ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse031022021.pdf
https://doi.org/10.30534/ijatcse/2021/031022021

A Comprehensive Analysis of Handling Imbalanced Dataset

Adeoti Babajide Ebenezer[1], Boyinbode O.K (PhD)[2], Oladunjoye Michael Idowu[3]
[1]The Federal University of Technology Akure, Ondo state Nigeria, jide022003@yahoo.com
[2]The Federal University of Technology Akure, Ondo state Nigeria, okboyinbode@gmail.com
[3]The Federal University of Technology Akure, Ondo state Nigeria, michaeloladunjoye.mo@gmail.com

ABSTRACT

Classification is a major obstacle in Machine Learning generally and also specific when tackling class imbalance problem. A dataset is said to be imbalanced if a class we are interested in falls to the minority class and appears scanty when compared to the majority class, the minority class is also known as the positive class while the majority class is also known as the negative class. Class imbalance has been a major bottleneck for Machine Learning scientist as it often leads to using wrong model for different purposes, this Survey will lead researchers to choose the right model and the best strategies to handle imbalance dataset in the course of tackling machine learning problems. Proper handling of class imbalance dataset could leads to accurate and good result. Handling class imbalance data in a conventional manner, especially when the level of imbalance is high may leads to accuracy paradox (an assumption of realizing 99% accuracy during evaluation process when the class distribution is highly imbalanced), hence imbalance class distribution requires special consideration, and for this purpose we dealt extensively on handling and solving imbalanced class problem in machine learning, such as Data Sampling Approach, Cost sensitive learning approach and Ensemble Approach.

Key words: Accuracy paradox, Cost sensitive, Dataset, Ensemble, Imbalanced, Machine Learning, Sampling.

INTRODUCTION

In recent times Learning form imbalanced datasets is another worrisome area for machine learning engineers and data scientist, certain critical real-world situations warrants class Imbalance particularly when gathering datasets to use for analysis, examples of these real-life situations includes facial recognition, Anomaly detections, oil spillage detection in satellite images, medical diagnosis, genetics and genome engineering and failure detection.

A dataset is said to be imbalanced if a class we are interested in falls to the minority class and appears scanty when compared to the majority class, the minority class is also known as the positive class while the majority class is also known as the negative class.

Imbalance refers to the ratio of disproportion or variance that exists in class categories in a dataset here the number of positively influenced examples in a particular datasets is less than the negatively influenced examples in a dataset. Imbalanced learning takes place whenever some types of data distribution frequently dominate the instance space compared to other distributions. Considering a

binomial classification problem, the class with larger percentage data is called the majority class while the other class with scanty data is called the minority class. [2]

The margin between the majority class and minority class is most times so wide and so magnanimous and this is called the level of imbalance. (Wu et al, 20008).

The level of imbalance is predominantly high in fraud detection systems with order of 100 to 1 while Other applications such as genetics and genome engineering has up to 100,000 to 1 level of imbalances. (Provost & Fawcet, 2001).

The high level of accuracy achieved when imbalanced datasets are directly used in predictive or classification tasks is meaningless because the minority class has been displaced with the majority class and so accuracy is always above 96%, meanwhile such accuracy is a fluke, for instance, in a typical fraud detection system where an accuracy of 99.9% is achieved, the solution is useless because no fraud will be detected because such trivial solution that predicts all unseen instances to belong to the majority class, this is a result of the high level of imbalance in fraud datasets [2]. This scene of fraud detection system with imbalanced dataset may land the financial institutions in trouble because wrong individuals will be accused of fraud because the False Negative rate will be high i.e. those that are wrongly accused of fraud.

The academic and industry community has done a thorough developmental work on the study of theory and practical applications of imbalanced datasets; in lieu of these several workshops, conferences and seminars has been organized. The first workshop organized to showcase the importance and needs to learn from imbalance dataset is the American Association for Artificial Intelligence workshop on Learning from Imbalanced dataset (AAAI'00) whose main aim includes observations of many application domains dealing with imbalanced datasets, and several important issues, such as how to evaluate learning algorithms, what evaluation measures should be used, one class learning versus discriminating methods, discussions over various re-sampling methods, discussion of the relation between class imbalance problem and cost-sensitive learning, the goal of creating classifiers that performs well across a range of costs and so on [17].

