International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

Comprehensive Assessment of Imbalanced Data Classification

Smita Nirkhi, Shashikant Patil

Abstract: ***This is an attempt to address the various challenges opportunities and scope for formulating and designing new procedure in imbalanced classification problem which poses a challenge to a predictive modelling as many of AI ML n DL algorithms which are extensively used for classification are always designed from the perspective of with majority of focus on assuming equal number of examples for a class. It leads to poor efficiency and performance especially in minority class. As Minority class is always very crucial and sensitive to classification errors and also its utmost important in imbalanced classification. This chapter discusses addresses and gives novel as well as deep insights with unequal distribution of classes in training datasets. Largely real time and real world classifications are comprising imbalanced distribution so need specialized techniques for more challenging and sophisticated models with minimal errors and improved performance.***

Keywords: Imbalanced Data; Class imbalance, Machine Learning; Algorithms; Data Mining .

I. INTRODUCTION

In several "supervised-learning as well as learning" solicitations, it is always a substantial variance among the "erstwhile likelihoods" of diverse "classes or distributions", i.e. among the likelihoods with that a specimen fits in the diverse "classes" of the categorization query. It is acknowledged as the "class imbalance" query. It is common in many real world challenges and queries from telecommunications, network, finance-world, ecosystem, natural science, medicine not only, and this can be well-thought-out and one of the topmost problem in data mining as of now. Additionally, it is worthy to check that the "minority distribution or class" is characteristically the solitary which have the utmost attention from a "learning and adaptive learning" argument of opinion and it is herewith suggesting a prodigious cost when improperly "categorized or classified".

A glitch associated with "imbalanced distributions of the datasets" is a customary "classification/categorization learning and adaptive learning" procedures are frequently prejudiced in the direction of the "majority or greater class" (acknowledged as the "undesirable and negative distribution or class") and consequently it is a sophisticated mis-classification rate for the "minority or smaller distribution or class" occurrences (termed as the "optimistic" or positive examples). During last two decades so many clarifications have been projected to pact with this tricky, both for customary "learning and advanced learning procedures or algorithms" and for collaborative practices.

They can be characterized into three main clusters:

- Datum sample
- Algorithmic alteration
- Cost-effective learning

Most of the researchers are of the opinion that the performance of numerous customary "classifiers in imbalance" areas are well exposed and shows that the noteworthy forfeiture of enactment is mostly owing to the crooked class or distribution spreading, given by the "imbalance or misbalance ratio (IR)", and it is a distinct as the proportion of the number of occurrences in the "majority or greater class" to the quantity of instances in the "minority or small class". Conversely, there are numerous surveys also propose that there are other aspects that underwrite to such performance deprivation.

There are six noteworthy glitches associated to statistics or data inherent features and it should be occupied as an account in order to deliver improved explanations on behalf of appropriately categorizing two of the classes of a query.

1. Identification of areas with small disjoints
2. The deficiency of compactness and evidence in the training statistics
3. Delinquent of overlying between the categorisation
4. Influence of bad and noisiest statistics of unwarranted purviews
5. Connotation of the doubtful instances to carry out a upright discernment between the +Ve and –Ve categorisation, and its association with noisy examples
6. Conceivable variances in the facts dissemination for the drill and assessment statistics, also acknowledged as the dataset transference.

A comprehensive assessment carried of this peculiar query or problem lets us know nearby the base where the hitches for "imbalanced and misaligned categorization" arises, concentrating on the scrutiny of noteworthy data intrinsic features. Explicitly, aimed at separately well-known situation studies show an investigational sample on in what way it distresses the performance of "erudition procedures", in order to strain its consequence.

