Handling Imbalanced Data: A Survey

Conference Paper · January 2018

CITATIONS
31

READS
5,269

1 author:

Neelam Rout
Institute of Technical Education and Research
6 PUBLICATIONS **74** CITATIONS

SEE PROFILE

Handling Imbalanced Data: A Survey

Neelam Rout, Debahuti Mishra and Manas Kumar Mallick

Abstract Nowadays, handling of the imbalance data is a major challenge. Imbalanced data set means the instances of one class are much more than the instances of another class where the majority and minority class or classes are taken as negative and positive, respectively. In this paper, the meaning of the imbalanced data, examples of the imbalanced data, different challenges of handling the imbalanced data, imbalance class problems and performance analysis metrics for the imbalanced data are discussed. Then different methods are summarized with their pros and cons. Finally, the examples of the imbalanced data sets having low-to-high imbalance ratio (IR) values are shown.

Keywords Imbalanced data · Performance analysis metrics · Different methods

1 Introduction

In recent years, the imbalanced data sets problem plays a key role in machine learning. The imbalance problem means the instances of one of the classes (majority class) are much more than the other class (minority class). The ratio between the majority and minority classes may be 100:1, 1000:1 and 10000:1; in short, the instances of majority class outnumber the amount of minority class instances. This problem is not only in binary class data but also in multi-class data. Usually, negative examples are defined as majority class and positive examples are called minority class [1, 2] Some applications utilizing the imbalanced data sets are the medical diagnosis, detection of oil spills and financial industry [3, 4]. Nowadays,

N. Rout (✉) · D. Mishra · M.K. Mallick
Siksha 'O' Anusandhan University, Bhubaneswar, India
e-mail: neelamrout@soauniversity.ac.in

D. Mishra
e-mail: debahutimishra@soauniversity.ac.in

M.K. Mallick
e-mail: mkmallick@soauniversity.ac.in

M.S. Reddy et al. (eds.), *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, Advances in Intelligent Systems and Computing 628, https://doi.org/10.1007/978-981-10-5272-9_39

this is a very challenging topic of research [5] where lots of points need focus at a time like the binary class problem, class overlapping, cost of misclassified class, multiple class problem, small size and small disjuncts of the imbalanced data sets. There are lots of efforts given to binary class imbalance problem, but the different types of issues related to the multi-class imbalance problem are not solved yet. In the multiple class imbalance problems, the number of the majority class may be one or more and for the minority class, the situation is also same. The multi-class problem might be solved by using decomposition or any other techniques, but it still needs focus. So, when the data are skewed in nature, then it is really very challenging to work with the minority class [6].

There are several solutions existed from earlier like data-level solutions and algorithmic-level solutions [7]. The accurate extraction of knowledge from the skewed data sets with different levels of noise is very difficult [8]. When the learning of instances for some classes is costly than other classes, then the misclassified costs are varied between classes and the degree of hardness between the classes is another issued [9, 10] which can be solved by using different techniques. Feature selection is a major challenge in the imbalanced data sets to deal with specific decision [11]. There are lots of publicly available sites from which imbalanced data sets might be found, e.g., UCI machine learning repository, Broad Institute, KEEL data sets repository (http://sci2s.ugr.es/keel/datasets.php), etc. Here, no specific area of the data set is used, and those which are imbalanced might be taken for the experiment, for example, cancers data, glass data and yeast data [3]. For the performance evaluation, the area under the curve (AUC), the receiver operating characteristic (ROC) graph, F-measure, G-mean, sensitivity, specificity are used instead of measuring accuracy using confusion matrix [12–16].

The rest of the paper is structured as follows. In Sect. 2, the imbalanced class problem in classification and performance evaluation is elaborated. In Sect. 3, the summary of different existing methods based on binary and multi-class imbalance problem is discussed. Section 4 introduces examples of the imbalanced data sets. Finally, we make our concluding remarks.

