OVERVIEW OF HANDLING IMBALANCED DATASETS IN MACHINE LEARNING

D. Kavitha[1] and Dr. R. Ramkumar[2]
1. Research Scholar, Computer Science, Nandha Arts and Science College, Erode.
E-mail: kavitha.d@nandhaarts.org
2. School of Computer Science, VET Institute of Arts and Science (Co-education) College, Erode

Abstract

Abstract— Data mining and Machine learning are the most important research areas which attracts researchers. Class imbalance problem plays a vital role in data mining. In real world, technology evolution makes increase in data in various problems. This is termed as data velocity and volume. In machine learning, the required accuracy level of imbalance data classification is not produced by various techniques. The number of observations differs for the classes in a classification dataset termed as imbalanced datasets which leads to inaccurate results. Machine learning algorithms are powered by data and it's important to have balanced datasets in a machine learning workflow. This paper describes the issues of imbalance datasets, differences between balanced and imbalanced datasets and various techniques for handling imbalance dataset problems. Of course, a single article cannot be a complete review of all the methods and algorithms, however hope that the references cited will cover the major theoretical issues which guiding the researcher in interesting research directions and suggesting conceivable combinations that have yet to be explored.

Keywords: balanced and imbalanced dataset, undersampling, oversampling and ensemble learning.

Why is imbalance an issue?

In current years, imbalance problems have got even more prominent. In many practical domains, imbalanced datasets occur, such as spotting unreliable telecommunication customers, detection of oil spills in satellite radar images, learning word pronunciations, text classification, information retrieval, detection of fraudulent telephone calls and filtering tasks, and so on. Machine learning algorithms are feed by data. Evan the best of the algorithms struggle to produce good results in the absence of a good quality dataset. The major differences in the distribution of the classes in the dataset is used to define an imbalanced dataset which means that a dataset is biased towards a class in the dataset. An algorithm trained on the same data will be biased towards the same class, if the dataset is biased towards one class. The model learns more from biased examples as opposed to the examples in the minority class. There may be circumstances where the model assumes any data fueled to it belongs to the minority class. As a result, the model may seem unmatured in its predictions, regardless of achieving high accuracy scores[2].

Balanced and imbalanced datasets

A dataset whose distribution of labels is roughly alike is referred as a balanced dataset. Labels in this context refer to a class associated with each data point. For instance, consider a dataset with two classes as female and male. If around half the distribution signifies the male class and the supplementary half represents the female class, then the dataset is balanced.

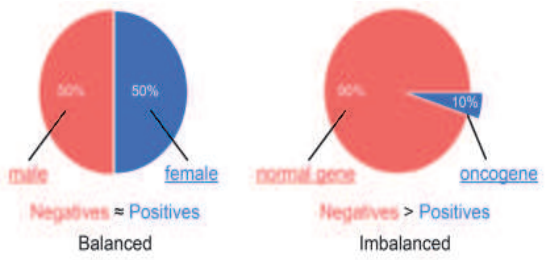


Fig.1. Example of balanced and imbalanced data

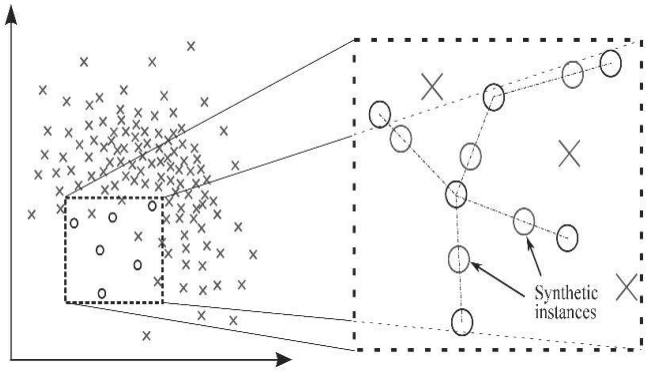
The distribution of an imbalanced dataset is pigeon-holed by very high differences amongst the classes involved. an imbalanced dataset may have a very high difference between the two classes, for example, the dataset of male and female classes mentioned above.

Techniques

There are methods to rectify the mentioned imbalance for to prevent the impact of imbalanced datasets. some are listed below.

Oversampling

Technique to amend unequal classes of data to create balanced datasets is called oversampling. This technique shot to rise the size of erratic samples to create a balance when the data is not sufficient. For example, study a classification problem with two classes and 100k data points. Positive class has 20k data points and the negative class has 80k data points. The minority class i.e., the positive class needs to be oversampled. For this the 20k data points are replicated four times to yield 80k. Hence, equal number of both positive and negative classes are created. 160K would be the size of the dataset as a result of oversampling. In the above cited example balance is accomplished no new or extra data is supplemented to the model. Oversampling is carried out by a technique Synthetic Minority Over-Sampling Technique (SMOTE) [1]. The approach is increased in the technique where the balance is obtained by duplicating the examples in a minority class. SMOTE synthesizes new examples as conflicting to duplicating examples. SMOTE opts for examples that happen to be in juxtaposition in a feature space. It then receipts a new example at a point along the segment, joining the adjacent examples. A classification approach where the likelihood of a data point belongs to one group or another depending on the data points in closest proximity to the data point is referred to as k-NN. SMOTE picks an instance of the minority class at random and computes its k-NN. A neighbour to it is then preferred aimlessly. After that, a synthetic example is formed at a point selected at random amongst the two examples. For a minority class to create balance, this process can generate as many synthetic examples as needed. A merit of oversampling is that there is no data loss from the original training dataset. All the data from both minority and majority classes are used. Hitherto the demerit of oversampling is that it causes overfitting



