STATE-OF-ART OF IMBALANCED DATA CLASSIFICATION METHODS

[1]K Venkata Nagendra, [2]Dr.Maligela Ussnaiah
[1]Research Scholar, [2]Assistant Professor,
[1]Department of Computer Science, VS University,
[1]SPSR Nellore, AP, INDIA.

Abstract: The problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation. In this paper, we provide a comprehensive review of the development of research in learning from imbalanced data. Our focus is to provide a critical review of the application domains, the state-of-the-art technologies and the current assessment metrics used to evaluate learning performance under the imbalanced learning scenario. This paper highlights the major opportunities and challenges for learning from imbalanced data.

Index Terms **- Classification, Class imbalanced problem, cost-sensitive learning, active learning, assessment metrics.**

I. INTRODUCTION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. A range of classification learning algorithms, such as decision tree, back propagation neural network, Bayesian network, nearest neighbor, support vector machines, and the newly reported associative classification, have been well developed and successfully applied to many application domains. However, imbalanced class distribution of a data set has encountered a serious difficulty to most classifier learning algorithms which assume a relatively balanced distribution. The imbalanced data is characterized as having many more instances of certain classes than others. In certain applications, the correct classification of samples in the small classes often has a greater value than the contrary case. For example, in a disease diagnostic problem where the disease cases are usually quite rare as compared with normal populations, the recognition goal is to detect people with diseases. Hence, a favorable classification model is one that provides a higher identification rate on the disease category.

Imbalanced or skewed class distribution problem is therefore also referred to as small or rare class learning problem. Research on the class imbalance problem is critical in data mining and machine learning. Two observations account for this point: (1) the class imbalance problem is pervasive in a large number of domains of great importance in data mining community. Reported applications include medical diagnosis,[1] detection of oil spills in satellite radar images [2], the detection of fraudulent calls [3], risk management [4], modern manufacturing plants [5], text classification [6] etc.; and (2) most popular classification learning systems are reported to be inadequate when encountering the class imbalance problem. Research efforts are addressed on three aspects of the class imbalance problem:

(1) The nature of the class imbalance problem (i.e. “In what domains do class imbalances most hinder the performance of a standard classifier?”); [7]

(2) The possible solutions in tackling the class imbalance problem; and

(3) the proper measures for evaluating classification performance in the presence of the class imbalance problem.

This paper provides an overview on the classification of imbalanced data. Each of the following sections studies one aspect of the class imbalance problem encompassing the different application domains, the learning difficulties with standard classifier modeling algorithms, the basic strategies for dealing with imbalanced learning, techniques to resolve imbalanced data, assessment metrics for imbalanced learning and in order to stimulate future research in this field, we also highlight the major opportunities and challenges. Finally, ends with the conclusion and references.

II. IMBALANCED DATA CLASSIFICATION APPLICATION DOMAINS

The class imbalance problem is pervasive in a large number of domains of great importance to the data mining community. This problem is intrinsic to some application domains. In some cases, It happens when the data collection process is limited due to certain reasons [8]. The following examples illustrate such cases.

- Fraud Detection
- Medical Diagnosis.
- Network Intrusion Detection
- Detection of oil spills from radar images of the ocean surface.

- Modern Manufacturing Plants
- Chemical and Bio medical engineering
- Energy management
- Security management
- Business management

In addition to these examples, other reported applications involve text classification and direct marketing. Some of these applications, such as fraud detection, intrusion detection, medical diagnosis, etc., are also recognized as anomaly detection problems. In anomaly detection, the goal is to find objects that are different from most other objects. Because anomalous and normal objects can be viewed as defining two distinct classes, a considerable subset of anomaly detection systems perceive anomaly detection as a dichotomous data partitioning problem in which data samples are categorized as either abnormal or normal. As anomalies are commonly rare as compared with normal observations, the class imbalance problem is thus intrinsic to the anomaly detection applications.

