View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 4 (2013) 201 – 207

Procedia IERI

www.elsevier.com/locate/procedia

2013 International Conference on Electronic Engineering and Computer Science

Cascade Quality Prediction Method Using Multiple PCA+ID3 for Multi-Stage Manufacturing System

Fahmi Arif[a]*, Nanna Suryana[a], Burairah Hussin[a]

[a]*Fac. of Information and Communication Tech., Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Melaka, Malaysia*

Abstract

Quality prediction model, as the key to realize the real-time online quality monitoring process, has been developed using various data mining techniques. However, most of quality prediction models are developed in single-stage manufacturing system, where the relationship between manufacturing operation and quality variables is straightforward. Previous studies show that single-stage quality system cannot solve quality problem in multi-stage manufacturing system due to the complex variable relationships. This study is intended to propose a data mining method to develop quality prediction model which is able to deal with the complex variable relationships in multi-stage manufacturing system. This method, named Cascade Quality Prediction Method (CQPM), is developed by considering the complex variables relationships in multi-stage manufacturing system. CQPM employs the combination of multiple Principal Component Analysis and Iterative Dichotomiser 3 algorithm. A case study in semiconductor manufacturing shows that the prediction model that has been developed using CQPM is performed better in predicting both positive and negative classes compared to others.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer review under responsibility of Information Engineering Research Institute

Keywords: Data mining; multi-stage manufacturing; quality prediction; PCA; ID3.

* Corresponding author. Tel.: +6063316023; fax: +6063316500.
E-mail address: fahmi_ra@utem.edu.my.

2212-6678 © 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer review under responsibility of Information Engineering Research Institute
doi:10.1016/j.ieri.2013.11.029

1. Introduction

Multi-stage Manufacturing System (MMS) refers to the system that involves more than one workstation to finish all required operation to manufacture the final product [1]. MMS is employed to produce complex products where raw materials are transformed into final product in a series of processing stages. Various data mining techniques have been implemented to develop quality prediction model for MMS in order to achieve faultless manufacturing operations. Quality prediction model is a formulation that explains the relationship between manufacturing operation and product quality level. It is the key to enable the ability in estimating product quality level during the manufacturing operation.

In every workstation within MMS there are two types of quality: partial and total quality [2]. Partial quality is the specific result from current operation, while the total quality is the result of current and preceding operation. The existence of partial quality and total quality in every workstation in MMS denotes that variable relationship in MMS is more complex than in Single-stage Manufacturing System (SMS). Considering the complex variable relationships in MMS, this study is intended to propose a data mining method, named Cascade Quality Prediction Method (CQPM), for developing quality prediction model in MMS.

2. Related Works

Previous studies explained that basically there are two different approaches in developing quality prediction model in MMS; single-point and multi-point approaches. Using single-point approach, one quality prediction model is developed for the whole manufacturing lines as illustrated in Fig. 1(a). On the other hand, using multi-point approach, one quality prediction model is developed for every workstation as illustrated in Fig. 1(b). Several studies show that various techniques have been applied to develop quality prediction model using single-point approach. Multi-PCA model [3], Multi-way Partial Least Square control chart [4], Directional Multivariate Exponential Weighted Moving Average [5], clustering [6], classification [7], and association rules [8-9] have been used to develop quality prediction in MMS using single-point approach.

Single-point approach assumed that each manufacturing workstation has an independent effect to the product quality level. It is ignoring the fact that each workstation has its own operation condition and behaviour [10]. Therefore, this model has difficulty to reveal the correlation between manufacturing operations from workstation to workstation. From the point of view of partial and total quality as explained by [3], this approach can only explain the partial quality at the last workstation.

Instead of using the single-point approach, some scholars developed quality prediction model for every workstation individually as illustrated in Fig. 1(b). Multi-point approach is applied by [11], [12] in fermentation process, and [10], [13] in injection moulding process. This approach produced the model that is able to explain the relationship among manufacturing operation variables in a workstation. However, [5] pointed out that this approach can be misleading and ineffective considering that the measurement of a workstation is probably confounded by the cumulative effect from the previous workstation.

