9th International Conference on Through-life Engineering Service
3-4 November 2020, Cranfield UK

Aircraft Predictive Maintenance Modeling using a Hybrid Imbalance Learning Approach

Maren David Dangut[a] Skaf Zakwan[b], Ian K. Jennions[c]

[a,c] *Integrated Vehicle Health Management (IVHM) Center, Cranfield University, Bedford, United*
[b] *Mechanical Engineering Department, Higher Colleges of Technology, United Arab Emirates*

* Corresponding author. E-mail address: maren.dangut@cranfield.ac.uk

Abstract

The continued development of the industrial internet of things (IIoT) has caused an increase in the availability of industrial datasets. The massive availability of assets operational dataset has prompted more research interest in the area of condition-based maintenance, towards the API-lead integration for assets predictive maintenance modelling. The large data generated by industrial processes inherently comes along with different analytical challenges. Data imbalance is one of such problems that exist in datasets. It affects the performance of machine learning algorithms, which yields imprecise prediction. In this paper, we propose an advanced approach to handling imbalance classification problems in equipment heterogeneous datasets. The technique is based on a hybrid of soft mixed Gaussian processes with the EM method to improves the prediction of the minority class during learning. The algorithm is then used to develop a prognostic model for predicting aircraft component replacement. We validate the feasibility and effectiveness of our approach using real-time aircraft operation and maintenance datasets. The dataset spans over seven years. Our approach shows better performance compared to other similar methods.

Keywords: prognostics; data-driven; data imbalance; predictive maintenance; aircraft

1. Introduction

The recent advancement in industrial technology, known as the fourth industrial revolution, has broken the barriers between physical and digital worlds. The technological revolution involves the integration of technologies such as the Internet of Things (IoT), the application of artificial intelligence (AI), the Application Programming Interface (API), and machine learning in the industrial process to enhance productivity. It is this collective force that has brought an increase in the generation and availability of industrial datasets. Businesses are leveraging the large available datasets generated by the modern industrial system to make a more informed decision.

One such application area is the assets predictive maintenance, which, instead of relying on component average life statistics, it uses direct condition monitoring data to forecast or estimate upcoming maintenance based on historical knowledge. Predictive maintenance has a comparative advantage in almost all industries compared to other forms of maintenance strategies. Application of equipment prognostics is vital, especially in a domain where the criticality of the system or component may affect health and safety, such as aircraft health monitoring, nuclear industries, and many more.

The increasing availability of datasets also comes along with more analytical challenges, which raises the necessity of applying an advanced algorithm to harness knowledge for better-informed decisions. This necessity is highlighted in equipment' predictive maintenance, where the monitoring system is expected to provide accurate prognostic alerts in advance to plan for maintenance ahead of time to avoid unexpected failure. One of the analytical challenges that inherently comes with raw asset operational datasets and affects the performance of data-driven predictive models is the data imbalance problem.

Nomenclature

API	Application Programming Interface
CBM	Condition-based Maintenance
CMS	Central Maintenance System
FIN	Functional Identification Number
MPG	Mix Gaussian Process

Data Imbalance problem is a well-known problem in machine learning and data mining communities[1]. It is a common problem faced by most of the real-world applications because industrial processes are designed to function normally with few faults recorded. Data imbalance occurs in industrial datasets as a result of a rare event failure, as compared to the healthy state of the monitoring system.

Rare failure occurs as a result of the infrequent occurrence of some unexpected equipment break down, causing unplanned maintenance. For example, in aircraft scheduled maintenance strategy, the failure that occurs in-between a scheduled maintenance -defined time intervals are proven to be rare, but their impact on business can be grave[2,3]. Therefore, the rare failures are often more critical to predict because its occurrence could have a potentially negative impact on society or business[4]. The majority of the data generated from the aircraft central maintenance system is highly characterized by a healthy majority, and the faulty minority (represents the rare failures).

