Data Balancing and Aggregation Strategy to Predict Yield in Hard Disk Drive Manufacturing

Nittaya Kerdprasop*, Anusara Hirunyawanakul, Paradee Chuaybamroong, and Kittisak Kerdprasop

***Abstract*—Hard disk drive manufacturing is complicated and involves several steps of assembling and testing. Poor yield in one step can result in fail product of the whole lot. Accurate yield prediction is thus important to product monitoring and management. This paper presents a novel idea of data preparation and modeling to predict yield in the process of hard disk drive production. Data balancing technique based on clustering and re-sampling is introduced to make the proportion of the pass and fail products comparable. Then, we propose a strategy to aggregate manufacturing data to be in a reasonable group size and efficient for the subsequent step of yield predictive model creation. Experimental results reveal that grouping data into a constant size of 10,000 records can lead to the more accurate yield prediction as compared to the intuitive idea of weekly grouping.**

***Index Terms*—Data balancing, data aggregation, yield prediction, hard disk drive manufacturing, machine learning**

I. Introduction

Data reliability and cost-efficient are two important factors that make hard disk widely used as the storage device to store big data in most organizations [1]. In the production process of hard disk drive (HDD), many small parts are assembled and being tested several times along the assembly line. The quality control process may take as long as three months per production lot [2, 3]. The HDDs that can pass all testing steps are called the pass units. Those that fail in any of the testing stages are called the fail units. The proportion of pass units to fail units is called yield [4, 5].

It is certain that HDD manufacturing industries require yield in the production process as high as possible. Accurate yield estimation is important for process engineers and product managers for proper planning in logistics and marketing. Yield estimation is traditionally performed by process engineers to rely on their own experience in calculating yields at each step of HDD manufacturing. Yield estimation is done manually and it is time consuming. We thus propose in this research work to apply a data-driven approach based on machine learning technology to automatically predict yield assisting engineers in the HDD manufacturing industry.

The difficult part of machine learning-based yield prediction is the excessive amount of data records and data attributes. The number of records can be higher than a million and the number of attributes can be more than hundreds. It is almost impossible to apply such high dimensionality data in the modeling step. Therefore, data pre-processing is an essential step to be applied prior to the deployment of machine learning technique [6–11].

We thus introduce a heuristic method to pre-process HDD manufacturing data. We firstly propose a novel idea based on cluster analysis to re-balance data. HDD data records contain two class of products: pass units and fail units. Normally, the number of pass units is much higher than the number of fail units. A high imbalance between the two classes can decrease significantly performance of the prediction model. Data improvement by making equal proportion among the two classes is essential. Reducing the number of data attributes is the next essential step of data-preprocessing. Before applying machine learning technique to create a model to predict yield, we also introduce a novel idea of data aggregation to group data records in order to reduce amount of data instances. Details of these techniques are explained in the next section.

II. Methodology

A. Data

Data used in the modeling and experimentation are real data collected from the HDD production in the three months period. The number of data records is 10,000,000 and the number of attributes (or features) is 125. Some important attributes are summarized in Table I.

B. Research Framework and Yield Prediction Steps

The four main steps of data-driven modeling to predict yield in the HDD manufacturing process is shown in Fig. 1.

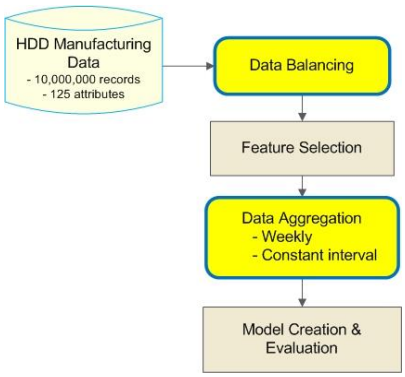

Fig. 1. Research framework for HDD yield prediction.

Data Balancing. The original dataset has high imbalance proportion between the pass and fail units (imbalance ratio is 28:1). Therefore, data re-balancing method is introduced. The

Manuscript received January 10, 2023; revised February 20, 2023; accepted April 27, 2023.

