World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:4, 2010

Machine Scoring Model Using Data Mining Techniques

Wimalin S. Laosiritaworn, Pongsak Holimchayachotikul

***Abstract*—** this article proposed a methodology for computer numerical control (CNC) machine scoring. The case study company is a manufacturer of hard disk drive parts in Thailand. In this company, sample of parts manufactured from CNC machine are usually taken randomly for quality inspection. These inspection data were used to make a decision to shut down the machine if it has tendency to produce parts that are out of specification. Large amount of data are produced in this process and data mining could be very useful technique in analyzing them. In this research, data mining techniques were used to construct a machine scoring model called 'machine priority assessment model (MPAM)'. This model helps to ensure that the machine with higher risk of producing defective parts be inspected before those with lower risk. If the defective prone machine is identified sooner, defective part and rework could be reduced hence improving the overall productivity. The results showed that the proposed method can be successfully implemented and approximately 351,000 baht of opportunity cost could have saved in the case study company.

***Keywords*—** Computer Numerical Control, Data Mining, Hard Disk Drive.

I. Introduction

HARD disk drive industry plays a significant role in Thailand's economic structure as Thailand is the one of the biggest manufacturer of hard disk drive head. The number of hard disk drive manufactured in Thailand has grown rapidly during the past few years. The case study company is the manufacturer of hard disk drive arm, one of the important parts in hard disk drive head. This part is manufactured in high volume using CNC machine. Sample of parts are taken from each machine during production run. They are inspected by Quality Control (QC) department and if samples are out of specification or show tendency to be, the machine need to be adjusted or shutdown before the damage was done. Currently, more than 120 machines are in use each day to manufacture hard disk drive arm. 6 parts per machine per shift were sample for quality inspection. The number of part to be inspected are currently exceeding QC department capacity, which results in large number of work in process each day. Inspection data were kept in the company database and large amount of data are produced every day. However, day to day job has keeping the company from analyzing that large amount of data to produce useful information.

Data mining (DM) is a set of tools that have been used widely for analyzing large amount of data to produce useful information. This paper aims to apple DM for machine scoring based on past quality inspection data. Machines are clustered into three groups which are low, medium and high risk to produce defective product based on their past performance. Then the Simple Additive Weight (SAW) was used to score each machine. The score is used to setup the order for QC inspection. This proposed method is named 'machine priority assessment model, MPAM'. With this model, machine with low score, ie. high risk to produce defective part, can be inspected before the lower risk machine. As a result, number of defective parts can be reduced.

This rest of the paper is organized as follows. Section 2 illustrates the reviews of related literatures. Section 3 proposes a research framework with detailed procedures for empirical study in hard disk drive manufacturing using data mining, decision making models and knowledge retrieval from machine database. Results and discussion of the proposed framework are mentioned and analyzed in section 4. Finally, conclusions and further research directions are provided in section 5.

W. S. Laosiritaworn is with the Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand (phone: +66-5394-4183; fax: +66-5394-4185; e-mail: wimalin@hotmail.com).

Pongsak Holimchayachotikul is with the Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand.

II. Literature Reviews

Data mining has been extensively used and received attention from the production field during the 1990s [1]-[3]. Presently, Data mining is employed not only in business but also in many different areas in manufacturing engineering to retrieve knowledge for use in process modeling predictive maintenance, fault detection, design, production, quality assurance, scheduling, and decision support systems. DM consists of five class of methods which are, predictive modeling, clustering, data summarization, dependency modeling, and change and deviation detection [4]. Data mining task used in this research are the clustering.

A. Data Clustering

Clustering is an unsupervised data mining technique to classify patterns into interested clusters. A number of algorithms for clustering have been proposed that can be divided into two main types: hierarchal and non-hierarchal approaches. k-means clustering is the most well-known and

frequently used in non-hierarchal clustering application. Gelbard et al.[5] compared 10 common clustering algorithms on four known data sets. The conclusion was to carefully choose clustering algorithm as the results is highly problem dependent. The three highest raking in the overall performance of those four data sets were the TwoStep, K-means and Binary-Positive algorithm respectively. In this research, k-means, EM and TwoStep algorithms have been applied for machine clustering.