Revised Manuscript Received on April 18, 2020.
* Correspondence Author
Smita Nirkhi1, Computer Engineering, Shri Ramdeobaba College of Engineering & Management, Nagpur. Email: smita811@gmail.com
Shashikant Patil*, EXTC Department,SVKMs NMIMS Shirpur Campus Email: sspatil@ieee.org

There is an argument that few of these areas have modern approaches allied, investigating their effective outcomes as well as contributions. Conversely, it is highlighted that they there is a need of an hour to address theme in more detail in an effort to make models with excellent superiority in specific classification situation. So, we have elaborated them as forthcoming developments of investigation of "imbalanced learning schemes". Conquering these kind of glitches be able to be the strategic key for emergent new approaches that progress the truthful identification of both the "smaller and greater classes". For the development of technology research and innovation efficiently, we must begin with a background and need of ground that what exactly the query of imbalanced data sets extant to AI ML and DL systems. How and when should imbalanced data sets be worrisome? At what time this issue is merely an erroneous how to sort out rectified in design adoptions? Precisely, we would like to address and discuss what the problem is not. Most of the studies and literatures illustrating that the imbalanced data set queries and problems are straight forward and easy to fix most of the algorithms and techniques are focusing mainly and broadly on two assumptions that maximizing accuracy and classifier works on same distribution of data. While designing the learning algorithms one must wisely think of weakening of assumptions and robustness of the same. Due to aforesaid two assumptions the results of the classifiers are unsatisfactory and very premature.

If you select the erroneous metric to appraise your prototypes, you are probably to select a poor model, or in the foulest case, be misinformed about the predictable performance of your model. Picking an apt metric is thought-provoking. Largely in pragmatic machine learning, but is predominantly tough for imbalanced classification glitches. Primarily, since most of the customary metrics that are extensively used adopt a well-adjusted class distribution, and since stereotypically not all classes, and consequently, not all forecast errors, are identical for imbalanced classification.

Choosing a model, and even the data preparation approaches together are an exploratory issue that is directed by the assessment metric. Tests accomplished with unalike models and the consequence of each trial can be enumerated with a metric.

There are customary metrics that are broadly used for estimating classification extrapolative models, such as classification accuracy or classification error.

Average metrics work fine on most glitches, which is why they are extensively embraced. Nonetheless all metrics make expectations about the delinquent or about what is significant in the problem. Consequently an assessment metric must be selected that efficiently captures what you believe. It is also significant about the model or forecasts, which makes selecting model assessment metrics thought-provoking

An imbalanced classification query is a specimen of a cataloguing p issues where the spreading of examples across the well-known classes is prejudiced or crooked. The spreading can diverge from a minor bias to a severe disparity where there is single specimen in the minority class for hundreds, thousands, or millions of specimens in the mainstream class or classes.

Imbalanced classifications posture a trial for extrapolative modelling as most of the machine learning processes used for cataloguing were premeditated around the postulation of an equal quantity of specimens for every class. This results in representations that have deprived extrapolative performance, explicitly for the minority class. This is a genuine query because normally, the minority class is more significant and consequently the issues are more subtle to cataloguing errors for the minority class than the majority class.

