ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Systematic Review on Dealing Imbalanced Data –A Machine Learning Approach

Supriya Indukuri, Suguna Kanumuru, Soundarya Dandu, Adinarayana Salina

B.Tech Student, Dept. of I.T., Shri Vishnu Engineering College for Women, Bhimavaram, India

B.Tech Student, Dept. of I.T., Shri Vishnu Engineering College for Women, Bhimavaram, India

B.Tech Student, Dept. of I.T., Shri Vishnu Engineering College for Women, Bhimavaram, India

Assoc Professor, Dept. of I.T., Shri Vishnu Engineering College for Women, Bhimavaram, India

ABSTRACT: Social Network Sites (SNS) have undergone a dramatic growth in recent years, the amount of data generation is growing like anything and it is not so easy for scaling and is of great importance. But many of these datasets are imbalanced in nature. For extracting useful information from such large dataset is challenging. This imbalanced nature of the datasets affects the performance of a classifier drastically. Different machine learning techniques are used to handle this issue and to deal with this data, it is necessary to understand the problem of imbalanced learning. There are various Under-sampling and oversampling techniques to resolve imbalanced learning problem are discussed in this paper. We have also discussed different SNS sources, challenges and research issues in handling imbalanced SNS data.

KEYWORDS: Imbalanced learning, Machine Learning, Oversampling, under sampling, performance, social networking.

I. INTRODUCTION

In today's era of machine learning and data mining, many real world applications work on datasets mainly for performing analysis and generating recommendations and predictions. For performing these calculations the dataset should be properly balance. But sometimes it is seen that these datasets are imbalance in nature which further leading to the problem of imbalanced data. The data which has an unequal distribution of samples among classes is known as imbalanced data. The class having more samples is generally a majority class and a class which contains very scarce samples is a minority class. Such type of data sets pose a great challenge to the classifier as it becomes problematic to classify the minority samples precisely because of their fewer amounts. Class imbalance occurs when there are significantly fewer training instances of one class compared to other classes. Imbalance is mainly caused by limitations in collecting data such as cost, privacy and the large effort required to obtain a representative data set. Classification of these imbalanced datasets is a very crucial task for the classifier as classifier may tend to favour the majority class samples. As a result of unequal distribution of data, majority class significantly dominates the minority class. Imbalance class presents several difficulties in learning, including imbalanced in class distribution, lack of data, and concept complexity. Standard classification algorithms fail to classify such form of imbalanced data accurately with least misclassification error.

II. LITERATURE REVIEW

In [1] author pointed that Social Networking sites(SNS) are varied and they incorporate a range of new information and communication tools such as availability on desktop and laptops, mobile devices such as computers and smart phones, digital photo/video/sharing and blogging . In [2] author discusses that online Community services are sometimes considered a SNS service, though in a broader sense, SNS service usually means an individual-centered service whereas online community services are group-centered. SNS allow users to share ideas, digital photos and videos, posts, and inform others about online or real world activities and events with people in their network. The unbounded

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

growth of content and users pushes the Internet technologies to its limits and demands for new solutions. Extracting intelligence from such unstructured and imbalanced data has becoming a quickly widening multidisciplinary area that demands the synergy of scientific tools and expertise. In a paper [4], authors have addressed the application of Text mining methods to harness large amounts of unstructured SNS data and transform it into structured information and then predict variety of customer trends and behavior. The rapid growth in popularity of SNS shown in table-1 has enabled large numbers of users to communicate, create and share content, give and receive recommendations, and, at the same time, it opened new challenging problems. In [9] author addresses describes the main issues that hinder the classifier performance in managing highly imbalanced datasets and the many factors that contribute to the class imbalance problems. Author suggested in [10] that ROSE package in R provides functions to deal with binary classification problems in the presence of imbalanced classes. In[11] author described imbalanced-learn is an open-source python toolbox aiming at providing a wide range of methods to cope with the problem of imbalanced dataset frequently encountered in machine learning and pattern recognition. In [12] authors have used Weka's SMO to Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier, This SMO replaces all missing values and transforms nominal attributes into binary ones. In a Paper [13] authors analyzed the performance of several techniques used to deal with imbalanced datasets in the big data scenario using the Random Forest classifier