2 Imbalanced Class Problem in Classification and Performance Evaluation

In this section, the problem of imbalanced data sets, the different approaches for dealing with class imbalance problem and performance evaluation in the imbalanced domains are discussed. There are several techniques by which the class imbalance problem may be solved. To solve the class imbalance problem, the concept of supervised classification should be known. The objective of classification is to predict categorical labels (e.g., loan application data; "yes" or "no"), where input and output are known. Data classification in which knowledge extraction is done by j features $b_1, b_2, \ldots, b_j \in B$ of n input instances, $x = (x_1, x_2, \ldots, x_n)$. The output class

labels (e.g., $Y_k \in C = \{c_1, \ldots, c_m\}$) of the supervised classification is known before and the mapping function is defined over the pattern $B^j \rightarrow C$.

The class imbalance problems are small sample size, class overlapping and small disjuncts. To deal with the class imbalance problem, different approaches are categorized like algorithm-level approaches, data-level approaches and cost-sensitive learning [3, 17]. Some of the existing techniques are random undersampling, oversampling, synthetic minority oversampling technique (SMOTE), selective pre-processing of imbalanced data (SPIDER) [3, 18], and some of the filtering-based methods are SMOTE-TL and SMOTE-EL [17] to eliminate noise in the imbalanced data sets. The ensemble-based method is another technique which is used to deal with the imbalanced data sets, and the ensemble technique is combined the result or performance of several classifiers to improve the performance of single classifier. The most common and effective ensemble algorithms are bagging and boosting (AdaBoost) [3]. The multiple class imbalanced data sets problems are difficult to handle than the binary class imbalanced data sets problems. Multi-class problem might be solved by one-versus-all (one-against-all) approach [19]. Till now, there are so many techniques, and modification of existing techniques is implemented for dealing with the imbalanced data sets. Some of the useful and powerful data-level techniques are SMOTE [18], modification of SMOTE, extension of SMOTE, B1-SMOTE and B2-SMOTE [20] to deal with imbalanced data. Ensemble learning [21] approaches are also very useful and fruitful.

Another important point is that to choose appropriately performance metrics for performance evaluation and analysis. Normally, confusion matrix [22] is used to calculate accuracy rate. But by using this method, the minority class data has not shown good result or ignored. This performance metrics has not given importance to the data of minority class and is usually reduced the global quantities such as error rate to give the best result. So, due to this reason for the performance evaluation of the imbalanced data, other metrics should be used rather than this one. Receiver operating characteristic (ROC) curves is a very workable visual tool which shows the trade-off in-between benefits or True Positive (TP rate) and costs or False Positive (FP rate) [12, 22–24]. Other metrics may be used like geometric-mean (Gm), F-measure (Fm), sensitivity, and specificity [22, 25].

3 Solutions of the Imbalanced Data Sets

3.1 Data-Level Approaches

To deal with the minority class of the imbalanced data, the authors [26] developed majority weighted minority oversampling technique (MWMOTE) method, and for experimental purposes 20 real-world data sets are used where G-mean, ROC, and AUC are taken as performance metrics. In [17], modification of the original sampling

technique is used for handling the imbalanced data with neighbourhood-balanced bagging (NBBag). In [16], the imbalanced data set is divided into two parts, i.e. training and testing sets; then different types of classification algorithms are taken for experiment, and finally the performance analysis is done by using different performance measures and ANOVA test is also done. As per their conclusion, SMOTE +PSO+C5 is the best classifier for 5 year survivability of breast cancer patients. In [19], the authors used one-versus-one binarization technique to decompose multi-class data into the two classes, and SMOTE+OVA algorithm is used to handle these data sets. Though Random Forests or Decision Trees is a successful and fast algorithm, it is used for classification purpose [19]. In [18] for handling the imbalanced data, Radial Basis Function Networks (RBFN) weights training methodology is used, and local and global terms are taken for designing of this method. In local weights training methods, the higher value of imbalance ratio (IR) gives better results and lower value of IR should be balanced with SMOTE or any other methods.