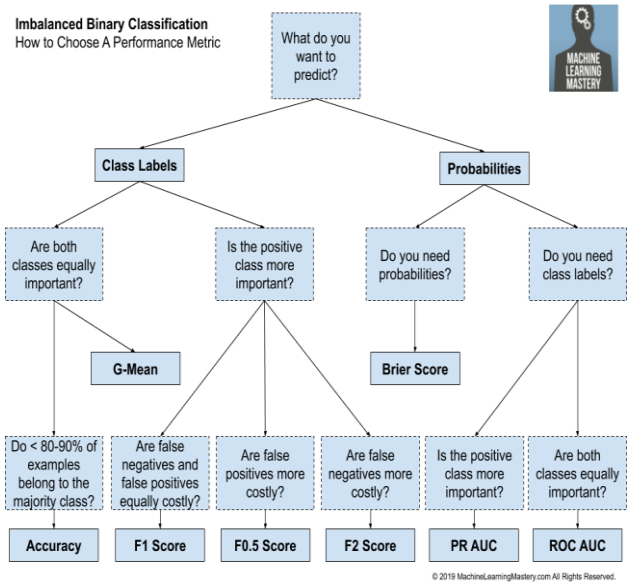


Fig 1 : Classification of Imbalanced Data

Source:https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

In many cases the algorithms are designed in such a way that are made for convenience and can be changed easily. Threshold and setting the threshold is crucial in designing machine learning algorithms. Different classifiers gives different estimates too. The best choice for setting the threshold is 50% on the output. While learning with imbalanced data distributions one or other assumptions will be definitely getting violated due to various reasons. So without adjusting threshold output designing will be definitely a critical mistake.

A common practice of designing algorithms with imbalanced datasets is artificially rebalancing data sets and reorganizing them. It is also called as up sampling and down sampling to replicate the cases and also to ignore the cases. Still such types of techniques are not solving the challenges effectively.

In case of extreme imbalances the situations and type of data sets are very crucial. Many studies shows that it is practically always unbearable to forecast how easy it will be to learn an everyday concept with a multifarious machine-learning algorithm. Fine-tuning the class balance by compulsion is inadequate either to duplicating the smaller class or to chucking away selected of the greater class. The previous does not add evidence and the latter essentially eradicates information. Bearing in mind this fact, the best research approach to focus on how machine learning procedures can pact most efficiently with any type of data given.

Machine learning of misbalanced datasets is a significant issue to be addressed, both practically and for investigation. Evolving a flawless thoughtful of this specific problem will have broader-ranging repercussions for investigation.

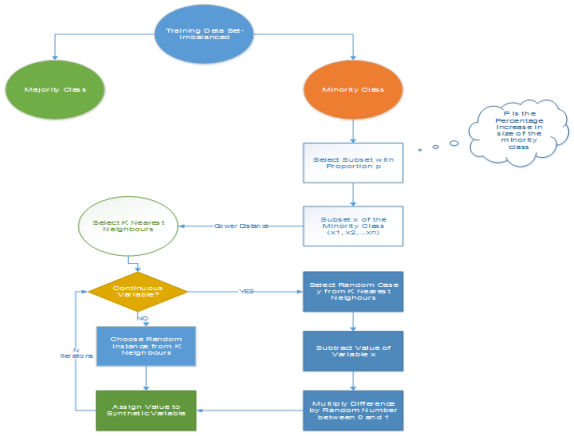


Fig 2 Imbalanced Classification Queries and Issues

The number of specimens that fit in to every class might be stated to as the class categorisation. Imbalanced class categorisation refers to a cataloguing predictive modelling queries and issues where the number of cases in the training dataset for each class label is not well-adjusted. That is, where the class spreading is not identical or nearby to equal, and is in its place prejudiced or crooked. Imbalanced Classification: A categorisation extrapolative modeling tricky where the scattering of specimens across the classes is not identical. For example, we may collect measurements of animals and have 75 examples of one animal species and 25 examples of a second flower species, and only these examples comprise our training dataset. This represents an example of an imbalanced classification problem. An imbalance arises as soon as one or additional classes have very little scopes in the training statistics as related to the new classes.

Encounters tackled with Imbalanced datasets

One of the prevalent uncertain chunks is the enormous statistics and its dispersal. The task is to advance identification of the sporadic minority class as contrasting to attaining sophisticated inclusive truthfulness. Machine Learning procedures have a habit of to yield disappointing results when classifiers confronted with imbalanced datasets. For every unprovoked dataset, if the occurrence to be foretold fits to the marginal class and the occurrence rate is fewer than five percent, it is typically stated to as a sporadic event.

Challenges and issues with typical Machine learning practices

The conservative model assessment procedures do not precisely quantity prototypical enactment when confronted with misbalanced data categorisations.

