Improving Quality of Products in Hard Drive Manufacturing by Decision Tree Technique

Anotai Siltepavet[1], Sukree Sinthupinyo[2] and Prabhas Chongstitvatana[3]

[1] Computer Engineering, Chulalongkorn University,
Bangkok, Thailand 10330
Anotai.Siltepavet@seagate.com

[2] Computer Engineering, Chulalongkorn University,
Bangkok, Thailand 10330
Sukree.S@chula.ac.th

[3] Computer Engineering, Chulalongkorn University,
Bangkok, Thailand 10330
Prabhas.C@chula.ac.th

Abstract

Hard drives manufacturing is a complex process. The quality of products is determined by a large set of parameters. However, there are defects in products that need to be removed. This work studied a systematic approach to find suitable parameters which can reduce the number of defected products. Our approach exploits decision tree learning and a set of algorithms to adjust decision parameters obtained from the decision tree. Moreover, because we cannot test the result in the real environment, we propose a trustable testing method which can predict the improvement obtained from the parameter adjustment system. The results from the experiments show that the quality of products in the dataset can be improved as much as 12%. This is significant in hard drive manufacturing.

Keywords: *Decision Tree, Data Mining, Hard drive Manufacturing, Improve Quality*

1. Introduction

Thailand's main electronics exports are hard disk drives (HDD) [1]. The manufacturing sector which produces good quality hard drive to market has been growing. The competition to produce ever better quality hard drives is intense. The production process of the hard drive consists of several stages, each of which is further split into sub-processes. The final product quality is subject to several controllable parameters used in the manufacturing. Normally, to measure the quality of the process control, we use *yield* value, number of good pieces per total pieces of production. In this study, we propose a new method which can increase the yield by properly adjusting the controllable parameters in the production. The results obtained from our approach will produce greater number of good pieces of produced works. The work focuses on the certain processes associated with the electrical measurements in manufacturing the hard drive. Parameters are classified into three categories, i.e. uncontrollable parameters, controllable parameters, and dependent parameters. Our work aimed to improve the yield and to reduce wastes by constructing a decision tree to find a set of parameters which affect the quality of work. However, we cannot test the results from the decision tree in the real manufacturing process. Hence, we also propose a testing method which can help predict the improvement of the parameter adjustment over the original setting. The main contribution of our approach is a new method which can find a set of controllable parameters which can improve the yield value in the production and a new testing method which can predict the improvement rate. Moreover, since adjusting some parameters may have an effect on other parameters, so that we also employ the standard linear regression model to refine the value of the dependent parameters which can improve the confidence of the results.

The rest of this paper is organized as follows. The related works is presented in Section 2. The method is explained in Section 3. The experiments and results are described in Section 4. The conclusion is summarized in Section 5.

2. Related Works

In this part, we review all theories and works which relate to our approach. In this study, we use the Decision Tree Learning as a classifier to categorize a product into two classes, *Fail* and *Pass*. The reason that we selected the Decision Tree in this study is the results obtained from this learning algorithm are in the form of a tree which is comprehensible to human. Hence, we can directly use the results to adjust the parameters in the production process.

2.1 Decision Tree

The Decision Tree Learning (DTL) is one of the most widely used and practical methods for inductive inference. The traditional DTL is a method for approximating discrete-valued target functions, in which the learned hypothesis is characterized in the form of a decision tree [2]. The internal nodes of the tree represent the attributes and the edge corresponding to each internal node denotes the decision made by the value of the attribute in the node. The goal is to construct a tree which is capable of correctly classifying data into different classes. A dominant feature of the predictive model is in the form of rules like IF-THEN, which is highly understandable. Moreover, because DTL employs greedy search based on the Entropy or Information Theory, learning time of DTL is very fast compared to other classification algorithms [3]. Therefore, it is suitable for using in an analysis of production data in manufacturing in which the amount of data is usually very large.

2.2 Weka

We use a well-known machine learning tool, Weka. It consists of a collection of machine learning algorithms for data mining tasks. Weka contains several tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [4]. The algorithms can either be applied directly to a dataset or called from user own Java code. Weka was developed by the University of Waikato in New Zealand since 1997.

