# Quality prediction modeling for multistage manufacturing based on classification and association rule mining

*Hung-An* Kao[1,2,*], *Yan-Shou* Hsieh[1], *Cheng-Hui* Chen[1], and *Jay* Lee[2]

[1] Central Industry Research & Service Division (CID), Institute for Information Industry, Nantou, 540, Taiwan
[2] NSF I/UCRC for Intelligent Maintenance Systems (IMS), University of Cincinnati, Cincinnati, OH 45221, USA

**Abstract.** For manufacturing enterprises, product quality is a key factor to assess production capability and increase their core competence. To reduce external failure cost, many research and methodology have been introduced in order to improve process yield rate, such as TQC/TQM, Shewhart Cycle、Deming's 14 Points, etc. Nowadays, impressive progress has been made in process monitoring and industrial data analysis because of the Industry 4.0 trend. Industries start to utilize quality control (QC) methodology to lower inspection overhead and internal failure cost. Currently, the focus of QC is mostly in the inspection of single workstation and final product, however, for multistage manufacturing, many factors (like equipment, operators, parameters, etc.) can have cumulative and interactive effects to the final quality. When failure occurs, it is difficult to resume the original settings for cause analysis. To address these problems, this research proposes a combination of principal components analysis (PCA) with classification and association rule mining algorithms to extract features representing relationship of multiple workstations, predict final product quality, and analyze the root-cause of product defect. The method is demonstrated on a semiconductor data set.

## 1 Introduction

For manufacturing enterprises, product quality is a key factor to assess production capability and increase their core competence. To reduce external failure cost, many research and methodology have been introduced in order to improve process yield rate, such as TQC/TQM, Shewhart Cycle、Deming's 14 Points, etc. Nowadays, impressive progress has been made in process monitoring and industrial data analysis because of the Industry 4.0 trend. Industries start to utilize quality control (QC) methodology to lower inspection overhead and internal failure cost. Currently, the focus of QC is mostly in the inspection of single workstation and final product, however, for multistage manufacturing system (MMS), many factors (like equipment, operators, parameters, etc.) can have cumulative and interactive effects to the final quality. When failure occurs, it is difficult to resume the original settings for cause analysis.

Currently, most state-of-art manufacturing lines are actually quite 'smart' in themselves, which means more and more build-in sophisticated sensors and computerized components while they are designed or increasingly various add-on sensors to monitor tools in real-time. The data delivered via both build-in and add-on sensors are highly related with process's status and performance. However, it is difficult for field engineers and management staff to get the information of product quality just through checking the big amount of mixed data, not mention being able to track the degradation trend, which will eventually lead to a catastrophic failure. Therefore, it is necessary to develop a data-to-information conversion tools that are able to convert process data into final quality and performance-related information. The output of these analytics is the real-time quality index after each workstation, which shows the current performance, for decision makers to effectively aware of the situation and make compensation decisions before potential failures occur. It can largely prevent waste in terms of time, spare parts and personnel.

To address these problems, statistical methods and machine learning technique are two main research target. Multivariate statistics was used to build quality prediction model for MMS, however, the correlation cannot be easily revealed for further verification and improvement [1, 2]. Some research treated variables from different stages as one single point and applied machine learning technique such as Naïve-Bayes, PCA, SVM, ANN, KNN, etc [3-5]. Naïve Bayes is a classifier it based on statistical knowledge, Naïve Bayes assumes independency between attributes and it attempts to maximize the posterior probability in determining the class. The advantage of using naive bayes classifiers is it have works well in many complex real-world situations. The k-nearest neighbor (k-NN) algorithm is a classifier based on closest data in feature space. It is amongst the simplest of all machine learning algorithms. But the

* Corresponding author: hakao@iii.org.tw

accuracy of classification will be effected by noisy or irrelevant features. It is one of the methods that are receiving increasing attention with remarkable results. SVM algorithm constructing an optimal separating hyper plane in the hidden feature space using quadratic programming to find a unique solution. The Artificial neural networks (ANNs) is based on human brain by making the right connections. It is a flexible mathematical structure. The structure is capable of identifying complex nonlinear relationships between input and output data sets. Like other machine methods, ANN have been used to solve a wide variety of tasks [5]. On the other side, some research consider the characteristics of MMS and conduct learning after extract related features [6-8].