Customary classifier procedures resembling Decision Tree and Logistic Regression have a prejudice to classes which have number of occurrences. They have a habit of to only foresee the majority class data. The features of the minority class are treated as noise and are regularly unnoticed. Thus, here remains a great likelihood of mis-classification of the smaller class as compared to the greater class.

Assessment of a categorization procedure enactment is restrained by the "Confusion-Matrix technique"which comprises evidence about the real and the forecasted class.

Table I : Confusion Matrix

Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Accuracy of a prototypical = (TN+TP) / (TN+FP+FN+TP)

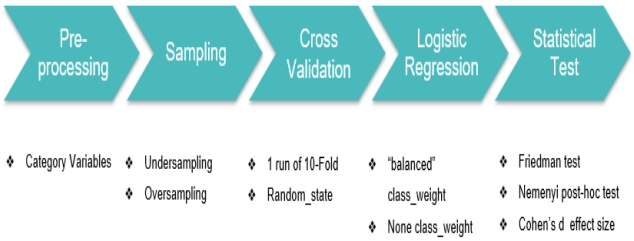
Conversely, when handling unfair classification territory accuracy is never a suitable extent to evaluate model enactment**.** For e.g. a particular classifier which accomplishes an accurateness of 97 % with an event rate of 3% is not truthful, if it categorizes all instances as the popular class and eliminates the 3 % smaller class interpretations as noise.

Samples of imbalanced classes

When we are trying to sort out explicit commercial experiments with imbalanced data sets, the classifiers created by customary machine learning algorithms may not give truthful outcomes. Apart from duplicitous transactions, other specimens of a common commercial problem with imbalanced dataset are:

Datasets to classify client churn where a massive majority of clients will remain same using the amenity. Precisely, Telecommunication corporations where Churn Rate is lesser than 2%. Datasets to recognize sporadic illnesses in therapeutic diagnostics etc. Expected Adversity like Tremors

II. METHODOLOGY



Pipeline for the experiments

In terms of algorithms, one can choose Logistic Regression as it has the hyper-parameter "balanced" class weight, which can be used as an algorithm-level method. Its simplicity and common application is also a big plus.

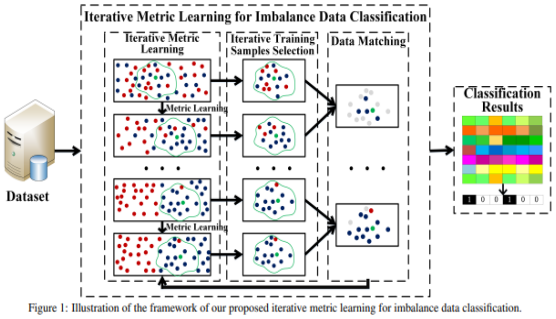


Figure 1: Illustration of the framework of our proposed iterative metric learning for imbalance data classification.

III. LITERATURE ANALYSIS:

In an arena of "Classification with Imbalanced Data and Statistics" has witnessed a blossoming evolution in last two decades. Numerous thought-provoking studies and schemes have been established under this domain; not only the novel approaches, but numerous survey documents, records and noteworthy methodologies for streamlining the "learning" aptitude of "classifiers" in this consequence.

Some of the studies presented "pre-processing", "cost-sensitive-learning" and "ensemble procedures, by deep studies and implementation and tried various methodologies in an "intra and inter-class valuation". Few of the literatures are discussing on concerns associated with usage of data inherent features in misbalanced cataloguing queries. It will definitely lead to advance the contemporary prototypes in regard to the existence of minor disjuncts, the deficiency of density in the training information, the overlying amongst classes, the recognition of data with noise, the implication of the marginal occurrences, and the dataset alteration among the training and the test dispersals. Lastly, numerous tactics and proposed ideas to sort out the query of data misbalance in aggregation with misbalanced information also claiming particular outcomes and results regarding to the behaving nature of the "learning and re-learning procedures and approaches for information with such type of intrinsic characteristics.

