Applying Machine Learning Technique to Detect Failures in Hard Disk Drive Test Process

ABSTRACT

This paper presents machine learning techniques to detect the failure in hard disk drive manufacturing test process. The data is high dimensionality and highly imbalance. Feature selection technique with filter method and embedded method with light gradient boost are used to reduce the dimension of data. We apply three techniques: SMOTE, Different Cost and SMOTE with Different Cost to handle imbalance data. Several machine learning methods are compared. The XGBoost with SMOTE and XGBoost with Different Cost (XGB DC) give the best performance with 91% ROC AUC and 73% PRC AUC. The SVM algorithm shows good performance on ROC AUC while low performance on PRC AUC. The XGBoost algorithm shows good performance of both ROC AUC and PRC AUC.

Keywords: SVM, imbalance data, high dimensionality, feature selection, hard disk drive

1 INTRODUCTION

Machine learning technique is widely used in manufacturing process. In hard disk drive manufacturing machine learning is used to improve the productivity, such as parameter improvement, anomaly detection and failure detection. There are several processes in hard disk drive manufacturing. Failure should be detected as early as possible in manufacturing process to reduce waste time and reduce cost. Each process has measured parameters and the data stored in database. The failure detection with machine learning technique is an opportunity to reduce the cost and increase productivity.

2 LITERATURE REVIEW

Support vector machine is one of popular machine learning algorithm for classification on two classes data. SVM is used to detect the failure in hard disk drive assembly process by using voice coil motor current to train the model (Simongyi & Chongstitvatana, 2018). The data set is imbalance, Fail case is only 3%. SVM algorithm is able to classify with 100% accuracy. Contrast to this previous work, this paper studies the failure during test process. Data set are the parameters collected from assembly and servo track write processes.

There are several research applying SVM and SMOTE technique to handle imbalance data such as Akbani et al. (2004). They applied SVM with SMOTE technique and used different error cost, called SDC (SMOTE with Different Cost). The SDC method gives the best performance compared to SVM and SVM with SMOTE.

SVM with different method to handle imbalance data is used (Tang et al., 2009). Several methods are studied: SVM-weight is a cost sensitive learning method, SVM-SMOTE oversampling minority class, SVM-RANDU under sampling and GSVM-RU random under sampling only the data which no vector supported. Comparing performance with 4 metrics, the GSVM-RU gives the best performance.

Tong and Daphne (2001) uses SVM with imbalance data by applying active learning to reduce the train data size and obtain better performance. Chakravarthy et al. (2019) studied C50, KNN, NN, RF and SVM with Random oversampling (ROSE) and SMOTE oversampling with different ratio. The SMOTE with 1:3 oversampling ratio gives best performance in all models. A fuzzy support vector machines is introduced (Ma et al., 2011) for class imbalance learning. There are several studies of the problem of class imbalance such as Guo et al., (2008), Chakravarthy et al., (2019), Wang and Japkowicz (2009).

Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is one of popular method to oversampling the minor class. Borderline SMOTE (Han et al., 2005) is another sampling method that develop from SMOTE to sample only the borderline of minor class. Sharma et al., (2018) studied oversampling minority class with consider on major class. Sampling with Majority (SWIM) creates minor class with similar Mahalanbois distance of the majority class. In Khire et al., (2019), Chandrashekar and Sahin (2014) the features input to the model is studied and the result showed that they are important to the performance of the model. The work in Zhang et al., (2017) uses SSVM-FS to select features. This method focuses on the imbalance class. The weight of SVM indicated the important features.

The Extreme Gradient Boost (XGBoost) is another machine learning algorithm which is a scalable tree boosting system. It is widely used in machine learning competition to achieve state-of-art result. It is an implement of gradient boost. XGBoost able to run beyond billions of examples on few resources (Chen & Guestrin, 2016).

Using accuracy to measurement the performance of machine learning algorithm on imbalance data is not appropriate. The model performance will show high accuracy when the model fails to predict the minor class. The performance measurement with area under curve (AUC) is not affected by the ratio of class (skew), while the accuracy, F1-score, Cohen's kappa and Krippendorf's are affected by skew (Jeni et al., 2013). The Receive Operation Characteristic (ROC) plot gives the overview performance. Precision-Recall plot (PRC) gives accurate prediction performance (Saito & Rehmsmeier, 2015).

This paper studies SVM and XGBoost algorithm to detect the failure in hard disk drives test process. We also applying SMOTE, Different Cost and SMOTE with Different Cost techniques to train the model. We measure the performance with area under curve of Receiver Operating Characteristics (ROC AUC) and area under curve of associated Precision/Recall (PRC AUC) values. The rest of the paper is as follows Section 3 explained hard disk drive manufacturing process. Section 4 described failure detection with machine learning techniques. Section 5 reported the experimental result. Conclusion is given in Section 6.