Table 1. SNS data sources

SNo	Social Networking	Features	Started	Users
1	Facebook	One of the best mediums for connecting people from all over the world with your business	February 4, 2004	More than 1.59 billion monthly active users.
2	Twitter	Interact with prospective clients, answer questions, release latest news and at the same time use the targeted ads with specific audiences.	March 21, 2006	More than 320 million active monthly users
3	LinkedIn	Great for people looking to connect with people in similar industries, networking with local professionals and displaying business related information and statistics.	May 5, 2003	24 languages and has over 400 million registered users
4	Google+	SEO value alone makes it a must-use tool for any small business	December 15, 2011	418 active million users as on December 2015.
5	YouTube	Largest and most popular video-based social media website	February 14, 2005	1 billion website visitors per month
6	Pinterest	Consists of digital bulletin boards where businesses can pin their content	March 2010	100 million users As on September 2015
7	Instagram	Visual social media platform to post information about travel, fashion, food, art and similar subjects.	October 6, 2010	More than 400 million active users
8	WhatsApp	A cross-platform instant messaging client for smart phones, PCs and tablets to send images, texts, documents, audio and video messages to other users	January 2010 but now acquired by facebook on February 19, 2004	More than 1 billion users
9	Snapchat	Image messaging application software	September 2011	average of 100 million daily active users as of May 2015
10	Digg	News aggregator with a curated front page that selects stories	November 2004	11 million active monthly users as on 2015
11	Viber	a Voice over IP (VoIP) and instant messaging app	December 2, 2010	As of April 2014, It has close to 600 million registered users and 230 monthly active users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

12	TELEGRAM	Telegram is a cloud-based mobile and desktop messaging app with a focus on security and speed. This instant messaging network is similar to WhatsApp and is available across platforms in more than eight languages	2013	In February 2016, Telegram announced that they had 100 million monthly active users, with 350,000 new users signing up every day, delivering 15 billion messages daily
13	LINKEDIN	Available over 20 languages to connect with different businesses, locate and hire ideal candidates.	2002	500 million users

III. ARCHITECTURE

In this architecture we have embedded the flow of handling the imbalanced SNS data shown in Fig-1. Collect the data from different SNS, transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre processing is a proven method of resolving such issues. After pre processing imbalanced data handling techniques are used to processed data. Performance of the techniques is evaluated to select the best approach to handle the imbalanced data. Finally conclusion remarks are represented.

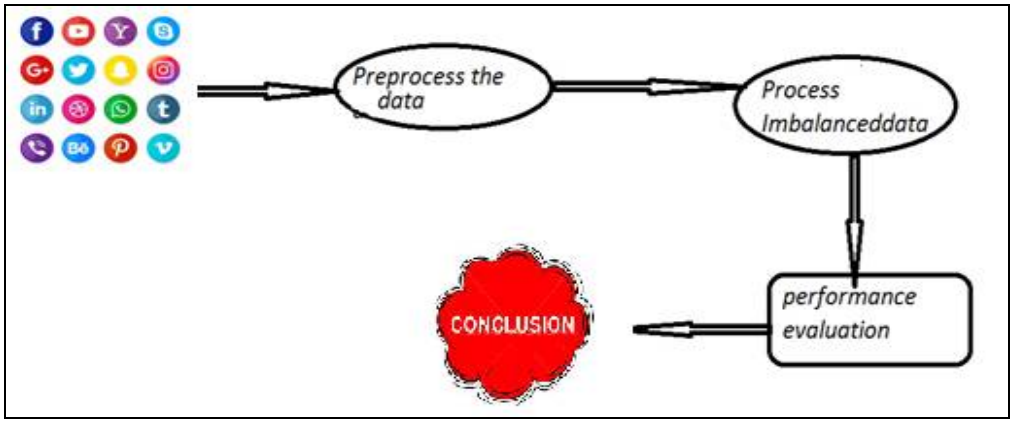

Figure 1: Architecture for Imbalanced data learning

IV. LEARNING FROM IMBALANCED DATA

We may distinguish three main approaches to learning from imbalanced data:

i. Data-level methods that modify the collection of examples to balance distributions and/or remove difficult samples.
ii. Algorithm-level methods that directly modify existing learning algorithms to alleviate the bias towards majority objects and adapt them to mining data with skewed distributions.
iii. Hybrid methods that combine the advantages of two previous groups. Let us now present a short overview of the mentioned approaches.

Data-level methods: concentrate on modifying the training set to make it suitable for a standard learning algorithm. With respect to balancing distributions we may distinguish approaches that generate new objects for minority groups (oversampling) and that remove examples from majority groups (under sampling). Standard approaches use random approach for selection of target samples for pre processing. However, this often leads to removal of important samples

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

or introduction of meaningless new objects. Therefore, more advanced methods were proposed that try to maintain structures of groups and/or generate new data according to underlying distributions .This family of algorithms also consists of solutions for cleaning overlapping objects and removing noisy examples that may negatively affect learners.

Algorithm-level methods: concentrate on modifying existing learners to alleviate their bias towards majority groups. This requires a good insight into the modified learning algorithm and a precise identification of reasons for its failure in mining skewed distributions. The most popular branch is cost-sensitive approaches. Here, given learner is modified to incorporate varying penalty for each of considered groups of examples. This way by assigning a higher cost to less represented set of objects we boost its importance during the learning process (which should aim at minimizing the global cost associated with mistakes). It must be noted that for many real-life problems it is difficult to set the actual values in the cost matrix and often they are not given by expert beforehand. Another algorithm-level solution is to apply one-class learning that focuses on target group, creating a data description.

Hybrid methods: concentrate on combining previously mentioned approaches to extract their strong points and reduce their weaknesses. Merging data-level solutions with classifier ensembles, resulting in robust and efficient learners is highly popular. There are some works that propose hybridization of sampling and cost-sensitive learning.

A. Real-life imbalanced problems

Developments in learning from imbalanced data have been mainly motivated by numerous real-life applications in which we face the problem of uneven data representation. In such cases the minority class is usually the more important one and hence we require methods to improve its recognition rates. This is closely related with important issues like preventing malicious attacks, detecting life-threatening diseases, managing a typical behavior in SNS or handling rare cases in monitoring systems.

V. CLASSIFICATION OF IMBALANCED DATA

We have two categories of classification for imbalanced data learning. They are Binary classification and multi classification. Binary classification problems can be considered as the most developed branch of learning from imbalanced data. In this classification, relationship between classes is well defined: one of them is a majority, while the other is a minority group. This allows for many straightforward approaches to balance the distributions or shift classifiers towards the minority class.

Simply changing the data distribution without considering the imbalance effect on the classification output may be misleading. Recent studies show that weighting or putting a threshold on continuous output of a classifier can often lead to better results than data re sampling and may be applied to any conventional classifier[6] .

Following directions should be taken in order to further develop classifier's output compensation for imbalanced data:

i. Currently the output is being adjusted for each class separately, using the same value of compensation parameter for each classified object. However, from our previous discussion we may see that the minority class usually is not uniform and difficulty level may vary among objects within. Therefore, it is interesting to develop new methods that will be able to take into consideration the characteristics of classified example and adjust classifier's output individually for each new object.

ii. A drawback of methods based on output adjustment lies in possibility of overdriving the classifier towards the minority class, thus increasing the error on the majority one. As we may expect that the disproportion between classes will hold also for new objects to be classified, we may assume that output compensation will not always be required. Techniques that will select uncertain samples and adjust outputs only for such objects are of interest to this field. Additionally, dynamic classifier selection between canonical and adjusted classifiers seems as a potential useful framework.

iii. Output adjustment is considered as an independent approach. Yet from a general point of view it may be fruitful to modify outputs even when data-level or algorithm-level solutions have been previously utilized. This way we may achieve class balancing on different levels, creating more refined classifiers. Analyzing the output

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

compensation may also bring new insight into supervising under sampling or oversampling to find balanced performance on both classes.