2.3 Related Works

There are many algorithms falling within the decision of the good property of a dataset such as ID3[5], C4.5[6], CART[7] and CHAID[8]. There is a related research using C4.5 in order to find the fail pattern of HGA Manufacturing [9] and to discover the source of the yield degradation by testing it with 20 attributes and 1000 records. Such research was conducted to compare C4.5 with other algorithms. The results showed that C4.5 performed better in dealing with continuous values and attributes with missing values than other algorithms. Furthermore, C4.5 can avoid overfitting issue that had been found when using large amounts of data such as manufacturing data. To apply CART algorithm, the leaf nodes must be continuous values, not discrete values. In such case, C4.5 algorithm is not useable. Finally, CHAID algorithm cannot cope with large dataset with many branching factors. Therefore it is not useful with the factory data. In addition, the work in [10] presented a model to detect the waste in the production of semiconductor and TFT-LCD with FDS (Fault Detection Scheme) in real time. The algorithm CART helped to create the structure of the model to determine the relationship between the parameters of the processes and product with in-spec process. The model adopted the concept of the minimum-cost for pruning the relatively insignificant rules in a tree model to avoid model overfitting. Another research relates to the use of data mining to solve problem in hard drive manufacturing processes was proposed in [11]. It presented analysis tools for improving the productivity of the hard drive manufacturing. It was found that a set of critical parameters and attributes might affect the yield. Using a decision tree algorithm and improved yield by shifting the mean of a certain parameters may produce a higher yield rate. Our study employs C4.5 to identify important parameters that can improve the quality of the product and the parameters obtained from the process will be adjusted to improve yield.

3. Method

The purpose of this work is to find the controllable parameters that affect the quality and to increase productivity of good pieces. First, we divide all parameters into three groups: Controllable, Uncontrollable, and Dependent Parameters. The first two groups are mutual exclusive, while the Dependent Parameters are the ones that depend on the Controllable Parameters. We will focus on adjusting only the parameter in the Controllable Parameters group.

3.1 Parameter Adjustment

The workflow of our method is shown in Fig.1. In order to increase the number of good products in the process, the controllable parameters are adjusted to reduce the number of failed product and gain the number of the passed product. However, adjusting some parameter in the process may affect other parameters. Hence, we use the linear regression method to find the relations between all pairs of parameters and adjust the dependent parameters according to the controllable parameters.

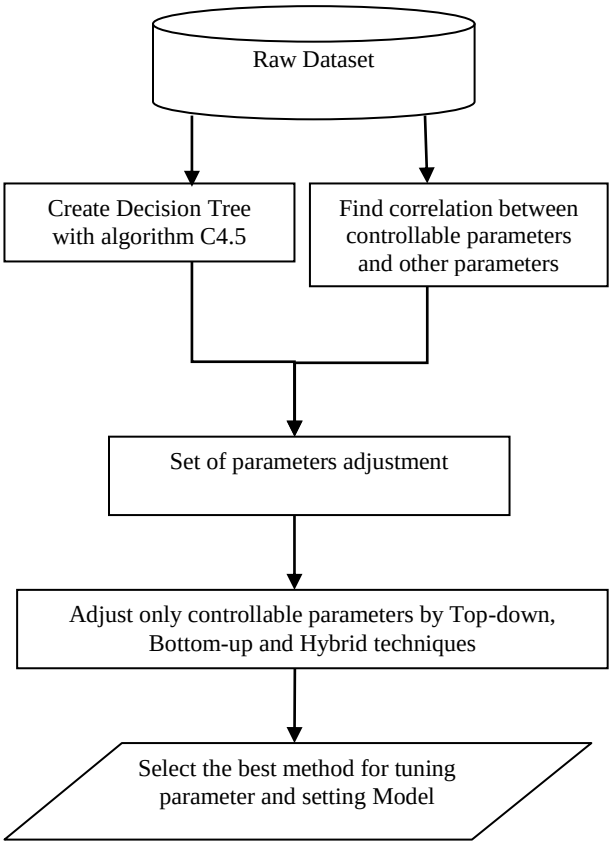

Fig. 1 The proposed method for improving yield in the production of hard drive

In the experiment, because we cannot test the results in the real production line, we divide the whole data set into two parts: training set and test set. In the training set, we randomly select 20% from the whole set as the validation set which will be used to check when to stop the adjustment process. The portion of all data is shown in Fig. 2

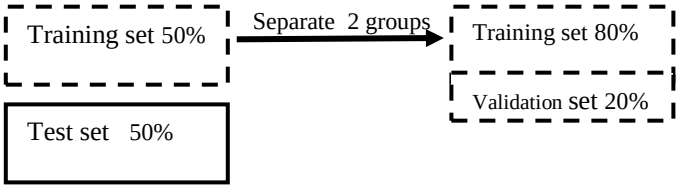

Fig. 2 The portion of raw dataset.

From this point, 80% of the training set is used as training set of the DTL. After the training session, we have a decision tree which shows us parameter that can be adjusted. From the decision tree, we select the parameters to be adjusted in both directions, top-down and bottom-up. All of them are described in Fig. 3, 4, and 5, respectively.