In this research, we propose a combination of principal components analysis (PCA) with classification and association rule mining algorithms to extract features representing relationship of multiple workstations, predict final product quality, and analyze the root-cause of product defect. The method is demonstrated on a semiconductor data set.

## 2 Methodology

In this research, classification and rule mining algorithms are implemented and compared for quality prediction. To consider the characteristics of MMS, single point approach and multi-point approach are both discussed. Firstly, the original manufacturing dataset will be pre-processed to remove noise and segmented into training and testing dataset. Secondly, we build the quality prediction model through classification and rule mining process. For classifier learning, the relationship among workstations and the impact to the final product are included in feature selection so the feature set can better represent MMS property. For rule mining, a priori algorithm is applied on the training dataset after minimum support and confidence threshold are defined. The output classifiers and association rules are evaluated by testing dataset. Finally, the model can be used online for quality prediction and defect cause analysis in multiple workstation scenarios. Figure 1 shows the flowchart of the proposed methodology.

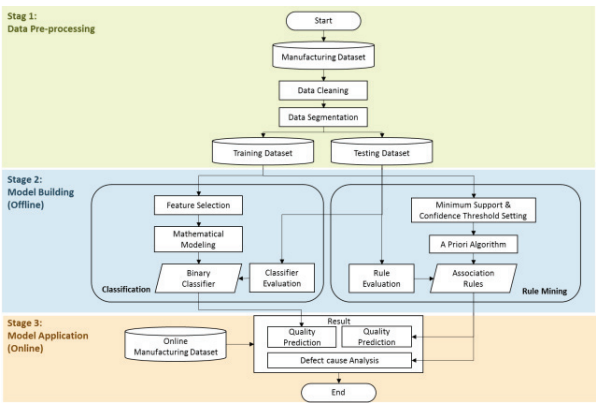


**Fig. 1.** The flowchart of proposed method.

### 2.1 Data pre-processing

Since manufacturing environment may have lot of anomaly events and situations, data pre-processing including cleaning and normalization is important. To deal with "Missing Value" issue, in this research we replace missing values with the mean value that calculated from all instances. Besides, if values from all instances of a certain variable are the same, the variable will be removed since it does not carry any discriminative information.

Another issue for quality prediction research is the insufficient instances for failed products. In recent manufacturing lines, typical shop floor is quite stable that the yield rate is usually higher than 85%. Hence, to balance the size of positive and negative datasets, over-sampling [9] and boosting technique [4] are introduced. In this research, we adopt the boosting technique by simply duplicate the fail dataset to the same amount of pass dataset.

### 2.2 Model building based on classification

To consider the characteristic of MMS, Cascade Quality Prediction Method (CQPM) [7, 10] is implemented to mine the hidden relationship. Figure 2 shows the scenario of MMS and idea of CQPM. In CQPM, three types of relationship are existed in MMS:

- R1: Relationship among manufacturing operation in a workstation.
- R2: Relationship among workstations.
- R3: Relationship between manufacturing operation variables and final product quality.

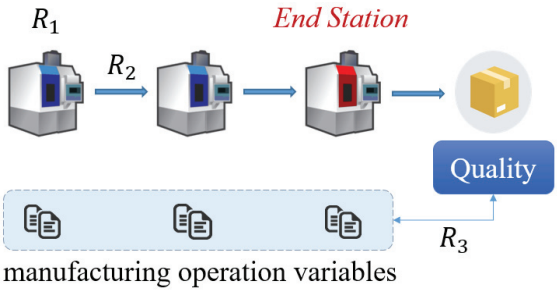


**Fig. 2.** Scenario of MMS and relationships identified with CQPM.

Based on CQPM definition, it is necessary to find the relationships of inter-correlated variables. PCA can then be used to transform latent variables into a new set of dimensions while they carry MMS characteristics [7]. After the feature set is selected, several classification algorithm, including decision tree, support vector machine (SVM), and Naive Bayes, will then be applied in order to train the binary classifier. For decision tree learning, iterative dichotomiser 3 (ID3), recursive partitioning and regression trees (RPART), and C50 are potential algorithms for decision tree learning.

