ANNALS of the ORADEA UNIVERSITY.
Fascicle of Management and Technological Engineering, Volume IX (XIX), 2010, NR2

ASPECTS REGARDING DATA MINING APPLIED TO FAULT DETECTION

DONCA Gheorghe, MIHĂILĂ V. Ioan
University of Oradea, donca.gheorghe@gmail.com

Keywords : FDD, maintenance, data-driven, database

Abstract : Systems health management includes fault detection, fault diagnosis (or fault isolation), and fault prognosis. We define prognosis to be detecting the precursors of a failure, and predicting how much time remains before a likely failure. Algorithms that use the data-driven approach, to prognosis, learn models directly from the data, rather than using a hand-built model based on human expertise. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools are used to identify correlations and to prognoses behaviors and future trends from data.

1. INTRODUCTION

There are several ways of classifying approaches to the problem of fault detection in an engineering system. Main mod of classifying fault detection techniques depends on whether the diagnosis assessment is based on deterministic information (e.g., one obtained from a model) or on stochastic information (e.g., historical, statistical data). We can think of existing solutions to the problems of performing fault detection as belonging to one (or perhaps even both) of two types: *data-driven* – also called *model-free* – techniques and *model-based* techniques, although other classifications exist. Data-driven techniques include data mining, machine learning and computational intelligence. Model-based techniques more commonly involve the description of a system through mathematical models of the physical laws governing its behavior.

The main differences between data and physical models lie in that the process information provided might be incomplete by the former but complete by the latter: the former need to go mining and learning to extract useful information contained in data, whereas the latter provides the useful information completely as the direct outcome.

Recently, due to the rapid development of computer related fields, real-time, abundant data acquisition has been made possible through the low cost and high performance of computing facility, measurement and storage equipments. Therefore the data analysis, characterized as the data mining methodology, is widely used to extract valuable information from the raw data by analyzing the correlation structure among process variables of the system.

The data mining methodology with the generalization and interpretation ability as its two main goals can produce satisfactory results for the system with abundant data but meager information of the system because it makes use of only the raw data from the system and doesn't require any rigorous theoretical knowledge of the system.

Even though most of production processes run according to the deterministic physical law, the theoretical process model based on the first principle has the possibility to deviate to a great degree from real process due to the characteristics of real processes, that is, the highly integrated unit processes, reflux system in consideration of economic advantage, complicated control loops for control purposes, measurement bias, equipment malfunction or other external disturbances. Therefore lots of recent researches are focusing on how to apply the data mining methodology to the abundant process raw data analysis, which, in the past, had been used as an auxiliary tool to help decide the proper control input. Three of the main application areas where the data mining methodology can prove efficient are process monitoring, process identification and property estimation.

Businesses use data mining to search and analyze databases for hidden patterns and may

find trends and predictive information that experts may miss because it lies outside their expectations or because they never thought to look for correlations there.
The four key fields of interests, from maintenance perspective are control, decision making, scheduling, and fault detection and diagnosis. The other four areas of data-driven methods for maintenance are computational intelligence, data-mining, control system theory, and machine learning methods, as shown in figure 1.
It is worthwhile pointing out that the possible combination of data mining and computational intelligence methods has been also explored, for instance integrating data mining with neural network, fuzzy logics, and evolutionary algorithms.

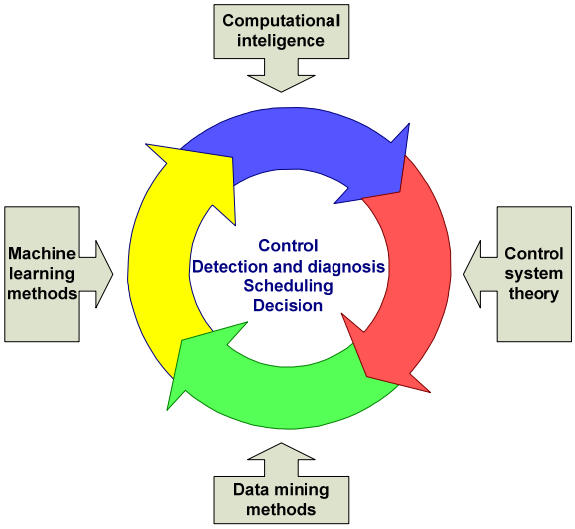

Figure 1. The 4+4 key fields [1]

2. DATA MINIG

Data mining can be defined as the science of extracting useful information from large data sets or databases. Data mining is used for building empirical models, which are based not on the underlying theory about the process or mechanism that generated the data. Data mining, as the name suggests it, is data-driven, and it provides a description of the observed data. Its fundamental objective is to provide insight and understanding about the structure of the data and its important features, and to discover and extract patterns contained in the data set. This discipline also referred to as knowledge discovery in databases (KDD), is a process of extracting implicit, previously unknown, and potentially useful information from data. Data mining brings together a multitude of disciplines, such as database systems, statistics, artificial intelligence, data visualization, and others.