The method we use to adjust the dependent variable is described in Fig. 5 (FIND-ALL)

```
TOP-DOWN[T,V,T']
Input: T, a decision tree
       V, a validation set
       T', a decision tree constructed from the
           validation set
Output: L, a list of pairs of parameter and its value
Initialize: i ← 0
1   WHILE(Accuracy is increasing)
2       N_i ← extract all nodes in level i^th of T
3       FOR EACH N_i
4           C_j ← all children of N_i
4           IF N_i is controllable parameter
5               IF C_j is PASS or FAIL
6                   Add a pair of parameter and its
                    value in the parent of C_j to L.
7       Accuracy ← EVALUATE[V,T',L]
8       i ← i+1
9   RETURN L
```

Fig. 3 TOP-DOWN Algorithm.

```
BOTTOM-UP[T,V,T']
Input: T, a decision tree
       V, a validation set
       T', a decision tree constructed from the
           validation set
Output: L, a list of pairs of parameter and its value
Initialize: i ← depth of T
1   WHILE(Accuracy is increasing)
2       N_i ← extract all nodes in level i^th of T
3       FOR EACH N_i
4           P_j ← all parent nodes of N_i
4           IF P_i is controllable parameter
5               IF P_j is PASS or FAIL
6                   Add a pair of parameter and its
                    value in P_j to L.
7       Accuracy ← EVALUATE[V,T',L]
8       i ← i – 1
9   RETURN L
```

Fig. 4 BOTTOM-UP Algorithm.

FIND-ALL[*T,V,T'*]
Input: *T*, a decision tree
V, a validation set
T', a decision tree constructed from the validation set
Output: *L*, a list of pairs of parameter and its value

1 $L_1 \leftarrow$ TOP-DOWN[*T,V,T'*]
2 $L_2 \leftarrow$ BOTTOM-UP[*T,V,T'*]
3 RETURN $L_1 \cup L_2$

Fig. 5 FIND-ALL Algorithm.

EVALUATE[*S,T,L*]
Input: *S*, sample set
T, a decision tree
L, a list of adjustable parameter
Output: Accuracy of *S* evaluated on *T*

1 *S'* ←Adjust the value in *S* using all dependent parameters from *L* and linear regression
2 *Accuracy* ← Evaluate the accuracy of *S* ' on *T*
3 RETURN *Accuracy*

Fig. 6 EVALUATE Algorithm.

The idea is to adjust only the controllable parameters and calculate other parameters which depend on such controllable parameters. We collect all adjustment that can increase the number of the good. Hence, we will collect all adjustment of each attribute that can change a class of the data from *Fail* to *Pass*. To find all adjustment on parameters, we traverse in the tree from both directions, top-down and bottom-up and test for the improvement if the parameter corresponding to that node can improve the yield.

3.2 Stopping Criterion

As shown in Fig. 3, 4, and 5, the algorithm tries to collect from both directions all adjustment that can improve the yield of overall production. However, another problem may arise when we adjust too many parameters. The result may be unpredictable when using in the real environment. So we decide to test it using the idea of general pruning algorithm, in which the algorithm will stop when there is no improvement on the validation set.

We use the validation set to test when we should stop. Hence, the validation set is required for testing the gained quality of each adjustment. The validation set is used as a training set for building another decision tree, namely *Validation Tree.*

After we obtain the tree from the original training set, the algorithm adjusts each parameter. In each node which corresponds to the controllable parameter, we change the value of that parameter to move all *Fail* pieces to *Pass* pieces. However, this change may affect other dependent parameters, so we use linear regression method to calculate the new value of all the dependent parameters. Then, we will have a new set of parameter adjustment. We apply all adjustment to all examples in the validation set and then use them as the example set to test with the *Validation Tree* to find the number of *Pass* and *Fail* pieces. We stop when the number of the *Pass* piece does not increase any further.

3.3 Testing Approach

Due to the problem that we cannot test our parameter adjustment in the real production process, we create a novel testing method to find the effectiveness of our adjustment algorithm.

From Fig.2, the original data set in our experiments were divided into two main portions, training set and test set. Albeit, the training set was further divided into two groups, training set for learning Decision Tree and validation set for stopping adjusting parameter. The test set, in our experiment, was used to construct a simulated factory, the characteristic of which was identified by another decision tree, *Test Tree.* This *Test Tree* was constructed by the original test set. Then the original test set was adjusted using the adjusted parameter set from the training data. Finally, the adjusted test set was applied to the *Test Tree* to measure the effectiveness of the parameter adjustment.

4. Experiments

The data of a product was collected from the production line in one day. There were 64,887 pieces in total. The data consisted of 12 separated parameters as shown in Table 1. The data were split in two halves as described in the previous section. The first group as a training set (A) consisted of 32,443 pieces. The other 32,444 pieces (B) were the test set.