1) K-means clustering

The k-means method is a clustering method, used to group records based on similarity of values for a set of input fields. The basic idea is to try to discover k clusters, such that the records within each cluster are similar to each other and distinct from records in other clusters. K-means is an iterative algorithm; an initial set of clusters is defined, and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit).

2) Expectation Maximization (EM) clustering

EM computes the probabilities of cluster membership based on one or more probability distributions in order to maximize the overall probability of the data to provide final cluster.

3) TwoStep clustering

TwoStep clustering algorithm was developed by Chiu et al. [5]. TwoStep algorithm use a likelihood distance measure that assumes the independent of variables in the cluster. Advantages of TwoSteps over the traditional methods are for example it automatically select appropriate number of cluster, can be used with both continuous and categorical variable, work efficiently with large amount of data.

B. Simple additive weight

This is also called the weighted sum method [7] and is the simplest and still the widest used Multiple Attribute Decision Making (MADM) method. Here, each attribute is given a weight, and the sum of all weights must be 1. Each alternative is assessed with regard to every attribute. The overall or composite performance score of an alternative is given by Equation 1.

$$P_i = \left[\sum_{j=1}^{M} w_{ij} (m_{ij})_{normal} \right] / \sum_{j=1}^{M} w_j \quad (1)$$

Where (m_{ij}) normal represents the normalized value of m_{ij}, and Pi is the overall or composite score of the alternative A_i. The alternative with the highest value of P_i is considered as the best alternative.

Many types of kernel functions can be applied. Equation (1) implies that the dot product in high dimensional space is equivalent to a kernel function of the input space. Mercer's condition is adopted to determine whether kernel function can be used.

III. Methodology

Research methodology is summarized in Figure 1. It Start from 'Objective definition' where the survey of current situation in QC department and preliminary data analysis was conducted. Then come the 'Data preparation' which data from company database was collected and clean to get rid of error such as missing data, incorrect data. Also in this stage, data were prepared to the format that can be compatible with the software used in 'Knowledge retrieval' stage