### 2.3 Model building based on rule mining

Dependency modeling is to mine the association rules between the target variable and the independent variables. This area of research is very active in financial data mining and is usually referred to as "association rule" mining. The most classical algorithm is the a priori algorithm proposed by Agrawal and Srikant in 1994 [11]. This approach finds the frequent items and builds relations aggressively. Then the algorithm will prune the rules by removing the less frequent itemsets. Huang et al. [12] proposed a method based on the physical facts. Dependency modeling was to build physics-based model to describe the relationship between important IC chip power parameters and temperature. Other approaches include principal component analysis contribution plot [13], or partial least squares analysis [14].

In this research, a priori algorithm is applied to the manufacturing dataset in order to mine the relationship model between operation parameters and quality outcomes. Firstly, assume there is a set T formed from a manufacturing dataset, a priori algorithm will find several association rules $X$ ($X \in T$) that can impact final quality $Y$ and meanwhile satisfy minimum property of support. The sets are called frequent item sets. Secondly, a support threshold $M$ is defined and each frequent item set consists at least $M$ items. Then, breadth-first search (BFS) is used to prune the items that do not meet the requirement. The equation of a priori algorithm to calculate support property of item sets is as follows:

$$Support(X) = \left| t \in T; X \in T \right| / \left| T \right| \quad (1)$$

*Support(X)* indicates the support degree of the association rule $X$, and $t$ stands for a single manufacturing rule. Therefore, the support degree can show the proportion of the frequency of each rule appears in the dataset. The confidence value can then be determined by the probability of $Y$ result once $X$ rule is true. The equation is as follows:

$$Confidence(X,Y) = Support(X \cup Y) / Support(X) \quad (2)$$

To describe the correlation degree between $X$ and $Y$, *Lift* is also defined as:

$$Lift(X,Y) = \frac{Support(X \cup Y)}{Support(X) * Support(Y)} \quad (3)$$

When *Lift(X,Y)* is 1, $X$ and $Y$ are independent and no association rules can be extracted. The correlation between $X$ and $Y$ is higher if *Lift* value is larger. Generally, a priori algorithm is a filtering process based on minimum support and minimum confidence, which are user-defined thresholds. Hence, the involvement of expert knowledge is encouraged in order to have a better filtering process.

## 3 Case study

### 3.1 Dataset description

To verify the proposed methodology, a semiconductor dataset SECOM (SEmiCOnductor Manufacturing) [15] is utilized. This dataset includes 1567 samples, and each consists of 590 manufacturing operation variables and 1 quality variable. Among the dataset, only 104 samples represent failure case. Each operation variable data is collected from a certain process control sensor from semiconductor manufacturing line and is named based on sensor ID. To balance positive and negative dataset, boosting is applied on the dataset so the number of failed instance is 1456. The training and testing dataset are segmented from the original dataset on a 3:1 basis.

For data pre-processing, firstly, we remove 115 redundant features during data pre-processing stage. Secondly, after feature extraction stage, 40 most important features are kept for the following classifier learning [15]. To simulate the scenario of MMS, the 40 features are separated into five groups (each representing a workstation) following the property of semiconductor monitoring process described in [16]. Before classification, PCA is applied on the dataset to acquire features with correlations. For each workstation, the features will be combined with previous workstations and through PCA we can get principle components (PC). The final data set includes 14 correlation features and one quality result label.

### 3.2 Results

Results of the six models learned from SECOM dataset are given in Table 1. To compare our methodology with state-of-the-art work, two references that have done research on SECOM dataset are listed. The evaluation metrics include: accuracy, true positive rate (TP rate), geometric mean (g-mean), balanced error rate (BER), F1 score (also F-score or F-measure), and false alarm rate (FAR). To definition of the confusion matrix for SECOM dataset and the equations for the above metrics are shown in Table 2 and equations (4) to (9).