In [20], it is stated that the noisy and borderline examples might create problems in the imbalanced data sets and the solution of this problem is re-sampling method with some filtering techniques, and the authors used the extension of SMOTE and also Iterative–Partitioning Filter (IPF) to tackle noisy and borderline examples problem. For the binary class imbalance problems, the authors [27] used RBF classifier (combined SMOTE and PSO) and to analyse results different types of performance metrics are taken, i.e. TP%, FP%, precision, G-mean and F-measures having different β% and ρ. In [28], the researchers tried to develop the solutions for the both imbalanced and noisy data by using 7 different sampling techniques. In [29], novel inverse random under-sampling (IRUS) method is proposed to tackle the problem due to the imbalanced data sets and the idea inverse (ratio of imbalance class cardinality) is used. It has also advantages on the multi-label classification; the IRSU is also compared with several other imbalance techniques, and for each method decision tree (C4.5) is taken as the base classifier. In Table 1, the sampling methods and its variant for handling the imbalanced data are summarized.

3.2 Algorithm-Level Approaches

Support Vector Machines (SVM): To balance the imbalanced data set, only cancer data sets (breast and colon) are taken [30] because nowadays this type of cancer-related diseases are very common. Here the input is protein sequences data having different dimensions of feature space (amino acid composition, split amino acid composition, pseudo-amino acid composition—series, and pseudo-amino acid composition—parallel). In this paper, the majority class is kept aside and megatrend diffusion (MTD) technique is used to synthesize some more data for the minority class for balancing the imbalance data. By developing SVM/KNN models and analysing computational cost, it is concluded that hybrid MTD-AVM is the best model as compared to RF, QPDR, NB, and KNN. The SVMPseAAC-S model has

Table 1 Theme: sampling method and its variants for handling the imbalanced data

S. No.	Algorithms/Methods/Approaches used	Findings	Limitations
1	SMOTE+PSO+C5 [16]	Estimate 5-year survivability of breast cancer patients	Breast cancer patients
2	Modification of sampling technique, i.e., Neighbourhood Balanced Bagging (NBBag) [17]	Better than existing over-sampling bagging extension and competitive to roughly balance bagging	Costly
3	SMOTE+OVA, random forest algorithm [19]	Useful for classification multi-class imbalanced data	Fix up the oversampling rate
4	SMOTE, radial basis function networks (Ribbons) [18]	Higher the dataset imbalance ratio tends to better result	More storage space
5	Extension of SMOTE, i.e., SMOTE-IPF [20]	Addressing the problem due to noisy and borderline examples in imbalanced datasets	Small sample size
6	Combined SMOTE and PSO-based RBF classifiers (SMOTE+PSO-RBF) [27]	To generate synthetic instances for the positive class, and RBF performs well	More storage space
7	Majority weighted minority over-sampling technique (MWMOTE) [26]	Generate the synthetic minority class samples according to euclidean distance	Multi-class problem
8	Data sampling techniques [28]	Robust with noisy imbalance data	Cannot compete to advance noise handling techniques
9	Inverse random under-sampling (IRUS) [29]	Beneficial for disproportionate training datasets sizes and improve the accuracy of the multi-label classification	Other different applications to multi-label classification

given accuracy 96.71% (C/NC data set), 96.50% (B/NBC data set) and 95.18% (B/NBC data set). From [31], it is known that Ant Colony Optimization (ACO) algorithm (where '0' indicates sample is eliminated and '1' indicates sample is kept for further use) is suitable to use with under-sampling method for classifying DNA microarray imbalance data and the classification technique support vector machine (SVM) is used for the imbalanced data sets. To tackle the problem due to the majority and minority class in the imbalanced data, in [32] the authors have experimented with the kernel scaling method to improve support vector machine and adjusted F-measure performance matrix is used to evaluate the performance of the classifier. Finally, the comparison of performance results of classifiers (C4.5, L2 Loss SVM, etc.) is seen in this research paper. Though SVM might not able to handle the imbalanced data properly, the authors have experimented with the base model EnSVM and selective ensemble EnSVM+ with additional re-sampling method [33].