The second organized workshop was the International Conference on Machine Learning workshop on Learning form Imbalanced datasets (ICML'03) whose objective was guided by the first workshop, here the ROC or cost curves were used as evaluation metrics, rather than accuracy [17].

Lately Association of computing Machinery (ACM) A Special Interest Group on Knowledge Discovery and Data Mining exploration (ACM SIGKDD Explorations'04) happens to the third workshop, whose main focus was on how to bring something meaningful from imbalance dataset,

which addresses data sampling feature extraction and one-class learning, using boosting algorithm combined with other over-sampling approaches to make a better model [17].

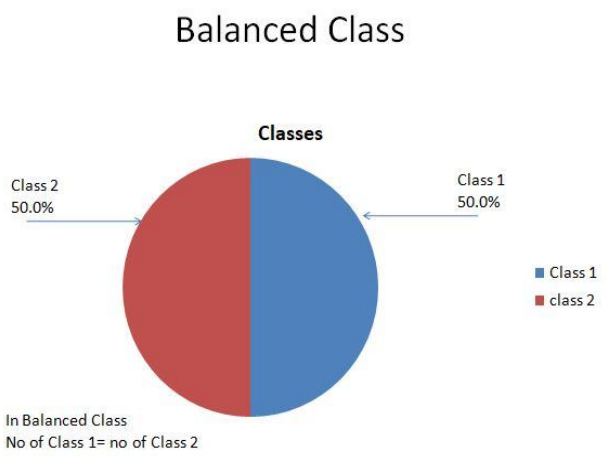

Figure 1: A visual description of a balanced dataset

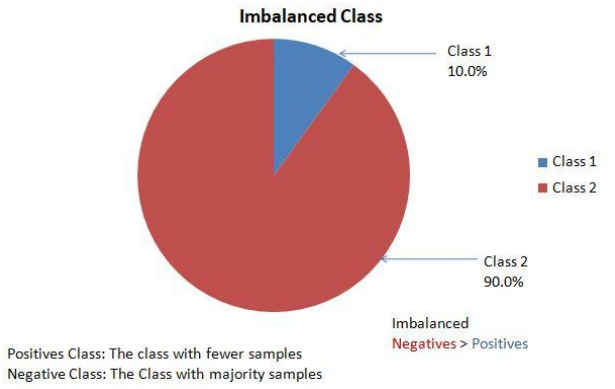

Figure 2: A visual description of Imbalanced dataset

Imbalanced dataset are risky to be subjected directly too machine learning models because the overpopulated samples could skew the outcome of the model in favor of the overpopulated samples, so for this reason imbalanced data are risky to be fed into model directly hence the outcome of such could be deceitful.

Consider a binary classification where one class has less than 3% of the dataset, while second class has 93% of the entire dataset, if this datasets is not resample the outcome from such model will be poor. Some red spots worthy to note are:

i. the use of classification accuracy

ii. its risky to fit the model on raw imbalanced data

iii. The use of standard Algorithm may leads to inaccuracy

Certain pitfalls to avoid when handling imbalanced dataset

Overfitting: Is a phenomenon in machine learning whereby a complex model is trained on too few datasets and model becomes acquitted with the training data, resulting into poor performance on other data.

It is crucial that learning task must suffer from class imbalance, if the majority class of the dataset is more important than the minority class, it is not a problem for the majority class to dominate the learning process [2]. It is when the minority class is significant or it cannot be sacrificed to the dominance of majority class, then it becomes necessary to learn from class imbalanced. Therefore, there is always an assumption in class imbalance learning that the minority class has higher cost than the majority class (Zhi-Hua Zhou) [20][21].