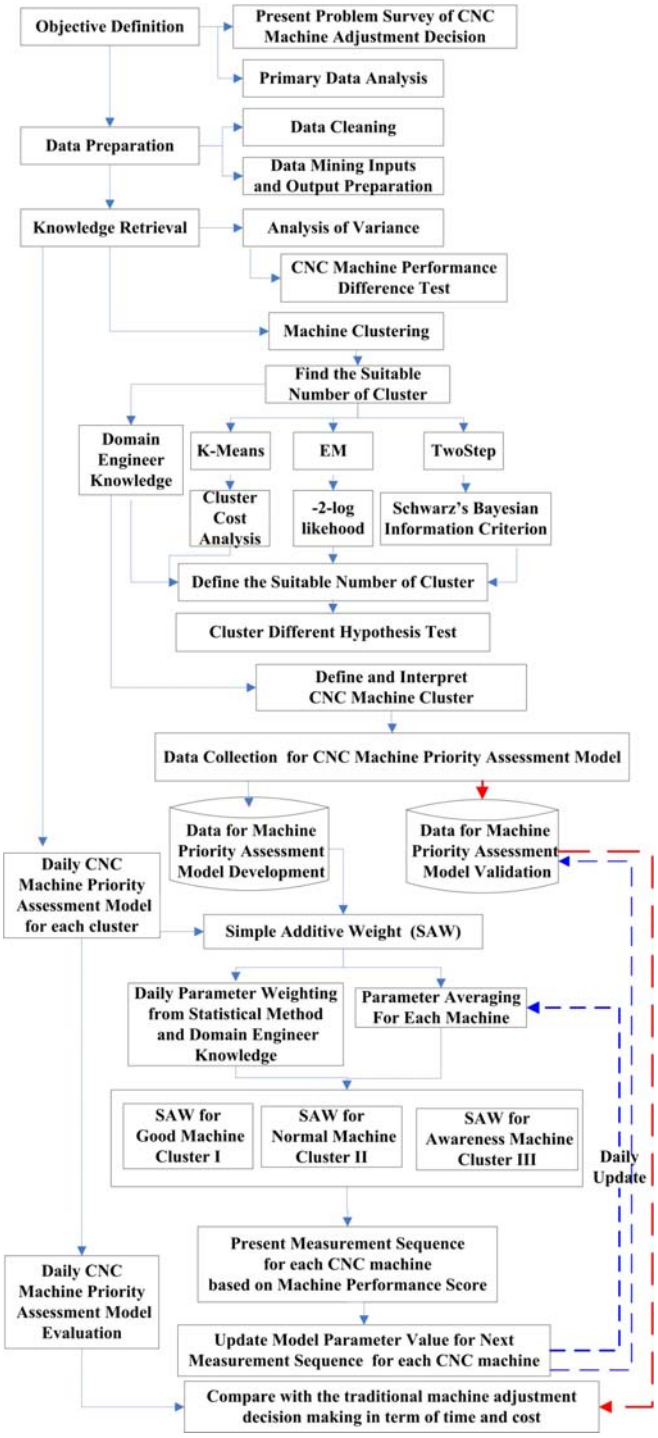

Fig. 1 Schematic diagram of the machine priority assessment model

In the 'Knowledge retrieval', Analysis of variance (ANOVA) was conducted first to ensure the statistically

difference of the performance of each CNC machine. Then clustering algorithms were implemented which are, K-means, Expectation Maximization (EM) and two step. Hypothesis test was conducted to identify the most appropriate cluster algorithm.

After clustering, daily data collected from CNC machine is stored in the database. Then the machine priority assessment model was developed by using SAW and the database of CNC clustering. The model construction starts from machine clustering, using only significant parameters. The results of the clustering were interpreted by domain engineering knowledge based on the different values of each factor according to each cluster. The average of each factor from each machine is calculated and also daily parameter weighting from statistical methods and domain engineering knowledge are conducted. These two components are then used in the SAW structure of the Daily CNC Machine Priority Assessment Model (MPAM). The results of this model are the machine performance scores, which score the performance of each machine. The quality inspection sequence can be established from those results. After practicing the results, the newly obtained ones from the current day will be updated in the processing for the rankings of the next day. The quality inspection sequence obtained from this model is compared with the actual inspection sequence (random inspection) using actual data to prove that the proposed model results in early detection of false machine.

TABLE I
ANALYSIS OF VARIENCE OF 22 PARAMETERS

Para-meter	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
A	12.9837	153	0.0849	48.8	0.0000
B	12.4915	153	0.0816	47.53	0.0000
C	890.626	153	5.82108	48.38	0.0000
D	889.449	153	5.81339	48.31	0.0000
E	186541	153	1219.22	47.78	0.0000
F	186597	153	1219.59	47.76	0.0000
G	11663.3	153	76.2307	47.71	0.0000
H	704797	153	4606.52	47.77	0.0000
I	521610	153	3409.22	47.75	0.0000
J	83996.6	153	548.997	47.77	0.0000
K	1.35E+06	153	8850.71	47.77	0.0000
L	99758.6	153	652.017	47.77	0.0000
M	8293.58	153	54.2064	47.81	0.0000
N	50529.7	153	330.259	47.75	0.0000
O	0.1440	153	0.0009	19.34	0.0000
P	0.1821	153	0.0012	24.46	0.0000
Q	0.0263	153	0.0002	41.05	0.0000
R	0.0272	153	0.0002	43.01	0.0000
S	0.0118	153	0.0001	11.03	0.0000
T	0.0026	153	0.0000	4.3	0.0000
U	0.5487	153	0.0036	81.17	0.0000
V	0.5214	153	0.0034	69.86	0.0000

IV. RESULTS AND DISCUSSION

A. Data collection

The inspection data of hard disk drive arm were collect from approximately 120 machines on service each day. 6 Samples were taken from each machine to perform inspection of about 22 parameters. Due to the commercial confidentiality, the name of variables are referred to as parameter A to parameter V. Analysis of Variance (ANOVA) was performed to ensure the mean difference of the 22 measured parameters produced from different machine. The results of f value from ANOVA are summarized in table 1. P value lower than 0.05 suggest the significant different in mean at 95% confidence, as a results it can be concluded from table 1 that all the machines produce part with significantly different in mean.

