TECHNIQUES AND TOOLS TO TACKLE IMBALANCED LEARNING

Rose Mary Mathew, R Gunasundari*

Abstract

Data is a prominent factor in machine learning. The quality and depth of data used for training determine the efficiency of our machine learning model. The skewness of data is viewed as one of the pain points in machine learning. This happens in datasets when one class outnumbers the other classes. For predicting better results, the model is trained with balanced data. Researchers had developed different techniques to reduce the skewness of the data. This paper points out the different techniques used for balancing the data. This paper also describes the software's like KEEL, WEKA, R, Python, Multi-Imbalance package and Spark that can be used for data processing and the different algorithms present to make the data balanced.

Keywords: Imbalanced data, majority, minority, sampling, SMOTE, packages, classifiers

I. INTRODUCTION

Analysing the data for the extraction of wonderful patterns can be done in data mining. Classification is a significant task that is performed by data mining methods. Classification is the process of grouping similar data points based on their characteristics, and if new data arrives with the help of some algorithm, it will foretell the new test data point [1]. Several algorithms are available in machine learning for performing classification

.

The prediction of a new test data point class label by an algorithm can happen by training the model with a dataset. Training a model means feeding the machine with past data to make the machine learn from this data to predict the class of new input data. The examples available for training the model is to be considered, and it should be balanced. A balanced data set means it contains an equal proportion of instances from each class. The proportion of all the classes present in the training dataset should be equal, and then the dataset can be termed as the balanced dataset, which can produce high accuracy models.

Imbalanced data means data present in the dataset which is used for training is not balanced. This means the proportion of class instances present in the training dataset is different [2]. That is, it can be like this for binary class classification problems. In binary class classification, one class outnumbers the other class. The former class is viewed as the majority class, and the latter is viewed as the minority class. In this scenario, our model tends to predict the majority class value for new input values, affecting the accuracy of the model. The same situation can occur for multi-class classification problems. In multi-class data, imbalanced means it is either a multi-majority case or multi-minority case [3]. Multi-majority mode is the majority of the classes have a higher proportion of the data in the training dataset and only a few have less proportion of data in the training dataset. Multi-minority mode is a large number of classes with a lower proportion of the data in the training dataset and only a few have a higher proportion of the data in the training dataset. Imbalanced data will affect the accuracy of the model.

So, to make good accuracy for the machine learning model, we should train the model with balanced data.

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

In machine learning, there are several approaches for balancing the dataset, namely data- level approach, algorithmic approach, cost-sensitive approach and the ensemble approach [4]. Several algorithms and methods are associated with each of these approaches. Resampling the data like oversampling and undersampling are done in the data level approach. The algorithmic level is aligned towards converting fundamental learning techniques to be more responsive to imbalanced class problems. Cost-sensitive learning solutions will use the data-level and algorithmic-level, considering high costs for misclassification and trying to minimise the cost. The ensemble approach modifies the learning algorithms with pre-processing data stage before applying each classifier or placing a cost-sensitive framework in the learning process [5].

In this paper, the second section discusses the different methods that can handle imbalanced data. The third section of this paper focuses on software tools/packages that deal with imbalanced data for processing and creating models. A list of software tools with their corresponding packages and methods to perform the imbalanced data classification are specified in this work.

II. METHODS TO SOLVE PROBLEMS

There was much research going into the classification of imbalanced data. Skewed data can be transformed to balanced then only the model should work correctly. There are several methods to balance the imbalanced data. Data pre-processing techniques minimises the effect of majority classes. Resampling of the data is happened by undersampling or oversampling the existing data.

The undersampling technique reduces the size of existing data, and these are used with majority classes to reduce the effect of skewness. Random Undersampling (RUS) is a technique to under-sample the instances of the majority class. One-Sided Selection (OSS) is an under-sampling technique [6]. Condensed nearest neighbour (CNN) is a method to reduce the size of the dataset in the k-NN method. CNN+Tomek Link for CNN, Class Purity Maximization, Neighbourhood Cleaning Rules are some of the methods used to reduce the size of the available data [7].