To handle the imbalance class, some researchers worked with feature extraction and selection methods like wrapper method [34]. In this paper, second-order cone programming support vector machines (SOCP-SVM) method is taken where linear programming support vector machine is used as formulation principle and the area under the curve (AUC) metric is used for performance analysis which worked well on six benchmark data sets. When parallel selective sampling (PSS) is combined with support vector machine (SVM), PSS-SVM method is produced and its performance is excellent than support vector machine (SVM) because it has no convergence [35]. In [36], the trained support vector machine (SVM) is taken as the preprocessor to improve the performance of MLP, LR, and RF intelligent algorithm. In this paper, two-phase balancing approach is taken, wherein phase one, the available unbalanced data is trained with SVM to get modified balance data; in the second phase, this modified data is the input to MLP, LR and RF. After that, the performance analysis is done. In [37], a new variant of SVM Near-Bayesian Support Vector Machine (NBSVM) is proposed and to reduce Bayes error, the authors developed NBSVM. This technique has many advantages over existing methods like it does not over-sample the minority class. This method is able to solve the problem due to overlapping between the target and non-target classes, small disjuncts in the imbalance data sets and noisy data sets. In Table 2, these methods are summarized in a tabular manner.

Clustering: The similarity-based hierarchical decomposition method is based on the outliers' detection and clustering techniques. In hierarchy construction, there are two parts: one is perfectly classified clusters and another one has misclassified clusters; the researchers have used this method to solve the problem classes overlapping and varieties of the majority and minority classes [38]. Fuzzy is the another useful technique for handling the imbalanced data to improve the performance of FRBCSs, and the authors have used 2-tuple genetic tuning where they have taken high and low imbalance ratio data sets [39], and also the pre-processing technique (SMOTE algorithm) is used to balance the imbalanced data. The summary of these methods are shown in Table 3.

3.3 Ensemble and Hybrid Methods

A new ensemble method is proposed in [40] to overcome the problem due to other different methods, that is bagging-based ensemble method. This method does not change the original class distribution like sampling methods, cost-sensitive learning methods. Firstly, the proposed method converts the imbalanced data (binary) into multiple balanced binary class data after that specific classification algorithm is applied to build multiple classifiers, and finally a specific ensemble rule is used like max distance, min distance, etc. and the max distance rule is performed the best among others. The hybrid method is the combination of modified back-propagation (MBP) and Gabriel graph editing (GGE) and used to handle the class imbalance and class overlapping for the multi-class problem [41]. This method is examined over

Table 2 Theme: support vector machine (SVM) and its variants for handling the imbalanced data

S. No.	Algorithms/Methods/Approaches used	Findings	Limitations
1	MTD-SVM and MTD-SVM [30]	To increase the sample of the minority class to predict human breast and colon cancers	Synthetic data generation is slightly expensive, and mega-trend diffusion (MTD) technique works well on small size data
2	ACO sampling, support vector machine [31]	To address imbalanced data classification problem by sample selection procedure based on ACO algorithm	Excessive computational and storage cost
3	Support vector machine with suitable kernel transformation, adjusted F-measure [32]	To handle the imbalance data with kernel scaling and also manage cost function	Efficient estimation strategy for parameters and different kernels
4	EnSVM and EnSVM^{+}: selective ensembles [33]	Effective than normal SVM	K value is not automatically determined
5	SOCP-SVM [34]	To improve classification performance and robust due to LP-SVM formulation	Only designed for imbalanced data
6	PSS-SVM [35]	Accurate statistical predictions and low computational complexity	For parallel and distributed computing
7	Support vector machine (SVM) [36]	Effectively balance the data and creates more number of instances for the minority class	Not so simpler and faster
8	Near-Bayesian support vector machine (NBSVM) [37]	Reduce misclassification cost due to rare class	Performance metrics

Table 3 Theme: clustering and its variants for handling the imbalanced data

S. No.	Algorithms/Methods/Approaches used	Findings	Limitations
1	Similarity-based hierarchical decomposition technique based on clustering and outlier detection [38]	Works effectively when data is highly overlapping, and classes are more imbalance	Computational complexity is more during training of this method
2	2-tuple-based genetic tuning, fuzzy rule-based classification systems (FRBCSs), genetic algorithms, genetic fuzzy systems [39]	To increase the performance of simple FRBCSs	Highly imbalanced datasets

Table 4 Theme: ensemble or hybrid methods and its variants for handling the imbalanced data

S. No.	Algorithms/Methods/Approaches used	Findings	Limitations
1	A new ensemble method includes three components, i.e., data balancing, modelling and classifying [40]	Prevent information loss or unexpected mistakes	Can handle only binary class imbalanced problem
2	Hybrid method (MBP+GGE) [41]	To deal with the class imbalance and class overlapping for the multi-class problems	Speed of the neural network convergence

seven strategies (SBP, MBP, SBP+GGE, MBP+GGE, SMOTE, SMOTE+GGE and RUS), and the authors have concluded that it is a very effective technique. The gist of these methods is shown in Table 4.