B. Clustering

For the TwoStep clustering algorithm, number of clusters were set to minimum of 2 and maximum at 15. Distance measure using Log-likelihood was used. Number of appropriated cluster determined by Silhouette coefficient was found to be 3. Percentage of data separated in to three groups were 22.6%, 38.1% and 39.4%.

TABLE II
MEAN AND C_{PK} FOR THE 22 PARAMETERS OF EACH CLUSTER FROM TWOSETP ALGORITHM

Para-meter	*Cluster 1*		*Cluster 2*		*Cluster 3*	
	Mean	*Cpk*	*Mean*	*Cpk*	*Mean*	*Cpk*
A	6.29	2.69	6.25	1.43	6.19	-0.84
B	6.29	2.71	6.26	1.28	6.19	-1.15
C	2.67	12.85	2.67	18.08	2.67	9.12
D	2.67	12.99	2.67	18.28	2.67	9.31
E	1.17	1.31	1.17	2.10	1.16	1.43
F	-1.15	4.43	-1.15	11.70	-1.15	6.07
G	-1.15	7.34	-1.15	30.85	-1.15	10.87
H	1.17	1.24	1.17	1.96	1.16	1.32
I	21.88	0.90	21.88	1.44	21.88	0.97
J	16.45	-0.08	8.07	-1.90	-6.38	-49.64
K	5.48	0.14	5.78	-1.85	6.30	-14.75
L	21.88	1.03	21.88	1.58	21.88	1.08
M	-6.31	-0.06	-2.00	-1.90	5.43	-48.95
N	-23.18	0.03	-15.97	-1.85	-3.54	-47.87
O	-6.31	-0.07	-2.00	-1.90	5.44	-46.12
P	6.14	0.08	3.89	-1.82	0.03	-49.29
Q	6.25	0.20	7.16	-1.78	8.72	-46.26
R	8.32	-0.08	-3.29	-1.91	-23.31	-50.08
S	8.83	-0.09	11.72	-1.91	16.71	-46.75
T	0.11	-0.04	3.27	-1.88	8.70	-49.68
U	-3.51	0.01	-4.59	-1.87	-6.44	-41.64
V	5.48	0.14	5.78	-1.83	6.30	-14.98

