World Academy of Science, Engineering and Technology
International Journal of Industrial and Manufacturing Engineering
Vol:3, No:12, 2009

Modified Data Mining Approach for Defective Diagnosis in Hard Disk Drive Industry

S. Soommat, S. Patamatamkul, T. Prempridi, M. Sritulyachot, P. Ineure and S. Yimman

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

***Abstract*—** Currently, slider process of Hard Disk Drive Industry become more complex, defective diagnosis for yield improvement becomes more complicated and time-consumed. Manufacturing data analysis with data mining approach is widely used for solving that problem. The existing mining approach from combining of the K-Mean clustering, the machine oriented Kruskal-Wallis test and the multivariate chart were applied for defective diagnosis but it is still be a semiautomatic diagnosis system. This article aims to modify an algorithm to support an automatic decision for the existing approach. Based on the research framework, the new approach can do an automatic diagnosis and help engineer to find out the defective factors faster than the existing approach about 50%.

***Keywords*—**Slider process, Defective diagnosis and Data mining.

I. Introduction

THIS article aims to modify the existing data mining approach from semiautomatic decision system in slider process of Hard Disk Drive Industry (HDDI) to be the fully automatic decision system for the yield improvement. Currently, the complexity in slider process and many manufacturing factors such as process stages, machine types, material types and methodology types are main factors that relate to the occurrence of the defective from final testing. Data mining approach are presently used to solve the problem for identifying the root causes of those defects based on manufacturing data. There are many related works to ensure that data mining can provide the best result for solving the problem in current manufacturing [1]. These works are described as follow.

C. Chen-Fu, W. Wen-Chih and C. Jen-Chieh (2007) [2] applied three data mining algorithms i.e. the K-Mean clustering, the Kruskal-Wallis (K-W) test and decision trees in semiconductor industry to identify the defective process stages and defective machines. There were three steps of working. Firstly, the final test yield was classified into two groups i.e. low and high group by using the K-Mean clustering. Secondly, the defective process stages were screened by using the machine oriented K-W test for statistical testing and the result came out as a P-Value for decision. Thirdly, the defective machine classification was undertaken by using decision trees. The defective process stage showed up when the P-Value of that process stage was less than specific alpha (Type I error). Once the defective process stage was defined, it was continued to classify defective machines by the decision trees. Each leave of the decision trees showed an average yield of each machine and classified which machines were defective.

S. Soommat, S. Patamatamkul, M. Sritulyachot, P. Ineure and S. Yimman (2008) [3] applied three data mining algorithms i.e. the K-Mean clustering, the machine oriented K-W test and the multivariate chart in slider process of HDDI to identify defective process stages, defective machines, defective materials and defective methodologies that impacted to final test yield. The procedures of analysis on the first and second step were the same as in [2] but in the third step, the decision trees was replaced by the multivariate chart to cover more defective factors i.e. machines, materials and methodologies [4], it has been proposed to improve the performance of [2].

C. We-Chou, T. Shian-Shyong, and W. Ching-Yao (2005) [5] applied an apriori association rule and continuity-based measurement function to capture defective machines that continually produced poor product quality. The same series of products work through the different machines along with the process flow. The pattern of poor machine showed up if it continuously performed with high rejected rate of products when compare to other machines. The decision system was an automatic system.

II. Recent Problem and Proposal

A. Recent Problem

Fig. 1 shows the defective diagnosis system of the existing approach in HDDI [3]. All interested manufacturing data were linked in the same database. The manual step was observed during the K-W decision and the multivariate chart interpretation.

S. Soommat is a student of Engineering Management Department, Faculty of Engineering, Vongchavalitkul University, Thailand (e-mail: ssoommat@yahoo.com).

S. Patamatamkul (PhD.) is an Assoc. Prof. of Engineering Management Department, Faculty of Engineering, Vongchavalitkul University, Thailand (e-mail: sanpat@kku.ac.th).

T. Prempridi is a Prof. of Engineering Management Department, Faculty of Engineering, Vongchavalitkul University, Thailand.

M. Sritulyachot (PhD.) is a lecturer of Engineering Management Department, Faculty of Engineering, Vongchavalitkul University, Thailand (e-mail: srimanop@yahoo.com.com).

P. Ineure (PhD.) is a Senior Engineering Manager of Six Sigma, Seagate Technology, Thailand (e-mail: Pijarn Ineure@Seagate.com).

