International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

INDUSTRIAL PIPELINES BREAKDOWN PREDICTION USING MACHINE LEARNING

N. Antony Sophia[1], M.Shrushti[2], A.Rinisha[3], R.Soundarya[4]

[1]Assistant Professor, Dept. of Computer Science and Engineering, Jeppiaar SRR Engineering College, Chennai.
[2,3,4] Final Year Student, Dept. of Computer Science and Engineering, Jeppiaar SRR Engineering College, Chennai.

Abstract - An imbalanced classification problem is an example of problem where the examples distributed across the known classes is biased or skewed. The distribution can vary from a slight to a severe imbalance where there is just one or two example for the minority class whereas millions of examples for the majority class or classes. Imbalanced classifications pose a challenge for predictive modelling as most of the machine learning algorithms used for classification were designed round the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. This can cause a problem as often stated, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class. We proposed a model which handles the imbalanced data and to predict the fault occurrences and their solution to rectify the fault as well as updating the new pipeline failure data in the model. In the industry fault rectification requires much amount of time and effort, finding the error is also a tedious process. This will reduce the cost and time efficiency in the oil and gas production industries.

***Key Words*: Dataset, Model View Controller framework.**

1. INTRODUCTION

It is a known fact that in a given dataset the number of examples given for a particular class may vary sometimes small and sometimes drastically and this can affect the accuracy of the result obtained. In Machine Learning where previous knowledge about a particular event is important needs more accuracy for prediction and for future knowledge. Considering an oil and gas industry the numerous faults occur in the pipelines so keeping track of each of those could be difficult and the number of reports for each type of fault differ.

The proposed model handles the imbalanced data and to predict the fault occurrences and their solution to rectify the fault as well as updating the new pipeline failure data in the model. In the industry fault rectification requires much amount of time and effort, finding the error is also a tedious process. This will reduce the cost and time efficiency in the oil and gas production industries.

2. RELATED WORKS

Feng Bao, Yue Deng , Youyong Kong , Zhiquan Ren, Jinli Suo , and Qionghai Dai "Learning Deep Landmarks for Imbalanced Classification", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 31, NO. 8, AUGUST 2020. [1]
Learning on imbalanced data is a common challenge faced in many practical problems ranging from image processing to biological analyses.
Due to the imbalanced class distribution, methods developed on balanced data can suffer from severe loss of power when handling imbalanced problems. To enable the analysis on imbalanced data, a number of works have been proposed to tackle this imbalance learning problem. These works can be categorized to sampling-based methods and cost-sensitive methods. Sampling-based methods focus on obtaining a balanced data set from the imbalanced data set using oversampling or undersampling techniques. The random oversampling and undersampling are two basic methods in this category. By randomly replicating the minor samples or removing the major samples, the methods derive balanced data sets from the original imbalanced data for the standard methods to process.

S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from 23 imbalanced data," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3573–3587, Aug. 2018. [2]
Proposed a cost-sensitive deep CNN to deal with the class-imbalance problem, which is commonly encountered when dealing with real-world datasets. The approach is able to automatically set the class-dependent costs based on the data statistics of the training set. Analyse was three commonly used cost functions and introduced class-dependent costs for each case. It showed that the cost-sensitive CE loss function is calibrated and guess aversive. Furthermore, the proposed was an alternating optimization procedure to efficiently learn the class-dependent costs as well as the network parameters. The results on six popular classification datasets show that the modified cost functions perform very well on the majority as well as on the minority classes in the dataset."

F. Wu, X.-Y. Jing, S. Shan, W. Zuo, and J.-Y. Yang, "Multiset feature learning for highly imbalanced data classification," in Proc. AAAI, 2017 [3]
In this paper, we are devoted to addressing the highly imbalanced learning problem from the perspective of feature learning. We propose a novel approach named UCML. This is the first attempt towards introducing the idea of MFL into imbalanced learning. We conduct experiments on five highly imbalanced datasets from various application fields. The results demonstrate that UCML outperforms state-of-the-art highly imbalanced learning methods. The experimental results indicate that three important components of our approach are effective. We also find that our approach is more robust to high imbalance ratio.