**Table 1.** Results of the proposed method and related work.

| | ID3 | CART | C5.0 | SVM | Naïve Bayes | A Priori | Rough Set | Ref. [7] | Ref. [4] |
|---|---|---|---|---|---|---|---|---|---|
| **Consider MMS?** | Y | Y | Y | Y | Y | N | N | Y | N |
| **Accuracy** | 0.682 | 0.818 | 0.954 | 0.683 | 0.681 | 0.773 | **0.984** | 0.900 | 0.859 |
| **TP rate** | 0.698 | 0.790 | 0.915 | 0.712 | 0.697 | 0.668 | **1.000** | 0.209 | **1.000** |
| **G-mean** | 0.682 | 0.820 | 0.957 | 0.687 | 0.682 | 0.765 | **0.992** | 0.445 | 0.916 |
| **BER** | 0.683 | 0.821 | 0.958 | 0.687 | 0.682 | 0.765 | **0.992** | 0.578 | 0.919 |
| **F1 score** | 0.666 | 0.826 | 0.956 | 0.659 | 0.665 | 0.746 | **0.992** | 0.209 | 0.641 |
| **FAR** | 0.333 | 0.148 | 0.000 | 0.338 | 0.333 | 0.124 | **0.0164** | 0.053 | 0.161 |

**Table 2.** Confusion matrix for quality prediction dataset.

| | | Predicted Condition | |
|---|---|---|---|
| | | Class = 1 (Fail) | Class = -1 (Pass) |
| True Condition | Class = 1 (Fail) | TP (True Positive) | FN (False Negative) |
| | Class = -1 (Pass) | FP (False Positive) | TN (True Negative) |

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$TP\ rate = \frac{TP}{TP + FN} \quad (5)$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (6)$$

$$BER = 1 - 0.5 \times \left( \left( \frac{FN}{TP + FN} \right) + \left( \frac{FP}{TN + FP} \right) \right) \quad (7)$$

$$F1\ score = \frac{2 \times TP}{(2 \times TP + FP + FN)} \quad (8)$$

$$FAR = \frac{FP}{FP + TN} \quad (9)$$

From the results shown in Table 1, Rough Set learning performs best among our seven models and two previous work. With a 100% TP rate, all failed product can be detected during manufacturing stage. Moreover, the false alarm rate is acceptable (about 1.64%). For quality prediction, the motivation is to relief the overhead of quality inspection and detect product failure at early stage. Therefore, high TP rate with low false alarm rate is the most important thing to consider. Compared with the work of Kerdprasop, K. et at. [4], who used decision tree with boosting technique on the same dataset, the proposed method gain better performance that indicate the introduction of MMS characteristics can benefit the model building process. The result from the work of Arif, F. et al. [7] has the good FAR, however, the very low TP rate is not applicable for quality prediction purpose. The confusion matrix of Rough Set result is shown in Table 3.

**Table 3.** Confusion matrix of Rough Set learning result.

| | | Predicted Condition | |
|---|---|---|---|
| | | Class = 1 (Fail) | Class = -1 (Pass) |
| True Condition | Class = 1 (Fail) | 1456 | 0 |
| | Class = -1 (Pass) | 24 | 1439 |

### 3.3 Discussion on Defect Cause Analysis

Beyond detect product quality, it is also a key to understand the root cause of product failure. In this case, association rule mining is a useful tool for mining the rules representing the relationship between manufacturing operation variables and final quality. The experiment set minimum confidence to 0.2 and minimum support to 2. The repetitive rules are eliminated, and there are 978,664 rules extracted. To avoid rules for exceptions and less frequently occurs, In training dataset, the minimum support is set to be 0.012 so only 483 valid rules remain. Table 4 shows the top 16 rules extracted.