Furthermore, an extreme data imbalance problem is a scenario where a dataset contains a high representation of samples in one class than other classes present in a dataset. Learning from an extremely imbalanced dataset is quite challenging for traditional machine learning algorithms, which often leads to undesired prediction outcomes. The class Imbalance problem has shown to degrades the performance of predictive modelling, causing imprecise prediction[5]. In a situation where the imbalance ratio is extreme, the learning algorithm may sometimes consider the minority class as an outlier or noise and end up dropping them, which will result in bias leaning that is learning from one class[6].

The class imbalance problem has recently drawn significant research attention. A lot of techniques and approaches for handling imbalance problems have been proposed in the literature. The majority of these techniques are based on the nature (distribution) of the dataset or its application domain. Although Imbalance learning has been extensively researched [7][1], the open literature lacks a unified solution to handling the imbalanced dataset for predictive maintenance modelling, especially in the aerospace domain and particularly the aircraft central maintenance system dataset. Hence, it is still an open area of research.

Therefore, in this paper, we proposed a hybrid technique to overcome the extreme imbalance problem in heterogeneous datasets. The proposed approach comprises the integration of boosting with divide and merge strategy and Mixed Gaussian Process (MPG). The technique is designed to enhance predictive maintenance modelling for aerospace applications. The focus is on enhancing the prediction of the minority class in the process of developing an aircraft components failure prognostic model.

This paper presents the following contributions.
1. A proposed hybrid technique for improving the prediction of the minority class in the imbalanced dataset is designed and implemented.
2. A predictive model for predicting component replacement is developed to improve predictive maintenance in aerospace.
3. The model is validated on real-time aircraft operational and maintenance dataset.

The remainder of this paper is organized as follows: Section two formally presents the related work to this study. The study methodology is presented in section three. The experimental setup is shown in section four. In section five, we discuss the results. Finally, we make our conclusion and future work in section six.

2. Related Work

Many approaches and methods for handling imbalanced datasets in the process of developing data-driven predictive modelling have been proposed in the literature. A comprehensive review of the existing methods can be found in [4,8,9]. The methods can be summarized into three main categories: 1. Data level approached:- It involves resampling the dataset before presenting it as input to the learning algorithm, and this can be achieved in different ways, some of which are under-sampling (that is randomly taking out some samples from the majority class to balance with the minority class). Over-sampling (involves adding more samples to the minority class to adjust with the majority class). A hybrid of the Under-sampling and Over-sampling is possible.
The algorithm level approach: - It involves modifying the learning algorithm so that it can respond favorably to both classes during learning. A typical example is cost-sensitive learning where the weight of classification is defined for each class; for example, a higher weight can be set for minority class so that during learning, the algorithm will focus more on the minority class, hence improving its prediction.
Another approach is the ensemble and hybrid methods; this approach involves the combination of two or more approaches to improve the predictive performance of the machine learning model.
The aforementioned approaches have their pros and cons. For instance, in the data under-sampling methods, since samples are reduced from the majority class, it makes it prone to losing informative data points, which could be used in defining a decision boundary during learning. Similarly, in the oversampling approach, since artificial synthetic data points are created, this can lead to the problem of generalization and also the original structure of the dataset is altered, which can affect the output of the model. Likewise, In the algorithm level, cost-sensitive methods, defining the cost of misclassification for each class is quite challenging.

Therefore because of the peculiarity of our dataset and the application domain, none of the out-of-box existing solutions was suitable.

2.1. Machine Learning

Machine learning is grouped into different types, such as supervised learning, unsupervised learning, semi-supervised learning, active learning and reinforcement learning. The use of more than one type of learning is referred to as hybrid learning. In this study, we are using classification, which is supervised learning, and clustering, which is unsupervised learning, hence the hybrid. In supervised learning, the algorithm builds a mathematical model from a dataset that contains input and known output (labels). The conventional approaches are classification and regression. In the case of unsupervised learning, the algorithm builds a mathematical model from the dataset that contains only input variables without labels. Unsupervised learning is mostly used to find structure in the dataset, such as grouping or clustering[10].
Many machine learning algorithms exist; their application depends on the nature and type of dataset and the problem at hand. For example, Support Vector Machine (SVM) and Decision Tree (DT) algorithms can be used for classification in supervised learning. Likewise, K-means, Gaussian process algorithm can be used for clustering in unsupervised learning. The combination of more one weak learns to form a robust classifier in order to achieve a better result is known as ensemble learning[11][12]. Many recent machine learning approaches have been designed based on ensemble learning to deal with various categories and dimensions of data imbalance challenges. Also, in many application domains [13], the most common ensemble learning techniques are bagging and boosting[14].

