

A Data-Driven Diagnostic System Utilizing Manufacturing Data Mining and Analytics

Weihong Guo and Shenghan Guo
Rutgers University

Hui Wang
Ford Motor Company

Xiao Yu
Optimal CAE

Annette Januszczak and Saumuy Suriano
Ford Motor Company

ABSTRACT

The wide applications of automatic sensing devices and data acquisition systems in automotive manufacturing have resulted in a data-rich environment, which demands new data mining methodologies for effective data fusion and information integration to support decision making. This paper presents a new methodology for developing a diagnostic system using manufacturing system data for high-value assets in automotive manufacturing. The proposed method extends the basic attributes control charts with the following key elements: optimal feature subset selection considering multiple features and correlation structure, balancing the type I and type II errors in decision making, on-line process monitoring using adaptive modeling with control charts, and diagnostic performance assessment using shift and trend detection. The performance of the developed diagnostic system can be continuously improved as the knowledge of machine faults is automatically accumulated during production. An example of the analysis at one Ford production plant is provided to demonstrate the implementation of this methodology.

Keywords: Feature selection, attributes control chart, Bayesian inference, Markov chain Monte Carlo (MCMC) algorithm, process monitoring, manufacturing system

CITATION: Guo, W., Guo, S., Wang, H., Yu, X. et al., "A Data-Driven Diagnostic System Utilizing Manufacturing Data Mining and Analytics," *SAE Int. J. Mater. Manf.* 10(3):2017, doi:10.4271/2017-01-0233.

INTRODUCTION

Among manufacturing industries, the automotive industry is the technological trend setter - it keeps evolving on fast pace and is now at the forefront of data analytics. Different kinds of information sources and sensor signals are increasingly used for quality control, prediction, decision making, and insight in the automotive industry. This creates a tremendous capability for researchers and engineers to access valuable information from various enterprise entities to improve their designs and operations. As the quantity of data has dramatically increased, thanks to the wide applications of low-cost and smart sensing devices along with fast and advanced computer systems, the challenges to fully realizing the value of data have increased and researchers are eager to meet those challenges.

There are a large number of assets and resources in an automotive manufacturing plant, among which are the high-value assets that this paper will focus on. High-value assets are those that are either

expensive or mission-critical. Take the gantry system for example. A gantry is commonly used for the pick and place process for the automated transfer of a part from a feed conveyor, a buffer, or an upstream process, to a downstream process, for example, one of the multiple parallel Computer Numeric Control (CNC) systems for a machining operation. The gantry is mission critical in the sense that down time in the gantry will cause the CNCs to idle while waiting for parts, which will result in lost production time and serious impact on the entire production line. Normal production is interrupted from the start of the failure to the moment the gantry is returned to production. The total maintenance time of an asset includes notification time, while the asset waits for the engineer to come, and the time for the actual diagnosis, fix, and test. While the latter varies from a quick reset to a major repair, it is essentially inevitable. The time spent waiting, however, has been recognized as one of the more serious wastes in lean production systems [1]. If we could extract useful information from manufacturing data and gain more insights about

the failures, it's very likely that we would be able to get a better understanding of the manufacturing operation, to do prognostics, to allocate maintenance resources [2-3] in a more efficient manner, and, eventually, to reduce lost production time and improve productivity.

This motivates our research to develop a data-driven diagnostic system that facilitates effective monitoring of the manufacturing process, early detection of system anomalies, quick diagnosis of fault root causes, and intelligent maintenance planning and resource allocation. The huge amounts of recorded manufacturing data enable us to study and explore data-driven methods to efficiently and effectively monitor the process and provide prognostics and diagnostics in manufacturing analytics. This research assumes that features or quality characteristics have already been defined and extracted based on the physical and mathematical understanding of the manufacturing process. Hence, feature extraction will not be discussed in this paper. Instead, this study focuses on what can be learned from those features - finding out the underlying distribution of a feature, determining monitoring thresholds, balancing false positives and false negatives, handling time series data with multiple features and/or correlated features, updating the monitoring strategy continuously as new observations become available, and integrating multiple decision rules for process monitoring and change detection.

Data-driven diagnosis and prognosis have been widely accepted in industrial practice [4, 5, 6, 7] including semiconductor manufacturing [8-9], the steel industry [10, 11, 12], and the chemical industry [13-14]. Many studies in this area, however, focus on multivariate statistics for extracting variables from complex raw data. The concept and application of statistical control charts have been widely studied [15], while (i) balancing false positives and false negatives and (ii) integrating control charts with feature selection and adaptive modeling are usually not handled in the areas of traditional control charts design.

