# On Applying Big Data to Transform the Inspection Lines

**JrJung Lyu, Chia-Wen Chen, and Hong Yu Chen**
Department of Industrial and Information Management
National Cheng Kung University
Tainan, Taiwan (R.O.C.)
jlyu@mail.ncku.edu.tw

**Abstract**

Inspection lines are a heavy duty yet expensive part in many industries such as wireless communication industry, auto industry. While inspection is always a necessary evil in a very complicated and competitive industry, there is a strong need to reduce the unnecessary inspection processes and resources to improve the competitiveness. This work develops a framework to apply big data analytics to improve the inspection processes based on case study method. A world class case company is selected and the data-warehouse is established for a specific lot of products. After the data pre-processing work, the relationships among the detection stations and the detection results could be found. Various meta-models are used, which including association rule method to find the correlation between different test items and to reduce the number of required tests. Decision tree is also applied to find the optimal controllable factor. Finally, the developed rules are further verified by empirical data. Based on the empirical results, the number of detection items could be reduced by around 16.2% - a huge cost saving. It is also found out that the critical control factors, recommended by the decision tree, are temperature and humidity, which is a way to improve its quality without extra cost. The proposed big data application framework is therefore feasible in this case and machine learning or other models could be further extended, which is the future research direction.

**Keywords:** Big data, Association rules, Decision tree, Wireless communications industry

## 1. Introduction

Inspection is a very complicated process for the information and communication industry (ICT), auto industry, airline industry, and many consumers-oriented industries. To main the quality, there are many specifications and standards are proposed and the companies have to spend many resources to verify the quality of the products. Many managers therefore argue that the manufacturing cost of each equipment could be dramatically reduced should some of the inspection process and/or related resources could be simplified. In practices, the settings of process parameters and the complexity of process combinations are difficult to predict and could simply by the experience of process engineers (Kudyba, 2014). In the modern era, Sarkar (2017) pointed out the potential of using big data, while Acharjya and Ahmed (2016) stated that there is currently no standard architecture that can solve all big data problems. This work proposes a big data analysis framework, considering the complex characteristics of manufacturing processes, and select a representative company in the WLAN industry to illustrate the feasibility of this framework.

## 2. Literatures Review

Acharjya and Ahmed (2016) explored the potential impact of big data challenges, open research issues, and various tools associated with it and provides a platform to explore big data at numerous stages. Caigny *et al.* (2018) proposed a new hybrid approach that is benchmarked against decision trees, logistic regression, random forests and logistic model trees with regards to the predictive performance and comprehensibility. Chien and Hsu (2014) proposed a novel approach to improve overall wafer effectiveness via data mining to generate the optimal integrated circuit (IC) feature designs that can bridge the gap between IC design and wafer fabrication by providing chip designer with the optimal IC feature size in the design phase to increase gross dies and reduce the required shots. Results have shown that the proposed approach can effectively enhance wafer productivity. Chu et al. (2016) developed a framework for a large amount of thin film transistor-liquid crystal display (TFT-LCD) manufacturing data and determine the possible causes of faults and manufacturing process variations. Results demonstrated that the practical viability of the framework. Sarkar (2017) provided a broad overview on big data and the effectiveness of healthcare big data for non-expert readers and builds a distributed framework of organized healthcare model for the purpose of protecting patient data. Ragab et al. (2018) applied the Logical Analysis of Data (LAD) to detect and diagnose faults in industrial chemical processes. Kashkoush and ElMaraghy (2017) proposed an integer programming model to discover association rules between product features and manufacturing system capabilities.

## 3. Methodology

Considering a world-class tel-communication company which is an outsourcer of various huge companies. This case company deliver high quality and high volume products for its customers which have to go through a series of inspection processes to ensure the consumers receive good quality products. The short-term objective of this company is to establish an analytical framework for reducing process inspection based on the concept of big data - through data collection, processing and analysis - to simplify its inspection processes.

Extended from the Standard Exploration Process (CRISP-DM) (Chapman et al., 2000), a research framework (see Figure 1) is proposed. This framework integrates with various methods - including association rules (Herzig and Nagappan., 2015), decision tree (Chu et al., 2017), Neural network (Chien and Hsu., 2014), and logistic regression (Caigny et al., 2018).

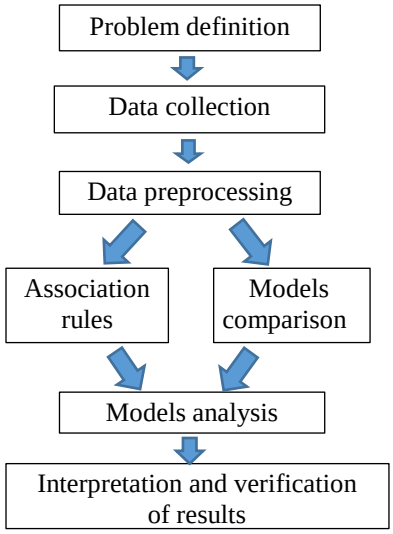


Figure 1. Research Framework

## 4. Results and Discussion

To elaborate the framework proposed in the previous section into case company, a six step detailed process is further illustrated. Note that some of the figures are not the same as in the shop floor due to the concern of confidential issue.