M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown", *Proc. ICML Workshop Learn. Imbalanced Data Sets II*, pp. 1-8, 2003. [4]
In this paper, we have examined how varying the decision threshold and roc analysis helped with the problem of imbalanced data sets. We also presented evidence suggesting that over-sampling and undersampling produces nearly the same classifiers as does moving the decision threshold and varying the cost matrix. We reported these results for only one data set and for only two classification methods, but the analysis of Breiman et al. (1984) implies that sampling and adjusting the cost matrix have the same effect. Adjusting the cost matrix, in turn, has the same effect as moving the decision threshold. roc analysis let us evaluate performance when varying any of these aspects of the learning method or its training. For future work, we hope to explore further the connections between sampling and cost-sensitive learning for imbalanced data sets. We are also interested whether weighting examples or concept descriptions produces classifiers on the same roc curve produced by moving the decision threshold or varying error costs. For instance, when boosting, are successive iterations producing classifiers on the same roc curve, or generating a series of curves of increasing area? Indeed, roc analysis may be a tool for developing a unified framework for understanding sampling, adjusting costs, moving decision thresholds, and weighting examples from underrepresented classes.

C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5375–5384, Jun. 2016[5]
Class imbalance is common in many vision tasks. Contemporary deep representation learning methods typically adopt class re-sampling or cost-sensitive learning. Through extensive experiments, we have validated their usefulness and further demonstrated that the proposed quintuplet sampling with triple-header loss works remarkably well for imbalanced learning. Our method has been shown superior to the triplet loss, which is commonly adopted for large margin learning but does not enforce the inter-cluster margins in quintuplets. Generalization to higher-order relationships beyond explicit clusters is a future direction to explore. Acknowledgment. This work is partially supported by SenseTime Group Limited and the Hong Kong Innovation and Technology Support Programme..

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002[6]

The SMOTE approach can improve the accuracy of classifiers for a minority class. SMOTE provides a new approach to over-sampling. The combination of SMOTE and under-sampling performs better than plain under-sampling. SMOTE was tested on a variety of datasets, with varying degrees of imbalance and varying amounts of data in the training set, thus providing a diverse testbed. The combination of SMOTE and under-sampling also performs better, based on domination in the ROC space, than varying loss ratios in Ripper or by varying the class priors in Naive Bayes Classifier: the methods that could directly handle the skewed class distribution. SMOTE forces focused learning and introduces a bias towards the minority class. Only for Pima — the least skewed dataset — does the Naive Bayes Classifier perform better than SMOTE-C4.5. Also, only for the Oil dataset does the Under-Ripper perform better than SMOTE-Ripper. For the Can dataset, SMOTE-classifier and Under-classifier ROC curves overlap in the ROC space. For all the rest of the datasets SMOTE-classifier performs better than Under-classifier, Loss Ratio, and Naive Bayes. Out of a total of 48 experiments performed, SMOTE-classifier does not perform the best only for 4 experiments. The interpretation of why synthetic minority over-sampling improves performance whereas minority over-sampling with replacement does not is fairly straightforward. Consider the effect on the decision regions in feature space when minority over-sampling is done by replication (sampling with replacement) versus the introduction of synthetic examples. With replication, the decision region that results in a classification decision for the minority class can actually become smaller and more specific as the minority samples in the region are replicated. This is the opposite of the desired effect. Our method of synthetic over-sampling works to cause the classifier to build larger decision regions that contain nearby minority class points. The same reasons may be applicable to why SMOTE performs better than Ripper's loss ratio and Naive Bayes; these methods, nonetheless, are still learning from the information provided in the dataset, albeit with different cost information. SMOTE provides more related minority class samples to learn from, thus allowing a learner to carve broader decision regions, leading to more coverage of the minority class.

