected in the dispatching center. The calculation and analysis of the dispatching indicators gives the power grid dispatchers a more comprehensive, clear, quantitative understanding. The data mining of the historic data can facilitate the understanding of some complex related issues in the dispatching operation.

B) Prediction of thermodynamic properties of refrigerants using data mining [2]

Hereafter more highlights will be provided about Küçüksille et al. work about *prediction of thermodynamic properties of refrigerants using data mining*. Although their research is not directly related to power systems, but it has been selected because they followed the CRISP-DM model to do their job, moreover the way they presented the results is very clear.

The halogenated refrigerants are a family of chemical compounds derived from the hydrocarbons by substitution of chlorine and fluorine atoms for hydrogen. The emission of chlorine and fluorine atoms present in halogenated refrigerants is responsible for the major environmental impacts, it has a direct negative impact on ozone layer. Because of this serious implications most of the developed countries prohibited the production and consumption of halogenated refrigerants, and encourage the use of alternative refrigerants like R134a, R404a, R407c and R410a .

E.U.Kucuksille et al studied thermodynamic properties as enthalpy, entropy and specific volume of alternative refrigerants using data mining method. In their research they studied four alternative refrigerants R134a, R404a, R407c and R410a. The results obtained from data mining have been compared to actual data from the literature. In their study they showed that, data mining is successfully applicable to determine enthalpy, entropy and specific volume values of refrigerants when using temperature and pressure as predictors. The thermodynamic properties of refrigerants used in data mining analysis were taken from Solkane Software.

1. Data mining process

To build their model E.U.Kucuksille et al followed CRISP-DM models which is composed of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [7].

2. Data sets

In their data set predictors are :temperature and pressure as predictors. To study three different targets :(enthalpy, entropy and specific volume). All these parameter were studied against four different refrigerants: R134a, R404a, R407c and R410a.

3. Regression Models and Results Evaluation

Researchers built 288 models, because their goal is to study thermodynamic properties of 4 different compounds (R134a, R404a, R407c and R410a) in both vapor and liquid states, to do this they used 12 algorithms to predict 3 different targets (enthalpy, entropy and volume) . So the total number of models is (4 X 2 X 12 X 3= 288 models). Selected algorithms are LR, MLP, PR, SMO, SVM, KStar, AR, RD, M5 model tree, RepTree, DT, M5'Rules.. Regression models were evaluated using three metrics: the correlation coefficient (R2-value), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). Equations (2,3,4) show how these metrics are calculated respectively:

$$R^2 = 1 - \frac{\sum_{m=1}^{n} (t_{m,m} - Y_{p,m})^2}{\sum_{m=1}^{n} (t_{m,m} - \bar{t}_{p,m})^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{m=1}^{n} |Y_{p,m} - t_{m,m}| \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{m=1}^{n} (y_{p,m} - t_{m,m})^2}{n}} \quad (4)$$

where **n** is the number of data patterns, $\mathbf{y_{p,m}}$ indicates the predicted class, $\mathbf{t_{m,m}}$ is the measured value of one data point m, and $\bar{t}_{m,m}$ is the mean value of all measure data points. I selected only R134a as a sample from their article, to show how they presented and evaluate their models. Equations (5,6,7) below show the selected models. Equation (5) represents the regression model for enthalpy using LR and PR algorithms. Equation (6) represents the regression model for entropy using M5' Rules algorithms. Equation (7) represents the regression model for volume using LR and PR algorithms.

$$h_{1=1.2395 * t + 0.9465 * p + 197.1715} \quad (5)$$

$$s_{t=0.0052 * t - 0.0022 * p + 1.0055} \quad (6)$$

$$v_{1} = 0.0012 * t + 0.0074 * p + 0.7499 \quad (7)$$

The best result for specific volume relating to vapor phase was obtained from MLP method. But, results are not very satisfactory. For this reason, equation obtained from this method for specific volume relating to vapor phase was not given. The following symbols are used in prediction models.

Targets : *h* :enthalpy (kJ/kg) , *s*: entropy (kJ/kg K),
v: specific volume (dm^3/kg)
Predictors: *t* : temperature (_C), *P*: pressure (bar)
Evaluation factors: *MAE* :Mean Absolute Error, R^2: correlation coefficient *RMSE*: Root Mean Squared Error,
Subscripts : *l* : liquid phase, *v*: vapor phase

C) Discussion about using Data Mining in Power Systems

It is very clear from above researches that data mining techniques could be used to solve many types of power systems' problems. Different prediction algorithms could be used, there is no restriction for selecting an algorithm, while it gives accurate results. Some researcher like E.U.Kucuksille [2] used many algorithms then selected the one that showed better results, while others selected only one algorithm for their problems like Hou et al. [32] who used neural network.

The research of E.U.Kucuksille et al. is really unique and could be a very good educational example of using data mining in engineering.

Their research is very clear because they followed the CRISP-DM model. Also they distinguish between analyst and customer responsibilities, by stating that their research deliverable is done by providing the analysis, but the deployment is the customer responsibility. However it could be much better if they provide more details about the data sets used for training and testing, is it separate or only one set. Also the authors mentioned that the best result for volume relating to vapor phase of R134a is obtained from MLP method. However, they stated that "results are not very satisfactory". However, they didn't show the model, and they didn't give any justification why they are not satisfied with these results. The same occurs when they presented volume relating to vapor phase of R404a, R407c and R410a, for three refrigerants the best result obtained when MLP method was used.