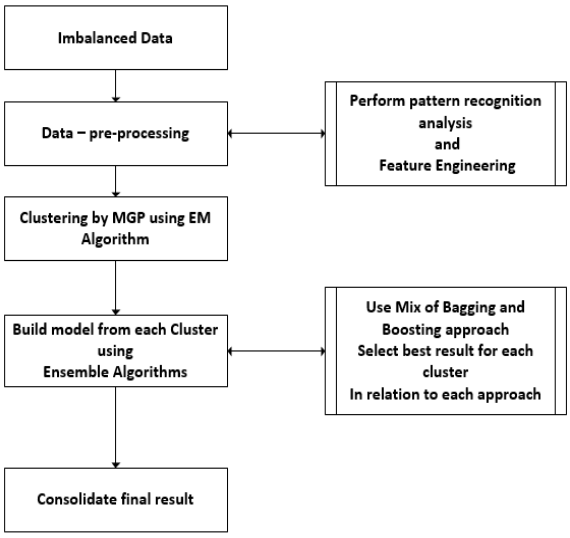

Figure 1. Machine Learning Hybrid Framework for enhancing class prediction

The framework is based on the hybrid approach, that is it combines a supervised and unsupervised machine learning methods to improve the prediction of the minority class.

2.2. Mixed Gaussian Process Methods

vandaplas et al. [15] and Fong et al. [16]. Shows that the clustering method, which is based on learning a mixture of Gaussians, involves the collection of a mix of k- component distribution to form a mixture distribution function. $f(x) = \sum_{k=1}^{k} \alpha_k f_{k(x)}$. (1)
α_k is the mixing weight for the Ith component in the construction of Gaussians distribution $f(x)$. K is the number of component distribution

The dataset used in this study is multi-variant, and some variables are latent, that is, unobserved, meaning the source distribution is not known. We use the Expected Maximization algorithm (EM) to minimize a likelihood function by iterating and guessing the distribution until convergence. K-means algorithm groups data using a hard clustering approach that is no overlapping of clusters.(Point belongs to a cluster, or it does not belong to) While EM algorithm computes the probability that it belongs to a cluster, which is referred to as soft clustering [17][18].

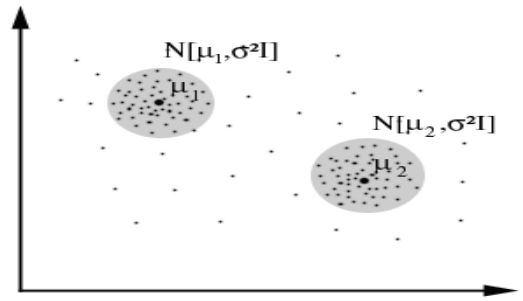

Figure 2. Clustering method based on learning a mixed of Gaussians

μ_i: is the mean; that is center of the mass
σ^2: is the variance; that is spread of the mass

Table 1. EM algorithm 1

Given an unknown observation of $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots \mathbf{x_n}$
1. Start with two randomly placed Gaussians ($\boldsymbol{\mu_1}, \boldsymbol{\delta_1^2}$), ($\boldsymbol{\mu_2}, \boldsymbol{\delta_2^2}$) in the space
2. E- Step: For each point: $\mathbf{P(1|x_i)}$ = does it look like it came from 1?
3. M-step: Adjust ($\boldsymbol{\mu_1}, \boldsymbol{\delta_1^2}$) and ($\boldsymbol{\mu_2}, \boldsymbol{\delta_2^2}$) to fit points assign to them
4. Loop until convergence

3. Methodology

Our proposed approach is similar to the hybrid method algorithm proposed by VanderPlas et al. [15]. However, our approach differs in the base learning algorithm. Instead of using hard K-means for clustering, we use a soft Mixed Gaussian Process with EM (MGP-EM).