3.4 Other Different Techniques

According to [42], density-based feature selection (DBFS) is a simple technique but effective for the high-dimensional and small sample size classes of the imbalanced data sets. For studying the problem of customer churn prediction, the authors [43] carried out the following steps.

1. First, the useless features are discarded from the original data sets, then missing values are replaced, and nominal to numeric conversion is done.
2. The under-sampling method is applied which is based on RUS and PSO.
3. The PCA, Fisher's Ratio, F-Measure and Mr techniques are used for the dimension reduction.
4. The KNN and RF classifiers are used for the model building, and finally the performance evaluation is done using the metrics like sensitivity, specificity and AUC.

The proposed Che-PmRF approach is a combination of PSO-based sampling, Mr-based feature selection and RF classifier, and it is very effective for the competitive telecommunication industry to tackle the problem of customer churn prediction. To get the best performance result for the imbalanced data sets, the LSI-based feature extraction is implemented to the information granulation-based data mining model, and the main advantage of this technique is that it could save storage space [44]. Distribution optimally balanced–stratified cross-validation (DOB-SCV) is used to deal with the imbalanced data sets to obtain better performance [45]. In [46], the authors proved that the weighted extreme learning machine (ELM) has the following advantages.

1. Simple and fast in implementation.
2. Directly apply to the multi-class classification tasks.
3. Different types of feature mapping functions or kernel methods are available.

Lastly, for the performance analysis, G-mean performance metric is considered for the unweighted ELM, weighted ELM W1 and weighted ELM W2 techniques.

In [47], the authors proposed two different approaches, i.e., decomposition-based and Hellinger distance-based methods, to solve the issues related to feature selection in the imbalance data sets. In the first method, large classes are decomposed into the pseudo-subclasses and later different strategies are used for the classification. The second one is used to measure the distributional divergence for handling the challenges of the imbalance class distribution. After that both methods are compared with the three traditional feature rank methods, which are correlation, Fisher and mutual information methods using the F-measure, AUC and ROC metrics. The basic information of these methods is given in Table 5.

Table 5 Theme: other different techniques for handling the imbalanced data

S. No.	Algorithms/Methods/Approaches used	Findings	Limitations
1	Density-based feature selection (DBFS) [42]	To handle the small sample size and high-dimensional problem in imbalanced datasets	Problems due to more than two classes
2	Chr-PmRF [43]	PSO-based sampling for balancing the datasets, mRMR for feature selection and RF classifier are used to handle customer churn prediction problem	High computational cost
3	Information granulation-based data mining approach, latent semantic indexing [44]	Accuracy is slightly better and faster than the numerical computing method	Overlapping class
4	Distribution optimally balanced stratified cross-validation (DOB–SCV) [45]	Obtaining a better performance estimation result for the classifier	Storage space
5	Weighted extreme learning machine (ELM) [46]	Simple theory, fast in implementation and apply directly to the multi-class classification tasks	Large variety in class distribution
6	Decomposition-based and Hellinger distance-based methods [47]	To solve the feature selection issues in the imbalanced datasets	Comparison with only three traditional feature selection methods