3 HARD DISK DRIVE MANUFACUTRING PROCESS

Hard disk drive process started in clean room to assembly all components such as base, spindle motor, head and disk together (Fig.1). The complete unit is sent out from clean room to write servo track pattern and test process. The servo track write is the process to write reference position signal on disk. The reference signal is written on control circular track and it is consistency space of track.

Figure 1. Hard disk drive components.

The test process consists of 4 main processes: Function test, Self Run Stress test (SRST), Final test and Featuring test. Function test starts with micro code download to functional the hard disk drive, measures and adjust parameters for best seek, read and write performance. SRST is the test to measure and analyze disk surface then records the defect location. This test is performed at high temperature. Also Re-adjust parameters and test customer functionality. It is the longest test process. Final Test is also called performance test. It tests the whole surface read write performance. Featuring test is the test to adjust parameters to meet customer requirement (Fig. 2).

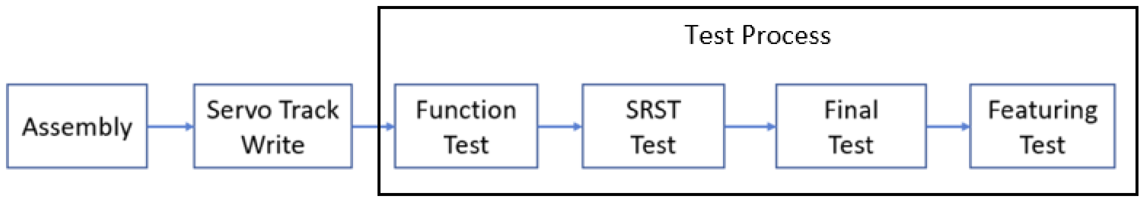

Figure 2. Hard disk drive manufacturing process.

4 FAILURE DETECTION USING MACHINE LEARNING

4.1 Data Collection

Data set are collected from hard disk drives manufacturing process which is separated in two parts. The first part is parameters collected from Assembly and Servo track write processes. Parameters are the measurement during assembly parts and write servo signal, such as motor current, head resistance, servo signal quality, distance between head and disk total 359 parameters. The second part is the target output collected from test process. The target output are 2 classes, negative class (Bad) and positive class (Good). Negative class data are very small compare to positive class. The ratio of failure to passer is 1:100. Data are collected with sampling passer and whole failure. Total data is 84325 rows. Each row represents for one hard disk drive (HDD) data. There are 79448 rows of passer and 4877 rows of failure. 70% of data is used for training and 30% for validation (Table 1.).

Table 1: Data set.

	Input Data	**Train Data**	**Validation Data**	**Unit**
Total Data	84325	59027	25298	HDD
Pass	79448	56613	22835	HDD
Fail	4877	3414	1463	HDD

4.2 Data Pre-processing

Data is pre-processed by eliminating parameters with missing value that are more than 40% and fill the missing data with mean value. Data is normalized with z-score.

4.3 Feature Selection

Feature selection is one of the methods to improve machine learning performance by reducing the redundant and unnecessary features. Filter method is the method to eliminate the feature by ranking according to importance with statistical measure such as Chi-square, ANOVA and correlation coefficient. Wrapper method selects subset of feature that give high performance on machine learning model such as Forward Selection, Backward Elimination and Recursive Feature Elimination. The wrapper method costs high computation time (Khaire et al., 2019).

There are 359 features in the data set which is very high dimension. To input all data to train the machine learning model will cost high computation time and low performance. The filter method is used to remove the constant features, duplicated feature and correlated features. The group of correlate features defined by Pearson's correlation coefficient more than 95% and select only one feature from each group with best AUC on random forest model. The embedded method with light gradient boost to rank the important of feature is used. To reduce the variation, feature important is rank with 10 iterations of accumulate values. 289 parameters are eliminated by feature selection. 70 parameters are input to train the model.

4.4 Imbalance Data Handling

The data set shows very small numbers of bad compare to good. Data set is highly imbalance. The SMOTE (Chawla et al., 2002) is one of the methods to handle imbalance data by oversampling the minority class. It uses the actual data to generate the synthetic sample of the minor class (Fig. 3). Another approach is assigning different cost of the learning class.