Moreover in results and discussion part authors just presented what is already mentioned in the comparison tables, they didn't discussed what is the meaning behind these numbers. Also the values calculated from obtained formulations were found to be in good agreement with actual values. The results of this work showed that data mining could be used for predicting accurately the thermodynamic properties of refrigerants for every temperature and pressure. The last point regarding conclusion section, although from models shown in the paper all algorithms were failed to give satisfied results for volume data sets in all refrigerants. However, authors generalize the good results, although it is better to be specific and mention that the models are not satisfactory in the case of vapor volume.

DATA MINING FOR POWER PLANTS

A) Using Data mining for Power Plant Optimization

Softstat which is recently acquired by Dell, proved that using data mining techniques in optimizing continuous processes of power plants, such as boiler performance in a coal-burning power plant, has superiority to traditional approaches like DOE (design of experiments), CFD (computational fluid dynamics) or statistical modeling. They published an article in 2007 about how to optimize a cyclone furnace of a coal power plant, for stable high flame temperatures and less undesirable slag to guarantee cleaner combustion. Although this goal could be achieved by traditional methods like DOE and CFD or even statistical modeling.

Computational fluid dynamics (CFD) is a branch of fluid mechanics that uses numerical analysis and algorithms to solve and analyze problems that involve fluid flows. One approach is to use CFD explicit theoretical complex and highly nonlinear models to understand how to set certain parameters, that optimize performance. Normally these methods are used to study parameter boundaries in a closed environment that is controlled by operators to ensure stable operations. However, in practice things may differ, so many obstacles faces applicability and effectiveness of CFD methods to optimize furnace performance of power plant, like: First : Theoretical pre-defined models will only study known parameters, so in a particular installation if there is any other parameter that affects performance, CFD will never look at it. Second, complexity of CFD models make it impossible to optimize. Data mining methods with its simplicity can summarize how parameters such as primary and secondary air flows, coal-flow, over fired air (OFA) distribution and so on will affect variability of flame temperatures.

Design Of Experiment (DOE) This approach is a trail-and-error testing of a real furnace, which is done by focusing in one parameter at a time and studying its effect on the target attribute (flame temperatures). Specific configurations for specific parameter settings are tried and their results are recorded, then linear model is built according to these results. There are some of shortcomings using DOE methods. First, the above model is very simple. In practice, the relationships between the parameters and target cannot be adequately represented by such an equation. Second, the specific DOE that is chosen to conduct the testing will limit, and some times "pre-ordain" the results. Moreover, the data collected during the parametric tests will not allow the analyst to verify that the results are wrong.

Data Mining to overcome CFD and DOE shortcomings

The powerful of data mining is the use of real data. Most of today's power plants are equipped with data gathering and storage systems that makes data available for on-line monitoring and future analysis. Data mining tools can deal with

this huge amount of data, thousands or even millions of instances for hundreds of attributes, it can perfectly detects useful patterns in those data that allow us to improve furnace performance.

Finding patterns.

First, researcher has to find which parameters affects the value of his target which is in this case the Flame Temperature. This is done by applying feature selection methods. These methods test whether combinations of parameters have any systematic relationship to flame temperature. Once important parameters have been selected, then various algorithms can be applied to build the model which shows how exactly these parameters have effect in specifying the value of flame temperature. After that the model could be deployed. The expected deliverables from such a project could be recommendations to make minor adjustments to specific parameters that dramatically improve the robustness of flame temperature [36].

Hao Zhou **s**howed a modelling Nox emission from coal fired utility boiler is critical to develop predictive emissions monitoring system (PEMS) and to implement combustion optimization software package for low Nox combustion. Hao Zhou presented an efficient NOx emissions model based on support vector Regression (SVR), and compares its performance with traditional modelling techniques ,like back propagation (BPNN) and generalized regression (GRNN) neural networks. Hao used NOx emissions data from an actual power plant, to train and validate the SVR model. Moreover ,an ant colony optimization(ACO)based technique was proposed to select the generalization parameter C and Gaussian kernel parameter g. The focus is on the predictive accuracy and time response characteristics of the SVR model [37].

Softsat is an specialized company in data analysis projects, in their research [36] they started by feature selection to identify only the parameters that have effects in their problem, then they built the prediction model. This practice is very important and better to be followed in such problems to reduce the parameters provided to the learning machine. Their research is very interesting, however they didn't nominate the team members who did this research.

B) Data-Driven Performance Optimization of Wind Farms

Wind energy is a green energy source and does not cause pollution, due to this fact it rreceives much attention recently. The generation of wind energy is relatively new. Therefore, the performance of wind power farms has not been thoroughly studied. One of the weakest points in wind power generation is the low predictive accuracy of the energy output. The Iowa Energy Center initiated a research project that supports the innovations in wind power. The proposed research offers a novel approach to modeling the performance of individual wind turbines as well as their collection (a wind farm) . A solution for prediction of wind farm performance should be able to predict the amount of energy to be produced on different time scales, e.g., 15 min, 2 hour, a day, and so on. The basic methodology used is this research is data mining, because this new industry generates huge amount of data that have not yet been explored. Professor Andrew Kusiak from Intelligent Systems Laboratory in The University of Iowa, was the Principal Investigator of this research project. Below is summary of some papers about wind power, more similar researches could be found in the research project final report [38].

1. Wind Speed and Power Prediction

Time series models for predicting the power of a wind farm at different time scales, i.e., 10-min and hour-long intervals have been developed with data mining algorithms. Five different data mining algorithms have been tested on various wind farm datasets. Two of the five algorithms performed particularly well. The support vector machine regression algorithm provides accurate predictions of wind power and wind speed at 10-min intervals up to 1 h into the future, while the multilayer perceptron algorithm is accurate in predicting power over hour-long intervals up to 4 h ahead. Wind speed can be predicted fairly accurately based on its historical values; however, the power cannot be accurately determined given a power curve model and the predicted wind speed. Test computational results of all

time series models and data mining algorithms are discussed. The tests were performed on data generated at a wind farm of 100 turbines [39].