III. LEARNING DIFFICULTIES WITH STANDARD CLASSIFIER MODELING ALGORITHMS

In this section, a subset of well-developed classifier learning algorithms is reviewed and discussed. To be appreciated by less knowledgeable readers to these learning algorithms, we have included a brief introduction on each classifier's learning methods and yielded an insight into the deficiency of each learning method when encountering the small classes.

i).Decision Trees : In the presence of the class imbalance problem, decision trees may need to create many tests to distinguish the small classes from the large classes. In some learning processes, the split action may be terminated before the branches for predicting small classes are detected. In other learning processes, the branches for predicting the small classes may be pruned as being susceptible to overfitting. The cause behind is that correctly predicting a small number of samples from the small classes contributes too little success to deduce the error rate significantly, as compared with the error rate increased by overfitting. Since the pruning is mainly based on the predicting error, there is a high probability that some branches that predicting the small classes are removed and the new leaf node is labeled with a dominant class.

ii) Back Propagation Neural Networks : The Multi-Layer Perceptron (MLP), trained by the Back Propagation (BP) algorithm [9], is one of the most widely used neural models for classification problems. The empirical studies also reported that the subsequent rate of decrease of net error for the small class was very low. It needed thousands of iterations to reach an acceptable solution. Usually, the training process was terminated before the net error for the small class could be decreased. The deficient performances of BP neural networks with imbalance data sets are also reported in Refs. [10].

iii). Bayesian Classification: Bayesian classification is based on the inferences of probabilistic graphic models which specify the probabilistic dependencies underlying a particular model using a graph structure [11]. In its simplest form, a probabilistic graphical model is a graph in which nodes represent random variables, and the arcs represent conditional dependence assumptions. For a given imbalanced data set, dependency patterns inherent in the small classes are usually not significant and hard to be adequately encoded in the networks. When the learned networks are inferred for classification, the samples of the small classes are most likely misclassified. Experimental results in Ref. [12] reported this observation.

iv). Support vector machines: Support Vector Machines (SVMs) are one of the binary classifiers based on maximum margin strategy introduced by Vapnik [13]. Originally, SVMs were for linear two-class classification with margin, where margin means the minimal distance from the separating hyper plane to the closest data points. SVMs seek an optimal separating hyper plane, where the margin is maximal. The solution is based only on those data points at the margin. These points are called as support vectors. SVMs are believed to be less prone to the class imbalance problem than other classification learning algorithms, since boundaries between classes are calculated with respect to only a few support vectors and the classes may not affect the class boundary too much.

Experiments were conducted on SVMs in Ref. [14] to draw boundaries for two data sets: the first data set with the ratio of the number of the large class instances to the number of the small class instances of 10:1, and the second data set with the ratio of 10000:1. It turned out that the boundary of the second data set was much more skewed towards the small class than the boundary for the first data set, and thus caused a higher incidence of classifying test instances to the prevalent class. The underlining reason for this phenomenon is that as the training data gets more imbalanced, the support vector ratio between the prevalent class and the small class also becomes more imbalanced. The small amount of cumulative error on the small class instances count for very little in the trade-of between maximizing the width of the margin and minimizing the training error. SVMs simply learn to classify everything as the prevalent class in order to make the margin the largest and the error the minimum.

v) Associative classifiers: Associative classification is a new classification approach integrating association mining and classification into a single system. For imbalanced data, association patterns describing the small classes are unlikely to be found since the combination of items characterizing the small classes occur too seldom to pass certain significance tests for detecting association patterns. Consequently, classification rules obtained from the discovered association patterns for predicting the small classes are therefore rare and weak. This observation is also discussed in Refs. [15],[16] and [17].

vi). K-nearest neighbor: K-Nearest Neighbor (KNN)is an instance-based classifier learning algorithm, which uses specific training instances to make predictions without having to maintain an abstraction (or model) derived from data. In the presence of the imbalanced training data, samples of the small classes occur sparsely in the data space. Given a test sample, the calculated k-nearest neighbors bear higher probabilities of samples from the prevalent classes. Hence, test cases from the small classes are prone to being incorrectly classified. Research works in Refs. [18] And [19] reported this observation. The below Table: 1 shows Classification Methods and learning difficulties with the imbalanced data.