Oversampling is the process of increasing the size of the dataset by producing artificial data points of the minority class, which leads to the reduction of skewness in the dataset. Random OverSampling(ROS) is a technique to oversample the instances of the minority class. Synthetic Minority Oversampling Technique (SMOTE) is a technique where artificial samples are produced from the minority classes to make the data balanced [1]. Variants of SMOTE like Borderline SMOTE, SMOTE+ENN, SMOTE-RSB, SMOTE+Tomek Link are also available for solving the imbalanced problems [8]. ADASYN, a technique used to minimise the effect of skewness. Selective Pre-processing of Imbalanced Data, SPIDER method is also used to balance the data [9].

Ensemble-based approaches can create robust classifiers that can handle imbalanced data problems well. SMOTEBagging, a combination of SMOTE method with bagging, can work with multi-class imbalanced data classification [10]. AdaBoost is a method where weak classifiers are combined to make a robust classifier that produces more accurate results. Variants of the AdaBoost algorithm can be used to treat imbalanced data situations [11]. Hybrid Ensemble for Classification of Multi-class Imbalanced (HECMI), a combination of boosting and bagging which concentrates on wrongly classified instances in multi-class imbalanced data problems [6].

III. SOFTWARE TOOLS AND PACKAGES FOR IMBALANCED LEARNING

For handling imbalanced data processing, several software tools and packages can be used. In this section, a

detailed description of the various tools is specified here.

i. KEEL (Knowledge Extraction and Evolutionary Learning)

KEEL is software that can perform data mining tasks. KEEL is software built upon a java platform, and it is the first tool that will be dealing with the evolutionary learning algorithms in machine learning [12]. KEEL caters for a user-friendly interface. KEEL is aimed to work with multiple types of data and algorithms to gain the anticipated result.

Models created with KEEL are implemented with an evolutionary algorithm for prediction; it also contains pre-processing and post-processing of the data. KEEL has a pre-processing data module that can deal with missing values, noise data, transformation and discretisation of data, and selection of features. A statistical library is included for data analysis. Using this statistical library, users can perform various statistical tests for analysing the data distribution. Java class library for evolutionary computation (JCLEC) is used to develop some algorithms. Knowledge Extraction Algorithms Library in KEEL incorporates multiple algorithms together with classical approaches for learning [13]. Algorithms for evolutionary rule learning models, fuzzy rule learning models, pruning neural networks, genetic programming, patterns subgroup discovery rules and data reduction have been included.

Imbalanced datasets are those datasets which are having skewness towards a class. The distribution of data among different classes is not uniform. In binary class skewed data classification, one class can be treated as the majority class, and the other is the minority class. KEEL Software Suite attends to this imbalanced data scenario by its imbalanced learning module. Most of the classification algorithms are biased towards the majority class label and having a high miscategorisation rate for instances of the minority class. KEEL Software Suite treats a dataset as imbalanced if the distribution of a class instance is lower than 40% of the number of samples that pertain to the other class, namely the proportion atween instances of the majority and the minority class data to be higher than 1.5.

To cope with the imbalanced data, KEEL is equipped with Over-Sampling techniques and Under-Sampling techniques. These oversampling and undersampling techniques can be applied during the pre-processing stage of the imbalanced data set [14]. TABLE.1 describes the different resampling techniques present in KEEL for attending to imbalanced problems.

KEEL Imbalanced Learning module proposed three different methodologies to cope with skewed data issues: algorithmic adaption for class skewness, Cost-sensitive classification, and Ensemble techniques for skewness [13]. TABLE.2 describes the different cost-sensitive and ensemble-based techniques present in KEEL for attending to imbalanced problems. The KEEL software suite provides a Visualisation module and Statistical test module, which can visualise the different data measures, and various statistical tests can be performed.

Over-Sampling Techniques	Under-Sampling Techniques
ADASYN	Condensed Nearest neighbour (CNN)
Adjusting the direction of the synthetic minority class examples (ADOMS)	
Agglomerative Hierarchical Clustering (AHC)	CNN + Tomek's modification of CNN
Borderline SMOTE	Class purity maximisation (CPM)
Random over-sampling	Neighbourhood cleaning rule
Safe Level SMOTE	One-sided selection
Synthetic minority over-sampling technique (SMOTE)	
SMOTE+ Edited Nearest Neighbours (ENN)	Random under-sampling
SMOTE-RSB	
SMOTE + Tomek Links	Undersampling based on clustering
SPIDER	
SPIDER2	Tomek's modification of CNN