2. Monitoring Power Curves

A data-driven approach to the performance analysis of wind turbines is discussed. Turbine performance is captured with a power curve. The power curves are constructed using historical wind turbine data. Three power curve models are developed, one by the least squares method and the other by the maximum likelihood estimation method. The models are solved by an evolutionary strategy algorithm. The power curve model constructed by the least squares method outperforms the one built by the maximum likelihood approach. The third model is non-parametric and is built with the k-nearest neighbor (k-NN) algorithm. The least squares (parametric) model and the non-parametric model are used for on-line monitoring of the power curve and their performance is analyzed [40].

3. Condition Monitoring and Fault Detection

The rapid expansion of wind farms has drawn attention to operations and maintenance issues. Condition monitoring solutions have been developed to detect and diagnose abnormalities of various wind turbine subsystems with the goal of reducing operations and maintenance costs. This research explores fault data provided by the supervisory control and data acquisition system and offers fault prediction at three levels: (1) fault and no-fault prediction; (2) fault category (severity); and (3) the specific fault prediction. For each level, the emerging faults are predicted 5-60 min before they occur. Various data-mining algorithms have been applied to develop models predicting possible faults. Computational results validating the models are provided [41].

4. Optimization of wind Turbine performance using data-Driven Models [24]

Andrew Kusiak et al proved that: using data driven models gives accurate result in wind power turbine optimization. Their goal is to model and optimize wind turbine performance, they did this by three objectives, maximization of the power produced by a wind turbine, and minimization of vibrations of the turbine's drive train and tower. Below is some details about this research which is highlighted by author because of its clarity.

Data Set

The data used in their research was obtained from a 150 MW wind farm. They used two data sets of wind turbine : 10-sec data and 1-min. The first set was collected directly from the SCADA system, while the 1-min set was calculated by taking the average of all parameters across one minute from the first set (the 10-sec set). Both sets were collected for two months, all instances were time stamped. Although more than 120 parameters were available in the SCADA system, only parameters related to wind turbine vibrations and their power output were selected. Feature selection is done according to the advice of domain experts.

Models

In their research Andrew et al. presented three models, two of which related to vibration while the third is related to the output power. Tow vibration sources were considered; vibrations due to the air passing through the wind turbine (this type is non controllable) and vibrations caused by forces originating with the control system that affect the torque and the blade pitch angle (this is the controllable vibration).

1) Drive Train Vibration Model:

2) Tower Vibration Model:

3) Power Output Model :

Figures 4,5,6 shows the predicted values compared to the actual values of the first 50 instances of each model. Using a histogram to present their results, makes the model evaluation easier and meaningful.

Evaluation of Models

Regression is used to predict the classes because all data sets and their classes are numeric. These models were evaluated using the following four metrics:

- **MAE:** Mean Absolute Error.
- **SD of MAE**: Standard Deviation of Mean Absolute Error.
- **MAPE**: Mean Absolute Percentage Error.
- **SDofMAPE**: Standard Deviation of Mean Absolute Percentage Error.

Equations (8,9,10,11) below show how these metrics are calculated respectively:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{8}$$

$$SDofMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(|\hat{y}_i - y_i| - \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i| \right)^2} \tag{9}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i}\right) \times 100\% \quad (10)$$

Where $\hat{\mathbf{y}}_\mathbf{i}$ and $\mathbf{y}_\mathbf{i}$ are the predicted and observed $\boldsymbol{i}$ th instance respectively, and n is the total number of instances.

$$SDofMAPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{\hat{y}_t - y_t}{y_t}\right| - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_t - y_t}{y_t}\right|\right)^2} \quad (11)$$

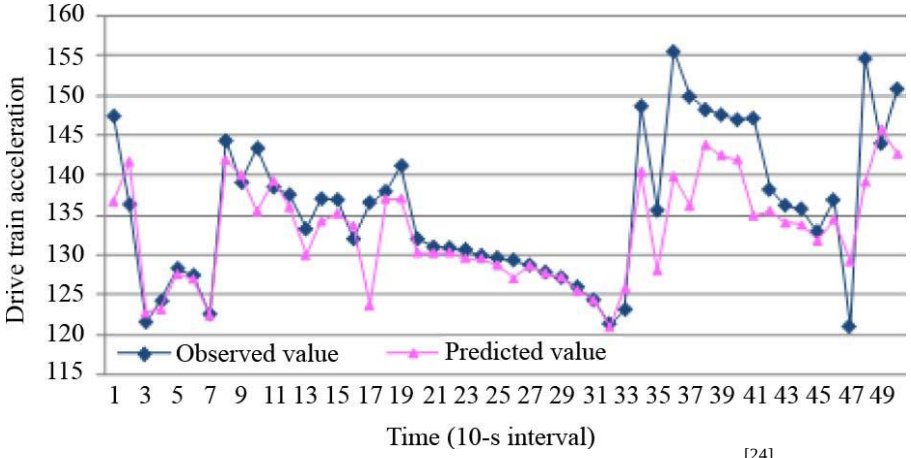

Fig. 4. First 50 test points of the drive train acceleration for 10-s data [24]