Table:1 Classification Methods and learning difficulties with the imbalanced data

S.No	Method	Learning difficulties with the Imbalanced
1	Decision Tree	a. Needs many splits to distinguish the small class b. Branches or leaves predicting the small class are prone to be pruned
2	Back Propagation Neural Networks	a. The gradient descent direction is dominated by the prevalent class b. Training error is minimized only for the prevalent class
3	Bayesian Classification	Dependency patterns inherent in small class are hard to be encoded
4	Support Vector Machines	a. Data points of the small classes are rare at the margin b. Decision boundary are skewed toward the small class
5	Association Classification	Patterns of the small class are unlikely found as they occur seldomly
6	K-Nearest Neighbor	The K-NN bear higher probabilities of samples from the prevalent class as samples of the small class occur sparsely

For classification with the class imbalance problem, accuracy is no longer a proper measure since the rare class has very little impact on the accuracy as compared to that of the prevalent class [20]. For example, in a problem where a rare class is represented by only 1% of the training data, a simple strategy can be one that predicts the prevalent class label for every example. It can achieve a high accuracy of 99%. However, this measurement is meaningless to some applications where the learning concern is the identification of the rare cases.

Given a data set with imbalanced class distribution, the identification performance on the small class is usually unsatisfactory. To remedy this, the learning objective can be: (1) to balance the identification abilities between the two classes; and/or (2) to improve the recognition success on the small class.

IV. BASIC STRATEGIES FOR DEALING WITH INMBALNCED LEARNING

i).Preprocessing Techniques: Preprocessing is often performed before building learning model so as to attain better input data. Considering the representation spaces of data, two classical techniques are often employed as preprocessor:

a). Resampling: Resmpling techniques are used to rebalance the sample space for an imbalanced data set in order to alternate the effective of skewed class distribution in the learning process. Resampling methods are more versatile because they are independent of the selected classifier. Resampling techniques fall into three groups depending on the method used to balance the class distribution.

i).Over Sampling Method: Eliminating the harms of skewed distribution by creating new minority class samples. Two widely used methods to create the synthetic minority samples are randomly duplicating the minority samples and SMOTE.

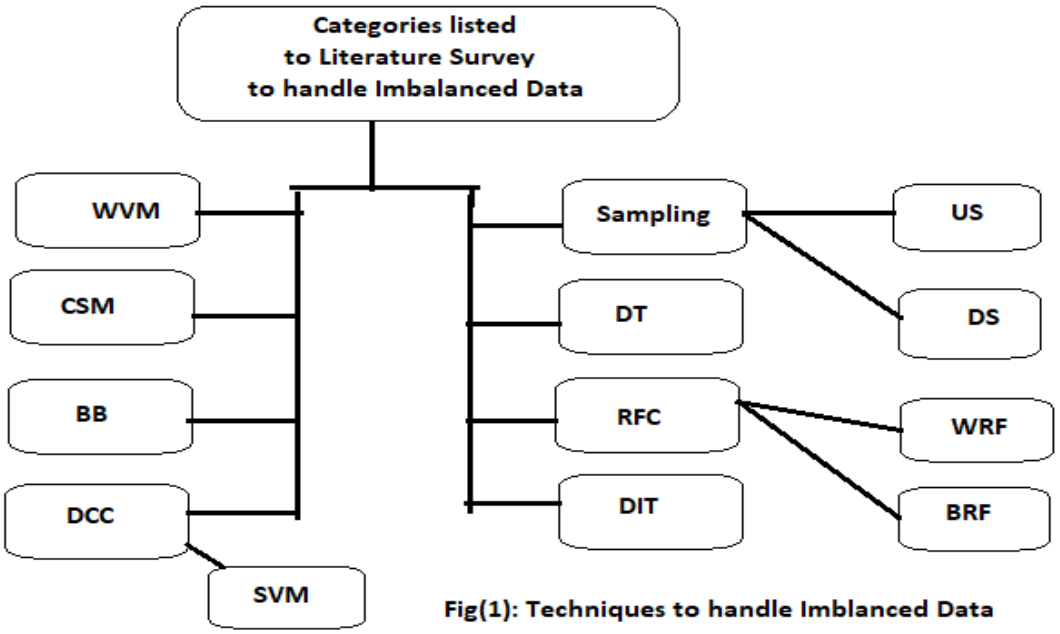
ii). Under Sampling Method: Eliminating the harms of skewed distribution by discarding the intrinsic samples in the majority class. The simplest yet most effective method is Random Under Sampling (RUS), which involved the random elimination of majority class example.