3. PROPOSED SYSTEM

Machine learning in oil & gas industries can be used to improve the capabilities of this growing competitive sector. The technology can also be used to optimize extraction and

deliver accurate models. The future scope includes analyzing the sensor data in the oil and gas industry to increase monitoring of each and every part of the machinery and machine learning approaches used in the predictive maintenance to enhance the life of the pipelines and machinery. Imbalanced data sets in day-to-day applications have a majority class with common instances and a minority class with abnormal or crucial instances. The Synthetic minority over-sampling technique is specifically designed for learning from imbalanced data sets which gives better machines.

4. PROPOSED ARCHITECTURE

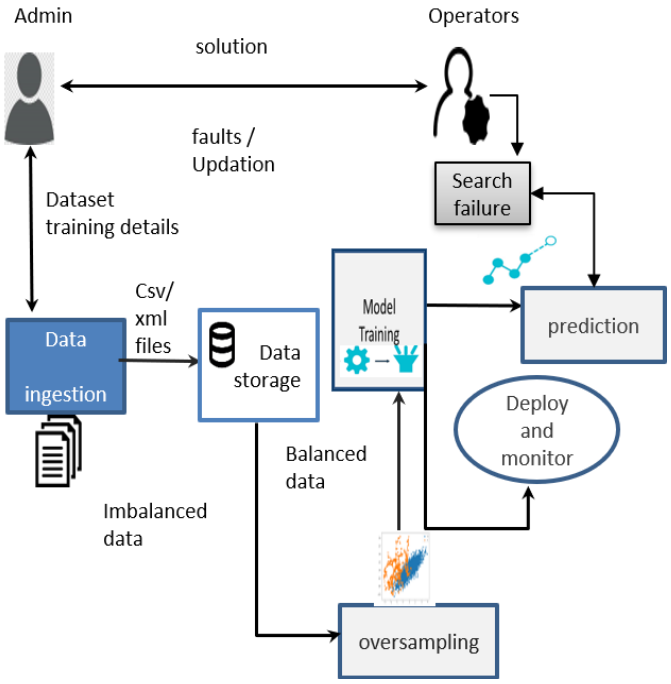

Fig –1: Proposed Architecture

5. MODULES AND DESCRIPTION

5.1 Oversampling

Oversampling technique is used to balance the imbalance dataset. An imbalanced classification becomes a problem when the dataset of the faults were imbalanced, it is difficult to predict the exact cause and therefore there occurs more variations. Thus leads to severe skew in distribution of training data. Due to the variable occurrences of faults which are the count of failures that happened in pipeline breakdowns, as a result training the model become tedious and increase in computational cost. So, the datasets are balanced by equalizing the occurrences from the major and minor set of faults occurred based on summing and reducing the counts by neglecting the minority classes leading to increase in overfitting the minority classes. This makes the predictions more efficient and accurate results.

5.2 Model Training

Model training involves achieving the correct target. In order to the reach the target, the data that are used undergo predictions. The model consist of pattern that maps the input data attributes to the target.

Here the MVC(Model view controller) pattern is used. In Model Training, The data to be trained are uploaded from the database stored in HeidiSql. The data are stored there in the form of CSV/XML files. Those data are preprocessed after fetching. The collected data are trained in such a way to predict the correct failure. The preprocessed data are then passed to balance the imbalanced dataset. Here the input used is the failures data report and new failures gets preprocessed to produce the solutions for errors in the form of text and video files.

5.2.1 Model View Controller

Model View Controller pattern is used here to integrate the application into separate elements: model, view, controller using users, admin and operators. In this scenario, the model represents the operator who works on the failures and solutions during breakdowns. The view represents the users interacting with the training model. Controller represents the important role of admin who maintains, controls and updates the database properly. The controller sends information and commands to the operators regarding the breakdown predictions.

5.3 Training new Data

Training new data involves uploading new failures and saving them. The Admin adds the new failure that occurred into the saved dataset that is present in the database. They update the current dataset. Once after the datasets are balanced, the new data of failure gets appended and trained for prediction. Thus model training is applied on the new fault data. The operator conveys the details such as location of new fault occurred and the type of failure to the admin and admin updates the report with the existing ones and use it for future purpose.