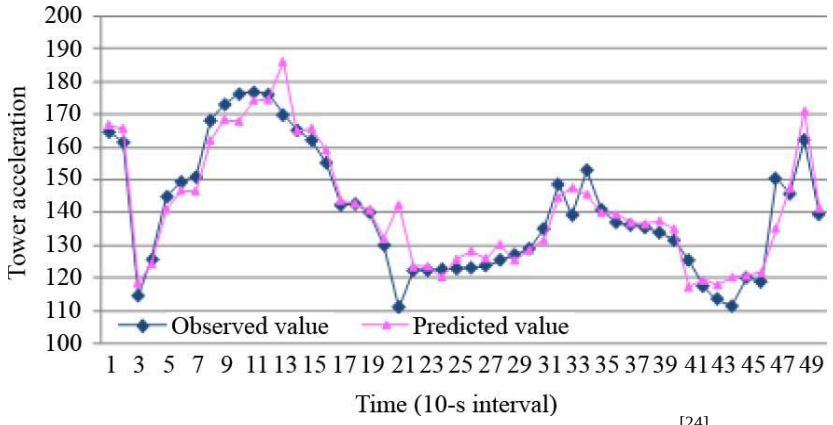

Fig. 5. First 50 test points of the tower acceleration for 10-s data [24]

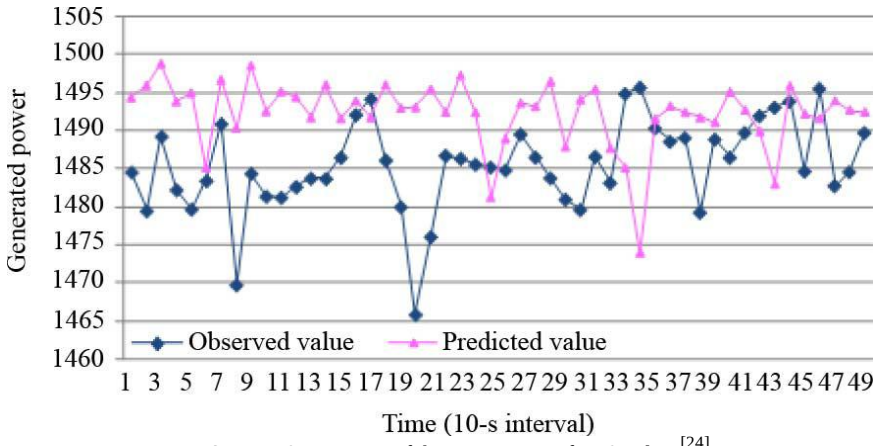

Fig. 6. First 50 test points of the power output for 10-s data [24]

5. Discussion about using Data Mining in Wind Power Plants

Professor Andrew et al are following clear methodology, they expressed clearly the description of their data sets and the preprocessing which was done to prepare data sets for machine learning tools. They presented their results in a very simple and direct forward way, using simple tables followed by comments to show the intuition behind numbers in these tables. Also they used histograms to compare between the actual and predicted values for each target separately. However, they didn't show why they choose NN in their models, may be if they selected another algorithm they could find better results.

C) Boiler Efficiency

Boiler is the primary source of energy in the power plants. Because the efficiency of combustion is crucial to the performance of boilers; many researchers focused in this area. The intelligent control approaches in combustion can be grouped into three categories: rule-based expert systems, soft computing, (i.e., neural networks, fuzzy logic, and evolutionary computation)and the the third one is hybrid systems which combines analytical modeling with the soft-computing methods. The intelligent control concept can be extended by incorporating data-mining algorithms.

1. Summary of Some Boiler efficiency Researches

Below is summary of some work related to the efficiency of boilers using data mining.

Zhe Song and Andrew Kusiak [42] applied a data-mining approach to develop a model for optimizing the efficiency of an electric utility boiler subject to operating constraints. Selection of process variables to optimize combustion efficiency is discussed. The selection of variables a coal fired boiler, is critical to control of combustion efficiency. Two schemes of generating control settings and updating control variables were evaluated. The first is based on both controllable and non controllable variables. While the second scheme merged response variables in clustering process. The process control scheme based on the response variables produces the smallest variance of the target variable due to reduced coupling among the process variables.

Chu *et al.* [43] applied a neural network to predict the performance index and some non-analytical constraints, thus speeding up the trial-and-error process of finding the optimal operating points, thereby optimizing the boiler's combustion process.

Büche et al. [44] applied an evolutionary computation algorithm to determine the optimal design of the burner reducing the emissions of NOx as well as the pressure fluctuation of the flame.

Cass et al. [45] combined the neural network and evolutionary computation techniques to determine the optimal fuel/air ratio.

Miyayama et al. [46] developed an expert system for combustion control using fuzzy logic and applied it to the coal-fired power plant.

Ogilvie et al. [47] applied the association-rule algorithm to mine relationships among the parameters of a power plant. The inducted rules were intended for an expert system.

Booth and Roland [48] developed a neural network model to optimize the boiler's operations and thus reduce the emission of NOx and improve the boiler's performance.

Chong et al. [49] applied a neural network model to identify the dynamic process of the nitrogen oxides and carbon monoxide emissions.

Burns *et al*. [50] used a genetic wrapper approach to select a subset of parameters for mining boiler data to improve combustion efficiency.

2. Discussion about Boiler Efficiency

All of above researches in this section are studying implementation of data mining in boiler efficiency. The majority of researches in this field are targeting to improve combustion efficiency or reduce emission of Nox. It is clear that most of researchers used neural networks.