b). Feature selection and extraction: the goal of feature selection is to select a subset of k features from the entire feature space that allows a classifier to achieve optimal performance. , where k is a user defined or adaptively selected parameter. Feature selection can be divided into filters, wrappers and embedded methods. Another way to deal with dimensionality is feature extraction. Feature extraction is related to dimensionality reduction which transforms data into a low dimensional space. Feature extraction creates new features from the original features using functional mapping, where as feature selection returns a subset of original features.

ii). Cost Sensitive Learning: By assuming higher costs for the misclassification of minority class samples with respect to majority class samples. Cost sensitivity learning can be incorporated both the data level and at the algorithmic level. The costs are often specified as cost matrix, where as Cij represents the misclassification of cost of assigning examples belong to class i to j.

V. TECHNIQUES TO RESOLVE DATA IMBALANCE

To resolve minority and majority of data samples between classes and equalize the distribution the techniques as depicted in Fig. (1) are categorized through the literature survey. The major focus is on sampling methods and WRF and BRF. Cost sensitive methods, weighted voting methods are also applied to balance the data samples. In some situations decision trees, boosting and bagging concept has been applied .There is need to focus on dynamic integration techniques [21], [22], [23].



Fig(1): Techniques to handle Imblanced Data

US: Under Sampling
DS: Down Sampling
RFC: Random Forest Classification
WRF: Weighted Random Forest
BRF: Balanced Random Forest
WVM: Weighted Voting Method
CSM: Cost Sensitive Method
DT: Decision Tree
BB: Boosting and Bagging
SVM: Support Vector Machine
DCC: Data Cleaning and Classification
DIT: Dynamic Integration Techniques

i).Sampling Methods:

A very first approach to comprise with imbalanced data set is sampling methods. The objective of sampling techniques is to update or convert unequal data distribution into the equal data samples. Provision of data balancing can be done by changing the original imbalanced data samples within the given class.

ii). Minority Sampling:

The situation which deals with excessive unbalanced classification can be handled by minority sampling (MS) method. For achievement of more perfect results the oversampled class information is reduced in the direction to balance under sampled data class. This step will lead more accurate results. Minority sampling models in RFC gives the accuracy without hammering original datasets. In this technique, an individual un pruned tree is developed and a huge count of training bootstrap sample is considered for every refereed data set. Diversion of consequential tree generation, this method deals with an additional important aspect of random selection of data samples.

iii). Unsystematic Under Sampling (UUS):