S. Yimman, (D. Eng.) is an Assoc. Prof. of Industrial Physics & Medical Instrument Department, Faculty of Applied Science, King Mongkut's University of Technology, Thailand. (e-mail: sym@kmutnb.ac.th).

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

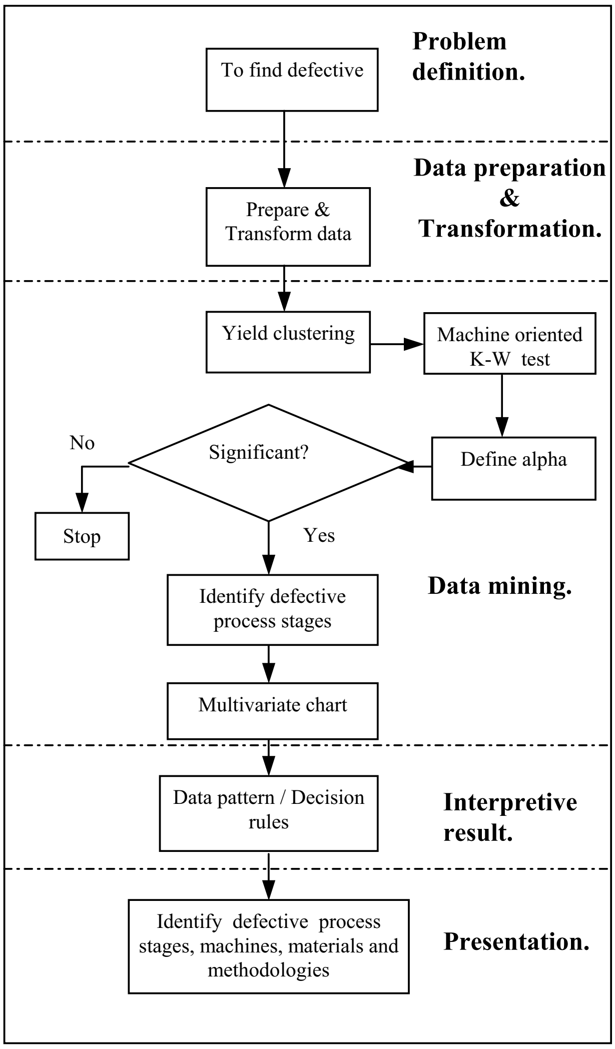
Problem definition.
To find defective
Data preparation & Transformation.
Prepare & Transform data
Yield clustering
Machine oriented K-W test
No
Significant?
Define alpha
Stop
Yes
Identify defective process stages
Data mining.
Multivariate chart
Data pattern / Decision rules
Interpretive result.
Identify defective process stages, machines, materials and methodologies
Presentation.

Fig. 1 Procedure of the existing data mining approach

B. Proposal

The proposed data mining approach for automatic decision on the K-W decision and the multivariate chart interpretation is shown in Fig. 2. The additional steps are as follow. The first step is combining the machine, material and methodology oriented approach during the K-W test to ensure that system can automatic covering all defective factors i.e. machines, materials and methodologies. The second step is providing the automatic decision on the multivariate chart by using the continuity-base measurement and the apriori association rules [5].

III. Theory

The proposed data mining approach includes the following steps: Yield clustering, define alpha (Type I error), the K-W test, identify defective process stages, the multivariate chart, the continuity-base measurement and the apriori association rules. The theory and functions of each algorithms are described below.

1) The yield clustering. It was classified by the K-Mean clustering [6]. This K-Mean clusetring is an unsupervised data mining algorithm for classification data into K groups. Many applications used this technique for instant grouping yield into two groups e.g. low and high for yield analysis [2], [3]. This research also uses it to classify the final test yield into two groups i.e. low yield and high yield group.