Table 6 Statistic summary of the 21 imbalanced datasets

ID	Dataset	#Attributes (R/I/N)	#Examples	%class (min., maj.)	Imbalance ratio (IR)
1	glass1	9 (9/0/0)	214	(35.46, 64.54)	1.82
2	glass0	9 (9/0/0)	214	(32.68, 67.32)	2.06
3	Ecoli1	7 (7/0/0)	336	(22.94, 77.06)	3.36
4	New-thyroid2	5 (4/1/0)	215	(16.29, 83.71)	5.14
5	Yeast 3	8 (8/0/0)	1484	(10.99, 89.01)	8.1
6	yeast-2_vs_4	8 (8/0/0)	514	(9.92, 90.08)	9.08
7	Glass 2	9 (9/0/0)	214	(7.94, 92.06)	11.59
8	ecoli-0-1-4-6_vs_5	6 (6/0/0)	280	(7.14, 92.86)	13
9	yeast-1_vs_7	7 (7/0/0)	459	(6.54, 93.46)	14.3
10	glass4	9 (9/0/0)	214	(6.07, 93.93)	15.47
11	Abalone9-18	8 (7/0/1)	731	(5.75, 94.25)	16.4
12	glass-0-1-6_vs_5	9 (9/0/0)	184	(4.89, 95.11)	19.44
13	shuttle-c2-vs-c4	9 (0/9/0)	129	(4.65, 95.35)	20.5
14	Glass 5	9 (9/0/0)	214	(4.21, 95.79)	22.78
15	yeast-2_vs_8	8 (8/0/0)	482	(4.15, 95.85)	23.1
16	Car-good	6 (0/0/6)	1728	(3.99, 96.01)	24.04
17	winequality-red-4	11 (11/0/0)	1599	(3.31, 96.69)	29.17
18	Winequality_red_8_vs_6	11 (11/0/0)	656	(2.74, 97.26)	35.44
19	Abalone_9_vs_10_11-12_13	8 (7/0/1)	1622	(1.97, 98.03)	49.69
20	kddcup-rootkit-imap_vs_back	41 (26/0/15)	2225	(0.99, 99.01)	100.14
21	abalone19	8 (7/0/1)	4174	(0.77, 99.23)	129.44

4 Examples of the Imbalanced Data Sets

In Table 6, different types of data sets are mentioned with their IR values in the ascending order. The data sets are taken from KEEL data set repository [48]. For each imbalanced data set, the number of attributes, number of examples, percentage of minority and majority of each class, and imbalance ratio (IR) are given in the table.

5 Conclusion

From the study, it is concluded that in the presence of the imbalance data sets, most of the standard classifier learning algorithms, such as nearest neighbour, decision tree, back-propagation neural networks, failed to give good results. In this paper, different types of existing techniques are discussed for tackling the imbalance class

problems but still improvement techniques are needed, necessarily. It is also known that the ensemble learning algorithms are the useful and powerful methods to deal with the imbalance class problem. Some of the imbalanced data sets have been shown with different IR values in the tabular manner. It is very important to balance the imbalance data with effective techniques and at the same time, cost factor should be given attention. The correct classifier techniques and performance evaluation metrics must be applied to achieve good results. There are many methods for handling the imbalanced data, but the main focus is to use the appropriate technique from the existing techniques or develop the new methods according to the need because it is not necessary that if one technique is worked well on an imbalanced data set, then the same method is worked for an another imbalanced data set.

References

1. He, Habib, and Edwardo Garcia. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284.
2. Van Pulse, Jason, and Tag hi Jehoshaphat. 2009. Knowledge Discovery from Imbalanced and Noisy Data. *Data and Knowledge Engineering* 68 (12): 1513–1542.
3. Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 42 (4): 463–484.
4. He, Habib, and Yunnan Ma. (eds.). 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley.
5. Yang, Anglia, and Wu Donning. 2006. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology and Decision Making* 5 (04): 597–604.
6. Wang, Shu, and In Tao. 2012. Multi Class Imbalance Problems: Analysis and Potential Solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 42 (4): 1119–1130.
7. Lakshmi, T. Jay, and C. Pradesh. 2014. A Study on Classifying Imbalanced Datasets. In *First International Conference on Networks and Soft Computing (ICNSC)*, IEEE.
8. Neapolitan, Ami. 2009. *Classification Techniques for Noisy and Imbalanced Data*. Dis. Florida Atlantic University.
9. Org Mennicke, J. 2006. Classifier Learning for Imbalanced Data with Varying Misclassification Costs.
10. Scrupulousness's, M.G., D.S. Antifascist, S.B. Konstantin, and P.E. Intelsat. Local Cost Sensitive Learning for Handling Imbalanced Data Sets. In *Mediterranean Conference on Control and Automation*, *2007*, *MED'07*, 1–6. IEEE.
11. Yin, Lithium, et al. Feature Selection for High-Dimensional Imbalanced Data. *Supercomputing* 105 (2013): 3–11.
12. Y, Wenona, Yuan-chin Ivan Chang, and Eunice Park. 2014. A Modified Area Under the ROC Curve and its Application to Marker Selection and Classification. *Journal of the Korean Statistical Society* 43 (2): 161–175.
13. Lou, Zen, Ruy Wang, Ming Tao, and Xian fa CAI. 2015. A Class-Oriented Feature Selection Approach for Multi-Class Imbalanced Network Traffic Datasets Based on Local and Global Metrics Fusion. *Supercomputing* 168: 365–381.