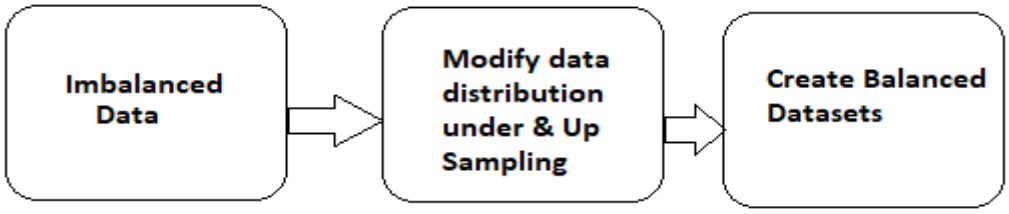
RUS techniques incorporate two main tasks to get balanced dataset. One deals with random observation selection from the under sampled class. Next step deals with of removal of particular data sample to resolve disturbances and maintain the equivalence.

iv).Instructive Under Sampling (IUS):

Instructive under sampling method deals with rejection or removal of observation from over weighted class .For removal strategy IUS follows precise criteria which has been decided in advance .This observation removal has been applied to the classes which contain comparatively more number of instances than the other class. For the production of achievable results an easy ensemble and balance cascade methods has been utilized. In IUS, at the beginning numerous subdivisions of datasets has been formed. These subdivisions of datasets apply replacement and extraction of data tuples samples from the over weighted class. In the second step of IUS for each subdivision of data tuple a separate classification has been used and finally grouping of all classifiers to deal with minority data class has been applied. Thus IUS methods are simple to deal and apply.

v).Oversampling:

This methods on statistical learning works in below mentioned as shown in Fig.(2). Support vectors from the minority concept may contribute less to the final hypothesis. Optimal hyper plane is also biased toward the majority class.



Fig(2): Balance creation through Sampling

vi).Balanced Random Forest:

Random forest encourages tree voting method for bootstrapped data sample and preparation of instructional data. While dealing with extreme data dispersion, there are chances of learning extremely imbalanced data. The probability of bootstrap trial data shows an undersized amount of data or still nothing of the underground class, and finally will result in classification tree with decreased performance .The prediction of minority class may result with decreased performance of class. Resolution of this above said problem can be handled by a binary way method. One of the solution deals with application of stratified data bootstrap sample and the other method deals with data sample replacement strategy within the data class. These two strategies are not sufficient to resolve data disturbances within class. Analytical Comparison of Classifiers on Sample Data set [25].

S.No	Classifier	Correctly classified Data Samples	Incorrectly classified Data Samples	Class
1	XGB-Extreme Gradient Boosting	78 %	22%	Positive/ Negative
2	RF-Random Forest	73%	27%	Positive/ Negative
3	SVM-Support Vector Machines	64%	36%	Positive/ Negative
4	DT- Decision Tree	70%	30%	Positive/ Negative
5	Functions Logistic	73%	27%	Positive/ Negative
6	Naïve Bayes	60%	40%	Positive/ Negative
7	Ada boost	73%	27%	Positive/ Negative
8	Attribute selected Classifier	72%	28%	Positive/ Negative

VI. ASSESSMENT METRICS FOR IMBALANCED LEARNING

As the research community continues to develop a greater number of intricate and promising imbalanced learning algorithms, it becomes paramount to have standardized evaluation metrics to properly assess the effectiveness of such algorithms. The following are the critical major assessment metrics for imbalanced learning.

- Singular Assessment Metrics
- Receiver Operating Characteristics (ROC) Curves
- Precision-Recall (PR) Curves
- Cost Curves

VII. OPPORTUNITIES AND CHALLENGES

The availability of vast amounts of raw data in many of today's real-world applications enriches the opportunities of learning from imbalanced data to play a critical role across different domains. However, new challenges arise at the same time. Here, we briefly discuss several aspects for the future research directions in this domain.

a).Understanding the Fundamental Problems : We believe that this fundamental question should be investigated with greater intensity both theoretically and empirically in order to thoroughly understand the essence of imbalanced learning problems. More specifically, we believe that the following questions require careful and thorough investigation:

1. What kind of assumptions will make imbalanced learning algorithms work better compared to learning from the original distributions?