Despite numerous works on this topic there are still many open challenges that need to be addressed. We identify the following future directions of research .

Multi-class imbalanced classification is not as well-developed as its binary counterpart. Here we deal with a more complicated situation, as the relations among the classes are no longer obvious. A class may be a majority one when it is compared to some other classes, but a minority or well-balanced for the rest of them. When dealing with multi-class imbalanced data we may easily lose performance on one class while trying to gain it on another. Considering this problem there are many issues that must be addressed by novel proposals. A deeper insight into the nature of the class imbalance problem is needed, as one should know in what domains does class imbalance most hinder the performance of standard multi-class classifiers when designing a method tailored for this problem. In the paper[5] authors had proposed an efficient learning approach for knowledge discovery from multi class imbalance datasets specifically designed for opinion mining which decomposes multi class into number of binary class samples followed by a unique technique for under sampling the instances from majority subset of each binary sample.

Following directions should be taken for introducing new methods dedicated to multi-class imbalanced pre processing:

i. It is interesting to analyze the type of examples present in each class and their relations to other classes. Here it is not straightforward to measure the difficulty of each sample, as it may change with respect to different classes. For example, a given object may be of borderline type for some groups and at the same time a safe example when considering remaining classes. Therefore, a new and more flexible taxonomy should be presented. Our initial works on this topic indicate that analysis of example difficulty may significantly boost the multi-class performance.

ii. New data cleaning methods must be developed to handle presence of overlapping and noisy samples that may additionally contribute to deteriorating classifier's performance. One may think of projections to new spaces in which overlapping will be alleviated or simple removal of examples. However, measures to evaluate if a given overlapping example can be discarded without harming one of classes are needed. In case of label noise it is very interesting to analyze its influence on actual imbalance between classes. Wrongly labeled samples may increase the imbalance or mask actual disproportions. Such scenarios require dedicated methods for detecting and filtering noise, as well as strategies for handling and relabeling such examples.

iii. New sampling strategies are required for multi-class problems. Simple re-balancing towards the biggest or smallest class is not a proper approach. We need to develop dedicated methods that will adjust the sampling procedures to both individual properties of classes and to their mutual relations. Hybrid approaches, utilizing more than one method seem as an attractive solution.

VI. IMBALANCED DATA HANDLING TECHNIQUES

A. Under sampling

Under-sampling methods work by reducing the majority class samples. This reduction can be done randomly in which case it is called random under-sampling or it can be done by using some statistical knowledge in which case it is called informed under-sampling. Some informed under-sampling methods and iteration methods also apply data cleaning techniques to further refine the majority class samples.

i. MLPUS

MLP-based under-sampling technique (MLPUS) which will preserve the distribution of information while doing under-sampling . The MLPUS involves three key mechanisms:

a) clustering of majority class samples.
b) selection of important samples using SM evaluation.
c) training of MLP using selected samples in SM evaluation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

ii. **EasyEnsemble**

In EasyEnsemble method, majority class is divided into several subsets and the size of each subset is equal to the size of a minority class. Then for each subset, it develops a classifier using whole minority class and majority class subset. Results generated from all the classifiers are combined to get the final decision. To develop a classifier Adaboost is used. As EasyEnsemble uses independent random sampling with replacement, it can be considered as an unsupervised learning algorithm .

iii. **BalanceCascade**

It is a supervised learning approach . BalanceCascade method works as follows: Subset of majority class is formed which contains a number of samples equal to the number of minority class sample. When C1 classifier is trained using the majority class subset and whole minority class, the samples from a majority subset which are correctly classified are removed. This new generated sampled set of majority class is given as an input toC2. The same procedure is iterated until final classifier is reached. At every classifier, the size of the majority subset gets reduced. In BalanceCascade there is a sequential dependency between classifiers. BalanceCascade differ from EasyEnsemble as it removes true majority samples in order to reduce redundancy.