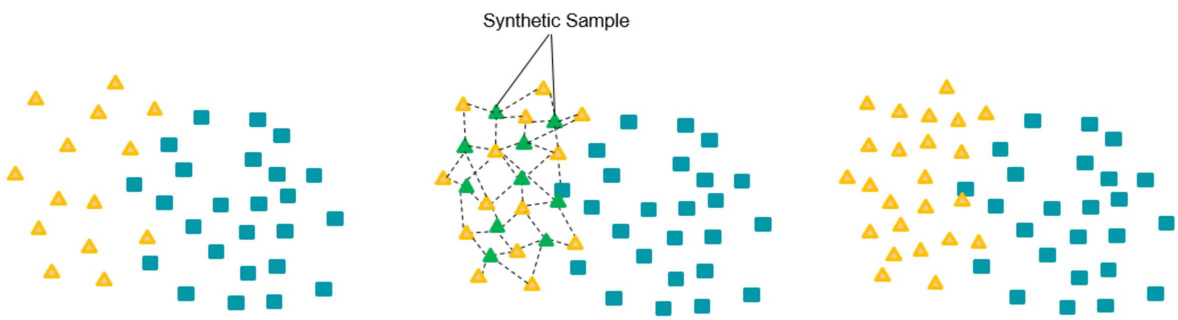

Figure 3. SMOTE: Synthetic Minority Over-sampling Technique

4.5 Training Model

Support Vector Machine (SVM) is a supervised learning algorithm. It is one of popular classification method. SVM algorithm finds the hyperplane that can separate the class of data with maximum margin. The closest data to hyperplane of each class is called support vector. The margin measures from support vector to hyperplane (Fig. 4).

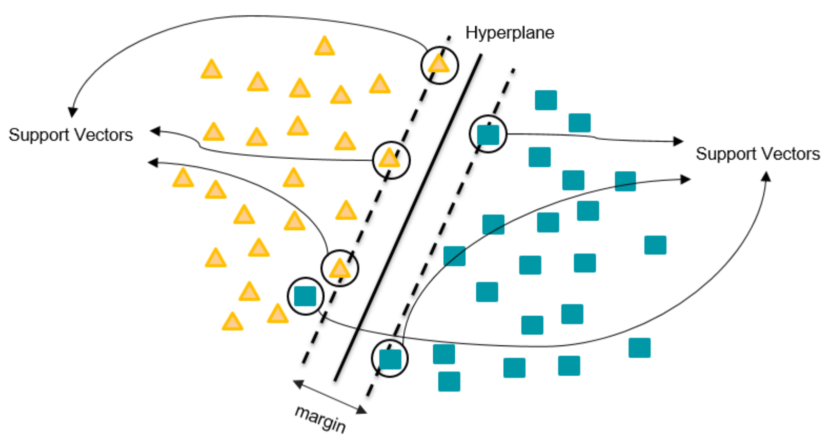

Figure 4. Support Vector Machine

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm. It is a scalable machine learning for tree boosting (Chen & Guestrin,2016). The boosting is sequential learner algorithm and use the error from previous learner to improve the accuracy of the next learner. The model is added and learn until the accuracy is not improved. XGBoost is a gradient boosting method which is fast and uses minimal resources. It is widely used algorithm and achieve state-of -art result.

We compare 6 methods. There are 3 methods training with SVM algorithm and another 3 methods training with XGBoost algorithm: 1.) SVM with SMOTE (SVM SMOTE), 2.) SVM with Different Cost (SVM DC), 3.) SVM with SMOTE and Different Cost (SVM SDC), 4.) XGBoost with SMOTE (XGB SMOTE), 5.) XGBoost with Different Dost (XGB DC) and 6.) XGBoost with SMOTE and Different Cost (XGB SDC). The SVM using 'rbf' kernel, class_weight equal to ratio of passer and failure and others hyper parameter setting are as follows. Degree = 3, gamma = 'scale' and max_iter = 1. The XGBoost using booster = 'gbtree', eta = 0.3, gamma = 0, max_depth = 6 and scale_pos_weight equal to ratio of passer and failure.

5 EXPERIMENT RESULT

The result of classifying the validation data of 25298 rows are as follows. The performance measurement with ROC AUC: SVM with SMOTE gives 80%, SVM with Different Cost (SVM DC) gives 89% and SVM with SMOTE and Different Cost (SVM SDC) gives 79% (Fig. 5). XGBoost with SMOTE (XGB SMOTE) and XGBoost with Different Cost (XGB DC) give the best performance at 91% and XGBoost with SMOTE and Different Cost (XGB SDC) gives 90% (Fig. 6). The performance measurement with PRC AUC: SVM SMOTE 32%, SVM DC 59% PRC AUC and SVM SDC has the lowest performance at 29% (Fig. 7). XGB SMOTE and XGB DC give the best performance at 73% while XGB SDC gives 71% (Fig. 8). The model with Different Cost gives the best performance on both SVM and XGBoost algorithm. The XGBoost algorithm in all methods give better performance than SVM (see Table 2).