(1). Problem definition: Find out the relationships among test items and the key factors that influence the test results.

(2). Data collection: Collect six-month process related data, including historical data such as machine parameters record, appearance model, work order number, ambient temperature and humidity, staff number, article number, and inspection items.

(3). Data pre-processing: Data is integrated through the definition of the product process by the domain experts and the data-warehouse is established. The fields of the unrecorded test results are deleted. Variables are deleted based on the experience of the domain engineers as shown in Table 1.

Table1. Data preprocessing

| **Data type** | **Category variable** | **Continuous variable** |
|---|---|---|
| Data unprocessed item | 11246 | 12561 |
| Data preprocessing item | 6223 | 6528 |
| Data unprocessed variable | 125 | 154 |
| Data preprocessing variable | 60 | 67 |

(4). Model establishment and comparison: In the first part, the association rule model is established, and the Apriori algorithm is used to evaluate the indicators for screening. Support indicates that the association rule must have certain significance with respect to all information. Confidence indicates whether the association rule has credibility indicators, and lift indicates the reliability of the rule and whether there is a positive relationship. The

second part conducts the comparison between the decision tree, the logistics regression and the neural network. As shown in Table 2, the minimum error indicates the best fit. The decision tree expresses the classification process in the state of the branch, has the best interpretation ability, and can clearly identify the factors that affect the detection results. Therefore, the second part adopts the decision tree for analysis and prediction.

Table 2. Model fitting comparison

| **Model** | **Misclassification Rate** | **Average Squared error** |
|---|---|---|
| Decision tree | 0.000911 | 0.000583 |
| Neural network | 0.0428 | 0.041287 |
| Logistic regression | 0.0428 | 0.041288 |

(5). Model analysis: The first part analyzes the dynamic association rules and are sorted based on the degree of confidence (see Table 3). For example, Rule 1 represents if the y10 detection pass, and y118 detection passes, then the y122 detection pass, one can find support is 0.97, confidence is 1 and lift is 1.0009.

Table 3. Sample association rule results

| Rules | Support | Confidence | Lift |
|---|---|---|---|
| 1.{y10=OK,y118=OK} => {y122=OK} | 0.97 | 1 | 1.0009 |
| 2.{y3=OK,y69=OK,y75=OK} => {y77=OK} | 0.97 | 1 | 1.0009 |
| 3.{y34=OK,y76=OK,y111=OK} => {y112=OK} | 0.97 | 1 | 1.0009 |
| 4.{y2=OK} => {y5=OK} | 0.995 | 1 | 1.0005 |
| 5.{y37=OK,y76=OK} => {y81=OK} | 0.947 | 1 | 1.0004 |
| 6. {y10=OK,y28=OK,y29=OK,y69=OK} =>{y71=OK} | 0.98 | 1 | 1.0004 |

CRAT is used as the decision tree algorithm in the second part. Gini dispersion is used as a basis for branching indicators (see Table 4). Take FT2 detection station as an example, six rules are established. It is clear that the machine type, maintenance, ambient temperature and humidity are the key factors affecting the detection results.

Table 4. Sample decision tree rules

| Rules | Results |
|---|---|
| 1.FT2_J2 ≠ E2，FT2_J1 ≠ E2、E7 | Y=PASS |
| 2.FT2_J2 ≠ E2，FT2_J1= E2、E7，FT1_REPAIRED ≠ Y，FT2_HUMIDITY< 0.52 | Y=PASS |
| 3.FT2_J2 ≠ E2，FT2_J1=E2、E7，FT1_REPAIRED ≠ Y，FT2_HUMIDITY≥ 0.52 | Y=FAIL |
| 4.FT2_J2 ≠ E2，FT2_J1=E2、E7，FT1_REPAIRED= Y， | Y=PASS |

| | |
|---|---|
| 5.FT2_J2=E2，SMTS_TEMPERATURE≥ 23 | Y=PASS |
| 6.FT2_J2=E2，SMTS_TEMPERATURE<23 | Y=FAIL |

(6). Interpretation and verification of results
The association rule results of the first part are provided to the domain experts to justify the validity of the rules. A total of 10 rules were extracted and verified by another empirical data. Comparison is shown in Table 5. The reliability is used as the verification standard. The error of rule 4 is 0, which means the reliability is consistent. It can be verified that proposed new rules can reduce the number of detections by around 16.2%.

After the above analysis is completed, the decision tree analysis, the decision tree model is put into the test data for verification, and to identify further operation parameters details. The output correctness rate of the binary confusion matrix is 0.95 (see Table 6) which is quite good.