2. To what degree should one balance the original data set?

3. How do imbalanced data distributions affect the computational complexity of learning algorithms?
4. What is the general error bound given an imbalanced data distribution?

b).Need of a Uniform Benchmark Platform: The limitation can create a bottleneck for the long-term development of research in imbalanced learning in the following aspects:

1. Lack of a uniform benchmark for standardized performance assessments;
2. Lack of data sharing and data interoperability across different disciplinary domains;
3. Increased procurement costs, such as time and labor, for the research community as a whole group since each research group is required to collect and prepare their own data sets.

c). Need of Standardized Evaluation Practices: The traditional technique of using a singular evaluation metric is not sufficient when handling imbalanced learning problems. Although most publications use a broad assortment of singular assessment metrics to evaluate the performance and potential trade-offs of their algorithms, without an accompanied curve-based analysis, it becomes very difficult to provide any concrete relative evaluations between different algorithms, or answer the more rigorous questions of functionality.

d) Incremental Learning from Imbalanced Data Streams: In regards to incremental learning from imbalanced data streams, many important questions need to be addressed, such as:

1.How can we autonomously adjust the learning algorithm if an imbalance is introduced in the middle of the learning period?
2. Should we consider rebalancing the data set during the incremental learning period? If so, how can we accomplish this?
3. How can we accumulate previous experience and use this knowledge to adaptively improve learning from new data?
4. How do we handle the situation when newly introduced concepts are also imbalanced (i.e., the imbalanced concept drifting issue)?

e). Semi supervised Learning from Imbalanced Data: The semi supervised learning problem concerns itself with learning when data sets are a combination of labeled and unlabeled data, as opposed to fully supervised learning where all training data are labeled. The key idea of semi supervised learning is to exploit the unlabeled examples by using the labeled examples to modify, refine, or reprioritize the hypothesis obtained from the labeled data alone. all of these methods have illustrated great success in many machine learning and data engineering applications, the issue of semi supervised learning under the condition of imbalanced data sets has received very limited attention in the community. Some important questions include:

1. How can we identify whether an unlabeled data example came from a balanced or imbalanced underlying distribution?
2. Given an imbalanced training data with labels, what are the effective and efficient methods for recovering the unlabeled data examples?
3. What kind of biases may be introduced in the recovery process (through the conventional semisupervised learning techniques) given imbalanced, labeled data?

VIII. CONCLUSION

This paper, beginning from a few precedents of use spaces that the class imbalance problem irritates, talks about the idea of the class imbalance problem; reviews most standard classifier learning algorithms, for example, decision tree, Back Propagation Neural Network, Bayesian network, nearest neighbor, support vector machines and associative classification, to pick up bits of knowledge into their learning troubles while experiencing imbalanced information; introduces an intensive review on announced research answers for this issue and examines both their focal points and limitations in an effort to prompt propelled investigate thoughts in future. Consequently the work outlined the different strategies used to manage the genuine situations with imbalanced data sets.

REFERENCES

[1]. P. M. Murph and D. W. Aha, UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California: Irvine (1991).
[2] M. Kubat, R. Holte and S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Mach. Learn. 30 (1998) 195–215.
[3] T. E. Fawcett and F. Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery 1(3) (1997) 291–316.
[4] K. Ezawa, M. Singh and S. W. Norton, Learning goal oriented bayesian networks for telecommunications risk management, Proc. Thirteenth Int. Conf. Mach. Learn., Bari, Italy (1996), pp. 139–147.
[5] P. Riddle, R. Segal and O. Etzioni, Representation design and brute-force induction in a Boeing manufactoring domain, Appl. Artif. Intell. 8 (1991) 125–147.
[6]. C. Cardie and N. Howe, Improving minority class predication using case-specific feature weights, Proc. Fourteenth Int. Conf. Mach. Learn., Nashville, TN (July 1997), pp. 57–65.
[7] N. Japkowicz and S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. J. 6(5) (2002) 429–450.
[8] Guo Haixiang, Li Yijing ;learning from class –imbalnced data:Review of methods and applications, Expert system applications 2017: pp220-239.
[9] J. Hertz, A. Krogh and R. G. Palmer, Introduction to the Theory of Neural Computation (Addison Wesley, 1991).
[10] K. Carvajal, M. Chac´on, D. Mery and G. Acuna, Neural network method for failure detection with skewed class distribution, INSIGHT, J. British Institute of NonDestructive Testing 46(7) (2004) 399–402.
[11]. J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Infernence (Morgan Kaufmann, 1988).