The objective of this paper is to develop a diagnostic system using manufacturing data for high-value assets in automotive manufacturing. The proposed method extends the basic attributes control chart with the following key elements: optimal feature subset selection considering multiple features and correlation structure, balancing the type I and type II errors in decision making, on-line process monitoring using adaptive modeling with control chart, and diagnostic performance assessment using process shift and trend detection. The performance of the developed diagnostic system can be continuously improved as the knowledge of machine faults is automatically accumulated during production.

The remainder of this paper is organized as follows. An overview of the proposed methodology is given in the next section, followed by detailed research procedures for the methodology development. After that, an example using data from one Ford production plant is provided to demonstrate the implementation of the proposed method. Finally, the last section concludes the paper and discusses future work.

METHODOLOGY OVERVIEW

The generic procedures of the proposed methodology are shown in Figure 1. The methodology starts with the basic control chart once a training dataset is collected. The width of control limits is determined by considering the tradeoff between the required type I error rate and expected type II error rate. To improve the effectiveness and efficiency of diagnostic system, multiple observations that are collected in time order of the production will be used in monitoring, and thus the next step is feature subset selection. As more and more knowledge about the process faults can be accumulated during the use of the process monitoring system, that knowledge is then used to further improve the diagnostic performance. For this purpose, adaptive modeling of the feature and adaptive control limits will be addressed.

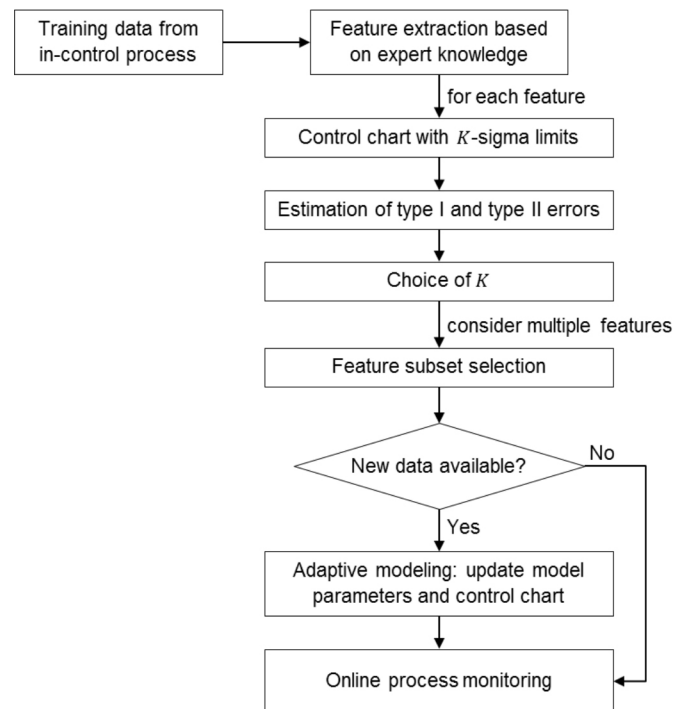


Figure 1. The framework of diagnostic system development with adaptive learning

In order to develop this new methodology, the following issues will be discussed in the paper: (1) Determining the underlying probability distribution of a feature using the goodness-of-fit test, based on which the basic control charts are designed; (2) Determining control limits, which balances the type I and type II errors; (3) Optimal feature subset selection by minimizing a weighted summation of the type I and type II errors associated with the features considered; two selection strategies are proposed; (4) Adaptive modeling of the feature distribution using Bayesian inferences; (5) Online process monitoring with adaptive control limits and sensitizing rules. The detailed procedures and algorithms will be discussed in the next section.

METHODOLOGY DEVELOPMENT

The Basic Attributes Control Chart for Fault Occurrences

Many quality characteristics in manufacturing can be conveniently represented as attributes, for example, the proportion of warped automobile engine connecting rods in a day's production, the number of nonconformities in a subassembly area, and the number of maintenance requests that require a second call to complete the repair. In this study, we focus on attribute features such as the occurrence of a machine fault or warning or repair within a certain period of time. Some examples of these attributes are the number of assembly errors in a day or in a rolling 24-hour period, the number of low oil level warnings in a week, and the number of repairs in a month. When the manufacturing process is stable and in statistical control, only chance causes of variation will be present and thus the quality characteristics are more likely to fall within a certain range [15]. If the process is operating in the presence of assignable causes, the process will be unstable, out of control, and requires actions; the quality characteristics are more likely to fall outside the control limits.