Generally and largely "class-misbalance-problem" does not only distress decision-tree structures but also disturbs other classifying and categorizing schemes like "Neural-Networks" and "Support- Vector-Machines".

Many of the contributions articles also focused on to deliver an acute assessment of the type and class of the query, the recent sophisticated techniques, as well as the contemporary investigation metrics used to gauge learning routine underneath the "imbalanced-learning" situation.

A taxonomy for ensemble-based approaches is also suggested to deal with "class-imbalance" where each suggestion shall be characterized subject to the innermost ensemble-procedure in which it is established. One can also think of exhaustive experiential assessment considering the greatest noteworthy available methods, within the families associated with the taxonomy projected, to display whether any of them makes an alteration. Few of the outcomes shows that empirically that "ensemble-based-algorithms" are sensible since they outdo the plain use of pre-processing methods prior to learning the "classifier", thus moderating the upsurge of complication by means of a substantial enrichment of the consequences.

In an attempt to challenges some of the shortcomings in the prevailing cataloguing methods, one of the study proposes a unique "class-specific-cost-regulation" ELM (CCR-ELM) for cataloguing glitches with imbalanced data dispersals, by presenting class-specific regulation cost for misclassification of every class in the performance table, which can diminish the special effects of the number of class trials and even the special effects of the dispersal gradation of the statistics. It must be noted that this tactic is appropriate for both binary classification and multiclass cataloguing unswervingly.

When LLM is used in some of the cases of imbalanced data classification jobs wherein one class (positive class) is extremely under represented related to the other class ("negative class"), it does not work well. It is essential to give a sophisticated reward to the accurate cataloguing of positive class configurations for sorting out imbalanced data grouping jobs. "Laplacian-matrix" centred "loss-function", "Laplacian-least- learning-machine" (L2MM) is also proposed to sort out the infeasibility of the prevailing "least-learning-machine" for "imbalanced-data-classification". This approach saves machine training time. One of the investigational query addressed and assess the performance of the miscellaneous procedures taken into account and subsequent outcomes found demonstrates that there is no single methodology to imbalanced big data classification which outdoes the others for all the data considered when using Random Forest. Furthermore, even in case of similar type of tricky, the finest kind of effective performance technique is reliant on the no. of "mappers" nominated to test the trials. Usually in many of the circumstances, when the "number-of-separations" is enlarged, an enhancement in the testing or run-time can be detected, conversely, this development in times is found at the expenditure of a minor drop in the correctness show. Such kind of decrement in the presentation is connected to the lack of concentration query/problem, which is assessed in some of the works from the imbalanced data point of outlook, as this concern worsens the efficiency of classifications and so of the "classifiers" in the "imbalanced-situation" more ruthlessly than in customary learning.

One of the studies carried a work resolve this problem with possible solution depicted that "weighted-extreme-learning-machine" established on the "ensemble-learning" technique enriches the categorisation efficiency while increasing the run-time of the procedure, beginning from the perception of "multi-core-learning", a "weighted-extreme-learning-machine" centred on combined "kernel-functions" and "reduced-kernel" procedure is projected. To diminish the time consuming approach of the algorithm due to the composite-kernel-approach", a "reduced-kernel" scheme established on the "sub-input-matrix" of the balance activity for class is discussed in one of the study. The suggested approach seems to be good in context to others in some cases such as G-Mean and AUC indexes still needs sincere efforts and solution to explore more proficient and precise parameter optimization will add new dimension to the field as this method has numerous parameters and their optimization needs attention.

In some of the applications wherein multi-mark grouping with imbalanced datum is always there in such cases different studies and researches are carried out and proposed schemes are there. One of the scheme in attempt to improve efficiency focusing on exploiting the "min-max particular" system. In "min-max" measured scheme halts a "multi-mark" query into an evolution of tiny 2 "class-sub-issues", which in turn are capable to be combined by two straightforward approaches. By introducing decay procedures some of the improvements in min-max functions can be seen. This scheme also claims superior to SVM in case of speed. "Very-fast-decision-tree"(VFDT) computing approach aims at reducing computing speed by taking consideration a small portion of data set as it does not required a large data set, a part or portion of data set is fine enough for getting desired outcomes.