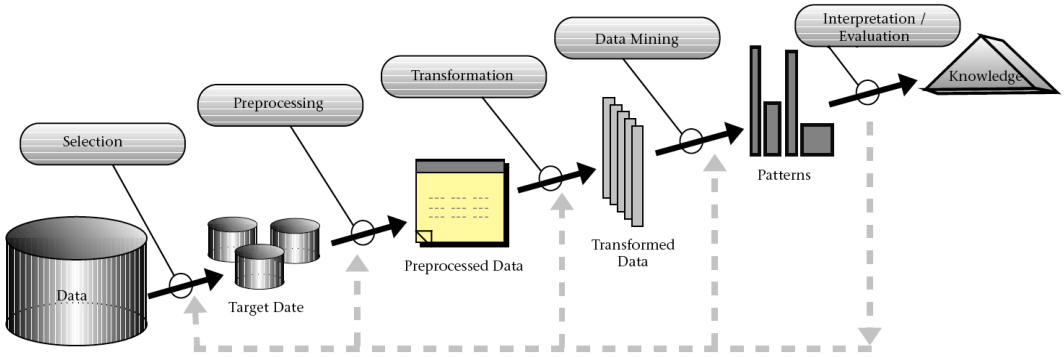

Figure 2. An overview of the steps that compose the KDD process [2]

3. FAULT DETECTION APPROACHES

A hierarchical structure of an advanced system based on figure 1 is shown in figure 3. There are four levels: decision making, scheduling, control and fault diagnosis, continuous process and batch process. On the left, four objectives associated with each level are the planning, optimization, design, and process modeling, respectively. On the right, from top-down the three jobs in the four levels are task generation, task assignment, and measurement, respectively. At each level, data are collected when the system is in operation. Thus, the vast amounts of data are available for processing, analyzing, and information acquisition.

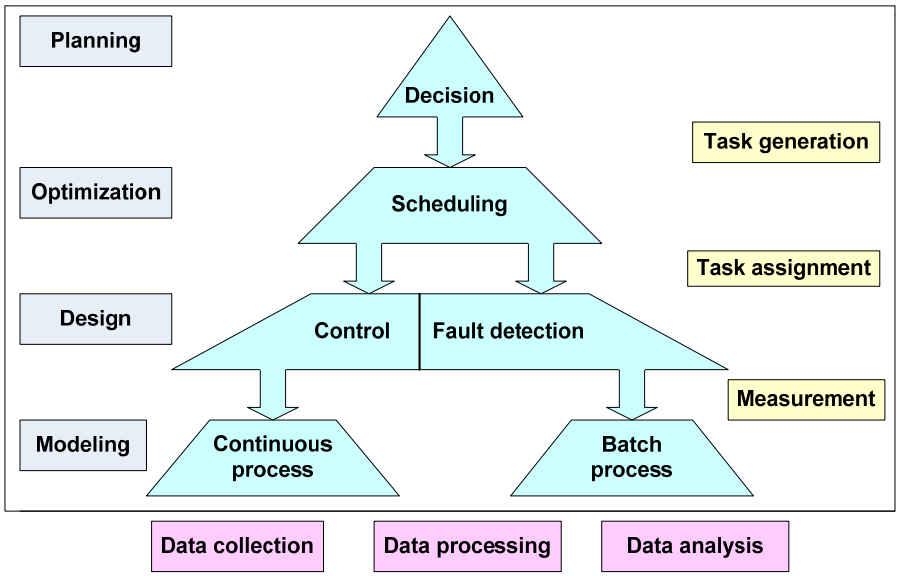

Figure 3*. Hierarchical structure of an advanced system *[10]

4. MAJOR ACHIEVEMENTS

Park, *et al*. [6] applied the BEAM (Beacon-based Exception Analysis for Multi-Missions) system to anomaly detection in Space Shuttle Main Engine data. BEAM has nine components that use nine different approaches to anomaly detection. The work described in [6] only used one of the nine components: the Dynamical Invariant Anomaly Detector (DIAD). DIAD is an unsupervised anomaly detection algorithm, which looks for anomalies in one variable at a time. Park, *et al*. trained DIAD using data from 16 nominal tests, and tested it using data from seven tests that contained known failures. It detected all of the major failures in these seven tests, although it missed some minor failures and had some false alarms.