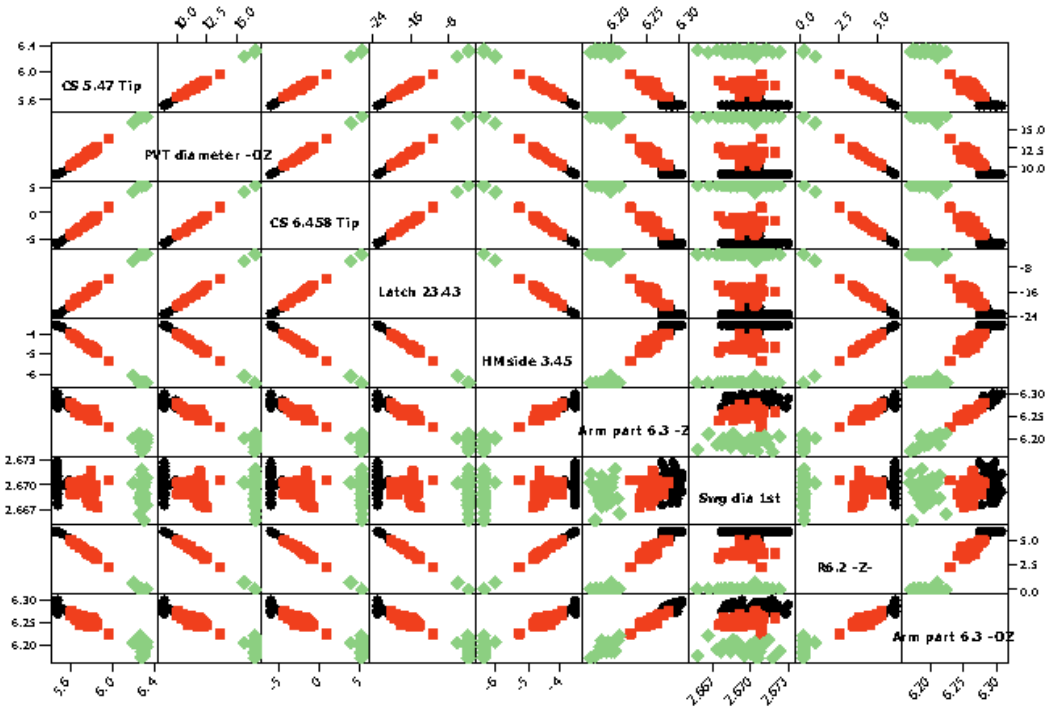

Fig. 2. Matrix plot of parameters clustered with TwoStep

For k-means and EM clustering algorithm, cluster costs were calculated to determine the appropriate number of cluster. It was found that 3 and 24 groups are appropriate for k-means and EM respectively.

In order to compare results from three algorithms, ANOVA analysis was carried. It was found that p-value of TwoStep algorithm are smallest. It can be concluded that the cluster obtained from this algorithm provide the most different in mean of each cluster. Cluster identified by TwoStep algorithm was then selected for further analysis

Table 2 shows cluster mean and C_{pk} for each parameter for the cluster obtained from TwoStep clustering. C_{pk} is a measure of process capability. It indicated the ability of the process to produce output within engineering tolerance and specification. Value near or below zero indicated that the process has high variation or off target. As a results, cluster 1 are machines with good condition (low risk of producing defective parts), cluster 2 are machines with normal condition (medium risk of producing defective parts), while cluster 3 are machines with poor condition (high risk of producing defective parts).

Figure 2. Show the example of matrix plot of parameters clustered with TwoStep. It shows that the clusters were formed very well with distinct different between cluster.

C. Simple Additive Weight

With the well defined cluster SAW were implemented to score machine in each cluster to score to identify order in which each machine will be inspected. The machine with low score, ie. prone to defect, will be inspected first.

After the MPAM was constructed, to validate the model, the random sequence of the traditional method was performed as usual. The intervals between the initial inspection and discovery of defected parts are then recorded. The CNC-MPAM usage example, compared the traditional method from real production, is demonstrated. For example machine number 139 and 141 is producing defective parts. The time from initial inspection to the inspection of machine number 139 and 141 is recorded. This time, the measured sequence of the proposed method will then be conducted. Since the inspection sequence is now ordered to inspect defective-prone machines in top priority, machine number 139 and 141 should be inspected earlier than the previous method. In this example, the difference of the two methods is then compared in terms of time and cost of defect also showed in Figure 3. The proposed method could have prevented 5,400 parts to be defected and results in opportunities cost saving of approximately 351,000 baht.

V. Conclusion

The CNC machine tooling setting process is a crucial process in hard disk drive arm production as the quality of finished arm depended on the accuracy and performance of CNC machine. The suitable time for CNC machine tooling setting is able to directly increase the accuracy and performance of CNC. Therefore, this research develop an application to provide the daily inspection order for CNC machine tooling adjustment of which the machines with higher risk of producing defective parts can be inspected and corrected before those with lower risk.