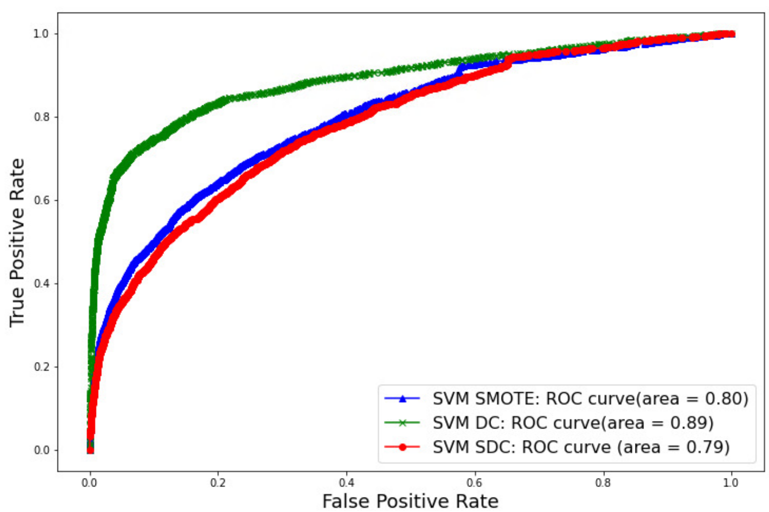

Figure 5. SVM ROC AUC plot

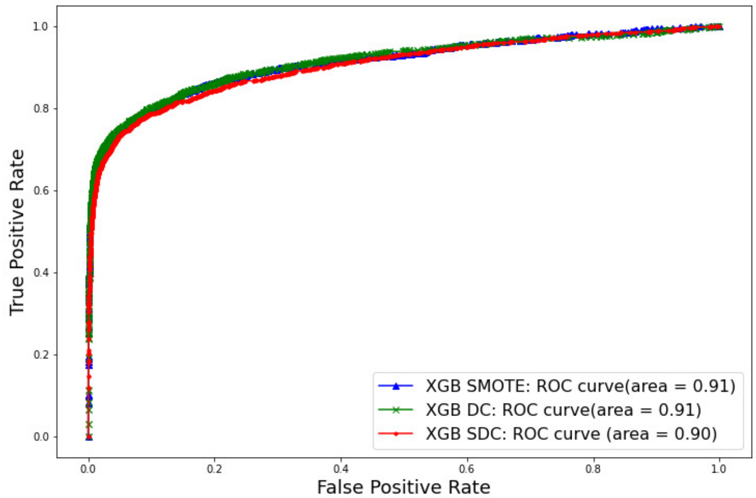

Figure 6. XGBoost ROC AUC plot

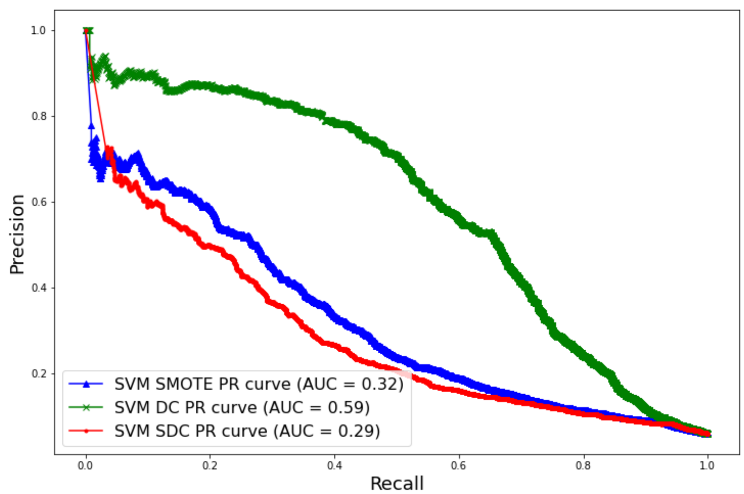

Figure 7. SVM PRC AUC plot

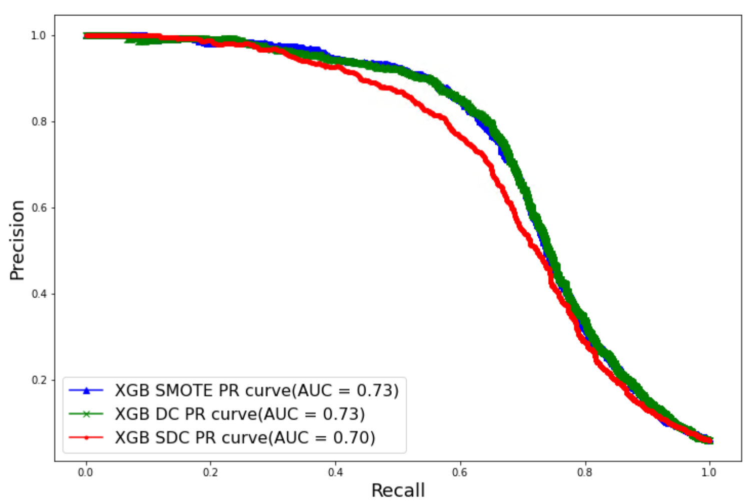

Figure 8. XGboost PRC AUC plot

Table 2: Result

Method	ROC AUC	PRC AUC
SVM SMOTE	80%	32%
SVM DC	89%	59%
SVM SDC	79%	29%
XGB SMOTE	91%	73%
XGB DC	91%	73%
XGB SDC	90%	70%

6 CONCLUSION

This study presents the method of failure detection with SVM and XGBoost algorithm. The proposed method employs feature selection and data imbalance handling. The experiment performs on real hard disk drive manufacturing data. The feature selection method are filter and embedded algorithm. The parameters are reduced from 359 to 70 to be input to the model. The model training with SVM and XGBoost algorithm with 3 different data imbalance handling methods: 1) SMOTE, 2) Different Cost and 3) SMOTE with Different Cost. The XGBoost algorithm has the better performance than SVM. The XGBoost with SMOTE and XGBoost with DC give the best performance with 91% ROC AUC and 73% PRC AUC. The SVM algorithm shows high performance on ROC AUC measurement while low performance on PRC AUC. XGBoost shows good performance on both of two measurements.

7 REFERENCES

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to Imbalanced datasets. Machine Learning: ECML 2004, 39-50. https://doi.org/10.1007/978-3-540-30115-8_7

Chakravarthy, A. D., Bonthu, S., Chen, Z., & Zhu, Q. (2019). Predictive models with Resampling: A comparative study of machine learning algorithms and their performances on handling Imbalanced datasets. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). https://doi.org/10.1109/icmla.2019.00245

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi:10.1145/2939672.2939785

Guo, x., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. 2008 Fourth International Conference on Natural Computation, 192-201. https://doi.org/10.1109/ICNC.2008.871

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: A new over-sampling method in Imbalanced data sets learning. Lecture Notes in Computer Science, 878-887. https://doi.org/10.1007/11538059_91