Take the occurrence of a machine fault in a high-value asset in a rolling 24-hour period as an example. In most occasions this fault can be fixed by a quick reset which takes no longer than a few seconds. However, the down times accumulate and thus too many resets would lead to lost production time. In the attributes control chart, we would like to detect high occurrences of this fault, while assuming the process is in control if the occurrences appear to be random and are inside the control limits. Let $X_1, X_2, \dots, X_n$ denote the occurrences observed in each 24-hour window when the process is under normal operating conditions. When the process is in control, it is usually assumed that the occurrence of nonconformities in samples of constant size is well modeled by the Poisson distribution. Without loss of generality, we model the data with a Poisson distribution and estimate the Poisson parameter λ using the Maximum Likelihood Estimation (MLE) [16]:

$$\hat{\lambda} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (1)$$

where $\bar{X}$ is the sample mean. We then apply the chi-square goodness-of-fit test [16] to test if the observed data $X_1, X_2, \dots, X_n$ indeed come from the Poisson distribution with parameter $\hat{\lambda}$, denoted as $\text{Poi}(\hat{\lambda})$.

H_0 : The sample data $X_1, X_2, \dots, X_n$ follow $\text{Poi}(\hat{\lambda})$ distribution.

H_a : The sample data $X_1, X_2, \dots, X_n$ do not follow $\text{Poi}(\hat{\lambda})$ distribution.

The p -value from the chi-square goodness-of-fit test is then examined to decide whether to reject the null hypothesis. If the p -value is less than or equal to a chosen significance level, for example, 0.05, the test suggests that the observed data is inconsistent with the null hypothesis and thus H_0 should be rejected. The control chart for fault occurrences will have control limits determined by the percentiles of the correct

underlying distribution. These percentiles could be obtained from a histogram if a large sample (at least 100 observations) is available, or from a probability distribution fit to the data.

If $p > 0.05$, the null hypothesis will not be rejected. It is then statistically reasonable to assume that $X_1, X_2, \dots, X_n$ follow the $\text{Poi}(\hat{\lambda})$ distribution. A control chart for fault occurrences can be constructed in a similar procedure as the control chart for nonconformities, also known as the c chart. The attributes control chart with K -sigma limits is defined as follows:

$$\begin{aligned} \text{UCL} &= \hat{\lambda} + K\sqrt{\hat{\lambda}} \\ \text{Center line} &= \hat{\lambda} \\ \text{LCL} &= \hat{\lambda} - K\sqrt{\hat{\lambda}} \end{aligned} \quad (2)$$

where UCL is the upper control limit, LCL is the lower control limit, $K\sqrt{\hat{\lambda}}$ is the distance from the center line to control limits. Should the LCL calculation in (2) yields a negative value, set $\text{LCL} = 0$.

Choice of Control Limits: Balancing Type I and Type II Errors

One of the critical decisions in designing the control chart for fault occurrences is specifying the control limits. As expressed in Eq. (2), the value K determines how far the control limits are from the center line. The choice of control limits will then determine the decision errors in the control chart monitoring system. The type I error occurs when the control chart shows a point falling beyond the control limits or points forming non-random patterns, when no assignable cause of variation is present; the type II error occurs when the control chart fails to detect an out-of-control process.

In automobile manufacturing, type I errors result in unnecessary maintenance efforts, manual inspections, and loss of production time. If a monitoring system has too many false positives, the plant engineers will tend to ignore the alarms, distrust the monitoring decisions, and eventually abandon the monitoring system. The risk of a type I error can be decreased if we select a larger K and move the control limits farther from the center line. However, widening the control limits will increase the risk of a type II error. Type II errors would result in misdetections of process shifts and delays in actions, potentially storing up trouble for the future. If we use a smaller K and move the control limits closer to the center line, the risk of type II error can be decreased, while the risk of type I error will be increased. Therefore, finding the optimal K that balances the type I and type II errors becomes a critical issue.

Assume that the fault occurrence x follows a Poisson distribution with parameter λ_0 when the process is in control and a Poisson distribution with parameter λ_1 when the process is out of control. The type I error from the control chart defined in Eq. (2) is expressed as:

$$\alpha = P(x \geq \text{UCL} | x \sim \text{Poi}(\lambda_0)) + P(x \leq \text{LCL} | x \sim \text{Poi}(\lambda_0)), \quad (3)$$

where λ_0 can be estimated by Eq. (1). The type II error is expressed as:

$$\beta = P(\text{LCL} < x < \text{UCL} | x \sim \text{Poi}(\lambda_1)), \quad (4)$$

where $\lambda_1 = \lambda_0 + \delta$ and δ is the magnitude of the process mean shift.