**Table 4.** Top 16 association rules that affect product quality.

| LHS | Support | Confidence | Lift |
|---|---|---|---|
| {V65=very-high,V81=low} | 0.0283 | 1 | 2.005 |
| {V29=normal-high,V65=very-high} | 0.0247 | 1 | 2.005 |
| {V65=very-high,V320=normal-high} | 0.0238 | 1 | 2.005 |
| {V386=high,V547=low} | 0.0224 | 1 | 2.005 |
| {V354=low,V576=high} | 0.0210 | 1 | 2.005 |
| {V67=very-low,V512=high} | 0.0206 | 1 | 2.005 |
| {V67=very-low,V344=normal-high} | 0.0206 | 1 | 2.005 |
| {V274=very-high,V384=normal-low} | 0.0206 | 1 | 2.005 |
| {V246=normal-low,V274=very-high} | 0.0206 | 1 | 2.005 |
| {V274=very-high,V518=normal-low} | 0.0206 | 1 | 2.005 |
| {V274=very-high,V383=normal-low} | 0.0206 | 1 | 2.005 |
| {V245=normal-low,V274=very-high} | 0.0206 | 1 | 2.005 |
| {V274=very- | 0.0206 | 1 | 2.005 |

| | | | |
|---|---|---|---|
| high,V517=normal-low} | | | |
| {V63=very-low,V78=low} | 0.0201 | 1 | 2.005 |
| {V67=very-low,V78=low} | 0.0201 | 1 | 2.005 |
| {V71=very-low,V78=low} | 0.0201 | 1 | 2.005 |

Right hand side (RHS) represents instances of failed product and left hand side (LHS) is the association rule set that can affect RHS. Among 483 association rules, {V65=very-high,V81=low} is the most frequent rule. All the rules extracted have confidence value 1. To visualize the relationship of top 16 rules, Figure 3 shows the balloon plot. Each balloon indicates a valid rule. The bigger size is related to the higher support degree, and the darker color means higher lift and confidence value. The arrows pointing to the balloon are the features that form the rule. If one feature connect to many balloons, the feature can be commonly seen in many different rules. From Figure 3, "V65=very-high" and "V274=very-high" are two important rules.

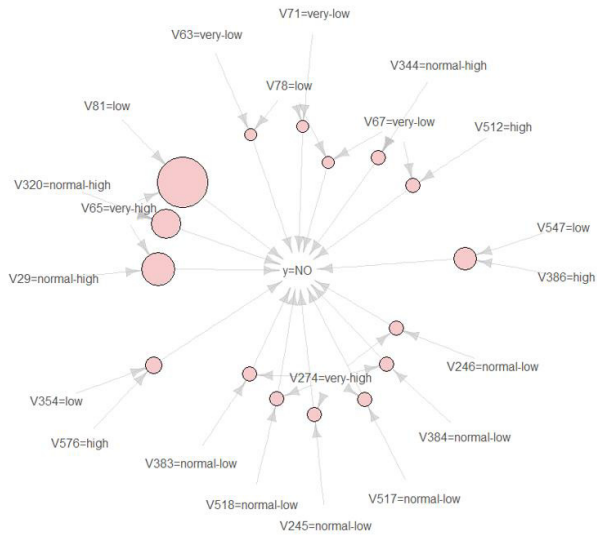


**Fig. 3.** Balloon plot for visualizing extracted association rules.

The quality prediction model is then build with the rules that pass the confidence and minimum support threshold. After applying rules on the testing dataset, 290 product failures are captured, while 74 failed instances are misclassified. Five rules with accurate prediction are listed in Table 5. From the results, many instances meet more than one rules. If the physical meaning of variables is known, the rules can be improved by expertise knowledge involvement and have more accurate prediction performance.