Nittaya Kerdprasop and Kittisak Kerdprasop are with the School of Computer Engineering, Suranaree University of Technology, Thailand.

Anusara Hirunyawanakul is with the Data Science and Computation School, King Mongkut's University of Technology North Bangkok, Rayong Campus, Thailand.

Paradee Chuaybamroong is with the Department of Environmental Science, Thammasat University, Thailand.

*Correspondence: nittaya@sut.ac.th (N.K.)

re-balancing strategy starts by grouping data into five main groups using k-means algorithm. After that, different data handling methods have been applied to each data group as illustrated in Table II.

Feature Selection. This step is for reducing number of data attributes. We experiment with several feature selection algorithms including decision tree (C5), classification and regression tree (CART), support vector machine (SVM), stepwise regression (SR), genetic algorithm (GA), chi-square (Chi^2), and information gain (IG). After experimentation, the best method is applied to extract important attributes to be used in the next step.

TABLE I: SOME ATTRIBUTES FROM THE HDD MANUFACTURING

Attribute	Meaning
Drive SN	Unique identification of each HDD lot
Week	Fiscal week that particular HDD had been assembled
Status	Status of test process (pass/fail) - "Pass" status indicates that this HDD passed the test process and be able to be input of the next operation step or ready to ship to customer. - "Fail" status means this HDD is rejected from the test process and must go to either "rework", "retest", "recycle" or "scrap" process according to the debug diagnostic failure symptom
HSA_PR	Head stack assembly status (prime/rework). - "Prime" means this HSA is the fresh new built component and never been installed in any other HDD before. - "Rework" means this HSA is a component that had been installed in another HDD, but that HDD had been rejected in the test process with the HSA labeled as rework. Thus, this HSA is recycled by being rebuilt again in this HDD.
Media_PR	Media status (prime/rework)
MBA_PR	Motor base assembled status (prime/rework)
VCM_PR	Voice coil motor status (prime/rework)
TC_PR	Top cover condition (prime/rework)
PCBA_PR	Printed circuit board assembled status (prime/rework)

Data Aggregation. This step is another contribution of this work. To decrease the number of data records and to improve performance of yield prediction, we propose data aggregation techniques using two main strategy: constant aggregation and weekly aggregation. Constant aggregation is the act of grouping data records with constant number such as a group of 500 records, whereas weekly aggregation is grouping by week. Example of grouping data as a constant interval of 10 records per group is shown in Fig. 2. Suppose data contain records of three weeks with selected five attributes (Fig. 3), the step of weekly aggregation is illustrated in Fig. 4.

TABLE II: DATA MANAGEMENT TO RE-BALANCE DATA

Group	Group characteristic	Re-balancing technique
1	Pass units > Fail units	Select representative of pass units. Then, apply k-nearest neighbor algorithm to reduce number of pass units to be equal to the number of fail units (k = number of fail - 1)
2	Pass units < Fail units	Apply re-sampling technique to increase number of pass units to be equal to number of fail units
3	Pass units = Fail units	Do nothing and add all data in this group to the dataset to be used in the next step
4	Only pass units	Discard this data group
5	Only fail units	Discard this data group

Drive SN	WEEK	STATUS	HSA PR	MEDIA PR
SN-001	WK01	Pass	Prime	Prime
SN-002	WK01	Pass	Prime	Prime
SN-003	WK01	Fail	Prime	RCY
SN-004	WK01	Fail	RCY	RCY
SN-005	WK01	Pass	Prime	Prime
SN-006	WK01	Pass	Prime	Prime
SN-007	WK01	Pass	Prime	Prime
SN-008	WK01	Pass	Prime	Prime
SN-009	WK01	Pass	Prime	Prime
SN-010	WK01	Pass	Prime	Prime
SN-011	WK02	Fail	Prime	RCY
SN-012	WK02	Pass	Prime	Prime
SN-013	WK02	Pass	Prime	Prime
SN-014	WK02	Pass	Prime	Prime
SN-015	WK03	Pass	Prime	Prime
SN-016	WK03	Pass	Prime	Prime
SN-017	WK03	Pass	Prime	Prime
SN-018	WK03	Pass	Prime	Prime
SN-019	WK03	Pass	Prime	Prime
SN-020	WK03	Fail	RCY	Prime

Fig. 2. Data sample after selecting main attributes.