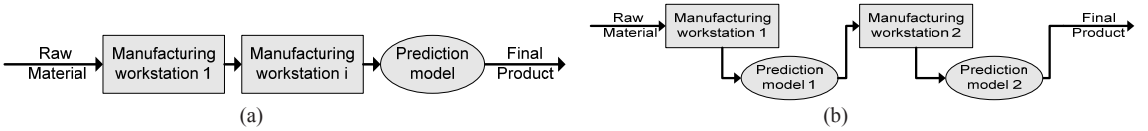

Fig. 1. (a) Single-Point Prediction Method; (b) Multi-Point Prediction Method

3. Framework of CQPM

Based on the concept of partial and total quality as described by [3], the characteristics of the output from a workstation are influenced by the manufacturing operation in that particular workstation and all preceding workstations. This concept can be illustrated in two-workstation MMS as shown in Fig. 2 (a). Considering the partial and total quality in a workstation as shown in Fig. 2 (a), the relationship among variables in MMS can be illustrated as shown in Fig. 2 (b).

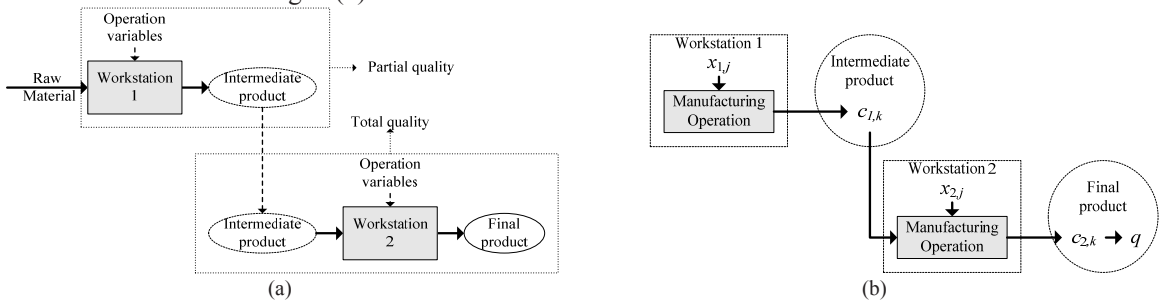

Fig. 2. (a) Partial and Total Quality in Two-Workstation MMS; (b) Illustration of Variable Relationships in Two-Workstation MMS

In the real world industrial setting, MMS probably consists of various numbers of workstations, every workstation consists of various numbers of manufacturing operation variables, and the intermediate product from each workstation has various numbers of characteristics. For this condition, the variables relationships in MMS can be expressed as follows:

$$c_{i,k} = f(c_{i\text{-}1,k}, x_{i,j}) \tag{1}$$

$$q = f(c_{n,k}) \tag{2}$$

where:

$x_{i,j} = j^{\text{th}}$ manufacturing operation variable in i^{th} workstation, $i = 1, 2, 3, \ldots, n$ and $j = 1, 2, 3, \ldots, m$

$c_{i,k} = k^{\text{th}}$ characteristics of the output from i^{th} workstation, $k = 1, 2, 3, \ldots$

q = final product quality level

The task of revealing relationship among operation variables as expressed in Eq. *(1)* is the process to investigate how the interaction of $x_{i,j}$ and $c_{i\text{-}1,k}$ can construct $c_{i,k}$. Without any underlying knowledge of the relationships among $x_{i,j}$, the process of finding the relationships of inter-correlated variables is the same with extracting those variables into some sets of new dimensions. This idea is exactly the same idea with Principal Component Analysis (PCA) technique. PCA can be employed for its ability in extracting the important information of several inter-correlated variables to be expressed as a set of new orthogonal variables [14].

Using PCA, the relationships between product characteristics and manufacturing operation variables in every workstation can be expressed as shown in Eq. *(3)*. $a_{k,j}$ is the amount of contribution of $x_{i,j}$ to $c_{i,k}$. In PCA, $a_{k,j}$ is the eigenvector of the covariance matrix of the variables.

$$c_{i,k} = \sum_{j=1}^{m(i-1)} a_{k,j} c_{(i-1),k} + \sum_{j=\left(m_{(i-1)}+1\right)}^{m_i} a_{k,j} x_{i,j} \tag{3}$$

In the last workstation, relationship between manufacturing operation variables and product quality level can be represented by the relationships between product characteristics ($c_{n,k}$) and the product quality level (q),

since the interaction among manufacturing operation variables has been represented by the product characteristics. This relationship is expressed in Eq. *(2)*. In the real world manufacturing setting, product quality level (q) is usually determined by the category either it is accepted or rejected. Therefore, the process of revealing Eq. *(2)* can be approached as the classification problem. In this study, Iterative Dichotomiser (ID3) algorithm is employed because of it advantages such as simple, high speed computation and easy-to-understand rules [7] which are offering more benefit in practical.