Large margin distribution machine many times hints to the inferior detection speed of the "minority class", which controverts to the requirements of high detection speed of the "minority class" in several tangible applications. One of the proposal presumes the association amongst cost-sensitive constraint and in-class finding rate, and projects LCSDM to find well-adjusted detection speed. Investigational outcomes confirms that LCSDM can progressively upsurge the margin dispersal of the minority class to attain an additional well-adjusted detection speed. As a customary learning approach, LCSDM is specifically pertinent to imbalanced statistics classification. Still such methods in near future needs more attention towards the relationship amongst the margin distribution and example distribution.

One more technique named as "enrichment" that uses the data (interpretations) from the exterior "dataset/information" is also projected by few researchers. In that scheme they have considered t tactics to experiment enrichment procedure: (1) by picking interpretations arbitrarily, (2) iteratively selecting interpretations that advance the classification outcome, (3) addition of interpretations that benefit the classifier to govern the boundary among classes better. Such procedure outdoes the prevailing approaches performing, on average, better than the others. The benefit is particularly perceptible for the tiniest data sets, for which prevailing approaches failed, whereas these techniques and approaches accomplished the superlative outcomes. Moreover, it smears equally to e multi-class and binary classification jobs. This also can be combined with other procedures dealing with the class imbalance tricky.

"Class imbalance" and "concept-drift-learning" are two diverse arenas in successive learning which have lately fascinated considerable attention in numerous arenas, "Class imbalance" and "concept drift" can occur either distinctly or simultaneously in a statistical information. Even though a lot of efforts has been done concentrating on the notion "drift problem", the "class-imbalance-problem" and in precise, the amalgamation of the twofold glitches in a statistics are largely uncultivated. MOS-ELM is the leading sequential learning technique to lighten the imbalance query for both binary class and multi-class data streams with concept drift. In MOS-ELM, a new adaptive window methodology is projected for "concept-drift-learning".

IV. DISCUSSIONS AND FINDINGS

Presently it is impossible to relate, compare, contrast, analyse and interpret the approaches as well as various studies carried out in the field, as they are assessed transversely with several kind of datasets with fluctuating intensities of class-imbalance, and consequences are testified with unreliable assessment metrics. Most of the researches and study showing conflicting outcomes, Moreover it is stating that performance is highly dependent on problem complexity, class representation, and the performance metrics reported. Largely, there is a dearth of substantiation which discriminates any single "learning-method" as effective and factual in regard to learning from "class-imbalanced-data", and supplementary trials are essential beforehand are extremely needed prior conclusions can be made.

The analysis shows that conventional "machine-learning" procedures for sorting out class imbalance can be stretched to deep learning replicas with attainment. The study also discovers that almost all exploration in this domain has been engrossed on Artificial Intelligence methods. In spite of a rising call for machine learning solutions, there is very pintsized investigation that accurately appraises machine learning in the perspective of class-imbalance and artificial intelligence. Machine learning from "class-imbalanced-data" is still predominantly understudied, and statistical evidence which compares newly published methods across a variety of data sets and imbalance levels does not exist.

Numerous domains for prospective studies are superficial. Relating the afresh projected approaches to a grander diversity of data sets and class imbalance intensities, equating outcomes with numerous balancing assessment metrics, and recording statistical substantiation will benefit to classify the chosen machine learning approach for future claims encompassing class imbalance. Investigating with machine learning approaches for sorting out class imbalance into machine learning and class scarcity will substantiate treasure to the forthcoming era of deep learning approaches. Finally the research in the area with deep learning is limited and needs attention with more sophisticated approaches and easier schemes. Still additional effectual methods must be explored for defining the parameters for CCR-ELM.