D) The Use of SVM in Thermal Power Plants

In recent years, support vector machines (SVM) , has been considered as one of the most effective supervised learning algorithms in many pattern recognition problems. It has been reported that SVM provides a better classification result than other methods such as neural networks or decision trees. Moreover, SVM has been widely applied to fault detection and diagnosis in production environment. For examples, Hsu et al. [51] integrated a feature extraction technique,

independent component analysis (ICA), into SVM to develop an intelligent fault detector for non-Gaussian in multivariate processes.

1. Summary of Some SVM Researches

Below is a summary of some selected researches that use SVM in power plants:

Li et al. [52] combined another dimension reduction method, partial least squares (PLS), with SVM to increase the performance of on-line fault detection in batch processes.

In the work of Zhang [53], both Kernel Independent Component Analysis (KICA, for non-Gaussian distribution) and Kernel Principal Component Analysis (KPCA, for Gaussian distribution) are used for fault detection in, named Tennessee Eastman process, which is a complex non-linear process created by Eastman Chemical Company.

Mahadevan and Shah [54] used one-class SVM for fault detection and diagnosis, they claimed that their approach outperformed Principal Components Analysis (PCA) and Dynamic Principal Components Analysis (DPCA).

Kai-Ying Chen et al [55] proposed a SVM based model to predict failures of turbines in a thermal power plant. In order to handle the huge amount of collected data, they started by feature selection techniques to eliminate irrelevant and noisy data, then they built their model based on the new clean dataset . To evaluate the effectiveness of their model, they used a real-world data from a thermal power company. Their SVM model can successfully detect the types of turbine faults with a high degree of accuracy (greater than 90%). Their method can assist the power plant engineers to find failure types without referring to the manufacturers, because normally manufacturers consultation is very expensive. Performance and effectiveness of their model which is based on SVM was compared to linear discriminant analysis (LDA) and back-propagation neural networks (BPN).

Although feature selection results in a slight loss of classification accuracy, but results within acceptable limits, and saving data processing time by eliminating the irrelevant attributes. However, if computational costs is not considered, Support Vector Machine Recursive Feature Extraction (SVM-RFE) algorithm and Genetic Algorithms (GA) might result in better performances. Their model can predict the fault types. To enhance the model to be able to discover the root cause of failure, they proposed further studies and suitable equipment such as sensors and data storage devices to provide more data for training.

2. Discussion about using SVM in Thermal Power Plants

Their research is very well prepared, it consists of feature selection and prediction. However to select SVM they depends on other researches, although according to datasets results that may differ. Moreover, they did feature selection using C4.5 algorithm then classification using SVM, according to Ian Witten [3] it is better to use same algorithms for both classification and feature selection evaluation.

From all above works, it is clear that SVM is one of the most useful fault detection approaches in industrial processes. Moreover SVM has been usually combined with feature selection techniques including ICA, PLS, PCA, KPCA, KICA, and DPCA.

E) Visual Data Mining: A case study in Thermal Power Plant

The basic idea of Visual Data Mining is to present the data in a visual format, to get better understanding of data, draw conclusions, and directly interact with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis, they also have a high potential for exploring large databases. Moreover these techniques provide a much higher degree of confidence in exploration findings.

Md Fazullula, et al applied in their research a visual data mining technique with parallel coordinates to a Thermal Power Plant data. AUTORULE, a visual analytics software, was used to process 744 records dataset which was collected at different loads. Sixteen predictors were considered for this research *(Coal Speed , Coal Flow, Primary Air Flow, Secondary Air Temperature, Burner Tilt, Induced Draught, Forced Draught, Flue Gas Temperature, Un burnt Oxygen, Sulphur Di Oxide, Nitrogen Di Oxide, Carbon Di Oxide , Date, Time, Total Air Flow, Primary Air Temperature)* to identify some of the possibilities where Nox could be high. Any record that had a Nox greater than 250 was considered to be as higher class and below 250 as lower class. As a result of their research, they found that when the total airflow is low, NOX is high. This was surprising as the NOX production is possible when

O_2 and N presence in air is high. i.e. when the total airflow is high. So, these results were submitted to domain expert for further investigation to identify the possible causes for this phenomenon [56].

Prasad et al. [57] proposed a histogram based method to monitor and maximize the performance of thermal power plants. Therefore, building an intelligent system for the fault prediction of any thermal power plants most valuable equipment, namely turbines, has become necessary. Since artificial intelligence techniques can improve prediction accuracy and decrease people involvement, it is very important to select an appropriate learning algorithm.

F) Using Data Mining in Thermal Power Plants

Huang et al. [58.] used principle component analysis (PCA) and T2 statistics to inspect different types of faults in a thermal power plant. They indicated that the efficiency and availability depend on reliability and maintainability, to raise efficiency, the equipment of thermal power plants is becoming larger and more complex. However, due to lack of manpower and information resources, the diagnosis and repair of failed equipment cannot be done immediately. Identifying the failure types of steam turbines and their root causes is time consuming, and requires professional knowledge in materials and mechanical engineering. Actually, thermal power plant engineers can only handle routine and direct maintenance tasks. Complex faults require intervention from technical support and equipment manufacturers. These types of tasks are very expensive and require special experiments, which leads to long downtime and causes production losses. The concept of e-maintenance has been introduced to mitigate all these problems and easily identify the root cause of failures, also it reduces the failures of production systems, eliminate unscheduled shutdown maintenances, and improves productivity [59].

1. Summary of Some Researches about using Data Mining in Thermal Power Plants

In an e-maintenance system, the intelligent fault detection system plays a crucial role for identifying failures. Data mining techniques are the core of such intelligent systems and can greatly enhance their performance. Recently, several data mining techniques such as artificial neural networks, fuzzy logic systems, genetic algorithms, and rough set theory have all been employed to assist the detection and condition monitoring tasks. For example,

Yang and Liu [60] presented a hybrid-intelligence data mining framework which involves an attribute reduction technique and rough set theory to diagnose the faults of boilers.