14. Mahmoud, Shani, Par ham Moravia, Cardin Highland, and Rasoul Moradi. 2014. Diversity and Separable Metrics in Over-Sampling Technique for Imbalanced Data Classification. In *4th International eConference on Computer and Knowledge Engineering (ICCKE)*, IEEE, 152–158.
15. Ghanavati, Mojgan, Raymond K. Wong, Fang Chen, Yang Wang, and Chang-Shing Perng. 2014. An Effective Integrated Method for Learning Big Imbalanced Data. In *IEEE International Congress on Big Data* (*Big Data Congress*), IEEE, 691–698.
16. Wang, Kung-Jeng, Bunjira Makond, Kun-Huang Chen, and Kung-Min Wang. 2014. A Hybrid Classifier Combining SMOTE with PSO to Estimate 5 year Survivability of Breast Cancer Patients. *Applied Soft Computing* 20: 15–24.
17. Błaszczynski, Jerzy, and Jerzy Stefanowski. 2015. Neighbourhood Sampling in Bagging for Imbalanced Data. *Neurocomputing* 150: 529–542.
18. Perez-Godoy, M.D., A.J. Rivera, C.J. Carmona, and M.J. del Jesus. 2014. Training Algorithms for Radial Basis Function Networks to Tackle Learning Processes with Imbalanced Datasets. *Applied Soft Computing* 25: 26–39.
19. Bhagat, Reshma C., and Sachin S. Patil. 2015. Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data Using Random Forest. *IEEE International Conference on Advance Computing (IACC)*, *2015*, IEEE.
20. Saez, J.A., J. Luengo, J. Stefanowski, and F. Herrera. 2015. SMOTE–IPF: Addressing the Noisy and Borderline Examples Problem in Imbalanced Classification by a Re-Sampling Method with Filtering. *Information Sciences* 291: 184–203.
21. Hu, Xiao-Sheng, and Run-Jing Zhang. 2013. Clustering-Based Subset Ensemble Learning Method for Imbalanced Data. *International Conference on Machine Learning and Cybernetics (ICMLC)*, *2013*, vol. 1. IEEE.
22. Han, Jiawei, and Micheline Kamber. 2001. Data Mining: Concepts and Techniques.
23. Subtil, Fabien, and Muriel Rabilloud. 2015. An Enhancement of ROC Curves Made Them Clinically Relevant for Diagnostic-Test Comparison and Optimal-Threshold Determination. *Journal of clinical epidemiology*.
24. Wang, Qihua, Lili Yao, and Peng Lai. 2009. Estimation of the Area Under ROC Curve with Censored Data. *Journal of Statistical Planning and Inference* 139 (3): 1033–1044.
25. Batuwita, Rukshan, and Vasile Palade. 2009. A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems. In *International Conference on Machine Learning and Applications*, *ICMLA'09, 2009*, IEEE.
26. Barua, S., M.M. Islam, X. Yao, and K. Murase. 2014. MWMOTE–majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering* 26 (2): 405–425.
27. Gao, Ming, Xia Hong, Sheng Chen, and Chris J. Harris. 2011. A Combined SMOTE and PSO Based RBF Classifier for Two-Class Imbalanced Problems. *Neurocomputing* 74 (17): 3456–3466.
28. Seiffert, C., T.M. Khoshgoftaar, J. Van Hulse, and A. Folleco. 2014. An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data. *Information Sciences* 259: 571–595.
29. Tahir, Muhammad Atif, Josef Kittler, and Fei Yan. 2012. Inverse Random Under sampling for Class Imbalance Problem and Its Application to Multi-label Classification. *Pattern Recognition* 45 (10): 3738–3750.
30. Majid, A., S. Ali, M. Iqbal, and N. Kausar. 2014. Prediction of Human Breast and Colon Cancers from Imbalanced Data Using Nearest Neighbor and Support Vector Machines. *Computer Methods and Programs in Biomedicine* 113 (3): 792–808.
31. Yu, Hualong, Jun Ni, and Jing Zhao. 2013. ACO Sampling: An Ant Colony Optimization-based Undersampling Method for Classifying Imbalanced DNA Microarray Data. *Neurocomputing* 101: 309–318.
32. Maratea, Antonio, Alfredo Petrosino, and Mario Manzo. 2014. Adjusted F-Measure and Kernel Scaling for Imbalanced Data Learning. *Information Sciences* 257: 331–341.