The MGP-EM approach helps in computing the probability of points belonging to the cluster, which deals with an in-between point to avoid ambiguity problems in clustering. The proposed method is designed to overcome the problem of class-overlapping or small-size sample, which is difficult for the classifier to learn, hence improving the prediction of a minority class. It is also to handle the problem of over-sampling using K-means clustering, which is sensitive to outliers and noise and unable to handle more massive datasets. Putting the data into lagging windows and bootstrapping it helps in the learning phase by keeping the statistics, which avoids processing the whole dataset; instead, it keeps only the statistics of the outcome of each window.

Bagging- based (i.e., divide and merge) is used to improve model performance that is increasing detection rate (True Positive) and reduce the false positive. The mixed Gaussian process is used as a based learner in the Boosting step.

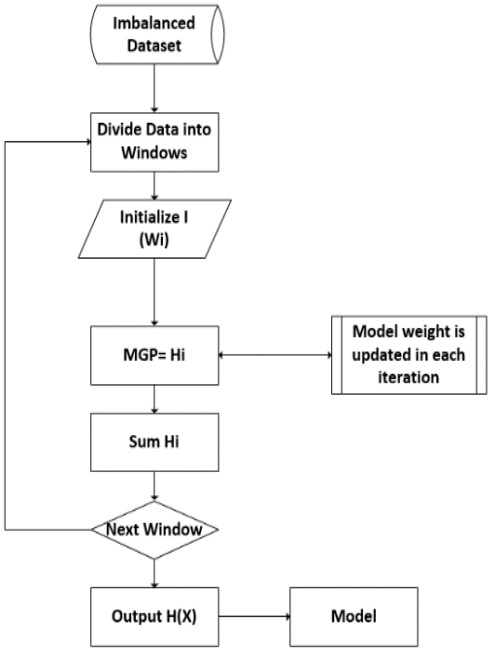

Figure 3. The Architecture of the proposed approach

We performed cross-validation during the training phase to avoid model over-fitting problems. We classify the model using the proposed hybrid method, using a cluster-based –Mixed Gaussian Process as weak learners. The result of MGP-EM is then combined with the Ensemble bagging method using the random forest as a based learner.

Table 2. Hybrid Algorithm 2.

STEP 1: Input the Imbalanced Dataset D = $x\epsilon \{x_1, x_2, \ldots x_n\}$
STEP 2: Divide the Data in Windows $W1, W2 \ldots Wn$
STEP 3: Initialization $x_1 = 1/n$
STEP 4: Then Mixed Gaussian Process (EM) is used as a base learner in the boosting = $\alpha_k f_k(x)$ and adjust weight
STEP 5: Calculate the True Positive and False Positive Rate
STEP 6: Iterate Until the end of windows
STEP 7: return final hypothesis $H(x) = \sum_{k=1}^{k} \alpha_k f_k(x)$
STEP 8: END

4. Experimental Setup

To validate the effectiveness of the proposed approach. The experiment uses a dataset obtained from a fleet of commercial aircraft, which has been recorded for over seven years. The data is a recorded component failure recoded as a log by aircraft cental maintenance computers. The data is heterogeneous in nature, meaning it comes from different aircraft sub-systems, and it contains numerical, textual, and symbolic.

As a first step, the data is preprocessed and transformed for machine learning, because to make use of the log-based dataset for developing a

predictive model, the log needs to be filtered and interpreted, and predictive feature extracted.

The data is then divided into two using the event date. Data from 2011 to 2015 is used for training the model and from 2016 to 2018 for testing the model.

In the experiment, we investigate the performance of the proposed method against existing ensemble learning methods.