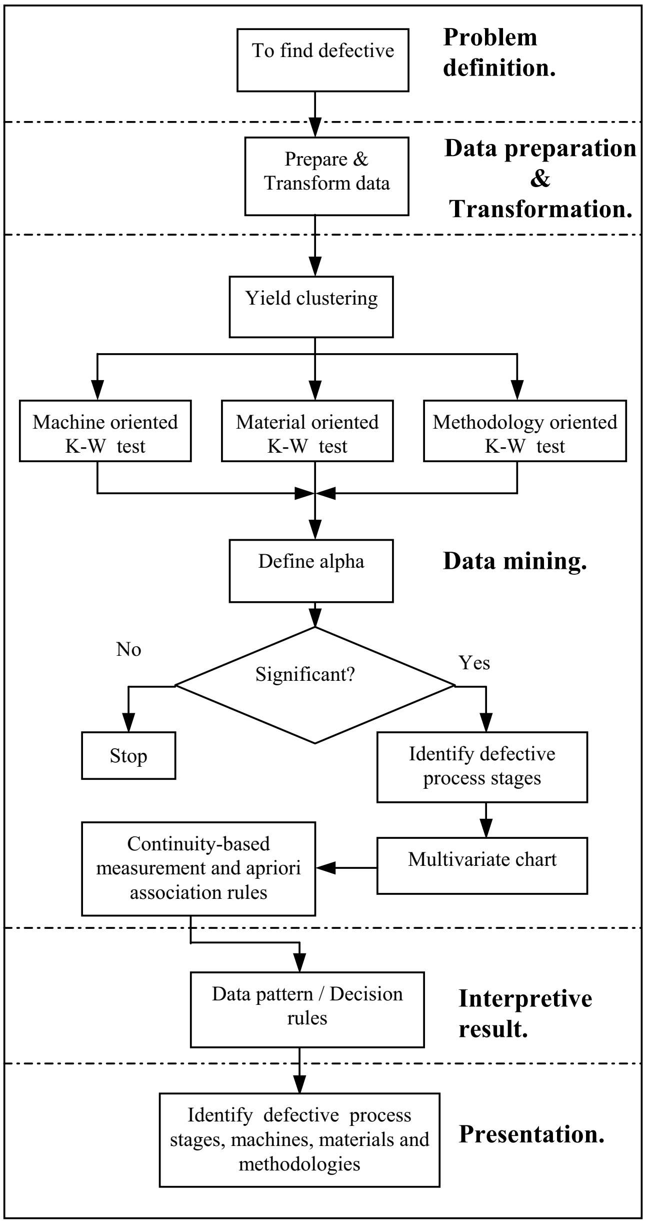
To find defective
Problem definition.
Prepare & Transform data
Data preparation & Transformation.
Yield clustering
Machine oriented K-W test
Material oriented K-W test
Methodology oriented K-W test
Define alpha
Data mining.
No
Significant?
Yes
Stop
Identify defective process stages
Continuity-based measurement and apriori association rules
Multivariate chart
Data pattern / Decision rules
Interpretive result.
Identify defective process stages, machines, materials and methodologies
Presentation.

Fig. 2 Procedure of the proposed data mining approach

2) Define alpha [2], [3]. The alpha is a type I error condition for statistical testing in order to define the risk level for problem detection.

3) The K-W test [7] is a non parametric statistical for null hypothesis testing of equality treatment mean for all continuous and independent variables. A decision of

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

null hypothesis testing uses the P-Value to identify a level of the significant difference that leads to reject of the null hypothesis with the given data. This research provides three combinations of testing approach i.e. machine, material and methodology oriented testing to replace the only single machine oriented in the existing data mining approach. This is to prevent an over look of defective diagnosis (i.e. material and methodology) in case of the single machine oriented testing shows no significant difference. The concept of machine, material and methodology oriented approach for statistical testing with K-W is shown in Table I, Table II and Table III, respectively.

TABLE I.
MACHINE ORIENTED APPROACH

Machines	Yield by time series				
	$Yield_1$	$Yield_2$	.	.	$Yield_m$
1	11	21	.	.	$m1$
2	12	22	.	.	$m2$
.	.	.	.	.	.
.	.	.	.	.	.
n	$1n$	$2n$	.	.	mn

The null hypothesis for the equality treatment mean testing of each machine numbers shows in (1), where μ_1, μ_2,……, μ_n are the mean or average yield of machine #1, #2,…...,#n, respectively. The mean of each machine is averaged from the $Yield_1$ to $Yield_m$ value. The $Yield_1$ value is the yield monitoring at time sequence 1, the $Yield_2$ value is the sequence 2 and the $Yield_m$ value is the sequence m, respectively. This concept is applied in the same manner for materials and methodology oriented approach.

$$H_0 = \mu_1 = \mu_2 = ... = \mu_n \tag{1}$$

H_1 is different mean at least one machine number.