[12]. H. K¨uck, Bayesian formulations of multiple instance learning with applications to general object recognition, Master's thesis, University of British Columbia, Vancouver, BC, Canada (2004).
[13]. V. Vapnik and A. Lerner, Pattern recognition using generalized portrait method, Automat. Remont Contr. 24 (1963) 774–780.
[14]. G. Wu and E. Y. Chang, Class-boundary alignment for imbalanced dataset learning, Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC (August 2003).
[15]. B. Liu, Y. Ma and C. K. Wong, Improving an association rule based classifier, Proc. 4th Eur. Conf. Principles of Data Mining and Knowledge Discovery, Lyon, France (September 2000), pp. 504–509.
[16]. Y. Wang, High-Order Pattern Discovery and Analysis of Discrete-Valued Data Sets, PhD thesis, University of Waterloo, Waterloo, Ontario, Canada (1997).
[17]. G. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets 6(1) (2004) 7–19.
[18]. G. E. A. P. A. Batista, R. C. Prati and M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets 6(1) (2004) 20–29.
[19]. J. Zhang and I. Mani, KNN approach to unbalanced data distributions: a case study involving information extraction, Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC (August 2003).
[20]. M. V. Joshi, V. Kumar and R. C. Agarwal, Evaluating boosting algorithms to classify rare classes: comparison and improvements, Proc. First IEEE Int. Conf. Data Min. (ICDM'01) (2001).
[21]. J. Gu, Y. Zhou, and X. Zuo," Making Class Bias Useful: A Strategy of Learning from Imbalanced Data", Chapter of State Power Economic Research Institute, pp 1-10, 2007.
[22] Napiera, J. Stefanowski, and S. Wilk ,"Learning from Imbalanced Data in Presence of Noisy and Borderline Examples", Springer RSCTC, 6086, pp. 158–167, 2010.
[23]. http:// sci2s.ugr.es/keel/imbalanced.php.
[24]. Napiera, J. Stefanowski, and S. Wilk ,"Learning from Imbalanced Data in Presence of Noisy and Borderline Examples", Springer RSCTC, 6086, pp. 158–167, 2010.
[25]. Mrs. A. S. More and Dr. Dipti P. Rana, Review of Random Forest Classification Techniques to Resolve Data Imbalance, 978-1-5090-4264-7/17 @ 2017 IEEE
[26].N. V. Chawla, N. Japkowicz and A. Kolcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets 6(1) (2004) 1–6.
[27]. S. Lavanya, S. Palaniswami, S.Sudha, "Efficient Methods To Solve Class Imbalance And Class Overlap", International Journal of Science, Engineering and Technology Research, vol. 3, no. 12, pp. 3298 - 3302 , 2014.
[28] H .He , E. A. Garcia, "Learning from Imbalanced Data," IEEE Transaction Knowledge and Data Engineering, Vol. 21, Issue 9, pp. 1263-1284, 2009.
[29]. Ying Mi, "Imbalanced Classification Based on Active Learning SMOTE", Research Journal of Applied Sciences, Engineering and Technology, vol.5 , no.3 ,pp. 944 - 949, 2013.
[30].D. Ramyachitra, P. Manikandan, "Imbalanced Dataset Classification And Solutions: A Review", International Journal of Computing and Business Research, vol. 5, no. 4, pp. 1- 29, 2014.