TABLE 1: Resampling Methods available in KEEL for generating balanced data

Cost-Sensitive Techniques for Classification	
C-SVM	Cost-Sensitive support vector machine
C4.5	Cost-Sensitive Decision Tree
Multilayer perceptron	Cost-Sensitive neural networks for classification problems
Ensembles for Class Imbalance	
Algorithm	**Base Classifier**
AdaBoost	C4.5 Decision Tree
AdaBoost.M1	C4.5 Decision Tree
AdaBoost.M2	C4.5 Decision Tree
Cost Sensitive Boosting	C4.5 Decision Tree
Bagging	C4.5 Decision Tree
BalanceCascade ensemble	C4.5 Decision Tree
DataBoost-IM	C4.5 Decision Tree
EasyEnsemble	C4.5 Decision Tree
IIVotes: SPIDER + IVotes	C4.5 Decision Tree
MSMOTEBagging	C4.5 Decision Tree
MSMOTEBoost	C4.5 Decision Tree
OverBagging	C4.5 Decision Tree
OverBagging2	C4.5 Decision Tree
RUSBoost	C4.5 Decision Tree
SMOTEBagging	C4.5 Decision Tree
SMOTEBoost	C4.5 Decision Tree
UnderBagging	C4.5 Decision Tree
UnderBagging2	C4.5 Decision Tree
UnderOverBagging	C4.5 Decision Tree

TABLE 2: Cost-sensitive and ensemble-based techniques present in KEEL

Package	Techniques
imbalance	SMOTE, ANSMOTE, RSLSMOTE, BLSMOTE, SLMOTE, DBSMOTE, RACOG, wRACOG, MWMOTE, RWO, PDFOS, ADASYN
ROSE	Random Oversampling
DmWR	SMOTE
unbalanced	Tomek, ubOver, ubNCL, ubUnder SMOTE, ubOSS, ubENN, ubCNN
ebmc	SMOTE Bagging, SMOTE Boost, RUS Boost, Under Bagging
smotefamily	SMOTE, ADASYN, Borderline SMOTE, SLS, RSLS, DBSMOTE, ANS

TABLE 3: Packages and algorithms present in R for making data balanced

DATA LEVEL	ENSEMBLE BASED
i. Oversampling Module	**i. BalanceBagging**
SMOTE	RUSBagging
MWMOTE	ROSBagging
ADASYN	SMOTEBagging
Random oversampling	RBBagging
ii. UnderSampling Module	**ii. BalanceBoost**
Random undersampling	RUSBoost
CLUS	SMOTEBoost
iii. Hybrid Sampling	AdaC2
SMOTE-ENN	**iii. Hybrid Ensemble**
SMOTETL	EasyEnsemble
SPIDER	BalanceCascade

TABLE 4: IRIC Library

ii. WEKA

WEKA is open-source software that is used for performing data mining problems. WEKA involves a variety of data pre-processing techniques, machine learning algorithms and visualisation tools [15]. An extensive software that drives to train models using machine learning algorithms. In addition to that, this software suite can assist big data.

Undersampling and Oversampling are the two essential tools offered by WEKA to perform imbalanced learning. Undersampling the majority class can be done with the aid of two filters, weka.filters.supervised.instance.resample and weka.filters.supervised.instance.spreadSubsample[16]. Oversampling might be attained with the aid of filter weka.filters.supervised. Resample. Furthermore, another filter weka.classifiers.meta—CostSensitiveClassifier can attain cost-sensitive classification of imbalanced data. SMOTE package in WEKA will give the provision to utilise SMOTE method in the form of a WEKA filter [17].

iii. R

R is a programming environment for doing statistical operations [18]. ROSE and DmWR are the two most used packages available in R to perform machine learning tasks. ROSE (Random Over Sampling Examples) package assists in spawning new data points based on sampling methods [19]. This package includes precise functions which help for completing the tasks quickly. A function oversample is available in this package to perform oversampling and undersampling data. This function can perform both tasks. Imbalance is a package in R which provides oversampling techniques. Another package called unbalanced offers different sampling techniques. Variants of SMOTE functions are available in a package called smote family. Ensemble-based techniques are available under the ebmc package of R. table 3 describes the different techniques present in R for attending imbalanced problems.

IRIC is a library in R language that has a collection of solutions for the imbalanced data problems [20]. IRIC addresses the three approaches for handling imbalanced data classification: data level, algorithmic level, and ensemble level. The following table 4 shows the details of methods available in the IRIC library.