ID3 algorithm that can only deal with categorical variable, therefore $c_{n,k}$ should be converted into categorical variables. Using the division of quality control chart area [15], $c_{n,k}$ is converted into categorical variables using its mean and standard deviation. As the result, there are 6 possible categories for $c_{n,k}$: *very low* ($c_{n,k} < \left(\overline{c_{n,k}} - 2\sigma\right)$), *low* ($\left(\overline{c_{n,k}} - 2\sigma\right) \le c_{n,k} < \left(\overline{c_{n,k}} - \sigma\right)$), *lower medium* ($\left(\overline{c_{n,k}} - \sigma\right) \le c_{n,k} < \overline{c_{n,k}}$), *upper medium* ($\overline{c_{n,k}} < c_{n,k} \le \left(\overline{c_{n,k}} + \sigma\right)$), *high* ($\left(\overline{c_{n,k}} + \sigma\right) < c_{n,k} \le \left(\overline{c_{n,k}} + 2\sigma\right)$), and *very high* ($c_{n,k} > \left(\overline{c_{n,k}} + 2\sigma\right)$), where $\overline{c_{n,k}}$ is the mean of $c_{n,k}$ and σ is it standard deviation. Hereinafter, the ID3 algorithm can be applied to extract the rules.

4. Case Study of CQPM Implementation

In order to evaluate CQPM performance, CQPM is applied into a semiconductor manufacturing dataset, namely SECOM dataset [16]. SECOM dataset consists of 1567 instances, and every instance has 590 manufacturing operation data and 1 quality data. The manufacturing operation data are collected from the continuous monitoring process using sensors and metrology equipments along the semiconductor manufacturing line. The manufacturing operation variables are named based on the sensors number, hence there are variables S_1 until S_{590}. At the end of manufacturing operation, functional testing was performed to ensure that the semiconductor meets the specification for which it is designed. If the result was meeting the expectation, then the semiconductor was classified as accepted product, otherwise it was rejected.

As a real life dataset, SECOM contain of some irrelevant variables and missing value data. A data cleansing procedure discards 452 instances with null and missing values. Regarding the irrelevant variables, since not all 590 sensors were used to gather quality-related data, [16] suggested the simple feature selection technique to select 40 variables that is highly related to the quality variables. These 40 variables are divided into five workstations based on the typical semiconductor manufacturing monitoring process as explained by [18]. The manufacturing operation variables are grouped into the each workstation as shown in Table 1.

Based on the framework of CQPM as explained previously, the process of developing the quality prediction model for MMS can be summarized as follows:

- Step 1: For $i = 1$ to $i = n$, reveal $c_{i,k} = f(c_{i\text{-}1,k}, x_{i,j})$ by applying multiple PCA into the dataset
- Step 2: Calculate the value of $c_{n,k}$ for every instance then transform this value into category
- Step 3: Using ID3, extract rules IF $c_{n,k} = a$ THEN q = b,

a = category of $c_{n,k}$ (*very low, low, lower medium, upper medium, high, very high*)

b = product quality level (*accepted, rejected*)

In this study, this process is implemented to SECOM dataset using MATLAB and RapidMiner 5 on 2.20 GHz computer with 2.00 GB memory.

As the result of applying multiple PCA, various number $c_{i,k}$ for every workstation are produced as shown in Table 2. In the last workstation, 22 mathematical models ($c_{5,1} - c_{5,22}$) are produced. These models are representing the cumulative effect of entire manufacturing operation variables to the final product characteristics. These models are used to calculate the value of $c_{5,1} - c_{5,22}$ for every instance hence the new dataset is produced. ID3 algorithm is then applied to this dataset after the value of $c_{5,1} - c_{5,22}$ for every instance has been transformed into categories. As the result, 219 if-then rules are extracted.