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced data--recommendations for the use of performance metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. doi:10.1109/acii.2013.47

Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2019.06.012

Ma, H., Wang, L., & Shen, B. (2011). A new fuzzy support vector machines for class imbalance learning. 2011 International Conference on Electrical and Control Engineering, 3781-3784. https://doi.org/10.1109/iceceng.2011.6056838

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on Imbalanced datasets. PLOS ONE, 10(3), e0118432. doi:10.1371/journal.pone.0118432

Sharma, S., Bellinger, C., Krawczyk, B., Zaiane, O., & Japkowicz, N. (2018). Synthetic Oversampling with the majority Class: A new perspective on handling extreme imbalance. 2018 IEEE International Conference on Data Mining (ICDM). https://doi.org/10.1109/icdm.2018.00060

Simongyi, M., & Chongstitvatana, P. (2018). Abnormality detection in hard disk drive assembly process using support vector machine. 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 612-615. https://doi.org/10.1109/ecticon.2018.8619935

Tong, S., & Daphne, K. (2001). Support Vector Machine Active Learning With Applications To Text Classification. The Journal of Machine Learning Research, 2(1), 45-66. https://doi.org/10.1162/153244302760185243

Wang, B. X., & Japkowicz, N. (2009). Boosting support vector machines for imbalanced data sets. Knowledge and Information Systems, 25(1), 1-20. https://doi.org/10.1007/s10115-009-0198-y

Yuchun Tang, Yan-Qing Zhang, Chawla, N., & Krasser, S. (2009). SVMs modeling for highly Imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1), 281-288. https://doi.org/10.1109/tsmcb.2008.2002909

Zhang, C., Wang, G., Zhou, Y., Yao, L., Jiang, Z. L., Liao, Q., & Wang, X. (2017). Feature selection for high dimensional imbalanced class data based on F-measure optimization. 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). https://doi.org/10.1109/spac.2017.8304290