Under-sampling	Over-sampling
Random Majority Undersampling with Replacement	Random Minority Oversampling with Replacement
Extraction of Majority-Minority Tomek links	SMOTE
Undersampling with cluster centroids	SMOTENC
NearMiss	SMOTEN
Condensed nearest neighbour	bSMOTE
One-sided selection	SVM SMOTE
Neighboorhood cleaning rule	ADASYN
Edited nearest neighbours	KMeans-SMOTE
Instance hardness threshold	ROSE - Random OverSampling Examples
Repeated edited nearest neighbours	
All-KNN	
Ensemble classifier aids samplers internally	**Over-sampling afterwards under-sampling**
Easy ensemble classifier	SMOTE + Tomek links
Balanced random forest	SMOTE + ENN
Balanced bagging	**Mini-Batch Resampling for Keras and Tensorflow**
RUS-Boost	

TABLE 5: Methods in Python for balancing the dataset

Modules	Algorithms
ADABOOST	Adaboost.M1, AdaC2.M1, PIBoost, SAMME, AdaBoost.NC
HDDT	MCHDDT, HDDTecoc, HDDTova
imECOC	imECOC+Dense, imECOC+Sparse, imECOC+OVA
Multi-IM	MultiIM+OVO, MultiIM+OAHO, MultiIM+OVA, MultiIM+A&O
FuzzyImb	FuzzyImbECOC
DOVO	DOVO
DECOC	DOVO+imECOC

TABLE 6: Modules and Algorithms in Multi-Imbalance

iv. Python

Python is a powerful language for scientific computing. Python provides a package named imbalanced-learn, which offers a couple of techniques to deal with imbalanced class problems [21]. This package offers four strategies to take on the imbalanced dataset. They are undersampling, oversampling, a combination of both and ensemble learning [22]. The following table 5 shows the details of methods available in Python for making the data balance.

v. Multi-imbalance

Multi-imbalance is software developed using Matlab. This can also be implemented in free software OCTAVE [23]. This software has seven modules that offer different algorithms to handle multi-class imbalanced problems. The seven modules and their included algorithms are shown in the following table 6.

vi. Spark

Spark is a framework that is extensive for handling multi-class imbalanced big data. This software supports resampling techniques like SMOTE, undersampling and oversampling. Furthermore, it provides two methods, informative resampling and novel partitioning for SMOTE in Spark nodes. Informative resampling conducted an analysis of difficulties in instance level and labelled it [24]. This can be used to select instances to mark their existence in balanced datasets. The second method includes a clustering-based partitioning in every Spark node that eases the absence of spatial coherence between examples from every class on account of random data splitting among nodes [25]. This helps the creation of the right instances for making the class distribution balanced. Informative resampling method and clustering-based partitioning in Spark nodes gain accuracy in the prediction process, and it is fitted for distributed environments.

IV. CONCLUSION

This paper discusses the issue of imbalanced data problems. Imbalanced data are always a burden for researchers. This paper points out the various methods of solving imbalanced data problems[26]. A detailed description of the different software and its various packages that can deal with imbalanced problems is discussed here. This paper will give students and researchers an idea about the algorithms used to solve imbalanced data and the tools to be used to solve the imbalanced scenario. Researchers can make use of these tools to solve their skewness problems.

REFERENCES

[1] Yoga. P, I. Pratama, and A. F. Nugraha, "Data level approach for imbalanced class handling on educational data mining multi-class classification," in 2018 Int. Conf. on Information & Communications Technology, ICOIACT 2018,vol. 2018 Jan, doi:10.1109/ ICOIACT. 2018. 8350792.

[2] V. S. Spelman and R. Porkodi, "A Review on Handling Imbalanced Data," Proc. 2018 International Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018, no. December, pp. 1–11, 2018, doi: 10.1109/ICCTCT.2018. 8551020.

[3] Shuo.W and X. Yao, "Multi-class imbalance problems: Analysis and potential solutions," IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 42, no. 4, pp. 1119–1130, 2012, doi: 10.1109/TSMCB.2012.2187280.

[4] Ranjana. S and R. Raut, "Review on Class Imbalance Learning: Binary and Multiclass," Int. J. Comput. Appl., vol. 131, no. 16, pp. 4–8, 2015, doi: 10.5120/ijca 2015907573.