Drive SN	WEEK	STATUS	HSA_PR*
SN0000001	WK01	Pass	Prime
SN0000002	WK01	Pass	Prime
SN0000003	WK01	Fail	Prime
SN0000004	WK01	Fail	Rework
SN0000005	WK01	Pass	Prime
SN0000006	WK01	Pass	Prime
SN0000007	WK01	Pass	Prime
SN0000008	WK01	Pass	Prime
SN0000009	WK01	Pass	Prime
SN0000010	WK01	Pass	Prime
SN0000011	WK02	Fail	Prime
SN0000012	WK02	Pass	Prime
SN0000013	WK02	Pass	Prime
SN0000014	WK02	Pass	Prime
SN0000015	WK03	Pass	Prime
SN0000016	WK03	Pass	Prime
SN0000017	WK03	Pass	Prime
SN0000018	WK03	Pass	Prime
SN0000019	WK03	Pass	Prime
SN0000020	WK03	Fail	Rework

Table: B

Group	COUNT	#PASS	#FAIL	Yield	HSA_PR=Prime	HSA_PR =Rework	HSA_Prime Ratio	HSA_Rework Ratio
Group1	10	8	2	80.0%	9	1	90%	10%
Group2	10	8	2	80.0%	9	1	90%	10%

Fig. 3. Example of constant aggregation by grouping data at a constant number of 10 records.

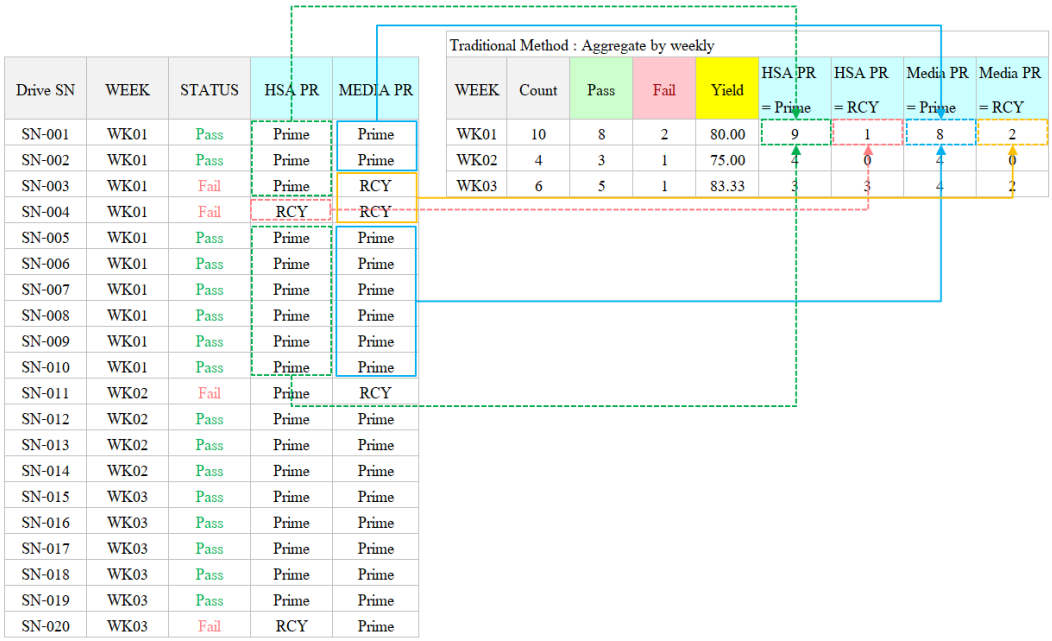

Fig. 4. Example of weekly data aggregation process.