Shu [61] established an interactive data mining approach based inference system to solve the basic technical challenge and speed up the discovery of knowledge in nuclear power plant. Besides, some related works designed data mining based models for failure inspections, but not for fault predictions, such as the work of Yang and Liu [60]. Vast amounts of data describing process variables for boilers and turbines have been used for monitoring, control and over-limit alarms.

Ilamathi P, et al [62] used ANN to build a prediction model that predicts nitrogen oxides emission from a 210 MW coal fired thermal power plant. The coal combustion parameters (oxygen concentration in flue gas, coal properties, coal flow, boiler load, air distribution scheme, flue gas outlet temperature and nozzle tilt) were used as inputs and nitrogen oxides as output of the model. Values predicted by ANN model were verified with the actual values.

2. Discussion about Using Data Mining in Thermal Power Plants

Due to the complexity of thermal power plants; it is very difficult to find the root cause of any failure. Because of this, the use data mining techniques in thermal power plants is very useful. Above researches shows that data mining is used for different purposes in thermal power plants like: identifying failures, predicts nitrogen oxides emission.

DISCUSSION

From the survey summarized in this article, it is obvious that data mining techniques could be used to solve many types of problems in electromechanical industry, power systems and power plants. Many prediction algorithms could be used. According to the available datasets; different results could be obtained.

Some researchers like E.U.Kucuksille [2] used many algorithms then selected the one that showed

better results, while others like Hou et al.[18] depends on one algorithm. E.U.Kucuksille [2] followed the CRISP-DM model; that makes their research very well organized. They clearly stated that their research deliverable is the analysis result, but the deployment is the customer responsibility. By this separation of duties they can guarantee smooth implementation for their work. Their research is very useful, but it would be great if they provide justification why they were not satisfied with some results.

Professor Andrew Kusiak and his team were able to solve a lot of problems in wind power plants using different data mining techniques. They followed a clear methodology, by properly describing their datasets, and presenting results by using simple tables and histograms, to compare between the actual and predicted values. In their article *Optimization of Wind Turbine Performance With Data-Driven Models* [10], they used NN. However they didn't mention why they selected NN to solve such a problem.

The boiler is the primary source of energy in the power plants. So many researchers focused in using data mining to improve the efficiency of combustion or reduce emission of Nox. It is clear that most of researchers used neural networks.

Many researcher used SVM in their researches. Kai-Ying Chen et al [38] proposed a SVM based model to predict failures of turbines in a thermal power plant. The research started by feature selection then prediction. However, to select SVM they depends on other researches, although according to datasets results that may differ. Moreover, they did feature selection using C4.5 algorithm then classification using SVM. But according to Ian Witten [3] it is better to use same algorithms for both classification and feature selection evaluation.

From all work summarized in this paper we can see that; if data mining techniques are properly applied they can solve various types of power plants problems. Datasets must be well prepared, and checked against all available algorithms to select the most suitable one for the problem. All these researches proved that this area needs a lot of exploration, and data mining is very poor full in solving power plant problems.

CONCLUSION

Recently data mining techniques is being applied in many areas of power plants. The reader can clearly noticed that, all above mentioned successful implementations of data mining techniques in power systems is just the beginning. Unlimited number of problems in power plants could be solved by various data mining techniques. From all above reviews we can conclude that: Whenever data is available in power plant and problem is clearly stated, answers could be obtained by data mining technique. That means to properly solve such a problem, there should be a domain expert and a data scientist, and they should work together to come up with a proper solutions.

Mainly, regression is the branch of data mining that is able to deal with various types of problems in power plants. The dataset is the main factor that determines the most suitable algorithm for a specific application.

REFERENCES

[1] Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau, An Overview of Data Mining Techniques Applied to Power Systems, Data Mining and Knowledge Discovery in Real Life Applications, (2009) , ISBN 978-3-902613-53-0, pp. 438.

[2] Ecir Ug˘ur Küçüksille, Res at Selbas , Arzu Sencan , Prediction of thermodynamic properties of refrigerants using data mining. ELSEVIER Energy Conversion and Management , (2011), 52: 836–848.

[3] Ian H. Witten, Frank Eibe, Mark A. Hall, (2011). *Data Mining : Practical Machine Learning Tools and Techniques 3rd ed.* Morgan Kaufmann New York

[4] http://www.saedsayad.com/data_mining_map.htm

[5] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499

[6] XindongWu ,Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach ,David J. Hand , and Dan Steinberg (2008) .Top 10 algorithms in data mining . Knowl Inf Syst (2008) 14:1–37.

[7] Colin Shearer , The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data warehousing,(2000), 16: 419 – 438

[8] Zhou ZH. , Three perspectives of data mining, Artif Intell, (2003), 143:139–46.