[5] Chaitra P. C and R. S. Kumar, "A review of multi-class classification algorithms," Int. J. Pure Appl. Math., vol. 118, no. 14, pp. 17–26, 2018, [Online].

[6] Vitor A. De and N. Do, "Techniques to deal with imbalanced data in multi-class problems: A review of existing methods," 2020.

[7] Alberto F, et al., "Analysing the classification of imbalanced datasets with multiple classes: Binarisation techniques and ad-hoc approaches," Knowledge-Based Syst., vol. 42, pp. 97–110, 2013, doi: 10.1016/j.knosys.2013.01.018.

[8] Ghorbani R. and Ghousi R., "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020. 2986809.

[9] V. López, et al., "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci. (Ny)., vol. 250, pp. 113–141, 2013, doi: 10.1016/j.ins.2013.07.007.

[10] M. Lango, "Tackling the Problem of Class Imbalance in Multi-class Sentiment Classification: An Experimental Study," Found. Comput. Decis. Sci., vol. 44, no. 2, pp. 151–178, 2019, doi: 10.2478/fcds-2019-0009.

[11] Shuo W. and X. Yao,"IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, PART B Multi-Class Imbalance Problems: Analysis and Potential Solutions," pp. 1–13, 2012, [Online].

[12] Issac T. et al., "KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining," Int. J. Comput. Intell. Syst., vol. 10, no. 1, p. 1238, 2017, doi: 10.2991/ijcis.10.1.82.

[13] Jesus A.F., et al., "KEEL: A software tool to assess evolutionary algorithms for data mining problems," Soft Comput., vol. 13, no. 3, pp. 307–318, 2009, doi: 10.1007/s00500-008-0323-y.

[14] Jesus A.F. et al., "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," J. Mult. Log. Soft Comput., vol. 17, no. 2–3, pp. 255–287, 2011.

[15] M. Imran, et al. "Data Mining of Imbalanced Dataset in Educational Data Using Weka Tool," Int. J. Eng. Sci.

Comput. IJESC, vol. 6, no. 6, pp. 7666–7669, 2016, doi: 10.4010/2016.1809.

[16] S. a I. Tools and R. Dimov, "WEKA: Practical Machine Learning Tools and," Seminar, no. April 2013, 2006.

[17] Mark H. et al, "The WEKA data mining software," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, 2009, doi: 10.1145/1656274.1656278.

[18] Max K., "Building predictive models in R using the caret package," J. Stat. Softw., vol. 28, no. 5, pp. 1–26, 2008, doi: 10.18637/jss.v028.i05.

[19] Nicola L, Giovanna M., and Nicola T., "ROSE: A package for binary imbalanced learning," R J., vol. 6, no. 1, pp. 79–89, 2014, doi: 10.32614/rj-2014-008.

[20] Bing Z., Zihan G. et al, "IRIC: An R library for binary imbalanced classification," SoftwareX, vol. 10, no. October, p. 100341, 2019, doi: 10.1016/j.softx.2019.100341.

[21] G. Lema. et al, "Journal of Machine Learning Research," J. Mach. Learn. Res., vol. 40, no. 2015, pp. 1–5, 2015.

[22] Gyorgy K., "Smote-variants: A python implementation of minority oversampling techniques", Neurocomputing, vol. 366, no. June, pp. 352–354, 2019, doi: 10.1016/j.neucom.2019.06.100.

[23] Cohengang Z. et al,"Multi-Imbalance: An open-source software for multi-class imbalance learning - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0950705119301042 (accessed Apr. 14, 2021).

[24] Alberto F., et al "An insight into imbalanced Big Data classification: outcomes and challenges," Complex Intell. Syst., vol. 3, no. 2, pp. 105–120, 2017, doi: 10.1007/s40747-017-0037-9.

[25] Erendira R. et al, "Data sampling methods to dealwith the big data multi-class imbalance problem," Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

[26] Rose Mary Mathew and R. Gunasundari, "A Review on Handling Multi-class Imbalanced Data Classification In Education Domain," 2021 Int. Conf. Advance Computing & Innovative Technologies in Engineering (ICACITE), 2021, pp. 752-755, doi: 10.1109/ICACITE51222. 2021. 9404626