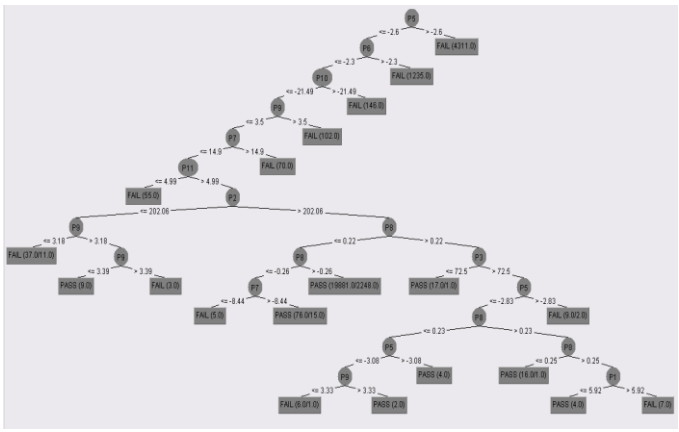

Fig. 7 An example of a decision tree

Table 1: Examples of Hard drive parameters

Name	**Detail**
P1,P2,P3,P4,P5,P6	Controllable Parameter
P7	Dependent Parameter on P2,P4
P8	Dependent Parameter on P3
P9,P10,P11	Uncontrollable Parameter

The results are reported according to the following methods: Top-Down, Bottom-Up, and Hybrid. The training set (A) consisting of 32,443 pieces, have the good parts 24,441 pieces and 8,002 bad pieces. The results of decision tree contain a rule set which can classify the examples. An example of classification rules is P6 <= -2.6. The result of classification is either *Pass* or *Fail*. The results of all methods: Top-Down, Bottom-Up, and Hybrid, are reported in Table 2.

A is the Training set consisted of 32,443 pieces. B is the Test set consisted of 32,444 pieces. A-adjusted is the number of *Pass* pieces of Training set after adjustment, similarly for the Test set B.

Table 2: Results of all methods

PASS Quantity	**Test on Training set A**		**Test on Test set B**	
	Group A	***Group A-adjusted***	***Group A***	***Group A-adjusted***
Top-Down	24441	26837	24607	27119
Bottom-Up	24441	25433	24607	25411
Find-All	24441	27589	24607	27544

Finally, Table 2 shows the quantity of good product before and after adjustment of Training set A and Test set B. The result of Find-All method shows that the number of good pieces are increased, from Training set A from 24441 to 27589, or 12.88 %, and from Test set B from 24607 to 27544, or 11.94 %. The average of two sets is 12.4 %.

5. Conclusions

This work proposes a method to improve the yield in hard drive manufacturing process by adjusting controllable parameters that affect the quality of product. A decision tree algorithm (C4.5 in particular) is applied to find important parameters from the product data. This decision tree is used to classify products into *Pass* and *Fail*. By adjusting the values of decisions variables appropriately the yield of the product can be improved. The experimental results show that our method achieves a significant improvement.

References

[1] The Board of Investment of Thailand (BOI), "Annual Report 2010," pp.52-53 , 2010.

[2] Mitchell, T.M., Machine Learning, McGraw-hill,1997.

[3] Kijsirikul, B., "Data Mining Algorithms", the final report on the Joint Government and Private Sectors, Chulalongkorn University 2004 (in Thai).

[4] Hall, M.E., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Vol.11, Issue 1, 2009.

[5] Quinlan, J. R., "Induction of decision trees," Mach. Learn., vol. 1, pp.81-106, 1986

[6] Quinlan, J. R., C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[7] Breiman, L., Friedman, J., Olshen, R., and Stone, C., Classification and Regression Trees. Wadsworth,1984.

[8] Kass, G., "An exploratory technique for investigating large quantities of categorical data. Applied Statistics," Appl.Statist., vol. 29, no. 2, pp. 119-127, 1980.

[9] Taetragool, U. and Achalakul, T., "Applying Decision Tree in Fault Pattern Analysis for HGA ManuFacturing", Int. Conf. on Complex, Intelligent and Software Intensive Systems, 2009, pp.83-89.

[10] Yi-Ting Huang, Fan-Tien Cheng and Min-Hsiung Hung, "Developing a Product Quality Fault Detection Scheme", IEEE Int. Conf. on Robotics and Automation, Japan, May 12-17,2009.

[11] Yamwong, W., Kaotien, J. and Achalakul, T.,"The Sampling-based Sensitivity Analysis Model for Yield Improvement in HDD Manufacturing", Int. Conf. on Complex, Intelligent and Software Intensive Systems, 2009, pp. 1211–1216.