We measured the performance of the model using precision, recall, F1-score.

$$\text{Precision:-}\frac{True\ Positive}{True\ Positive+False\ Postive} \quad (2)$$

$$\text{Recall:-}\frac{True\ Positive}{True\ Positive+False\ Negative} \quad (3)$$

$$\text{F1-Score:-}\ 2*\frac{Precision*Recal}{Precision+Recall} \quad (4)$$

We presented the experimental results in Table 3. In the experiment, we select out of many, the aircraft components identified by Functional Item Number (FIN) that are replaced as a result of an unplanned breakdown. We focused on the aircraft component with the highest number of replacements in the dataset. The components considered are:
4000KS - Electronic Control Unit, 4000HA - High-Pressure Bleed Valve, 4001HA – Pressure Regulating Valve, 5RV1 – Satellite Data unit.

5. Result and Discussion

As seen in Table 3. The result of the proposed method is compared against the baseline algorithm, which is the Random Forest algorithm (RF). The result shows that our approach outperformed the baseline method both in precisions and recall. Similarly, the F1-score indicates that the proposed approach is able to detect both classes with less bias. The high recall score shows that the proposed approach is able to detect the minority class better, which is our class of interest (the rare faults). However, the model includes some points of the majority (false negatives). This can be considered acceptable in this context, as we are more interested in reducing the false-positive rate than a false negative. It can also be observed that the imbalance ration has an effect on the result. In the cases with higher IR, the model is able to learn better while in the cases with the lower performance, the performance is drop. Despite the extreme imbalance ratio in all the cases considered, the proposed method was able to predict more than 80% of the rare equipment failure. The result also shows the effectiveness of the model in handling extreme class imbalance problems in big data.

6. Conclusion

This paper proposes a hybrid framework for data-driven predictive maintenance. We focus on enhancing the prediction of the minority class in the data Imbalance classification problem. Data imbalance problem is a data analytics challenge that degrades the performance of data-driven predictive models. Our approach is based on a hybrid ensemble method, which improves the prediction of the minority class during learning. The proposed MGP-EM approach helps in computing the probability of points belonging to the cluster, which deals with an in-between point to avoid ambiguity problems in clustering. The proposed method overcomes the problem of class-overlapping or small-size sample, which is difficult for the classifier to learn, hence improving the prediction of a minority class. It also overcomes the problem of over-sampling using K-means clustering, which is sensitive to outliers and noise and unable to handle more massive datasets. In the feature, we will try to improve the performance of the aircraft predictive model by including other aircraft related datasets such as environmental and weather data. We will also work on improving the detection of extremely minority in a multi-class context by applying the deep-learning approach.

Table 3.The result showing the performance of the proposed Framework

	Ensemble method -Random Forest (Baseline)				**Proposed Hybrid Approach**			
Components	4000KS	4000HA	4001HA	5RV1	**4000KS**	**4000HA**	**4001HA**	**5RV1**
Precision	0.77	0.70	0.71	0.79	**0.94**	**0.90**	**0.92**	**0.96**
Recall	0.60	0.59	0.60	0.63	**0.85**	**0.80**	**0.82**	**0.89**
F1-Score	0.67	0.64	0.65	0.70	**0.89**	**0.85**	**0.87**	**0.93**
AUC	0.60	0.65	0.66	0.72	**0.90**	**0.86**	**0.88**	**0.95**
IR	0.0031	0.0024	0.0028	0.0039	**0.0031**	**0.0024**	**0.0028**	**0.0039**

Acknowledgments

Appreciation goes to the Integrated Vehicle Health Management Center (IVHM), Cranfield University, for giving me the opportunity to carry out this study and PTDF Nigeria for sponsoring the study.

