Detection of Machines Anomaly from Log Files in Hard Disk Manufacturing Process

Thanatarn Pattarakavin [a], Prabhas Chongstitvatana [b]

Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University, Bangkok, Thailand 10330

[a] thanatarn.pattarakavin@wdc.com, [b] prabhas@chula.ac.th

Keywords: hard disk manufacturing; head stack assembly; log files; machine learning;

Abstract. Hard disk manufacturing is an important industry in Thailand. The production line of its manufacturing process is highly complex and consists of hundreds of automated machines running a continuous flow production. When an anomaly event occurs the production line has to be stopped and the diagnosis engineering team must identify and locate the source of error amongst those machines and correct them quickly. In an automated production line, all machines are monitored and their log files are sent to a server continuously. Engineers use these log files to diagnose the causes of the errors. This work proposes to use machine learning method to identify the anomaly events in the log files. The experimental results show that it is very accurate and it can help the team of engineers to perform diagnosis quickly and effectively.

Introduction

Hard disk manufacturing (Fig. 1) is an important industry in Thailand. The industry reached 15,000 million US dollars of export value and over 110,000 personnel employed [1]. The production of hard disk is a highly automated process. Over thousand machines worked in a tightly synchronised network producing a continuous flow of outputs [2]. These machines need to communicate and collaborate to keep the production line running smoothly. To monitor the status of these machines in real-time, many reports which are generated automatically from the machines are send to a server as log files. A large number of records flow to the server in real-time. When an anomaly occurs and the production line has to stop, it is critical to locate the cause of disruption and correct it quickly. Team of engineers who support the diagnosis of the production line needs to analyse those large number of log files.

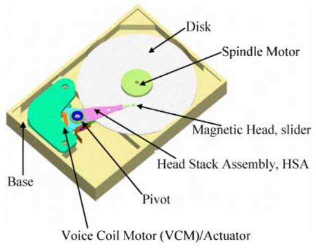


Figure 1: Diagram of a computer hard disk drive [13]

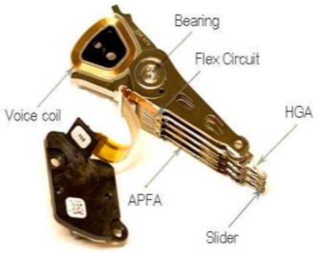


Figure 2: Head Stack Assembly (HSA) [9]

Log files from various machines have different formats [3]. They also contain many unimportant messages. This made analysis of the machine problems complicated and human experts are required. Therefore many errors are caused by the operators of the machines, for example, operators ignored console log messages for a new X-ray machine because error messages were common. Some of the ignored messages involved X-ray power settings. In some situations the X-ray machine delivered full power doses of radiation, killing 4 patients because the critical messages were missed [4].

To reduce the diagnosis time, this work proposes an automated detection of faulty events in the log files. The automated system can monitor the log files in real-time and notify the engineering team when an anomaly occurs [5]. It can identify the location and the cause of anomaly operations. This tool can help the supporting engineers to quickly locate the causes that disrupt the production line and correct it quickly. Machine learning techniques [6,7] are used in this work to learn the pattern of anomaly events in the log files. It also has been used to improve productivity [8]. As log files come from many different types of production machines, a pre-processing is done to extract the important information which is used as input. Four techniques in machine learning are investigated, namely, Support Vector Machine (SVM), Nearest Neighbour (k-NN), Naive Bayes, and Decision Tree. The data from the Head Stack Assembly (HSA) production line (Fig. 2) is used. The log files from 12 October 2015 to 13 January 2016 contain 61 period hours are analysed. The effectiveness of this approach is reported.

Head Stack Assembly Process

The Head Stack Assembly (HSA) [9,10,11] is an arm that traverses inside the hard drive with an induction slider (read/write head) attached to the end. With multiple storage surfaces, the set of heads are used to perform read/write operation simultaneously on all surfaces. The assembly process is highly skill operations using a number of specialised machines to perform the assembly operations. Human operators are manned at each station and operate or inspect the instruments. All machines are connected to a central server that monitors all operations to keep records of the events as part of statistical control of the production line.

Main HSA components are HGA, Slider, APFA, Flex Circuit, Bearing and Voice coil. There are seven steps in HSA assembling. The flow process of the HSA assembling can be explained as following:
Loading -- The APFA assembly of HGA is fixed to the shuttles by 2 operators.
Swaging -- The shooting the ball steel binds the HGA and the APFA together.
Unloading -- Take the HSA out of the shuttles. After that the HSA is fixed into the flow fixture by 2 operators.
Bonding -- The electrical cable is connected between the HGA and the flexible circuit on the APFA by 2 operators.
Tacking -- The epoxy glue has been dropped in order to hold the long tail of HGA and the slot of the APFA by 2 operators.
VMI -- The inspection for quality detects any physical defects of HSA by an operator.
Quasi testing -- The test for electrical performance of the HSA is operated by an operator.