The Performance of machine learning algorithms is mostly evaluated using predictive accuracy, however this is inappropriate in imbalance class learning, there's this assumption that the imbalanced learning that the minority class has higher cost than the majority class. A simple predictive accuracy is not enough in situation where there is imbalanced class.[20][2]

STRATEGIES OF HANDLING CLASS IMBALANCED DATASET

1. **Data Sampling Approach:** Subjecting imbalanced dataset to machine learning models directly without been manipulated could cause a misleading performance inference, the reason being that classification dataset has an imbalanced class distribution. Class distributions plays a vital role on machine learning process, once the class distribution is not balanced the minority distribution will be ignored to favor the majority class especially in a scene of high level of imbalance. To avoid a misleading performance when working on imbalanced dataset Data Sampling Approach became necessary [5][2].
   Data Sampling Approach makes use of either under-sampling or oversampling method or both methods to even-up the class differences, in most times sampling approach combine techniques used in transforming a training dataset in order to balance-up the class distribution. Once this dataset is balanced standard machine learning models can be used to train the artificially balanced datasets.

Note Data Sampling can be performed on the training dataset and not on the testing datasets, it's worth knowing that the reason for data sampling is to address the problem of relative class imbalanced in the training dataset and ignoring the underlying cause of the imbalance from the problem domain.
There are two ways to data sampling are:

a. Under-sampling
b. Over-sampling

a. Under-sampling approach is a sampling methodology that selectively depopulate the majority class, but ensure that certain dataset holding sensitive information about the majority class are retained and preserved while the depopulation is takes.
   There are several ways in which under-sampling can be implemented so as to improve the performance of this sampling approach.

i. **Random Under-sampling** is a non-heuristic way of re-sampling whereby the majority class is depopulated at random until it is equal to that of the minority class, the major disadvantage of this approach is it could result into loss of valuable data [27].

Having listing various flavors of under-sampling approaches to imbalanced dataset, under-sampling has its pro's and con's.

Merits of using under-sampling approach:

- It is effective for large-scale applications where the distribution of the majority class is massive.
- It reduces the duration for training dataset and storage [2].

Demerits of Under-sampling are:

- Few sensitive data can be lost when reducing the majority class and trying to balance the class distribution [2].

b. **Over-sampling** is a sampling approach where samples of the minority class are frequently replicated till a balanced class distribution is achieved [22].

This methods has two disadvantages [7].

i. **Random Over-sampling.:** is an approach to oversampling where the minority class are randomly replicated so as to have a balanced class distribution.

Disadvantages of Random Over-sampling

a. It will increase the likelihood of occurring overfitting, since it replicate d samples of the minority class [7].
b. Over-sampling makes learning process more time-consuming if the original dataset is already fairly large but imbalanced [7].

c. **Synthetic Minority Oversampling Technique (SMOTE).** Is an over-sampling approach to data sampling that works by creating synthetic

data samples in relation to the feature space to over-sampled the minority class distribution.

In SMOTE the minority class is over-sampled by taking each minority class datasets and replacing them with synthetic datasets.

The synthetic data sets are close in the feature space by drawing a line between the samples in the feature space and drawing new samples as a point along that line. SMOTE approach has been inspired by a technique that proved successful in handwritten character recognition [4].

According to (Chawla et al, 2002) SMOTE generates synthetic examples by randomly interpolating between a minority class examples and one of its neighbors from the same class, Data cleaning techniques such as the Tomek-link can be applied further to remove the possible noise introduced in the interpolation process[5].

Several methods have been developed lately to improve the performance of SMOTE algorithm such as dealing with nominal features, they are: **SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous)** and **SMOTE-N (Synthetic Minority Over-sampling Technique Nominal)**. These methods can be considered as a generalization of the original SMOTE algorithm to handle data sets with mixed features (Continuous and nominal) [11].

d. **Adaptive Synthetic Sampling (ADASYN):** is a better alternative to SMOTE that generates synthetic samples, inversely proportional to the density of the datasets in the minority class. It is designed to create synthetic samples in regions of the feature space where the density of minority samples is low, and fewer or none where the density is high [6].

