Machine Learning Techniques for Quality Monitoring and Prediction

Lily He

Background

Product quality is a key factor for manufacturing companies to assess their production capability. According to (ISO 9000:2015), quality can be defined as "the degree to which a set of inherent characteristics fulfills requirements". Quality control seeks to monitor and predict the quality of products during the manufacturing processes. Traditional statistical process control methods have been widely used in the industrial environment due to their simplicity and applicability, this success is attributed to stable and low complexity processes, which thus limits the current application fields. Today's manufacturing processes have become more complicated with high dimension variables, uncertain and dynamic environments, and multistage of manufacturing processes. Consequently, traditional statistical process control methods are not sufficient in dealing with current problems. Therefore, a more efficient method is needed to tackle quality monitoring problems. Machine learning (ML), as a computational engine for data mining and pattern recognition, is capable of dealing with complex, high-dimensional multistage manufacturing processes. ML techniques can be used to transform the massive amount of data that had been collected by build-in or add-on smart sensors into valuable information that can explain the uncertainties and assist in making more informed decisions.

Motivation

A complex modern semi-conductor manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be applied to identify the most relevant signals. The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs.

The analysis will primarily focus on:

- What is the potential problem of current models?

Currently, most quality monitoring models focus on single-stage manufacturing or address the manufacturing chain as a single point. However, for the manufacturing chain, many factors (e.g., equipment,

manufacturing variables, operators) may have interactive and cumulative effects on the final product quality. Also, most of the models need to wait until the end of the process to make the prediction, which might be problems in the product quality analysis and prediction.

- **Is there a way to to reduce wastes of time and resources?**

The purpose of this case study is to introduce an intelligent real-time quality monitoring framework, which is a strategy that most of the models wait until the end of the process to make the prediction, thus it is capable of predicting and identifying the quality deviations for multistage manufacturing systems as early as possible to reduce wastes of time and resources.

- **How to to improve the performance of the quality monitoring process?**

To enhance current business improvement techniques, the application of feature selection as an intelligent systems technique is going to be investigated.

Description of the Data

The SECOM (Semiconductor Manufacturing) datasets, consists of 2 files the dataset file SECOM consisting of a 1567(examples) x 590(features) matrix and a labels file containing the classifications and date time stamp for each example. The datasets consist of manufacturing operation data and the semiconductor quality data. 1567 observations were taken from a wafer fabrication production line. Each observation is a vector of 590 sensor measurements which represents a single production entity with associated measured features, plus a label of pass/fail yield for in house line testing. Also, there are only 104 fail cases which are labeled as positive (1 corresponds to a fail), whereas much larger amount of examples pass the test and are labeled as negative (encoded as -1). This is a 1:14 proportion. In this work not only a feature selection method for extracting the post discriminative sensors is proposed, but also boosting and data generation techniques are devised to deal with highly imbalance between the pass and fail cases. The null values are represented by the 'NaN' value as per MatLab. This needs to be taken into consideration when investigating the data either through pre-processing or within the technique applied.

Proposed Analysis

Our analysis of the SECOM data will comprise of different unsupervised and supervised machine learning techniques such as principal component analysis, support vector machine, neural network and random forest to consider the accumulative effect of different workstations and to construct the quality monitoring model.

Unsupervised Learning Methods

- PCA
- K-Means

Unsupervised learning is used when the data set consists of unlabeled training examples, where X is a feature vector without any corresponding target value Y. It is used for information extraction, dimensionality reduction, density estimation, data visualization, outlier detection, and process monitoring. Principal component analysis (PCA) and K-Means algorithms are used in this work to represent the accumulative effect for each workstation in the manufacturing chain and identify operational patterns for each workstation.

Supervised Learning Methods

- LR
- SVM
- K-NN
- RF

Supervised learning is used when the data set consists of labeled training. Where X is a feature vector with the corresponding target value y, the target value y can be discrete or continuous. If categorical output variables are set, the problem is formulated as a classification problem. Otherwise, the problem is formulated as a regression problem. Supervised learning can be used for fault diagnosis, process monitoring, quality prediction, and remaining useful life estimation. Supervised learning methods in the manufacturing processes
Include logistic regression, support vector machine, K-nearest neighbors and random forest. They used association rule mining for root cause analysis of defective products.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Mandel J, S. P. (2015). A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, *06*(01). https://doi.org/10.4172/2155-6180.1000224

Cortes, C.; Vapnik, V.N. Support-Vector Networks. Mach. Learn. 1995, 20, 273–297. [CrossRef]

Platt, J.C. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Advances in Kernel Methods:

Support Vector Learning; Schölkopf, B., Burges, C.J.C., Smola, A.J., Eds.; MIT Press: Cambridge, MA, USA, 1998.

Hamel, L.H. Knowledge Discovery with Support Vector Machines; Wiley: Hoboken, NJ, USA, 2009.

Platt, J.C. Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods. In Advances in

Large Margin Classifiifiers; Smola, A.J., Barlett, P., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 2000.
Dietterich, T.G. Ensemble Methods in Machine Learning. In Proceedings of the First International Workshop on Multiple Classififier
Systems; Springer: Berlin, Germany, 2000.

Appendix

The data is represented in a raw text file each line representing an individual example and the features separated by spaces.
The data link: https://archive.ics.uci.edu/ml/datasets/SECOM