Method

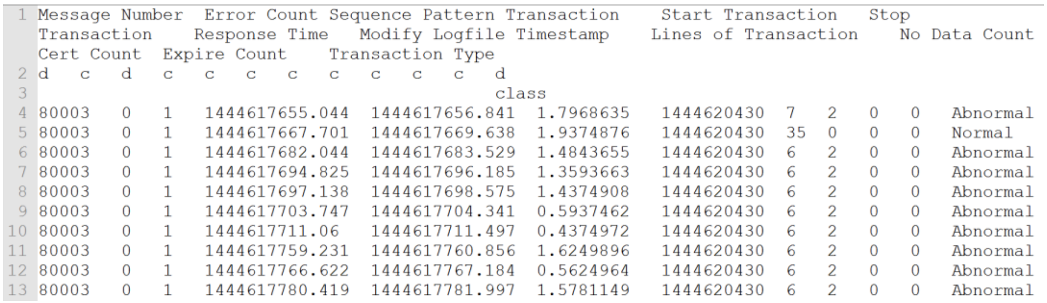
There are many causes for the abnormal behaviour of machines in the production line. Some errors are caused by human, for example, operators enter the wrong machine ID or have the obsoleted authorisation. Many errors are caused by the changes in software configuration on the servers and the controller of production machines, for example, by the update of software and firmware. Most of physical errors such as power outage, mechanical failure of production machines, can be detected easily by the operators. Most critical errors come from the *invisible* software side as the effect will not be immediately perceptible. Therefore, the effort is concentrated on the detection of these software related errors.

In the experiment, the tool, Orange canvas [12], is employed. It is a machine learning and data mining suite for data analysis. A pre-processing program is written to extract critical information from log files. All log files are separated into individual transactions. Each transaction is defined as an event. The machine learning methods learn the pattern of event and classified them into *normal* or *anomaly*. The input to machine learning process consists of 11 variables (Fig. 3). The following paragraph describes those variables.

1) First variable is the message number. It is discrete and represents type of request.
2) Second variable is the count of error string. It represents error frequency.
3) Third variable is the sequence of pattern of each transaction. If it is a generic message the value is set to 1 otherwise it is set to 0.

4) Forth variable is the starting time of each transaction. It represents start time of the event.
5) Fifth variable is the ending time of each transaction. It represents stop time of the event.
6) Sixth variable is the response time of each transaction. It represents performance of machines.
7) Seventh variable is last time of this log file modification.
8) Eighth variable is the length of each transaction. This is the number of lines of transaction.
9) Ninth variable is the count of *no data* keyword. It represents mistake frequency.
10) Tenth variable is the count of *cert* keyword. It represents mistake frequency.
11) Eleventh variable is the count of *expire* keyword. It represents lack of quality control.

The identification of these variables is done by the experts who are responsible for diagnosis of the HSA production line. They investigate the log files looking for these kinds of events. These variables can be used to cover a large class of anomaly (but not all). The study is limited to this class of anomaly as some of the exceptions events are too difficult to obtain as they are rarely occur (and they are not occurring during the period of study).



```
Message Number  Error Count Sequence Pattern Transaction    Start Transaction   Stop
Transaction     Response Time   Modify Logfile Timestamp    Lines of Transaction    No Data Count
Cert Count  Expire Count    Transaction Type
d   c   d   c   c   c   c   c   c   c   c   d
                                                class
80003   0   1   1444617655.044  1444617656.841  1.7968635   1444620430  7   2   0   0   Abnormal
80003   0   1   1444617667.701  1444617669.638  1.9374876   1444620430  35  0   0   0   Normal
80003   0   1   1444617682.044  1444617683.529  1.4843655   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617694.825  1444617696.185  1.3593663   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617697.138  1444617698.575  1.4374908   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617703.747  1444617704.341  0.5937462   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617711.06   1444617711.497  0.4374972   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617759.231  1444617760.856  1.6249896   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617766.622  1444617767.184  0.5624964   1444620430  6   2   0   0   Abnormal
80003   0   1   1444617780.419  1444617781.997  1.5781149   1444620430  6   2   0   0   Abnormal
```

Figure 3: Records in log files

Experiments

The data from the Head Stack Assembly (HSA) production line is used. The log files from 12 October 2015 to 13 January 2016 are collected. They contain 70 files with the total size of 74.7 Mbytes. From all log files, the pre-processor extracts 104,582 records of important information, then computes the 11 variables. This is used as input to the machine learning process. Each record is manually labelled as *normal* or *anomaly* with the knowledge of the actual event occurred in the production line. Of all records, 104,445 records are *normal* and 137 records are *anomaly*. The reason for the small amount of *anomaly* data is that the hard disk manufacturing is a highly developed system with the quality level at the six sigma. That means the process must not produce more than 3.4 defects per million opportunities.

This data is used to train and test the machine learning method using 10-fold validation. All records are divided into 10 groups with mixed of *normal* and *anomaly* events. Nine groups are used for training, one group is withhold and is used for testing. The training and testing is carried out ten times, each time with different testing group. The measurement of accuracy is average over these ten experimental runs. Four machine learning methods are compared: SVM, k-NN, Naive Bayes and Decision Tree.