In ADASYN, the amount of synthetic samples (of the minority class) to be generated is determined by the density distribution r^i.

ADASYN principle works by generating the minority data samples intelligently with respect to their ratio of distributions from the entire datasets [6].

ADASYN method does not only reduce the learning bias introduced by the original imbalance data distribution, but can also intelligently move the decision boundary to focus on those difficult to learn samples. [6]

2. **Cost-Sensitive Learning Approach:** Is a different approach used for solving class imbalance problems in machine learning, In Inductive Learning approach to classification, all the classification algorithms have different misclassification errors; or they implicitly assume that all misclassification errors cost equally. In real-world scenario, this assumption is wrong, for instance in deciding whether to grant a Bank customer a Bank Loan or not. Consider a situation where bank wants to know whether to give his/her customer loan or not. Denying a loan to a good customer is not as bad as giving a loan to a bad customer that may never repay it.

Cost-sensitive learning makes use of cost-matrix for different types of errors or instances to facilitate learning from imbalanced datasets. Cost-sensitive learning does not modify the imbalanced data distribution directly; instead, it targets this problem by using different cost-matrices that describes the cost for misclassifying any particular data samples [3]. A cost-sensitive learning technique takes costs, such as misclassification cost into consideration during model construction and produces a classifier that has the lowest cost.

Decision making by using a Cost Matrix.

Let $C(i,j)$ denote the cost of estimating an example from class i as class. In a two class problem, $C(+ve,-ve)$ denotes the cost of misclassifying a positive sample as the negative

sample, and $C(-ve, +ve)$ denotes the cost of the contrary case. Cost-sensitive learning technique take advantage of the fact that it is more expensive to misclassify a true positive instance than a negative instance, that is, $C(+ve, -ve) > C(-ve, +ve)$. For a two-class problem, a cost sensitive learning method assigns a greater cost to false negatives than to false positives, hence resulting in a performance improvement with respect to the positive class [14].

Mathematically, assuming that $(\boldsymbol{i}, \boldsymbol{j})$ entry in a cost matrix $\boldsymbol{C}$ be the cost of predicting class $\boldsymbol{i}$ when the true class is **j**. if $\boldsymbol{i = j}$ then the prediction is right, while if $\boldsymbol{i \neq j}$ the prediction is wrong. The optimal prediction for an example x is the class $\boldsymbol{i}$ that minimizes

$$\boldsymbol{L(x, i) = \sum_{j} P(j|x) C(i, j)}.$$

Costs are not necessarily monetary. A cost can be measured timewisely, or the severity of an illness for each $\boldsymbol{i}$, $\boldsymbol{L(x, i)}$ is a sum over the alternative probabilities for the true class of $\boldsymbol{x}$. in this framework, the role of a learning algorithm is to produce a classifier that for any example $\boldsymbol{x}$ can estimate the probability $\boldsymbol{P(j|x)}$ of each class $\boldsymbol{j}$ being the true class of $\boldsymbol{x}$. For an example ***x,*** making the prediction $\boldsymbol{i}$ means acting as if $\boldsymbol{i}$ is the true class of ***x.*** [13]

Table 1: Cost Matrix

	Actual Negative	**Actual Positive**
Predicted Negatives	True Negative $\boldsymbol{C(0,0)}$	False Negative $\boldsymbol{C(0,1)}$
Predicted Positives	False Positive $\boldsymbol{C(1,0)}$	True Positive $\boldsymbol{C(1,1)}$

True Positive (TP) or F++ refers to the positive sets of data in the dataset correctly predicted as positive data by the classifier

False Negative (FN) of F+- refers to the positive sets of data in the dataset wrongly predicted as negative data by the classifier

False Positive (FP) or F-+ refers to the negative sets of data in the dataset that is wrongly predicted as positive data by the classifier

True Negative (TN) or F-- refers to the negative sets of data in the dataset that is truly predicted as negative data by the classifier.