TABLE II
MATERIAL ORIENTED APPROACH

Materials	Yield by time series				
	$Yield_1$	$Yield_2$	.	.	$Yield_m$
A	$1A$	$2A$	.	.	mA
B	$1B$	$2B$	.	.	mB
.	.	.	.	.	.
.	.	.	.	.	.
Z	$1Z$	$2Z$	.	.	mZ

The null hypothesis for the equality treatment mean testing for each material type shows in (2), where μ_A, μ_B,……, μ_Z are the average yield of material #A, #B,……,#Z, respectively. The mean of each material type is averaged from the $Yield_1$ to $Yield_m$ value.

$$H_0 = \mu_A = \mu_B = ... = \mu_Z \tag{2}$$

H_1 is different mean at least one material type.

TABLE III
METHODOLOGY ORIENTED APPROACH

Methods	Yield by time series				
	$Yield_1$	$Yield_2$	.	.	$Yield_m$
1	11	21	.	.	$m1$
2	12	22	.	.	$m2$
.	.	.	.	.	.
.	.	.	.	.	.
r	$1r$	$2r$	.	.	mr

The null hypothesis for the equality treatment mean testing of each methodology type shows in (3), where μ_1, μ_2,……, μ_r are the average yield of methodology #1, #2,……,#r, respectively. The mean of each methodology type is averaged from $Yield_1$ to $Yield_m$ value.

$$H_0 = \mu_1 = \mu_2 = ... = \mu_r \tag{3}$$

H_1 is different mean at least one methodology type.

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

4) The multivariate chart [8] is the graphical and data vector generator tool that separates the mean of dependent variables (i.e. yield) to view the mean of the data in each category (i.e. methodologies, materials and machines) to see how they change across among them.

5) The continuity-based measurement [5] is used to verify the continual function of suspected root cause once the pattern from the multivariate chart is observed. In general, manufacturing process characteristics, the root cause observations (i.e. machine sets, material sets and methodology sets) often produce defective products continuously. High continual function indicates high probability of being the root cause. The continuity–based measurement defines as $\phi' = \phi$ * Continuity and the continuity equation shows in (4).

$$Continuity = \frac{1}{\sum_{i=1}^{|X|-1} d(\alpha(x_i)), \alpha(x_{i+1}))/|X|-1} \quad if\ |X|>1 \qquad (4)$$

$$Continuity = 0 \qquad if\ |X| \leq 1$$

Where:

$X\ (x_1, x_2, ...x_m)$ is time sequence 1 to m.

$\alpha\ (x_i)$ is order of time sequence.

$d(\alpha(x_i))$ and $d(\alpha(x_{i+1}))$ are distance of $\alpha(x_i)$ and $\alpha(x_{i+1})$ which can easily be calculated by $\alpha(x_{i+1})$ - $\alpha(x_i)$

6) The apriori association rule [9], [10] is an expression $A \Rightarrow B$, where A and B are sets of items. If given a database D of transactions (Trans.) where each transaction $T \in D$ is a set of items $A \Rightarrow B$ is that whenever a transaction T contains A then T also contains B. The strong rule is the % support and % confidence are met the minimum threshold from users. The % support and % confidence are shown as (5) and (6).

$$\%\ Support(AB) = \frac{Number\ of\ Trans.\ contains\ (AB)}{Total\ number\ of\ Trans.} \qquad (5)$$

$$\%\ Confident(AB) = \frac{Support\ \ (AB)}{Support\ \ (A)} \qquad (6)$$

In order to confide the discovered rule of $A \Rightarrow B$, a domain-independent interestingness measurement is proposed by [4] as shown in (7), where N is total number of tuples, $|A|$ is number of tuples that contains the antecedent A, $|B|$ is number of tuples that contains the antecedent B and $|A \& B|$ is number of tuples that contains both A and B. Then, the domain-independent interestingness is

$$\phi = \frac{|A \& B|1 - |A||B|/N}{\sqrt{|A||B|(1-|A|/N)(1-|B|/N)}} \qquad (7)$$

IV. Results

Following the proposed framework of data mining, it conducted an empirical study in a slider process of a HDDI in Thailand. There was a critical need to find the root causes of low yield problem to reduce the manufacturing cost. The diagnosis framework were performed in five major steps: Problem definition, Data preparation & Transformation, Data mining, Interpretative result and Presentation [9], [10].