Table 1. Semiconductor Manufacturing Operation Variables in Every Workstation

Workstation	Number of Variable	Original Name of Variables	Variable Name for CQPM
1	4	S_{15}, S_{27}, S_{33}, S_{36}	$x_{1,1}$, …., $x_{1,4}$
2	9	S_{48}, S_{60}, S_{62}, S_{64}, S_{118}, S_{122}, S_{124}, S_{125}, S_{131}	$x_{2,1}$, …., $x_{2,9}$
3	10	S_{134}, S_{145}, S_{153}, S_{184}, S_{201}, S_{206}, S_{288}, S_{342}	$x_{3,1}$, …., $x_{3,10}$
4	8	S_{421}, S_{426}, S_{427}, S_{430}, S_{435}, S_{454}, S_{461}, S_{470}	$x_{4,1}$, …., $x_{4,8}$
5	9	S_{478}, S_{492}, S_{511}, S_{520}, S_{525}, S_{560}, S_{569}, S_{572}, S_{574}	$x_{5,1}$, …., $x_{5,9}$

Table 2. Involved Variables in Every Workstation

Workstation	Input Variable		Output Variable	
	Number	Name	Number	Name
1	4	$x_{1,1}$ until $x_{1,4}$	3	$c_{1,1}$ until $c_{1,3}$
2	12	$c_{1,1}$ until $c_{1,3}$ and $x_{2,1}$ until $x_{2,9}$	9	$c_{2,1}$ until $c_{2,9}$
3	19	$c_{2,1}$ until $c_{2,9}$ and $x_{3,1}$ until $x_{3,10}$	13	$c_{3,1}$ until $c_{1,13}$
4	21	$c_{3,1}$ until $c_{1,13}$ and $x_{4,1}$ until $x_{4,8}$	17	$c_{4,1}$ until $c_{4,17}$
5	26	$c_{4,1}$ until $c_{4,17}$ and $x_{5,1}$ until $x_{5,9}$	22	$c_{5,1}$ until $c_{5,22}$

The implementation of CQPM to SECOM dataset produces the quality prediction model which consists of 22 mathematical model and 219 if-then rules. In order to evaluate the performance of this model, 10-fold cross validation is performed. For the comparison, two single-point prediction methods are also applied. The first comparison model is developed using single-point approach with ID3 algorithm (SP-ID3) whereas the other is developed using single-point approach with PCA+ID3 (SP-PCA+ID3). The result of validation result is shown in Table 3.

In Table 3, it is shown that the prediction model that has been developed using CQPM involves fewest numbers of variables than others. It leads to the lowest computation time to build the model. In term of accuracy, highest accuracy is obtained by SP-PCA+ID3 method. However, this measurement is inappropriate to explain the performance of decision tree algorithm which is applied in imbalanced dataset. Alternatively, [19] suggest the measurement of G_{mean} that indicates the ability of the model in classifying both positive and negative classes. Since CQPM achieve the highest G_{mean}, it can be concluded that the model that has been developed using CQPM is performed better in classifying both accepted and rejected classes compared to SP-ID3 and SP-PCA+ID3.

Table 3. Comparison of Cross Validation Result

Prediction Method	Number of Variables	Computation Time	Accuracy	G_{mean}
CQPM	22	1.045 seconds	0.9002	0.4448
SP-ID3	40	3.884 seconds	0.8808	0.2123
SP-PCA+ID3	26	2.374 seconds	0.9113	0.2643

5. Discussion

Based on the 10-fold cross validation result as explained previously, it can be concluded that the

implementation of CQPM in MMS are able to produce a quality prediction model with better performance compare to the model that are developed using SP-ID3 and SP-PCA+ID3. SP-ID3 treats all manufacturing operation variable as having equal contribution to the final product quality level. It assumes that every manufacturing operation variable has individual effect to the final product quality. As the result, final product quality can be directly estimated by evaluating the value of manufacturing operation variable using the extracted rules.