A Cost Matrix is a table with rows and columns that assigns cost to each cells, just like a confusion matrix, cost matrix assigns a cost to each cell. $\boldsymbol{C()}$ indicates the cost

A learning process may involve various costs such as the test cost, teacher cost, intervention cost (Turney, 2000). The popular used cost-sensitive learning is the misclassification cost.

Misclassification cost can be broadly divided into two types.

i. Example dependent cost

ii. Class dependent cost

Example dependent cost: assumes that the costs are associated with examples, that is, every example has its own misclassification cost; Example dependent cost can also be called Test Cost, it assumes that acquiring a certain feature is connected with a given cost.

The example dependent cost approach to cost-sensitive learning aims at creating a classifier that obtains the best possible predictive performance, while utilizing features that can be obtained at lowest possible cost [18]. It can be seen as a multi-

objective learning, where we try to strike a balance between performance and cost of used features. In most time, the more costly features offers, the higher the predictive power, leading to a problem of whether to use several cheaper features or few more expensive ones. This can also be viewed as a feature selection task, but many cost-sensitive classifiers (e.g., decision trees) have the cost optimization procedure inbuilt [17].

Class dependent cost: assumes that the costs are associated with classes, that is, every class has its own misclassification cost. Its cost-sensitive learning approach that aims to train a classifier in such a way that it will focus on classes that have higher costs assigned to them. They can be seen as priority ones and we want to influence the training procedure by treating them differently.

Over the years cost-sensitive learning has gained most attention for problems with skewed class distributions [19], it is also often used in balanced scenarios, where incorrect classification outcomes may lead to severe consequences. [19]

3. Ensemble Methods for solving Imbalanced Problems

Ensemble learning is another unique way for solving class imbalance problems in machine learning, where several classifiers are trained with the imbalance dataset and the outcome from the multiple classifiers are combined to draw a new result

Traditionally there are two approaches to ensemble learning they are:

i. Boosting

ii. Bagging

Ensemble leverage on power of multi-modular base classifiers that learned on different subsets of the training data to improve on the classification performance over traditional classification algorithm. Few advantages of using ensemble approach are

i. The more the classifier, the less the rate of error, which implies the model becomes more accurate

ii. Multiple classifiers perform better than a single classifier.

Few ways of implementing Ensemble approach are: Boosting, AdaBoost, Random Subspaces, Random Forests and Bagging

AdaBoost is one of the effective boosting algorithm ensemble methods in Machine Learning, where several chunks of weak classifiers are summed together so as to create a better and effective classifier with high accuracy.

How AdaBoost works goes thus:

- AdaBoost picks subsets training data in a stochastic way.
- The selected models are repeatedly trained with a more accurate result from the pool of previous predictions from the model
- AdaBoost focus on sets of samples that were previously misclassified and such samples are assigned with greater weights than the rightly classified samples, Note: misclassified samples will get highest classification probability in the next iteration.
- weights are assigned to individual base learners with respect to overall predictive performance at each iteration
- The higher the accuracy of a classifier the higher the weight assigned to it
- The entire process gets terminated when all the training data are classified correctly or when it gets to a certain threshold of a maximum number of estimators.

- Lastly a consensus vote is taken from all participating classifiers [28].

MetaCost (Domingos, 1999). Meta Cost is an ensemble method that makes use of groups of decision tree classifiers by bagging to estimate the posterior probability p(y|x). It begins to learn an internal cost-sensitive model then estimates class probabilities using bagging and then re-labels the training examples with their minimum expected cost classes, and finally relearns a model using the modified training set. [16]

Boosting.

Boosting: The boosting Algorithms (Leskovec and Shawe-Taylor, 2003), improve performance of weak classifiers by forcing the learners to focus more on difficult examples. Boosting algorithms have been adapted to address the problem with minority classes. [7]

Bagging.