**Table 5.** Results of defect cause analysis.

| Instance ID | Cause |
|---|---|
| 367 | {V441=high,V520=normal-high},{V170=high,V520=normal-high},{V80=very-high,V170=high},{V7=high,V432=high},{V60=normal-high,V432=high},{V80=very-high,V497=high},{V424=high,V426=very-high},{V295=normal-high,V426=very-high},{V160=normal-high,V426=very-high},{V426=very-high,V490=normal-high},{V426=very-high,V471=normal-high}, |
| 368 | {V157=high,V456=low},{V157=high,V184=low},{V157=high,V320=low},{V356=normal-high,V386=high},{V2=low,V430=high}, |
| 370 | {V355=low,V586=high},{V317=low,V568=low},{V217=low,V586=high},{V355=low,V584=high}, |
| 384 | {V438=high,V563=very-low},{V167=high,V563=very-low},{V81=high,V438=high},{V13=low,V201=low},{V60=normal-high,V201=low},{V201=low,V286=normal-high}, |
| 390 | {V124=low,V561=very-high}, |

## 4 Conclusion

While information communications technologies become mature and easily adopted, factories tend to have more data from add-on sensors, machine controllers, metrology, maintenance history, and all kinds of e-manufacturing systems. How to process the data for a deep understanding and insight on product quality is definitely a key for manufacturing industries to be competitive in this Industry 4.0 revolution.

In this paper, a framework for quality prediction modeling in MMS environment is introduced. The proposed methodology with the capability on predictive analysis can benefits manufacturers to acquire the real-time quality information of their manufacturing lines earlier and improve current process immediately. Eventually, it can optimize the manufacturing process and also can extend to the collaboration among manufacturers, suppliers and customers.

This study is conducted under the "III Innovative and Prospective Technologies Project (1/1)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

## References

1. Y. Qi, P. Wang, and X. Gao. Enhanced batch process monitoring and quality prediction using multi-phase dynamic PLS, in *Control Conference (CCC), 2011 30th Chinese*. 2011. IEEE
2. C. Zhao et al., Stage-based soft-transition multiple PCA modeling and on-line monitoring strategy for batch processes. *Journal of Process Control*, 2007. **17**(9): pp. 728-741
3. M. McCann et al., Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control, in *NIPS Causality: Objectives and Assessment*. 2010
4. K. Kerdprasop, N. Kerdprasop, Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process, in *Proceedings of the International MultiConference of Engineers and Computer Scientist*. 2011
5. S. Munirathinam, B. Ramadoss, Predictive Models for Equipment Fault Detection in the Semiconductor Manufacturing Process, *International Journal of Engineering and Technology*, 2016. **8**(4): p. 273
6. D. Djurdjanovic, J. Ni, Stream of variation based analysis and synthesis of measurement schemes in multi-station machining systems, Ann Arbor, 2001. **1001**: pp. 2109-2125
7. F.Arif, N. Suryana, and B. Hussin, A data mining approach for developing quality prediction model in multi-stage manufacturing, *International Journal of Computer Applications*, 2013. **69**(22)
8. P. Jiang et al., Real-time quality monitoring and predicting model based on error propagation networks for multistage machining processes, *Journal of Intelligent Manufacturing*, 2014. **25**(3): pp. 521-538
9. K. Chomboon, K. Kerdprasop, and N. Kerdprasop, Rare class discovery techniques for highly imbalance data, in *Proc. International multi conference of engineers and computer scientists*. 2013
10. K. Salahshoor, H.K. Alaei, and H.K. Alaei. A new on-line predictive monitoring using an integrated approach adaptive filter and PCA, in *Soft Computing Applications (SOFA), 2010 4th International Workshop on*. 2010. IEEE
11. R. Agrawal, R. Srikant. Fast algorithms for mining association rules, in *Proc. 20th int. conf. very large data bases, VLDB*. 1994.
12. H. Huang, G. Quan, and J. Fan, Leakage temperature dependency modeling in system level analysis, in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*. 2010. IEEE
13. K. Friston et al., Functional connectivity: the principal-component analysis of large (PET) data sets, Journal of Cerebral Blood Flow & Metabolism, 1993. **13**(1): pp. 5-14
14. H. Wold, *Partial least squares.* Encyclopedia of statistical sciences, 1985
15. P. Murphy, D. Aha, UCIML repository secom dataset
16. G.S. May, C.J. Spanos, *Fundamentals of semiconductor manufacturing and process control*., 2006: John Wiley & Sons