Table 5. Comparison of original data and verification data

| Rules | Original | Test | Error |
|---|---|---|---|
| 1.{y10=OK,y118=OK} => {y122=OK} | 1 | 0.9936 | 0.0064 |
| 2.{y3=OK,y69=OK,y75=OK} => {y77=OK} | 1 | 0.9997 | 0.0003 |
| 3.{y34=OK,y76=OK,y111=OK} => {y112=OK} | 1 | 0.9934 | 0.0064 |
| 4.{y2=OK} => {y5=OK} | 1 | 1 | 0 |
| 5.{y37=OK,y76=OK} => {y81=OK} | 1 | 0.9995 | 0.0005 |
| 6. {y10=OK,y28=OK,y29=OK,y69=OK} => {y71=OK} | 1 | 0.9999 | 0.0001 |
| 7.{y8=OK,y12=OK} => {y14=OK} | 0.9999 | 1 | 0.0001 |
| 8.{y14=OK} => {y16=OK} | 0.9999 | 1 | 0.0001 |
| 9.{y2=OK,y10=OK,y18=OK} => {y19=OK} | 0.9999 | 1 | 0.0001 |
| 10.{y37=OK} => {y38=OK} | 0.9999 | 0.9995 | 0.0004 |

Table 6. Confusion matrix

| Actual/Predictive | PASS | FAIL |
|---|---|---|
| PASS | 2104 | 5 |
| FAIL | 84 | 1 |

## 5. Conclusion
In this work, a big data analysis framework is proposed to solve the complicated empirical problem in a case company. Various methods, including association rules and decision tree method are used to develop an improved production model. Through a six step process, association rules are developed and can roughly reduce the detection time by 16.2%. The model is further verified by another empirical data to illustrate its feasibility. This study also finds out that the critical control factors affecting the detection results are temperature and humidity in the

case company. The proposed big data application framework can integrate with machine learning or other models in the future study.

## References

Acharjya, D., & Ahmed, K. P., A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications*, 7(2), page. 511-518, 2016.

Caigny, A. D., Coussement, K., & Bock, K. W. D., A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. *European Journal of Operational Research*, 269(2), page. 760-772, 2018.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R., CRISP-DM 1.0: Step-by-step data mining guide. New York, USA: SPSS, 2000.

Chien, C.-F., & Hsu, C.-Y., Data Mining for Optimizing IC Feature Designs to Enhance Overall Wafer Effectiveness. *IEEE Transactions on Semiconductor Manufacturing*, 27(1), page. 71-82, 2014.

Chu, P.-C., Chen, C.-C., & Chien, C.-F., Analyzing TFT-LCD Array Big Data for Yield Enhancement and an Empirical Study of TFT -LCD Manufacturing in Taiwan. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 23(5), page. 318-331, 2016.

Herzig, K., & Nagappan, N., Microsoft- Empirically Detecting False Test Alarms Using Association Rules. Paper presented at the ICSE '15 Proceedings of the 37th International Conference on Software Engineering, Florence, Italy, 2015.

Kashkoush, M. and ElMaraghy, H., An Integer Programming Model for Discovering Associations between Manufacturing System Capabilities and Product Features, *Journal of Intelligent Manufacturing*, 28, pages. 1031-1044, 2017.

Ragab, A. El-Koujok, M., Poulin, B., Amazouz, M., and Yacout, S., Fault Diagnosis in Industrial Chemical Processes Using Interpretable Patterns based on Logical Analysis of Data, *Expert Systems With Applications*, 95, pages. 368-383, 2018.

Sarkar, B. K., Big Data for Secure Healthcare System: A Conceptual Design. *Complex & Intelligent Systems*, 3(2), page. 133-151, 2017.

## Biographies

**JrJung Lyu** is a professor in the Department of Industrial and Information Management at National Cheng Kung University since 1989. He obtained a PhD degree in industrial engineering from the University of Iowa, USA. Dr. Lyu has participated in many projects, public services, and reviewing committees since 1989. He is the founder of CQI (Center for Quality & Innovation) at National Cheng Kung University and serving as the president of e-Business Management Society (EMBS) and many public services. Dr. Lyu has published over a hundred journal papers, several textbooks (including e-Business strategy, global quality management, healthcare quality management), and earned the Personal Award of the National Quality Award, Taiwan, in 2002. He is a fellow of CSQ and currently appointed as the LMC chair in Taiwan for ASQ, USA. His current research interests include strategy for innovative services, big data application in healthcare quality, biz model of EV, genetic diagnostics applications.

**Chia-Wen Chen** is a research associate in CQI (Center for Quality & Innovation) at National Cheng Kung University since she earned her Ph.D. degree. She has participated in many projects and has published several papers in the international journals. She is also serving as the deputy chief secretary of EBMS (e-Business Management Society) now.

**Hong Yu Chen** is a graduate student in the Department of Industrial and Information Management at National Cheng Kung University. He was serving as a RA in a big data application project stated above.