The word Bagging is derived from Bootstrap and Aggregating, Breiman brought the idea of bootstrap & aggregating to build an ensemble of classifiers that is stronger than individual comprising units.

In Bagging each classifier is trained using a generating sample of datasets usually called (bootstrap samples) taken from the original dataset as a replacement so as to ensure a enough amount of instances per classifier each samples usually contains the same number of instances as the original dataset. Hence, diversity is obtained with the re-sampling procedure by the usage of different data subsets.

Some of the original instances are likely to appear more than once when training the same classifier while other instances may not appear at all [28].

Finally, when an unknown instance is presented to each individual classifier, a majority or weighted vote is used to infer the class.

There are many variations of bagging algorithms.

IBA Improved bagging algorithm. The re-sampling process is redefined by marking each sample with information entropy (IBA; Jiang, Li, Zheng, & Sun, 2011), empirical study revealed that IBA can improve bagging generalization in certain machine learning challenges.

The Online Bagging and Boosting algorithm is another variation of boosting that execute quickly and shield bagging's level of accuracy (Oza, 2005)

Wagging also known as weight aggregation assigns weight randomly to classifier. (Weight aggregation; Bauer & kohavi, 1999)

Double Bagging is a variation of boosting that train two classifiers in each iteration using out-of bag examples (Hothorn & Lausen)

Bagging algorithm is good for reducing over-fitting in order to create strong learners for generating accurate predictions, unlike boosting, bagging allows replacement in the bootstrapped sample. [16]

CONCLUSION

In this paper we've looked into Imbalanced dataset and how it can be handled whenever encountered in the course of solving machine learning problems, however we specifically delve into strategies of handling class imbalanced from the data sampling approach, to cost sensitive and finally to ensemble methods.

Working directly with Imbalanced dataset may cause the model to suffer Overfitting; a situation whereby the interested class; usually the minority class lacks enough data to train the model. To avoid overfitting during data preparation stage, there are several data sampling strategies that can

be deployed, the selection depends on the nature of the project and preferred model to train the dataset. Generally data sampling method handles the imbalance dataset either by using undersampling or oversampling strategies so as to balance up the class distribution.

The Synthetic Minority Oversampling Technique (SMOTE) and ADASYN (Adaptive Synthetic Sampling) are advance sampling strategies that adjust the class imbalance in a more intelligent way by creating a synthetic samples, in ADASYN the amount of synthetic samples generated is determined by the density distribution while in SMOTE oversampling is done by synthetic generating data samples with respect to the feature space in order to oversample the minority class distribution.

The cost sensitive learning approach takes misclassification cost into consideration during model construction and is goal is to reduce the total cost and it modifies class imbalance at the algorithm level.

Lastly the ensemble method is an efficient way of handling class imbalance, an approach that leverages on multi-modular base classifiers that learned on different subsets of the training data its benefit is because multiple classifiers performs better than a single classifier.

REFERENCES

1. M. Palt, M. Bach, and A. Werner. (2019). **The Proposal of Under-sampling Method for Learning from Imbalanced Datasets**. *23rd International Conference of Knowledge-Based and intelligent Information & Engineering Systems*
2. https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/ (Accessed: 5 October, 2020).
3. S. Lam, Phung, and A. Bouzerdoum. (2009) "**Learning Pattern classification tasks with imbalanced data sets**". *Faculty of Engineering and Information Sciences, University of wollongong.*
4. N. V. Chawla, L. O. Hall, K.W. Bowyer, and W. P. Kegelmeyer. **SMOTE: Synthetic Minority Over-sampling Technique**. Journal of Artificial Intelligence Research, 16 2002, pp. 321-357.
5. Machine Learning & Pattern Recognition Series, Series Editor Ralf Herbrich and Thore Graepel (2012) Taylor & Francis Group , LLO.
6. H. Haibo, B.Yang, A. Edwardo, Garcia, and L. Shutao. (2008) **ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning** International Joint Conference on Neural Networks (IJCNN 2008). 2008, pp. 1322-1328
7. S. Kotsiantis, P. Pintelas, and D. Kanellopoulos. **Handling imbalanced datasets: A review** GESTS International Transactions on Computer Science and Engineering, Vol. 30, 2006, pp. 25-36.
8. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/ (Accessed: 17 October, 2020).
9. https://aidevelopmenthub.com/tour-of-evaluation-metrics-for-imbalanced-classification/ (Accessed: 23 September, 2020).
10. J. Laurikkala. **Improving Identification of Difficult small classes by Balancing Class Distribution.** University of Tampere, Department of computer & Information Sciences series of Publications A A-2001-2, April 2001.
11. A.T. Elhassan. M. Aljourf. Al-Mohanna, & M. Shoukri. **Classification of Imbalance Data Using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method**. 2017
12. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-