[9] Chiu S-H, Chen C-C, Lin T-H. Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. Artif Intell (2008), 44:221–31

[10] R K Kapooria, S Kumar, K S Kasana, An analysis of a thermal power plant working on a Rankine cycle:

a Theoretical Investigation. Journal of Energy in Southern Africa, (2008) 19:77-83.
[11] www.learnengineering.org (retrieved 11 Nov 2015).
[12] Ramos, S.; Vale, Z. (2008). Data mining techniques application in power distribution utilities. Transmission and Distribution Conference and Exposition, pp. 1-8. ISBN. 978-1-4244-1903-6, Chicago, April 2008.
[13] Saibal Chatterjee, Sivaji Chakravorti, Chinmoy Kanti Roy and Debangshu Dey. (2008) Wavelet network-based classification of transients using dominant frequency signature, Electric Power Systems Research, Vol. 78, No. 1, (January 2008) 21-29.
[14] Figueredo, V.; Rodrigues F.; Vale, Z.; Gouveia, J. B. (2005). An eletric energy Consumer characterization Framework based on data mining techniques. IEEE Transactions Power Systems, Vol. 20, No. 2., May 2005 , 596- 60), ISSN 1558-0679
[15] Dola, H.M.; Chowdhury, B.H. (2005). Data mining for distribution system fault classification. Power Symposium, 2005. Proceedings of the 37th Annual North American,pp. 457 – 462, ISBN 0-7803-9255-8, October 2005.
[16] Mori, H.; Kosemura, N.; Kondo, T.; Numa, K.; (2002). Data mining for short-term load forecasting. Power Engineering Society Winter Meeting. pp. 623 – 624, ISBN 0-7803-7322-7, Jan 2002
[17] Hagh, M.T.; Razi, K.; Taghizadeh, H. (2007). Fault classification and location of power transmission lines using artificial neural network, International Power Engineering Conference, pp. 1109 – 1114, ISBN 978-981-05-9423-7, Singapore, Dec 2007.
[18] Silva, K. M.; Souza, B. A.; Brito, N. S. D. (2006). Fault detection and classification in transmission lines based wavelet transform and ANN. IEEE Transaction on Power Delivery, Vol 21 , No. 4, (October 2006) 2058-2063, ISSN 0885-8977
[19] Costa, F. B.; Silva, K. M.; Souza, B. A.; Dantas, K. M. C.; Brito, N. S. D. (2006). A method for fault classification in transmission lines bases on ANN and wavelet coefficients energy. International Joint Conference Neural Networks. pp. 3700 – 3705, ISBN 0-7803- 9490-9, Vancouver, July-2006.
[20] Vasilic, S.; Kezunovic, M. (2002). An Improved Neural Network Algorithm for Classifying the Transmission Line Faults. IEEE Power Engineering Society Winter Meeting, pp.918 – 923. ISBN 0-7803-7322-7, Jan 2002.
[21] Dash, P.K.; Samantaray, S.R.; Panda, G. (2007). Fault Classification and Section Identification of an Advanced Series-Compensated Transmission Line Using Support Vector Machine. IEEE Transactions on Power Delivery, Vol 22. No. 22, (January 2007) 67 – 73, ISSN 0885-8977.
[22] Vasilic, S.; Kezunovic, M. (2005) Fuzzy ART Neural Network Algorithm for Classifying the Power System Faults. IEEE Transactions on Power Delivery, Vol. 20, No. 2, (April 2005) 1306-1314, ISSN 0885-8977.
[23] Huisheng Wang; Keerthipala, W.W.L. (1998). Fuzzy-neuro approach to fault classification for transmission line protection. IEEE Transactions Power Delivery, Vol.13, No. 4, 1093-1104, ISSN 0885-8977.
[24] Andrew Kusiak, Zijun Zhang, and Mingyang Li, Optimization of Wind Turbine Performance With Data-Driven Models. IEEE Transactions On Sustainable Energy, (2011), 1:66-76.
[25] Himani Tyagi and Rajat Kumar (2014). Optimization of a Power Plant by Using Data Mining and its Techniques. International Journal of Advances in Science Engineering and Technology 2:83-87
[26] www.statsoft.com (retrieved 29 Sep 2015).
[27] G. K. F. Tso, K. K. W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural Networks, Energy, 32 (2007) 1761–1768.
[28] A. Şencan, Modeling of thermodynamic properties of refrigerant/absorbent couples using Data Mining Process, Energy Conversion and Management, 48 (2007) 470– 480.
[29] V. Figueiredo, F. Rodrigues, J. G. Gouveia, An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques, IEEE Transactions on Power Systems, 20(2005) 596-602.
[30] A. Kusiak, A. Burns, F. Milster, Optimizing combustion efficiency of a circulating fluidized boiler: A data mining approach, International Journal of Knowledge Based Intelligent Engineering Systems, 9(2005) 263-274.
[31] X. Lu, Z. Y. Dong, X. Li, Electricity market price spike forecast with data mining techniques, Electric Power Systems Research, 73 (2005) 19–29.
[32] Zhijian Hou, Zhiwei Lian, Ye Yao, Xinjian Yuan, Data mining based sensor fault diagnosis and validation for building air conditioning system, Energy Conversion and Management 47 (2006) 2479–2490.
[33] Zhun Yu, Fariborz Haghighat, Benjamin C.M. Fung, Hiroshi Yoshino, Energy and Buildings 42 (2010) 1637–1646.
[34] Andrew Kusiak, Mingyang Li, Fan Tang, Modeling and optimization of HVAC energy consumption, Applied Energy 87 (2010) 3092–3102.
[35] Haiming Zhou1, Dunnan Liu2, Dan Li1, Guanghui Shao3, Qun Li, Operating Analysis and Data Mining System for Power Grid Dispatching, Energy and Power Engineering, 5,(2013), 616-620.
[36] www.softstat.com (retrieved 21 dec 2015).
[37] Zhou H, Zhao J, Zheng L , Wanga C, Fa k, Cen F, Modeling Nox Emissions From Coal- Fired Utility Boilers Using Support Vector Regression With Ant