The results report Classification Accuracy (CA), F1, Precision and Recall. F1 is harmonic mean of precision and recall. Recall and precision are two widely used metrics employed in applications of classification problem related to information retrieval context.

Discussion and Conclusions

Table 1 shows the comparison of the performance of four machine learning methods on the data. There is no significant difference between all methods. Because the data of *normal* and *anomaly* is highly imbalance, the investigation goes further into the measurement of each group. Table 2 shows the performance of learning the *anomaly* data. Two methods, SVM and Decision Tree perform better than k-NN and Naive Bayes. Table 3 shows the *normal* group which has much larger data. It shows that indeed, with that large amount of data, all four methods perform well without significant difference in performance between them.

In conclusion, all four methods can be used for the problem of identifying the anomaly events occur in the production line of HSA. The SVM is the best method from this experiment as its measurement is the best in both *normal* and *anomaly* cases. The accuracy is very high and it can be applied with confidence. The proposed method is shown to be highly accurate and it can help the support engineering team in reducing the time for diagnosis the cause of disruption of production line and correct it quickly.

Table 1 Comparing the performance of four machine learning methods

Overall score from all data				
	CA	F1	Precision	Recall
SVM	1.000	1.000	1.000	1.000
k-NN	0.999	1.000	1.000	1.000
Naïve Bayes	0.990	0.995	1.000	0.990
Decision Tree	1.000	1.000	1.000	1.000

Table 2 Performance of four machine learning methods on *anomaly* data

Score from Anomaly group				
	CA	F1	Precision	Recall
SVM	1.000	1.000	1.000	1.000
k-NN	0.999	0.751	0.890	0.650
Naïve Bayes	0.990	0.195	0.109	0.956
Decision Tree	1.000	0.996	1.000	0.993

Table 3 Performance of four machine learning methods on *normal* data

Score from Normal group				
	CA	F1	Precision	Recall
SVM	1.000	1.000	1.000	1.000
k-NN	0.999	1.000	1.000	1.000
Naïve Bayes	0.990	0.995	1.000	0.990
Decision Tree	1.000	1.000	1.000	1.000

Acknowledgment

This project was sponsored by Western Digital in support of its university education and outreach programs; however the preparation of this report represents the student's own authorship.

References

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

[2] T. Gao, C. Du, W. Sun and L. Xie, "An integrated plant / control design method and application in hard disk drives," *International Journal of Systems Science,* vol. 47, no. 3, pp. 644-651, 2016.

[3] S. G. Eick, M. C. Nelson and J. D. Schmidt, "Graphical Analysis of Computer Log Files," *Communications of the ACM December 1994,* vol. 37, no. 12, pp. 50-56, 1994.

[4] A. Fabio, "Killed by a Machine: The Therac-25," HACKADAY, 26 October 2015. [Online]. Available: http://hackaday.com/2015/10/26/killed-by-a-machine-the-therac-25/. [Accessed 2016 May 31]

[5] F. Schuster, A. Paul and H. Konig, "Towards learning normality for anomaly detection in industrial control networks," in *Emerging Management Mechanisms for the Future Internet,* Cottbus, Germany, International Federation for Information Processing, 2013, pp. 61-72.

[6] T. M. Mitchell, “The Discipline of Machine Learning,” *CMU-ML-06-108,* pp. 1-7, July 2006.

[7] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Addison Wesley Boston, 2006.

[8] A. Siltepravet, S. Sinthupinyo and P. Chongstitvatana, “Improving Quality of Products in Hard Drive Manufacturing by Decision Tree Technique,” *International Journal of Computer Science Issues,* vol. 9, no. 3, pp. 29-34, 2012.

[9] A. Dokmai and A. Kengpol, “The Improvement for Optimization of Head Stack Assembly (HSA) Assembling Process by Using the Virtual Reality 3D Simulation Model,” *AIJSTPME,* vol. 3, no. 4, pp. 47-56, 2010.

[10] W. Kaewka and K. Tangchaichit, “The characterization in the Head Stack Assembly (HSA) During the Swaging Process: Optimization of Actuator Arm Material,” *KKU Res J,* vol. 15, no. 10, pp. 910-918, October 2010.

[11] K. Tangchaichit and W. Kaewka, “A Study of the Head Stacks Assembly (HSA) during the Swaging Process: Optimization of the Ball Velocity.,” *KKU Res J,* vol. 15, no. 2, pp. 104-112, Feburary 2010.

[12] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik and B. Zupan, “Orange: data mining toolbox in Python,” *The Journal of Machine Learning Research,* vol. 14, no. 1, pp. 2349-2353, 2013.

[13] Y. Wang, Q. Maio and M. Pecht, “Health monitoring of hard disk drive based on Mahalanobis distance,” in *2011 Prognostics & System Health Management Conference,* Shenzhen, 2011.

Authors’ background

Your Name	Title*	Research Field	Personal website
Thanatarn Pattarakavin	Master Student	Machine learning	N/A
Prabhas Chongstitvatana	Full Professor	Optimization method	http://www.cp.eng.chula.ac.th/~piak/