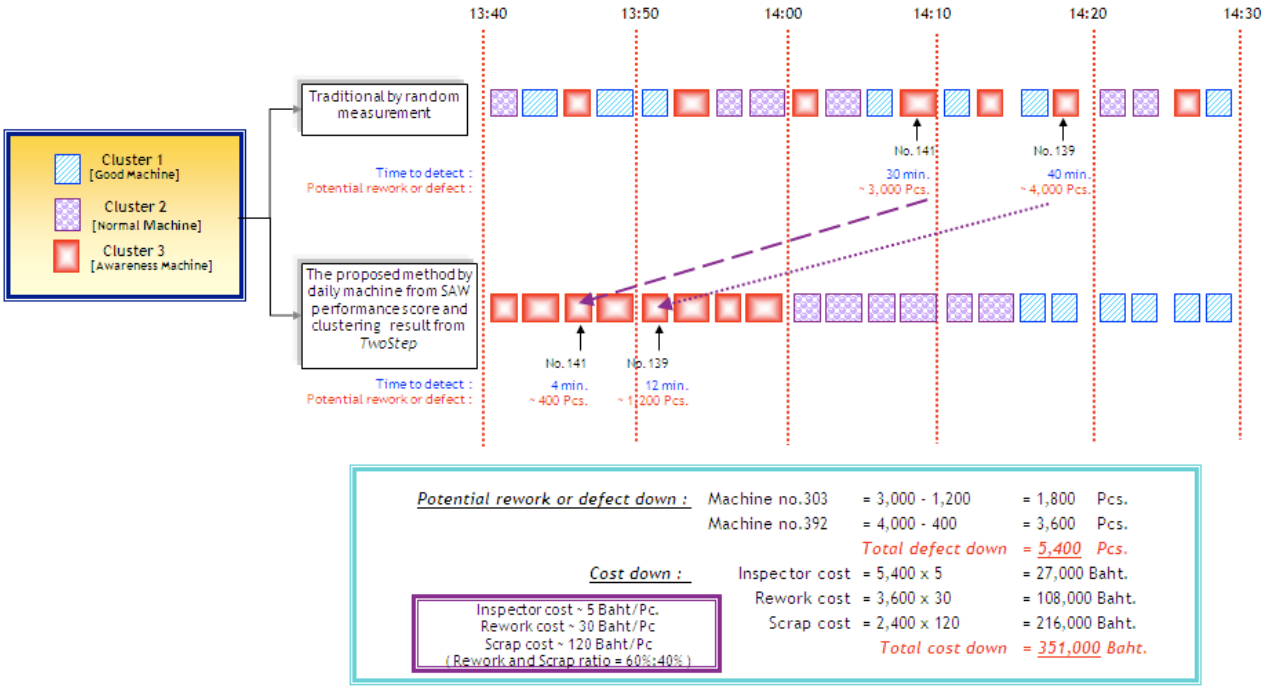

Fig. 3. Model evaluation result showing the comparison of 'random' inspection vs. the order obtained from MPAM.

Most engineers are familiar with the SAW method; nevertheless, this approach has some concerning point such as the normalization method might be complex enough to handle some problem. One key of knowledge discovery and retrieval in database in practice is data preparation. The data selection and data preprocessing, taken very time-consuming, cannot be ignored and needs much patience. Based on the empirical results, MPAM was validated. The results suggested that MPAM are capable of defect reduction by reordering inspection sequence.

However, there are certain limitations of this research such as the time to detect of abnormal machine depends significantly on the parameter weighting. From stochastic process analysis with markov chain, MPAM excludes probability of defect occurrence state and the first time to shut down for each machine. Further potential research may be embraced by applying other MADM methods such as Fuzzy-SAW Fuzzy-AHP instead of SAW or improve the parameter normalization in SAW. Moreover, we can develop the model related with the real process natural in term of stochastic process to increase the rate of time to detect abnormal machine.

Acknowledgment

The authors would like to thank the Industry/University Cooperative Research Center)I/UCRC (in HDD Component, the Faculty of Engineering, Khon Kaen University and National Electronics and Computer Technology Center, National Science and Technology Development Agency, Thailand for financial support.

References

[1] Han, J., & Kamber, M., *Data mining: concepts and techniques*. Morgan Kaufmann Publishers., 2001

[2] Irani, K. B., Cheng, J., Fayyad, U. M., and Qian, Z., "Applying Machine Learning to Semiconductor Manufacturing," *IEEE Expert*, vol. 81, pp. 41–47. 1993.

[3] Lee, M. H., "Knowledge Based Factory," *Artificial Intelligence in Engineering*, vol. 8, pp.109–125, 1993.

[4] Fayyad, U., and Stolorz. P., "Data Mining and KDD: Promise and Challenges," *Future Generation Computer System*,vol.13, pp. 99-115. 1997.

[5] Gelbard, R., Goldman, O., and Spiegler, I., "Investigating Diversity of Clustering Methods: An empirical Comparison," *Data & Knowledge Engineering.*,vol. 63, pp. 155-166, 2007.

[6] Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris. C., "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Databases Environment." In: *Proc. 7th ACM SIGSDD International conference on Knowledge Discovery and Data Mining*, 2001, pp.263-268.

[7] Fishburn P.C. Method for estimating additive utilities, Management Science, vol.13-17, 1997, pp.435-453