Colony Optimization, Engineering Applications of Artificial Intelligence 25, (2012), 147–158.
[38] http://www.iowaenergycenter.org/data-driven-performance-optimization-of-wind-farms/?golargetext=true 0 (retrieved 13 Nov 2015).
[39] A. Kusiak, H.-Y. Zheng, and Z. Song, Short-Term Prediction of Wind Farm Power: A Data-Mining Approach, *IEEE Transactions on Energy Conversion*, Vol. 24, No. 1, (2009), 125-136.
[40] A. Kusiak, H.-Y. Zheng, and Z. Song, On-line Monitoring of Power Curves, Renewable Energy, Vol. 34, No. 6, (2009) 1487-1493.
[41] A. Kusiak and W.Y. Li, The Prediction and Diagnosis of Wind Turbine Faults, *Renewable Energy*, Vol. 36, No. 1, (2011), 16-23.
[42] Zhe Song and Andrew Kusiak, Constraint-Based Control of Boiler Efficiency: A Data-Mining Approach, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, (2007) , VOL. 3, NO. 1.
[43] J. Z. Chu, S. S. Shieh, S. S. Jang, C. I. Chien, H. P. Wan, and H. H. Ko, "Constrained optimization of combustion in a simulated coal-fired boiler using artificial neural network model and information analysis," FUEL, vol. 82, no. 6, (2003), pp. 693–703.
[44] D. Büche, P. Stoll, R. Dornberger, and P. Koumoutsakos, Multiobjective evolutionary algorithm for the optimization of noisy combustion processes, IEEE Trans. Syst., Man, Cybern. C, vol. 32, no. 4, (2002), pp. 460–473, Nov.
[45] R. Cass and B. Radl, Adaptive process optimization using functional link networks and evolutionary optimization, Control Eng. Practice,(1997) vol. 4, no. 11, pp. 1579–1584.
[46] T. Miyayama, S. Tanaka, T. Miyatake, T. Umeki, Y. Miyamoto, K. Nishino, and E. Harada, A combustion control support expert system for a coal-fired boiler, in Proc. IEEE Industrial Electronics, Control, Instrumentation, Kobe, Japan, (1991), pp. 1513–1516.
[47] T. Ogilvie, E. Swidenbank, and B. W. Hogg, Use of data mining techniques in the performance monitoring and optimization of a thermal power plant, in Proc. Inst. Elect. Eng. Colloq. Knowledge Discovery Data Mining, 1998, pp. 7/1–7/4.
[48] R. C. Booth and W. B. Roland, Neural network-based combustion optimization reduces NOx emissions while improving performance, in Proc. IEEE Industry Applications Dynamic Modeling Control Applications Industry Workshop, (1998), pp. 1–6.
[49] A. Z. S. Chong, S. J. Wilcox, and J. Ward, Neural network models of the combustion derivatives emanating from a chain grate stoker fired boiler plant, in Proc. Inst. Elect. Eng. Seminar Advanced Sensors Instrumentation Systems Combustion Processes, (2002), pp. 6/1–6/4.
[50] A. Burns, A. Kusiak, and T. Letsche, Mining transformed data sets, in Knowledge-Based Intelligent Information and Engineering Systems, R. Khosla, R. J. Howlett, and L. C. Jain, Eds. Heidelberg, Germany: Springer, (2004), vol. I, LNAI 3213, pp. 148–154.
[51] C.-C. Hsu, M.-C. Chen, L.-S. Chen, Intelligent ICA–SVM fault detector for non-Gaussian multivariate process monitoring, Expert Systems with Applications 37 (4) (2010) 3264–3273.
[52] Y. Li, Z. Wang, J. Yuan, On-line fault detection using SVM-based dynamic MPLS for batch processes, Chinese Journal of Chemical Engineering 14 (6) (2006) 754–758.
[53] Y. Zhang, Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM, Chemical Engineering Science 64 (5) (2009) 801–811.
[54] S. Mahadevan, S.L. Shah, Fault detection and diagnosis in process data using one class support vector machines, Journal of Process Control 19 (10) (2009) 1627.
[55] Kai-Ying Chen, Long-Sheng Chen, Mu-Chen Chen, Chia-Lung Lee, Using SVM based method for equipment fault detection in a thermal power plant, Elsevier Computers in Industry 62 (2011) 42–50.
[56] Md Fazullula s, Mr. Praveen M P, S.S. Mahesh Reddy, Visual Data Mining: A case study in Thermal Power Plant. IJISET - International Journal of Innovative Science, Engineering & Technology, (2014). Vol. 1 Issue 6: 110 – 115.
[57] G. Prasad, E. Swidenbank, W. Hogg, A novel performance monitoring strategy for economical thermal power plant, IEEE Transactions on Energy Conversion 14 (3), (1999) 802–809.
[58] X. Huang, H. Qi, X. Liu, Implementation of fault detection and diagnosis system for control systems in thermal power plants, in: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, June 21–23, 2006.
[59] B. Iung, A.C. Marquez, Editorial: special issue on e-maintenance, Computers in Industry 57 (2006) 473–475.
[60] P. Yang, S.S. Liu, Fault Diagnosis for boilers in thermal power plant by data mining, in: Proceedings of Eighth International Conference on Control, Automation, Robotics and Vision, Kunming, China, December 6–9, 2004.
[61] Y. Shu, Inference of power plant quake-proof information based on interactive data mining approach, Advanced Engineering Informatics 21 (3) (2007) 257–267.
[62] Ilamathi P, Selladurai V, Balamurugan K., Predictive modelling and optimization of Nox emission from power plant, IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 1, No. 1, (2012), pp. 11~18 ISSN: 2252-8938.