1) Problem definition was defined as interested defective factors i.e. process stages, machines, materials and methodologies in slider process of HDDI that impacted to slider yield.

2) Data preparation & Transformation was undertaken by linking all interesting attributors i.e. process stage names, machine numbers, material types and methodology types from each process stage in the same database as shown in Fig. 1.

3) Data mining.

3.1 Yield clustering. The slider yield was clustered into two groups i.e. low and high by using the K-Mean clustering. There were totally of 9964 slider lots that manufactured in two weeks, the K-Mean clustering classified the yield into "High group" with an average yield at 95.57% (5520 lots) and "Low group" with an average yield at 86.06% (4444 lots). The yield cutting point that separated those two yield groups was located at 90%.

3.2 Defining the alpha. It was maintained the same as [1], [2] at 0.03.

3.3 The K-W test performed a statistical test with the automatic K-W testing under a combination of machine, material and methodology oriented. The P-Value was verified under Chi-Square distribution [6] and the result from each process stage showed in Table IV.

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

TABLE IV
P-VALUE OF K-W TEST

Order	P-Value (Ordered from low to high)		
	Process Stages	P-Value	Impact
1	Stage#4	0.001	Yes
2	Stage#3	0.05	No
3	Stage#5	0.50	No
4	Stage#8	0.65	No
5	Stage#7	0.75	No
6	Stage#1	0.80	No
7	Stage#2	0.95	No
8	Stage#6	0.95	No

3.4 Identifying the defective process stages. The defective process stage showed up on stage# 4 because of the P-Value of this process stage was shown less than the specific alpha (0.03).

3.5 The multivariate chart generated a graph and data vector. The graph showed the average yield from each lot at y-axis (dependent variables) across each category at x-axis (i.e. methodology, material types and machine numbers). As the result, it discriminated the defective machines from other factors based on visualization. The data in Table V showed the data vector that generated from the graph. Since the presentation space is limited, it showed only the low yield portion after comparing to K-Mean clustering result and some of high yield portion as examples. This data vector provided the information to the continuity function analysis and continuity-base measurement in the next step of diagnosis.

3.6 The continuity-based measurement verified a continual level of suspected defective factors (i.e. machine #1, #4, #5, #6 and #15) from the data vector. This research used manufacturing date to be the key index for continual testing. If the suspected factors shows the continual function (i.e. continue low performing) with the continuity – based measurement (ϕ') > 80%. This meant that suspected factors is high potential to be a cause of defect. The result of continual function of machine #1, #4, #5, #6 and #15 is shown in Table VI and the continuity-based measurement (ϕ') result is shown in Table VII. As the result, the continuity-based measurement of those machines showed the value to be 1. This meant that all those suspected factors provided a very high confident to be the real root cause of defect.

TABLE V
DATA VECTOR

Trans.	Machines	Methods	Materials	Yield	Cluster
1	1	1	A	82.0	Low
2	1	1	B	82.3	Low
3	1	1	C	82.0	Low
4	1	2	A	83.7	Low
5	1	2	B	84.0	Low
6	1	2	C	84.0	Low
7	4	1	A	89.2	Low
8	4	1	B	88.3	Low
9	4	1	C	88.9	Low
10	4	2	A	88.9	Low
11	4	2	B	89.5	Low
12	4	2	C	89.8	Low
13	5	1	A	89.2	Low
14	5	1	B	89.2	Low
15	5	1	C	89.7	Low
16	5	2	A	89.9	Low
17	5	2	B	89.8	Low
18	5	2	C	89.4	Low
19	6	1	A	85.1	Low
20	6	1	B	85.4	Low
21	6	1	C	85.4	Low
22	6	2	A	86.0	Low
23	6	2	B	85.4	Low
24	6	2	C	85.1	Low
25	15	1	A	84.3	Low
26	15	1	B	83.7	Low
27	15	1	C	82.9	Low
28	15	2	A	82.9	Low
29	15	2	B	84.3	Low
30	15	2	C	83.1	Low
31	3	1	A	95.1	High
32	3	1	B	96.3	High
33	3	12	C	94.2	High
34	3	2	A	95.7	High
35	3	2	B	96.1	High
36	3	2	C	95.3	High

Due to the limit space, only the low yield group and some of high yield group are listed as examples