B. **Oversampling**

In oversampling method, new samples are added to the minority class in order to balance the data set. These methods can be categorized into random oversampling and synthetic oversampling[8]. In random oversampling method, existing minority samples are replicated in order to increase the size of a minority class. In synthetic oversampling technique, artificial samples are generated for the minority class samples. These new samples add the essential information to the minority class and prevents its instances from the misclassification.

i. **SMOTE** Synthetic Minority Over-sampling Technique (SMOTE) works[7] consider some training data which has s samples, and f features in the feature space of the data. Note that these features, for simplicity, are continuous. As an example, consider a dataset of birds for clarification. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight (all continuous). To then oversample, take a sample from the dataset, and consider its k nearest neighbours.

ii. **RAMOBoost** Ranked Minority Oversampling in Boosting (RAMOBoost) is a technique which systematically generates synthetic samples depending on sampling weights. It adjusts these weights of minority samples according to their distribution. This method works in two stages. In first stage decision boundary is shifted towards the samples which are difficult to learn from both majority and minority classes. In the second stage to generate synthetic samples a ranked sampling probability distribution is used.

iii. **Borderline-SMOTE .**The main objective of Borderline-SMOTE is to identify minority samples located near decision boundary. Then these samples are used further for oversampling. This method focuses on borderline samples because classifier may misclassify them. In two methods borderline-SMOTE1 and borderline-SMOTE2 has been proposed. Both methods give better results on TP rate and F-value as compared to SMOTE.

iv. **ADASYN.** HaiboHe, E.A. Garcia, proposed a novel approach adaptive synthetic sampling to handle imbalanced data set. In synthetic sample generation process, there is no need to consider all minority samples as there may be problem of overlapping. ADASYN uses the weighted distribution of minority samples. It assigns weight to minority sample depending on importance of minority sample. Samples which are difficult to classify got higher weight than others. More samples are generated for the sample having a higher weight. ADASYN

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

v. **MWMOTE** Majority Weighted Minority Oversampling Technique (MWMOTE). The objective of MWMOTE is twofold: to improve the sample selection process and to improve the synthetic sample generation process. MWMOTE involves three key phases.

a. In the first phase, MWMOTE identifies hard-to-learn and the most important minority class samples from the original minority set, Smin and construct a set, Simin by the identified samples.
b. In the second phase, each member of Simin is given a selection weight, Sw, according to its importance in the data.
c. In the third phase, using the clustering approach, MWMOTE generates the synthetic samples from Simin using Sws and produces the output set, by adding the synthetic samples to Smin.

C. **Challenges in Imbalanced learning**

Despite intense works on imbalanced learning over the last two decades there are still many challenges in existing methods and problems yet to be properly addressed w.r.t imbalanced learning problems. They are:

i. Focus on the structure and nature of examples in minority classes in order to gain a better insight into the source of learning difficulties.
ii. Develop methods for multi-class imbalanced learning that will take into account varying relationships between classes.
iii. Propose new solutions for multi-instance and multi-label learning that are based on specific structured nature of these problems.
iv. Introduce efficient clustering methods for unevenly distributed object groups and measures to properly evaluate and select partitioning models in such scenarios.

VII. PERFORMANCE MEASURES

In machine learning, the classifier of imbalanced learning is basically evaluated by a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model [4] on a set of test data for which the true values are known. The confusion matrix for a binary classification is shown in table-2 .

Table 2Confusion matrix for binary classification

	Predicted Positive	Predicted Negative
Actual positive	TP (number of True Positive)	FN (number of False Negative)
Actual Negative	FP (number of False Positive)	TP (number of True Positive)

Among various performance measures[3], the measures that are most relevant to imbalanced data learning are precision, recall, F-measure, sensitivity ,geometric mean, ROC curve, AUC and precision recall curve.
Precision of a classifier is the percentage of positive predictions made by the classifier that are correct.