5.4 Predict Failures

In prediction the model is trained model is predicted to produce the desired result. Prediction of failures is that the failure that occurs at any place and type of the failure occurs are recorded and stored. In order to resolve the issues, the operator interact with the model to fetch the error. The corresponding solution for the particular error will be displayed and viewed in the form of text and video files. The sequential steps are shown in the form of text files and demo of the prediction is in the form of videos. The trained dta of

failures are finally predicted to obtain the solution for the failure occurred.

6. RESULTS

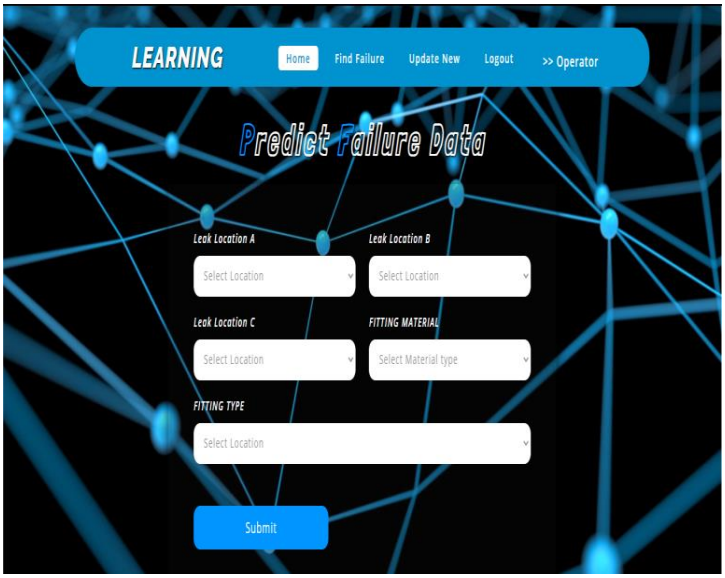

Fig –2: Predict Failure

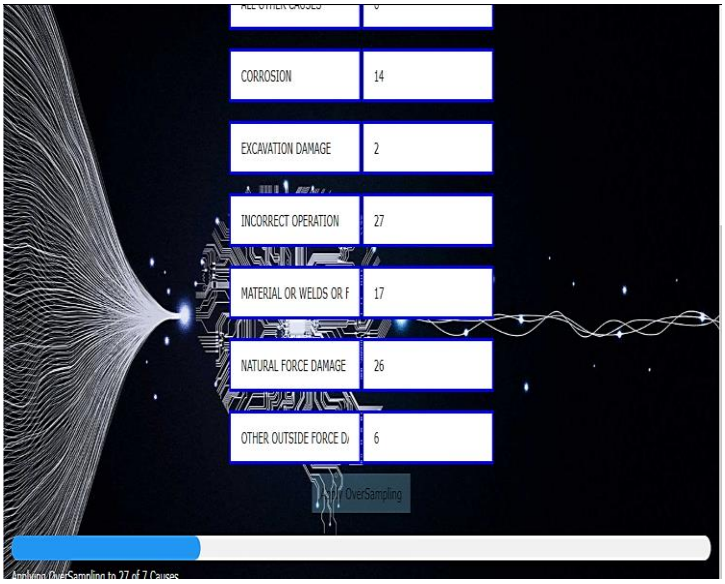

Fig –3: Oversampling

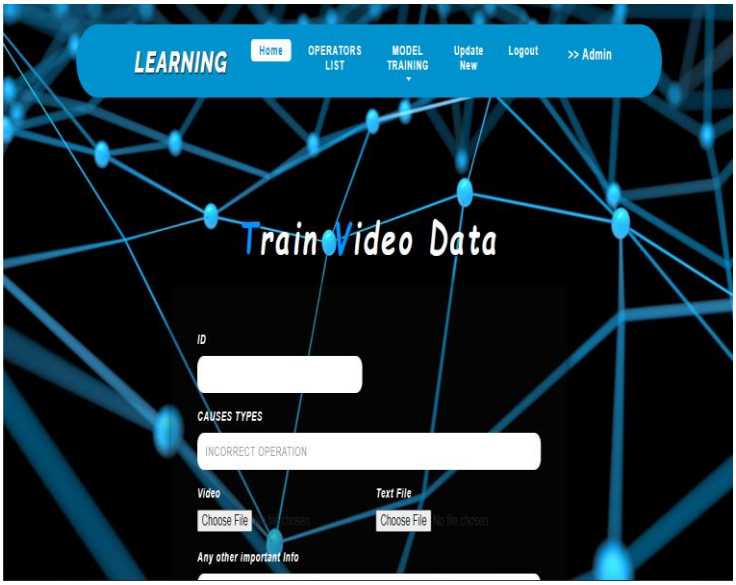

Fig –4: train data solution

7. CONCLUSION

Using machine learning in oil & lamp gas industries will allow skilled workers to become more efficient. It also saves them a lot of time from conducting unwanted tasks. However, workers must adapt these skills very soon. Workers who can properly master all these skills will become stable and permanent. They will also be better served by machine learning algorithms that are run by standardized and quality data. This will yield the best possible results.

REFERENCES

[1] Feng Bao, Yue Deng , Youyong Kong , Zhiquan Ren, Jinli Suo , and Qionghai Dai "Learning Deep Landmarks for Imbalanced Classification", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 31, NO. 8, AUGUST 2020.

[2] Y. Deng, F. Bao, Q. Dai, L. F. Wu, and S. J. Altschuler, "Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning," Nature Methods, vol. 16, pp. 311–314, Mar. 2019.

[3] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[4] F. Bao, Y. Deng, Y. Zhao, J. Suo, and Q. Dai, "Bosco: Boosting corrections for genome-wide association studies with imbalanced samples," IEEE Trans. Nanobiosci., vol. 16, no. 1, pp. 69–77, Jan. 2017.