Here we have focused very much on usage of unbalanced information in executing binary classification. To be precise, an assessment of the foremost causes of the disaster of both parametric and nonparametric customary classifiers has been conveyed, and certain novel perceptions have been offered roughly with the things of class imbalance. Undeniably, literature dealing with skewed binary classification has grown-up at a fiery percentage in modern ages, but it has primarily engrossed on recommending refined learning approaches or substitute assessment metrics. In its place, the tricky of high inconsistency of the correctness's estimator has been entirely overlooked. In fact, when the dispersal of the classes is skewed, the predictable replicas accomplish very poorly but, bad approximations of the classifier's performance may lead to ambiguous assumptions about the superiority of the forecast. The necessity to concurrently deal with both the glitches of model estimate and model assessment has ascended, and an integrated and efficient outline has been proposed, established on a levelled bootstrap system of data re-sampling. The projected method embraces the prevailing results centered on oversampling as a distinct instance; it is reinforced by a theoretic outline and lessens the danger of model over fitting. The claim of the projected method to factual and replicated data has revealed outstanding concert, compared with other comparable approaches previously recognized in the literature. The method might also be efficaciously used for an enhanced approximation of the learner's correctness and, if one is keen to bear an improved computational intricacy, it may be shared with snaring philosophies, thus refining the concert of classification even more.

V. CONCLUSION

Here we have experienced that imbalanced proportion does not have the extreme substantial consequence on the classifiers' efficiency,

but other parameter and issues that should be considered wisely. There some cases such as conjunction with a skewed information dispersal, inflict a heavy-duty handicap for attaining a high cataloguing performance for both of the classes of this particular problem, i.e., the existence of trivial disjuncts, the absence of concentration or trivial sample dimension, the class overlying, noisy datum, the precise controlling of marginal specimens, and the dataset swing. It highlights that there is a up-to-date necessity to investigate it i.e. intrinsic features of the datum, so that in upcoming studies on classification with imbalanced information must emphasize on discovering and gauging the utmost noteworthy data possessions, in an attempt to be capable to delineate upright elucidations as well as substitutions to overcome the glitches.