Differently, SP-PCA+ID3 and CQPM are considering the interaction effect of the manufacturing operation variables to the final product quality. SP-PCA+ID3 assume that all manufacturing operation variables are interacted each other in the same time as in single manufacturing system. On the other hand, CQPM employ multiple PCA from workstation to workstation hence the cumulative effect of manufacturing operation variables to the final product quality can be captured.

In this study, the model that is developed using CQPM has been proved that it performs better than others. However, considering the relatively high accuracy (0.9002) and the relatively low G_{mean}, (0.4448), it can be concluded that the probability of miss-classification in negative class is still high. Further improvement in technical level to increase the performance of this method is still possible. Additional technique might be combined to improve the performance of the prediction model.

References

[1] Li J, Freiheit T, Hu SJ, Koren Y. A Quality Prediction Framework for Multistage Machining Processes Driven by an Engineering Model and Variation Propagation Model. Journal of Manufacturing Science and Engineering. 2007:129: 6 1088–1100.
[2] Xu X. Fuzzy Control for Manufacturing Quality Based on Variable Precision Rough Set. Fifth World Congress on Intelligent Control and Automation. 2004 2347–2351.
[3] Zhao C, Wang F, Lu N, Jia M. Stage-Based Soft-Transition Multiple PCA Modeling and On-Line Monitoring Strategy for Batch Processes. Journal of Process Control. 2007:17:9 728–741.
[4] Chang Y, Wang J, Tan S, Wang F, Chen W. Quality Prediction of Strip Steel Based on Windows-Mean MPLS. International Journal of Iron and Steel Research. 2010: 17:7 28–33.
[5] Zou C, Tsung F. Directional MEWMA schemes for multistage process monitoring and diagnosis. Journal of Quality Technology. 2008: 40:4 407–427.
[6] Kim D, Kang P, Cho S, Lee H, Doh S. Machine Learning-Based Novelty Detection for Faulty Wafer Detection in Semiconductor Manufacturing. Expert Systems with Applications. 2012:39:4 4075–4083.
[7] Jingjun F, Shuting Y. Alumina Production Operations Management Information System Based on Data Mining Technology. Int Forum on Computer Science-Technology and Applications. 2009 . 315–318.
[8] Chen W, Tseng S, Wang C. A Novel Manufacturing Defect Detection Method Using Association Rule Mining Techniques. Expert Systems with Applications. 2005: 29:4 40–52.
[9] Lau H, Ho G, Chu K, Ho W, Lee C. Development of An Intelligent Quality Management System Using Fuzzy Association Rules. Expert Systems with Applications. 2009: 36:2 1801–1815.
[10] Guo X, Wang F, Jia M. A Stage-Based Quality Prediction and Control Method for Batch Processes. International Conference on Machine Learning and Cybernetics. 2005:4 2044–2049.
[11] Ma L, Jiang Y, Wang F, Gao F. Multi-PCA Models for Process Monitoring and Fault Diagnosis. International Sysmposium on Advance Control of Chemical Process, 2003.
[12] Qi Y, Wang P, Gao X. Enhanced Batch Process Monitoring and Quality Prediction Using Multi-Phase Dynamic PLS. 30th Control Conference (CCC). 2011: 5258–5263.
[13] Ge Z, Zhao L, Yao Y, Song, Gao F. Utilizing Transition Information in Online Quality Prediction of Multiphase Batch Processes. Journal of Process Control. 2012 1–13.

[14] Abdi H, Williams LJ. Principal Component Analysis. Wiley Interdiciplinay Reviews: Computational Statistics, New York: John Wiley & Sons; 2010.
[15] DeVor RE, Chang T, Sutherland JW. Statistical Quality Design and Control: Contemporary Concepts and Methods. 2nd ed. Prentice-Hall; 2007.
[16] McCann M, Johnston A. Secom Dataset. UCI Machine Learning Repository. 2008. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html.
[17] May GS, Spanos CJ, Fundamentals of Semiconductor Manufacturing and Process Control. New Jersey: John Wiley & Sons; 2006.
[18] Jolliffe IT. Principal Component Analysis. New York: Springer; 2002
[19] García S, Fernández A, Herrera F. Enhancing the Effectiveness and Interpretability of Decision Tree and Rule Induction Classifiers with Evolutionary Training Set Selection Over Imbalanced Problems. Applied Soft Computing. 2009: 9:4 1304–1314.