[5] S. Ando and C. Y. Huang, "Deep over-sampling framework for classifying imbalanced data," in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases. Springer, 2017, pp. 770–785.

[6] F. Wu, X.-Y. Jing, S. Shan, W. Zuo, and J.-Y. Yang, "Multiset feature learning for highly imbalanced data classification," in Proc. AAAI, 2017

[7] Mary Harin Fernandez .F and Ponnusamy .R, "An Efficient Technique for the Automatic Generation of Ontology Based Students Behaviour Using Optimal Fuzzy-GSO Algorithm", Journal of Computational and Theoretical Nanoscience, ISSN: 1546- 1955 (Print), EISSN: 1546-1963 (Online), Vol. 14, No. 9, pp. 4488-4495, 2017.

[8] Mary Harin Fernandez .F and Ponnusamy .R, "Data Preprocessing and Cleansing in Web Log on Ontology for Enhanced Decision Making", Indian Journal of Science and Technology, ISSN (Print): 0974-6846, ISSN (Online):0974-5645, DOI: 10.17485/ijst/2016/ v9i10/88899,Vol.9,No.10, pp. 1-10, 2016.

[9] Mary Harin Fernandez .F and Ponnusamy .R., "A Novel Analysis and Prediction of Students‟ Behaviour using Semantic Similarity-Based Improved J48 IL Algorithm in Personalized Library Ontology", International Journal of Intelligent Engineering and Systems, Print ISSN : 2185310X ,Online ISSN : 2185-3118, Vol. 11, No.5, pp.173-182, 2018.

[10] Mary Harin Fernandez .F and Ponnusamy .R, "Ontology-based modeling student learning behaviour analysis in digital library domain knowledge using markov chain and GUHA", Seventh International Conference on Advanced Computing, Organized by MIT campus, Anna University, Chennai, Tamil Nadu, India, ISBN No.978-1-5090-1933-5, DOI: 10.1109/ ICoAC.2015. 7562789, pp. 1-6,2015.

[11]Mary Harin Fernandez .F and Ponnusamy .R., "Decision Making And Analyzing Ontology From Ontology Log Data Using Description Logic" , International Conference on Advanced Communication, Control and Computing Technologies, Organized by Syed Ammal Engineering College, Ramanathapuram. ISBN No. 978-1-4799-3914-5, pp. 629-633, 2014.