algorithm-in-python/ (Accessed: 17 September, 2020).

13. C. Elkan, "**The Foundations of Cost-Sensitive Learning."** Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI'01), 2001, pp. 973-978.
14. Jason Brown. **Imbalanced Classification with Python, Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning**. 2020 Edition v1.2, pp.35-95
15. B. Mohammed, K.D Hassiba, and A.A Taklit, (2013). **Evaluation Measures for Models Assessment over Imbalanced datasets**. Journal of Information Engineering and Applications ISSN 2224-5782 2013, Vol.3, No. 10
16. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. (2011). "**A Review of Ensembles for the Class Imbalance Problem: Bagging, Boosting and Hybrid-Based Approaches.** *IEEE Transactions on Systems, Man and Cybernetics Part C:* Applications and Reviews. 2011.
17. X. Gou., Y. Yin. C. Dong. G. Yang. and G. Zhou. (2008). **On the Class Imbalance Problem**., China Fourth International conference on Natural Computation.
18. Q. Zhou, H. Zhou, and T. Li. **Cost-sensitive feature selection using random forest: selecting low cost subsets of informative features**. Knowledge. Based System. 95, 2016, pp. 1–11.
19. X. Liu, and Z. Zhou. **The influence of class imbalance on cost-sensitive learning: an empirical study.** In*:* Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), Hong Kong, pp. 970–974, 18–22 Dec 2006.
20. Z.H. Zhou. **Ensemble Methods Foundation and Algorithms** *CRC Press, Taylor & Francis Group. 2012, pp. 159-200.*
21. Z.H. Zhou, and X. Y. Liu. **On Multi-Class Cost-Sensitive Learning**, *Computational* Intelligence, 2010, Volume 26, Number 3.
22. H.T Ryan, and N.V Chawla. **Imbalanced Dataset: From Sampling to Classifiers**. Department of Computer Science and Engineering, The University of Notre Dame, Notre Dame IN USA 2014.
23. Z. H. Zhou. (2011). **Ensemble Learning**. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China.
24. S. Giovanni, and J. Elder. **Ensemble Methods in Data Mining Improving Accuracy through combining Prediction** *A Publication in the Morgan & Claypool Publishers series Synthesis Lectures On Data Mining And Knowledge Discovery.*
25. A. Fernandez, S. Garcia, M. Galar, R.C. Prati, B. Krawczyk, and F. Herrera. **Cost-Sensitive Learning from Imbalanced Datasets**. Springer Nature Switzerland AG 2018, 63-78. doi:10.1007/978-3-319-98074-4_4.
26. N. Japkowicz. **Assessment Metrics for Imbalanced Learning. Imbalanced Learning: Foundations, Algorithms, and Applications** The Institute of Electrical and Electronics Engineers, John Wiley & Sons, Inc. 2013, Ch 8.
27. M. Galar, A. Fernandez, E. Barrenechea, and H Bustince. **A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approaches** IEEE Transactions on Systems, man and Cybernetics PART C: Applications and Reviews. 2011.
28. O. Sagi, and L. Rokach. **Ensemble learning: A survey. Wiley Interdisciplinary Reviews:** Data Mining and Knowledge Discovery, 8(4), e1249. doi:10:1002/widim. 1249