It can be noticed from Figs. 3 and 4 that the new attribute named yield has been created. Value of yield can be computed from the number of pass units in each data group divided by all units in a group and multiply by 100 to be yield percentage at each new aggregated data record. The four new data attributes (HSA_PR = Prime, HSA_PR = RCY, Media_PR = Prime, Media_PR = RCY) are also created to be used later in the modeling step.

Model Creation & Evaluation. The last step of this research is the use of re-balanced are aggregated data to create model for predicting yield in the HDD manufacturing. Two learning algorithms are applied: multiple linear regression (MLR) and artificial neural network (ANN).

III. Experimentation and Results

At the first step of data balancing that data have been clustered into five groups, imbalance ratio between the pass and fail units is illustrated in Table III. This imbalance ratio has been managed by the proposed method resulting in the equal proportion as shown in Table IV.

TABLE III: Imbalance Ratio of The Original Data

Data	# Pass Units	# Fail Units	Imbalance Ratio
Cluster 1	2,064,114	71,181	29:1
Cluster 2	1,981,558	70,550	28:1
Cluster 3	1,850,018	65,561	28:1
Cluster 4	1,924,591	70,846	27:1
Cluster 5	1,834,891	66,690	27:1

After data balancing, 7 methods to feature selection have been applied and then tested with the two learning algorithms (MLR and ANN). Results of feature selection are presented in Table V. Performance of feature selection practiced by engineers is also presented as a baseline for comparison. It can be seen from the results that feature selected with genetic algorithm to create model using the algorithm multiple linear regression is the best technique for yield prediction.

TABLE IV: Pass and Fail Units after Applying The Data Balancing Technique

Data	# Pass Units	# Fail Units	Imbalance Ratio
Cluster 1	71,181	71,181	1:1
Cluster 2	70,550	70,550	1:1
Cluster 3	65,561	65,561	1:1
Cluster 4	70,846	70,846	1:1
Cluster 5	66,690	66,690	1:1

TABLE V: Comparative Results of Feature Selection Methods

Feature Selection Method	Modeling Algorithm	Model Error (RMSE)	Model Error (MAE)
Human engineers		1.700	1.400
C5	MLR	0.866	0.605
	ANN	1.707	1.263
CART	MLR	24.105	5.913
	ANN	1.630	1.251
SVM	MLR	2.037	1.247
	ANN	1.864	1.384
SR	MLR	10.326	2.842
	ANN	1.851	1.306
GA	MLR	**0.732**	**0.559**
	ANN	1.706	1.269
Ch^2	MLR	0.821	0.690
	ANN	1.707	1.262
IG	MLR	0.821	0.690
	ANN	1.707	1.262

We then applied GA and MLR to test the two data aggregation methods: constant aggregation and weekly

aggregation. For constant aggregation, seven sizes of data aggregation have been tested. The results are shown in Table VI. It can be clearly seen that data aggregation of constant size perform better than weekly aggregation method and the errors are the same for data of sizes 10K up to 40K. The non-decreasing errors also occur with other two feature selection methods as shown in Fig. 5.

TABLE VI: YIELD PREDICTION ACCURACY TESTED WITH DIFFERENT DATA AGGREGATION METHODS

Data Aggregation Method	Data Size	Yield Prediction Error (RMSE)
Weekly		0.958
Constant	1K	1.163
	2K	1.017
	5K	1.017
	10K	**0.732**
	30K	0.732
	35K	0.732
	40K	0.732