In order to find the optimal K that balances the type I and type II errors, we use Monte Carlo simulation with N iterations to estimate the type I and type II errors. Specifically, in the j^{th} iteration, $j = 1, 2, \dots, N$, we generate a sample of n independent and identically $\text{Poi}(\lambda_0)$ distributed random values representing the fault occurrences in a 24-hour window when the process is in control, that is, $X_{ij}^0 \sim \text{Poi}(\lambda_0)$, $i = 1, 2, \dots, n$; we also generate n independent and identically $\text{Poi}(\lambda_1)$ distributed random values representing the fault occurrences in a 24-hour window when the process is out of control, that is, $X_{ij}^1 \sim \text{Poi}(\lambda_1)$, $i = 1, 2, \dots, n$. The estimated type I error from the j^{th} iteration is

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n [1 - I(\text{LCL} < X_{ij}^0 < \text{UCL} | X_{ij}^0 \sim \text{Poi}(\lambda_0))], \quad (5)$$

where $I(\cdot)$ represents an indicator function. The estimated type II error from the j^{th} iteration is

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^n I(\text{LCL} < X_{ij}^1 < \text{UCL} | X_{ij}^1 \sim \text{Poi}(\lambda_1)). \quad (6)$$

With N iterations in Monte Carlo simulation, we obtain an unbiased estimate for α as

$$\hat{\alpha} = \frac{1}{N} \sum_{j=1}^N \hat{\alpha}_j \quad (7)$$

and an unbiased estimate for β as

$$\hat{\beta} = \frac{1}{N} \sum_{j=1}^N \hat{\beta}_j. \quad (8)$$

We illustrate how $\hat{\alpha}$ varies with K and how $\hat{\beta}$ varies with K and δ in the operating-characteristic (OC) curves for the attributes control chart for fault occurrences. Based on the OC curves, we select an interval for K where $\hat{\alpha}$ is smaller than or equal to a chosen threshold α_0 . The endpoints of the interval are

$$\begin{aligned} K_L &= \min\{K | \hat{\alpha} \leq \alpha_0\}, \\ K_U &= \max\{K | \hat{\alpha} \leq \alpha_0\}. \end{aligned} \quad (9)$$

The optimal K will then be selected as the value that minimizes $\hat{\beta}$:

$$K^* = \arg \min_{K \in [K_L, K_U]} \hat{\beta}. \quad (10)$$

Feature Subset Selection

One of the standard assumptions in conventional control charts is that the data generated by the process when it is in control are independently distributed. Conventional control charts may not work well if the quality characteristic exhibits even low levels of correlation over time. Unfortunately, this independence assumption is not even approximately satisfied in many manufacturing processes. It is also well acknowledged that all manufacturing processes are driven by inertial elements, and when the interval between samples becomes small relative to these forces, the observations on the process will be correlated over time.

While the underlying correlation structure may be too complicated to be modeled, we explore the possibility of monitoring multiple observations that are taken in time order of the production to improve the effectiveness and efficiency of the diagnostic system. We would like to use the features observed prior to time t to predict if the process will be in control at time t . An example of the feature set is the fault occurrences observed in the l 24-hour windows prior to the decision time t . In order to remove redundant or irrelevant features, reduce overfitting, and simplify the monitoring system, we need to select an optimal feature subset that gives the best predictive power.

Let t denote the time when a prediction on process status needs to be made. Consider the features from the W windows prior to t . We propose two methods in selecting the optimal feature subset from the W features.

Method 1

Selecting w consecutive windows out of the W windows

In this method, we want to select the most recent w consecutive windows out of W such that the feature subset will minimize the cost function in Eq. (11), in which w is to be determined. This cost function is a weighted summation of the type I and type II errors associated with the features considered:

$$\min_{w \in \mathcal{W}} c_1 \alpha_w + c_2 \beta_w, \quad (11)$$

where $\mathcal{W} = \{1, 2, \dots, W\}$ is the set of all possible values for w , $c = (c_1, c_2)$ is the cost vector, $c_1 + c_2 = 1$. The cost vector is used to balance the type I and type II errors since they may not be equally important in real applications. We recommend a large value of c_1 if false alarms are more expensive and a large c_2 if misdetections are more expensive. α_w and β_w denote the type I and type II errors calculated for the most recent w consecutive windows, respectively. They can be computed using conditional probabilities as:

$$\hat{\alpha}_w = P(\cup_{l=1}^w f_l \geq \text{UCL} | \text{process in control})$$

$$+ P(\cup_{l=1}^w f_l \leq \text{LCL} | \text{process in control}), \text{ and}$$