TABLE VI
CONTINUITY FUNCTION ANALYSIS

Machines	Continuity Analysis		
	Manufacturing date	*Observed low yield date*	*Continuity*
1	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	1
4	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	1
5	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	1
6	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	1
15	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	Date 1,2,3,4,5,6,7,8,9, 10,11,12,13,14	1

Open Science Index, Industrial and Manufacturing Engineering Vol:3, No:12, 2009 publications.waset.org/4174.pdf

TABLE VII
CONTINUITY –BASED INTERESTINGNESS ANALYSIS

Machines	Continuity-Based Measurement		
	ϕ	*Continuity*	ϕ'
1	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
15	1	1	1

3.7 The apriori association rules generated a set of A (defective machine sets) that conducted B (Low yield group). This research set the % minimum support and % minimum confident at 20% and 80%, respectively. The result of %supporting and %confident for 1-itemset (machines) is shown in Table VIII whereas 2-itemsets (machines, methods) and 3-itemsets (machines, methods, materials) were ignored because of mismatching in the minimum supporting condition for this case study.

TABLE VIII
% SUPPORTING AND CONFIDENT OF 1-ITEMSET

Machines	%Support and% Confident		
	Number of Support	*%Support*	*%Confident*
1	6	20	100
4	6	20	100
5	6	20	100
6	6	20	100
15	6	20	100

4) Interpretive result. The result was automatic reported when the system was turn on.

5) Presentation. Report of defective factors i.e. process stages and machines numbers were presented to responsible team for proving a corrective action.

V. CONCLUSION

The proposed fully automatic data mining approach for defective diagnosis can deliver the same result as the existing semiautomatic approach when applying with the same case study. The lists of defective machines were #1, #4, #5, #6 and #15. The approximation of delivery time was an hour for the proposed fully automatic approach whereas the existing semiautomatic approach delivered in 2 hours or more due to manual data extraction and decision. So the time reduction to get the conclusion reduced by 50%. The key work for the proposed data mining approach required various cases of problem for training the system, especially the continuity-base measurement and the apriori association for the rule generation in order to support more various cases coming in the system.

ACKNOWLEDGMENT

The authors would like to thank Seagate Technology (Thailand) for supporting this study i.e. Basic knowledge of a HDD manufacturing process and analysis software.

REFERENCES

[1] A.Harding, M. Shahbaz, Srinivas and A. Kusiak, “Data mining in manufacturing: A review,” *J. Manufacturing Science and Engineering*, vol. 128, Nov. 2006, pp.969-976.

[2] C. Chen-Fu, W. Wen-Chih and C. Jen-Chieh, “Data mining for yield enhancement in semiconductor manufacturing and an empirical study,” *ELSEVIER, J. Expert system with application*, vol. 33, 2007, pp. 192-198.

[3] S. Soommat, S. Patamatamkul, M. Sritulyachot, P..Ineure and S. Yimman, “Applying data mining approach for slider yiled diagnosis in HDD manufacturing,” presented at the IQC2008 Inter Conference Bangkok, Thailand , November 26-28, 2008., Paper C10.

[4] S. Soommat, S. Patamatamkul, M. Sritulyachot,P. .Ineure and S. Yimman, “ Defective diagnosis using deciction trees and multivariate chart: A case study in hard disk dive industry,” *in Proc. 14st National Grauate Conf*., KMUTNB, Bangkok, Thaiand , 2009, pp. 25-35.

[5] C. Wei-Chou, T. Shian-Shyong and W. Ching-Yao, “A novel manufacturing defect detection method using association rule mining techniques,” *ELSEVIER J. Expert system with application*, vol. 29, 2005, pp. 807-815.

[6] J. Mecqueen, “Some methods for classification and analysis of multivarate observations,” *in proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp.281-297.

[7] D.C.Montgomery and G C. Runger, *Applied Statistcs and Probability for Engineers.* 2nd ed., USA., New York: Addison-Wesley, 1999, ch.9 and ch.14.

[8] SAS Institute Inc, *JMP Statistics and Graphics guide*. 5th ed., USA: SAS Institue Inc., 2002, ch.37.

[9] I.H.Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed., Morgan Kaufmann Publis.USA, San Fan : Elsevier Inc., 2005, ch.3-ch.5.

[10] R.J. Roiger and M. W. Geatz. *Data Mining: A Tutorial-Based Primer*. Int. ed., USA., New York : Addison-Wesley, 2003., ch.1-ch.13.