REFERENCES

1. Sara Belarouci, Mohammed Amine Chikh. "Medical imbalanced data classification", Advances in Science, Technology and Engineering Systems Journal, 2017
2. Victoria López, Alberto Fernández, Salvador García, Vasile Palade, Francisco Herrera. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics", Information Sciences, 2013.
3. Wendong Xiao, Jie Zhang, Yanjiao Li, Sen Zhang, Weidong Yang. "Class-specific cost regulation extreme learning machine for imbalanced classification", Neurocomputing, 2017
4. Dafei Wang, Wujie Xie, Wenhan Dong. "Composite reduced-kernel weighted extreme learning machine for imbalanced data classification", IOP Conference Series: Materials Science and Engineering, 2019
5. Mirza, Bilal, and Zhiping Lin. "Meta-cognitive online sequential extreme learning machine for imbalanced and concept-drifting data classification", Neural Networks, 2016.
6. Justin M. Johnson, Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance", Journal of Big Data, 2019 [35] https://www.ele.uri.edu
7. Haibo He, Yang Bai, Edwardo A. Garcia, ShutaoLi. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on
8. Computational Intelligence), 2008 www.science.gov
9. Haibo He; Garcia, E.A. (2009). "Learning from Imbalanced Data". IEEE Transactions on Knowledge & Data Engineering. 21 (9): 263–1284. doi:10.1109/TKDE.2008.23
10. Chawla, Nitesh V. (2010) Data Mining for Imbalanced Datasets: An Overview doi:10.1007/978-0-387-09823-4_45 [12] Maimon, Oded; Rokach, Lior (Eds) Data Mining and Knowledge Discovery Handbook, Springer ISBN 978-0-387-09823-4 (pages 875–886)
11. Ling, Charles X., and Chenghui Li. "Data mining for direct marketing: Problems and solutions." Kdd. Vol. 98. 1998.
12. Rahman,M.M. Davis,D.N. (2010) Addressing the Class Imbalance Problem in Medical Datasets, International Journal of Machine Learning and Computing vol. 3, no. 2, pp. 224–228, 2013.
13. Lopez, A. Fernandez, S. Garcia, V. Palade, F. Herrera, "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics," Information Sciences, 2013.
14. S. Wang, Z. Li, W. Chao, and Q. Cao, "Applying Adaptive Oversampling Technique Based on Data Density and Cost-Sensitive SVM to Imbalanced Learning," IEEE Int'l Joint Conf. on Neural Networks IJCNN-2012, doi: 10.1109IJCNN.2012.6252696, 2012.
15. P. Yang, P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications," IEEE Trans. on Cybernetics, 2014.
16. B. Das, N. C. Krishnan, and D. J. Cook, "RACOG and wRACOG: Two Probabilistic Oversampling Techniques," IEEE Trans. On Knowledge and Data Engineering, 2015.
17. B. Krawczyk, "Cost-Sensitive One-vs-One Ensemble for Multi- Class Imbalanced Data," IEEE Int'l Joint Conf. on Neural Networks IJCNN-2016, doi: 10.1109IJCNN.2016.7727503, 2016.
18. C. Zhang, K. C. Tan, and R. Ren, "Training Cost-sensitive Deep Belief Networks on Imbalance Data Problems," IEEE Int'l Joint Conf. on Neural Networks IJCNN-2016, doi:10.1109IJCNN.2016.7727769, 2016.
19. Y. Fong, S. Datta, I. S. Georgiev, P. D. Kwnong, and G. D. Tomaras, "Kernel-based Logistic Regression Model for Protein Sequence without Vectorialization," Biostatistics, 2015.
20. X. Wang, E. P. Xing, and D. J. Schaid, "Kernel Methods for Largescale Genomic Data Analysis," Briefings in Bioinformatics, 2015.
21. V. Balasubramanian, S. S. Ho, and V. Vovk, "Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications," Elsevier, 2014.
22. https:// www.analyticsvidhya.com
23. https://www.sci2s.ugr.e

AUTHORS PROFILE

Dr. Smita Nirkhi has completed M.Tech in Computer Science Engineering, PHD in computer science & Eng. She has Published 39 papers in various international conferences & 7 papers in reputed and peer reviewed international journals. Also presented paper at International Conference at Singapore and IIT, Kanpur, India. She has received grant of 8 lakhs from AICTE for her Research under RPS. She has attended and organized STTP workshops along with other training programs. She has total 15 years of professional experience. She worked as a reviewer for various conferences. Her Area of interest include Soft computing, Data mining, web mining, pattern recognition, MANET, Digital Forensics ,Machine Intelligence, Pattern Recognition, Authorship Analysis, Data Science.ements, with photo that will be maximum 200-400 words.

Prof. Shashikant Patil is a Fellow of Institution of Engineers; IETE and IRED as well as Senior Member ACM & Senior Member IEEE with 20 years' Teaching and Research experience. He is a recipient of Best Researcher Award 2014 of SVKMs NMIMS Shirpur. He has shouldered the responsibilities as Section Ambassador; Regional Lead Ambassador for IEEE Day 2014 2015 & 2016. He is member of IEEE RFID; SIG Member of IoT; Managing Editor for IEEE SDN; IEEE RFID SC & EiC for IEEE CRFID Newsletter; Jury Member IEEE SIGHT; Travel Grant Chair for IEEE DySPAN 2017; TPC Member for more than 300 Conferences with more than 60 publications. He is Potential Reviewer and Journal Referee for SCI/SCIE/Scopus indexed Journals and associated with teaching fraternity all over the globe.