Precision(P) = TP/(TP+FP) ----------(1)

where TP = true positive
FP = false positive

Recall is the percentage of true positive patterns that are correctly detected by the classifier.

Recall (R) = TP(TP+FN) ----------(2)

F-Measrue is defined as the harmonic mean of recall and precision.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

F-Measure= (2*R*P)/(R+P) ----------(3)
Sensitivity, specificity. These measures are utilized when performance of both the classes is concerned and expected to be high simultaneously.
Geometric mean(G-Mean) is the balance between classification performances on the majority and minority classes. This metric takes into account both sensitivity and specificity where sensitivity is on positive examples and specificity on negative examples then
Sensitivity= Recall ----------(4)
Specificity= 1- FP/(Tot_Negetivies) ----------(5)
G-Mean= √(Sensitivity X Specificity) ----------(6)
ROC and AUC: Receiver operating characteristic (ROC) and the area under ROC (AUC) are the two most common measures for assessing the overall classification measures. ROC is a graph showing the relationships between benefits and costs. AUC is used to summarize the performance of the classifier into a single metric. Larger the AUC , the better is the performance of the classifier.
Precession-Recall Curve(PR Curve) PR curve is used in imbalanced data learning similar to ROC curve to depict the relationship between precision and recall as the classification threshold varies.

VIII. CONCLUSION

In this paper, we have discussed current research challenges in learning from imbalanced data that have roots in contemporary real-world applications. We analyzed different aspects of imbalanced learning such as classification, clustering, regression, mining data streams and big data analytics, providing a thorough guide to emerging issues in these domains. Despite intense works on imbalanced learning over the last two decades there are still many shortcomings in existing methods and problems yet to be properly addressed.

REFERENCES

1. Liu, Y. L., and X. X. Ying. "A review of social network sites. Definition, experience and applications." The Conference on Web Based Business Management. Scientific Research Publish-ing, USA, 2010.
2. Adinarayana Salina, E. Ilavarasan," Mining Usable Customer feedback from Social Networking data for Business Intelligence", GJMS Special Issue for Recent Advances in Mathematical Sciences and Applications-13 GJMS Vol 2. 2 1–9, ISSN: 2164-3709
3. Phung, Son Lam, Abdesselam Bouzerdoum, and Giang Hoang Nguyen. "Learning pattern classification tasks with imbalanced data sets." (2009): 193.
4. Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. "Facing imbalanced data--Recommendations for the use of performance metrics." Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013.
5. Salina Adinarayana, E. Ilavarasan, "Two stage Decision Tree Learning from Multi-class Imbalanced Tweets for Knowledge Discovery", June 17 Volume 5 Issue 6 , International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, PP: 15 – 19
6. Friedland, Lisa, Amanda Gentzel, and David Jensen. "Classifier-Adjusted Density Estimation for Anomaly Detection and One-Class Classification." Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014.
7. Gutiérrez, Pablo D., et al. "SMOTE-GPU: Big Data preprocessing on commodity hardware for imbalanced classification." Progress in Artificial Intelligence (2017): 1-8.
8. López, Victoria, et al. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." Information Sciences 250 (2013): 113-141.
9. Vanhoeyveld, Jellis, and David Martens. "Imbalanced classification in sparse and large behaviour datasets." Data Mining and Knowledge Discovery(2017): 1-58.
10. Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. "ROSE: A Package for Binary Imbalanced Learning." R Journal 6.1 (2014).\
11. Lemaıtre, Guillaume, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning." Journal of Machine Learning Research 18.17 (2017): 1-5.
12. Thai-Nghe, Nguyen, Zeno Gantner, and Lars Schmidt-Thieme. "Cost-sensitive learning methods for imbalanced data." Neural Networks (IJCNN), The 2010 International Joint Conference on. IEEE, 2010.
13. Del Río, Sara, et al. "On the use of MapReduce for imbalanced big data using Random Forest." Information Sciences 285 (2014): 112-137.