References

1. He H., Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(9): 1263–1284. Available at: DOI:10.1109/TKDE.2008.239
2. Dangut M David., Skaf Z., Jennions IK. An integrated machine learning model for aircraft components rare failure prognostics with log-based dataset. ISA Transactions. Elsevier Ltd; 2020; (xxxx). Available at: DOI:10.1016/j.isatra.2020.05.001
3. Wang Y. Strategies for Aircraft Using Model-Based Prognostics. 2018;
4. Shang J., Mingyun G., Yijing L., Bing G., Yuanyue H., Haixiang G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications. Elsevier Ltd; 2016; 73: 220–239. Available at: DOI:10.1016/j.eswa.2016.12.035
5. Dangut MD., Skaf Z., Jennions IK. An integrated machine learning model for aircraft components rare failure prognostics with log-based dataset. ISA Transactions. Elsevier Ltd; 2020; (xxxx). Available at: DOI:10.1016/j.isatra.2020.05.001
6. David Dangut M., Skaf Z., Jennions I. Rescaled-LSTM for Predicting Aircraft Component Replacement Under Imbalanced Dataset Constraint. 2020 Advances in Science and Engineering Technology International Conferences (ASET). IEEE; 2020. pp. 1–9. Available at: DOI:10.1109/ASET48392.2020.9118253
7. Branco P., Torgo L., Ribeiro RP. A Survey of Predictive Modelling under Imbalanced Distributions Adapting Resampling Strategies for Dependency-Oriented Data in Imbalanced Domains View project International Workshop on Cost-Sensitive Learning View project A Survey of Predictive Modelling . 2015. Available at: https://www.researchgate.net/publication/275968092 (Accessed: 13 September 2018)
8. He H. Imbalanced Learning. Self-Adaptive Systems for Machine Intelligence. New Jersey: John Wiley & Sons, Inc.,Hoboken, New Jersey.; 2011. 44–107 p. Available at: DOI:10.1002/9781118025604.ch3
9. Rout N., Mishra D., Mallick MK. Handling imbalanced data: A survey. Advances in Intelligent Systems and Computing. 2018. 431–443 p. Available at: DOI:10.1007/978-981-10-5272-9_39
10. Abraham A., Pedregosa F., Eickenberg M., Gervais P., Muller A., Kossaifi J., et al. Hands-On Machine Learning with Scikit-Learn and TensorFlow.pdf. O'Reilly Media; 2014. 568 p. Available at: DOI:10.3389/fninf.2014.00014
11. Camacho-Navarro J., Ruiz M., Villamizar R., Mujica L., Moreno-Beltrán G. Ensemble learning as approach for pipeline condition assessment. Journal of Physics: Conference Series. 2017. Available at: DOI:10.1088/1742-6596/842/1/012019
12. Zhang D., Jiao L., Bai X., Wang S., Hou B. A robust semi-supervised SVM via ensemble learning. Applied Soft Computing Journal. 2018; 65: 632–643. Available at: DOI:10.1016/j.asoc.2018.01.038
13. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine. 2006; 6(3): 21–44. Available at: DOI:10.1109/MCAS.2006.1688199
14. Zhou ZH. Ensemble methods: Foundations and algorithms. Ensemble Methods: Foundations and Algorithms. 2012. 1–218 p. Available at: DOI:10.1201/b12207 (Accessed: 31 January 2019)
15. VanderPlas J. Python Data Science Handbook. O'Reilly. 2016. p. 541. Available at: http://shop.oreilly.com/product/0636920034919.do%0Ahttps://jakevdp.github.io/PythonDataScienceHandbook/05.01-what-is-machine-learning.html
16. Fong Chun Chan. Using Mixture Models for Clustering. Available at: http://tinyheero.github.io/2015/10/13/mixture-model.html (Accessed: 26 May 2019)
17. Batista GEAPA., Prati RC., Monard MC. Balancing Strategies and Class Overlapping. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010. 24–35 p. Available at: DOI:10.1007/11552253_3
18. Visa S., Ralescu A. Learning imbalanced and overlapping classes using fuzzy sets. Workshop on Learning from Imbalanced Datasets II (ICML '03). 2003; (0): 91–104. Available at: https://link.springer.com/chapter/10.1007/978-3-540-24694-7_32

2020-10-26

Aircraft predictive maintenance modeling using a hybrid imbalance learning approach

Dangut, Maren David

SSRN

Maren DD, Zakwan S, Ian KJ (2020) Aircraft predictive maintenance modeling using a hybrid imbalance learning approach. In: TESConf 2020 - 9th International Conference on Through-life Engineering Services, Online, 3-4 November 2020

http://dx.doi.org/10.2139/ssrn.3718065