Schwabacher [7] used two unsupervised anomaly detection algorithms, Orca and GritBot, to look for anomalies in data from two rocket propulsion systems, the Space Shuttle Main Engine and rocket engine test stand E-1 at NASA Stennis Space Center. These algorithms support both discrete and continuous variables, and look for anomalies in the relationships among the variables, in addition to looking for anomalies in the individual variables. The algorithms detected some anomalies that were already known to the experts, and some others that were not known to the experts but were not considered to be significant.

Iverson's Inductive Monitoring System [4] is another unsupervised learning system for fault detection. It uses a clustering algorithm to cluster the nominal training data into clusters representing different modes of the system. When new data fails to fit into any of the clusters, it signals an anomaly, using the distance from the nearest cluster as a measure of the strength of the anomaly. After the STS-107 Space Shuttle Columbia disaster,

Iverson applied IMS to some relevant data. He trained it using data from five previous Space Shuttle flights, and then tested it using STS-107 data. It detected an anomaly in data from temperature sensors on the Shuttle's left wing shortly after the foam impact, suggesting in retrospect that with the aid of IMS, flight controllers might have been able to detect the damage to the wing much sooner than they did.
Oza *et al.* [5] used neural nets and ensembles of neural nets for helicopter fault detection. Their method of detecting a fault is to assume that a fault has occurred when an actual maneuver fails to match a predicted maneuver. The data they used included vibration data from the gearbox, angular velocity and torque of the planetary gear, and altitude, velocity, and orientation of the helicopter, from a set of experimental flights in which the pilot always performed a predetermined maneuver. They obtained very high accuracy rates at predicting the maneuver, especially when using ensemble methods. It remains to be seen whether failure of their method to predict a maneuver will be highly correlated with faults, as they have hypothesized.
Srivastava [9] presents algorithms based on envelope detection and dynamic hidden Markov models for detecting anomalies in time series data with large numbers of discrete and continuous variables. He tests the algorithms using synthetic data motivated by a fleet of aircraft.

5. CONCLUSIONS

The data analysis characterized as the data mining methodology, is increasingly used to extract valuable information from the raw data by analyzing the correlation structure among process variables of the system. Due to progress of fault detection process with data mining Several authors have stated their intention to do fault prognostics with data mining methodology.

REFERENCES

[1] Donca G., Mihăilă I., Ganea M., Hirţe D., Nica M., *Maintenance role in life cycle management*, în Analele Universităţii din Oradea, Fascicola Management şi Inginerie Tehnologică, volumul XVI, Editura Universităţii din Oradea, ISSN 1583 – 0691, pp. 2158-2163, 2007
[2] Donca G., Mihăilă I., Ganea M., Hirţe D., Nica M., *Aspects of condition based maintenance*, în Analele Universităţii din Oradea, Fascicola Management şi Inginerie Tehnologică, volumul XV, Editura Universităţii din Oradea, ISSN 1583 – 0691, pp. 973-978, 2006
[3] Donca G., Mihăilă I. M., *Aspects of maintenance strategy selection process*, în Analele Universităţii din Oradea, Fascicola Management şi Inginerie Tehnologică, volumul XVIII, Editura Universităţii din Oradea, ISSN 1583 – 0691, pp. 1654-1659, 2009
[4] Iverson David L., *Inductive System Health Monitoring*. Proceedings of the International Conference on Artificial Intelligence, IC-AI '04, Volume 2 & Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications, MLMTA '04, Las Vegas, USA, June 21-24, 2004
[5] Oza N., Tumer K., Tumer I. & Huff E., *Classification of Aircraft Maneuvers for Fault Detection.* Lecture Notes in Computer Science, Volume 2709, pp. 375 – 384, 2003
[6] Park H., Mackey R., James M., Zak M., Kynard M., Sebghati J., and Greene W., *Analysis of Space Shuttle Main Engine Data Using Beacon-based Exception Analysis for Multi-Missions.* Proceedings of the IEEE Aerospace Conference, IEEE, New York, Vol. 6, March 9-16, 2002, pp 6-2835 - 6-2844
[7] Schwabacher M., *Machine Learning for Rocket Propulsion Health Monitoring*. Proceedings of the SAE World Aerospace Congress, Dallas, 2005
[8] Schwabacher M., *A Survey of Data-Driven Prognostics*, AIAA, Arlington, USA, 2005
[9] Srivastava A., *Discovering System Health Anomalies Using Data Mining Techniques.* Proceedings of the Joint Army Navy NASA Air Force Conference on Propulsion, Charleston, June 2005
[10] XU J., HOU Z., *Notes on Data-driven System Approaches*, Acta Automatica Sinica nr. 6, 2009