33. Liu, Y., X. Yu, J.X. Huang, and A. An. 2011. Combining Integrated Sampling with SVM Ensembles for Learning from Imbalanced Datasets. *Information Processing and Management* 47 (4): 617–631.
34. Maldonado, Sebastian, and Julio Lopez. 2014. Imbalanced Data Classification Using Second-Order Cone Programming Support Vector Machines. *Pattern Recognition* 47 (5): 2070–2079.
35. D'Addabbo, Annarita, and Rosalia Maglietta. 2015. Parallel Selective Sampling Method for Imbalanced and Large Data Classification. *Pattern Recognition Letters* 62: 61–67.
36. Farquad, M.A.H., and Indranil Bose. 2012. Preprocessing Unbalanced Data Using Support Vector Machine. *Decision Support Systems* 53 (1): 226–233.
37. Datta, Shounak, and Swagatam Das. 2015. Near-Bayesian Support Vector Machines for Imbalanced Data Classification with Equal or Unequal Misclassification Costs. *Neural Networks* 70: 39–52.
38. Beyan, Cigdem, and Robert Fisher. 2015. Classifying Imbalanced Data Sets Using Similarity Based Hierarchical Decomposition. *Pattern Recognition* 48 (5): 1653–1672.
39. Fernandez, Alberto, Maria Jose del Jesus, and Francisco Herrera. 2010. On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Datasets. *Information Sciences* 180 (8): 1268–1291.
40. Sun, Z., Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou. 2015. A NOVEL Ensemble Method for Classifying Imbalanced Data. *Pattern Recognition* 48 (5): 1623–1637.
41. Alejo, R., R.M. Valdovinos, V. García, and J.H. Pacheco-Sanchez. 2013. A Hybrid Method to Face Class Overlap and Class Imbalance on Neural Networks and Multi-class Scenarios. *Pattern Recognition Letters* 34 (4): 380–388.
42. Alibeigi, Mina, Sattar Hashemi, and Ali Hamzeh. 2012. DBFS: An Effective Density Based Feature Selection Scheme for Small Sample Size and High Dimensional Imbalanced Data Sets. *Data and Knowledge Engineering* 81: 67–103.
43. Idris, Adnan, Muhammad Rizwan, and Asifullah Khan. 2012. Churn Prediction in Telecom Using RANDOM Forest and PSO Based Data Balancing in Combination with Various Feature Selection Strategies. *Computers and Electrical Engineering* 38 (6): 1808–1819.
44. Chen, M.C., L.S. Chen, C.C. Hsu, and W.R. Zeng. 2008. An Information Granulation Based Data Mining Approach for Classifying Imbalanced Data. *Information Sciences* 178 (16): 3214–3227.
45. Lopez, Victoria, Alberto Fernandez, and Francisco Herrera. 2014. On the Importance of the Validation Technique for Classification WITH Imbalanced Datasets: Addressing Covariate Shift When Data is Skewed. *Information Sciences* 257: 1–13.
46. Zong, Weiwei, Guang-Bin Huang, and Yiqiang Chen. 2013. Weighted Extreme Learning Machine for Imbalance Learning. *Neurocomputing* 101: 229–242.
47. Yin, L., Y. Ge, K. Xiao, X. Wang, and X. Quan. 2013. Feature selection for High-Dimensional Imbalanced Data. *Neurocomputing* 105: 3–11.
48. http://sci2s.ugr.es/keel/datasets.php.