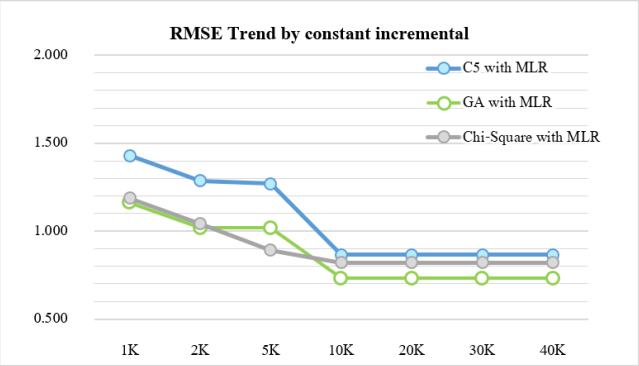

Fig. 5. The trend of prediction errors modeling with MLR that are trained with varied sizes of constant data aggregation and performed feature selection with C5, GA, and Chi^2.

IV. CONCLUSION

This research presents a methodology to prepare data for modeling with machine learning technique in order to predict yield in the hard disk drive (HDD) manufacturing process. The data preparation steps used in this work are data balancing, feature selection, and data aggregation. The prepared data are then modeled with two algorithms: multiple linear regression and artificial neural network.

The focus of this research is the data preparation techniques. We propose a technique to re-balance data to contain the equal amount of the two data classes: pass and fail HDD units. The proposed data balancing is based on data clustering. We also introduce the idea of data aggregation based on weekly time-frame and aggregation at constant size. Experimental results reveal that data aggregation at the constant size of 10,000 records incorporated by the genetic algorithm for feature selection and then modeling with multiple linear regression yield the best predictive model for the specific task of HDD yield prediction.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The first author is responsible for manuscript preparation and submission. The second author is the main contributor in data collection and experimentation. The third author helps revising the manuscript. The fourth author contributes the research idea.

REFERENCES

[1] R. Wood, "Future hard disk drive systems," *Journal of Magnetism and Magnetic Materials*, vol. 321, no. 6, pp. 555–561, 2009.
[2] V. Kasavajhala, "Solid state drive vs. hard disk drive price and performance study," in *Proc. Dell Tech. White Paper*, pp. 8–9, 2011.
[3] H. Lee, C. O. Kim, K. H. Ko, and M. K. Kim, "Yield prediction through the event sequence analysis of the die attach process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 4, pp. 563–570, 2015.
[4] J. Li, X. Ji, Y. Jia, B. Zhu, G. Wang, Z. Li, and X. Liu, "Hard drive failure prediction using classification and regression trees," in *Proc. 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pp. 383–394, 2014.
[5] T. Yuan, S. Z. Ramadan, and S. J. Bae, "Yield prediction for integrated circuits manufacturing through hierarchical Bayesian modeling of spatial defects," *IEEE Transactions on Reliability*, vol. 60, no. 4, pp. 729–741, 2011.
[6] K. N. Malitski and S. E. SAP, *Method for Shipment Planning/Scheduling*, U.S. Patent 8244645, 2012.
[7] A. K. R. Katta and R. Allgor, *Heuristic Methods for Customer Order Fulfillment Planning*, U.S. Patent 8352382, 2013.
[8] M. Braglia, D. Castellano, M. Frosolini, and M. Gallo, "Overall material usage effectiveness (OME): a structured indicator to measure the effective material usage within manufacturing processes," *Production Planning & Control*, vol. 29, no. 2, pp. 143–157, 2018.
[9] J. Shi, G. Zhang, and J. Sha, "Optimal production planning for a multi-product closed loop system with uncertain demand and return," *Computers & Operations Research*, vol. 38, no. 3, pp. 641–650, 2011.
[10] A. P. Rastogi, J. W. Fowler, W. M. Carlyle, O. M. Araz, A. Maltz, and B. Büke, "Supply network capacity planning for semiconductor manufacturing with uncertain demand and correlation in demand considerations," *International Journal of Production Economics*, vol. 134, no. 2, pp.322–332, 2011.
[11] S. M. Liozu and A. Hinterhuber, "Industrial product pricing: a value-based approach," *Journal of Business Strategy*, vol. 33, no. 4, pp. 28–39, 2012.