. *Fig.2.* **Generation of Synthetic Instances with the help of SMOTE**

Undersampling

When a class that exists is in copiousness, to balance the dataset undersampling focuses to reduce the size of copious class. For instance, study a similar situation of the undersampling technique where there is a classification problem with two classes and 100K data points. The positive class is of 20K data points, the negative class is of 80K data points. To balance the dataset undersampling the majority class is required. This results in choosing 20K data points aimlessly from the 80K available. Hence 20K positive and 20K negative data points are obtained, adding up to the total dataset size of 40K data points. Tomek links is an existing method for classifications problems which shots to improve the accuracy of data classification. Removal of as much as possible class label noise is done. Class noise is a type of label noise that changes the instances of labels assigned. By considering an example of assigning a positive label for a negative instance when a higher probability of being incorrect instances could be removed by tomex links even borderline examples referred as tomex link removal. Tomex link are points of different class labels which are closest neighbors to each other. Subsequently this technique makes it conceivable to identify points near different class labels, it is easy to get rid of them until none are left. In terms of undersampling, tomex link removal technique gets rid of unwanted overlap between classes. Neighbors of the same class in close proximity is considered as the result. The dispute of support vector machine algorithm can function similarly to the Tomek links method. SVM algorithm is very effective at classification tasks since it finds a hyperplane decision boundary that separates examples into two categories. For both classes the hyperplane is placed equidistantly. To maximizing the distance between the boundary and nearest examples from each class by which the hyperplane margin is defined. However, a distinguishing aspect between Tomek links and SVMs is that SVMs are ineffective at imbalanced classification also very sensitive in imbalanced datasets and produce a less than optimal performance[6].

Ensemble learning

An ensemble-based method plays a vital role to deal imbalanced datasets. The acceptance is that multiple learning methods are more operative than a single one. It's an approach that combines the performance or results of many classifiers to better the performance of a single classifier. Let's briefly define few of ensemble methods:

Bagging– This method attempts to apply similar learners on tiny sample populations which takes the mean of all predictions made [3][4].

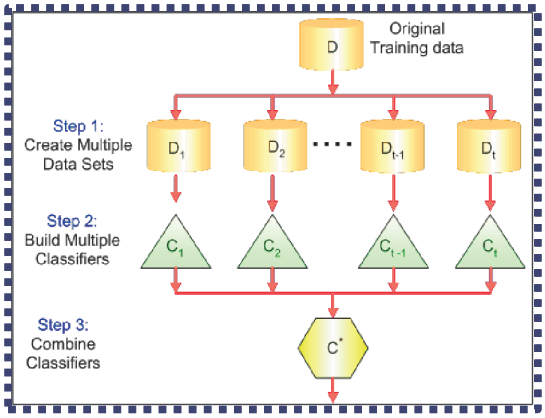


Fig. 3. Ensemble with Bagging

Boosting

A technique that adjusts an observation's weight based on the most recent classification. Boosting attempts to rise the weight of an observation that has been incorrectly classified. This is an iterative process, since new models are then trained on the inefficiencies of prior models. This makes the newer models better and stronger than the previous ones. The resulting ensemble has many machine learning models. These models boast different accuracies and can provide better accuracies when used together. Boosting also reduces bias error of the models[3][4].

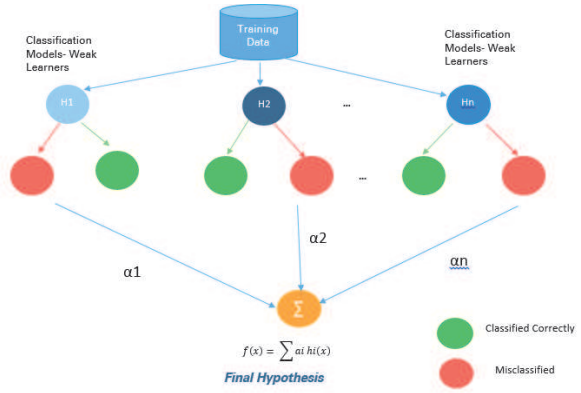


Fig.4. Ensemble with Boosting

Stacking

It is an attractive way of combining models which works on the basis of a learner to combine output from different learners and it can lead to decrease in either bias or variance error depending on the combining learner could be used.

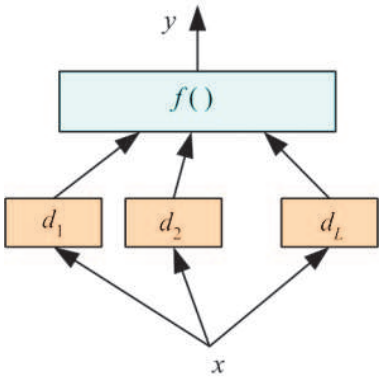


Fig.5. Ensemble with Stacking

The right evaluation metrics

Whenever there is an imbalance in class labels, that means that the classification accuracy metric is not ideal for model performance[5]. What do paper mean by this classification accuracy is a good metric when the samples belonging to each class are equal in number. Consider a scenario with 93% of samples from class C and 7% from class F in a training set. A model can simply achieve 93% training accuracy by predicting each training sample in class C. This is even assuming that it fails to predict any samples in class F correctly. Therefore, when dealing with imbalanced datasets, it's wise to use the correct evaluation metrics. Accordingly, it's not advisable to use accuracy as a measure of performance and may consider the metrics such as F1-score, precision, and recall.

Precision is the number of true positives against the total positive results predicted by a classifier.

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives}+\text{FalsePositives}}$$

The recall is the number of true positives divided by all the samples that should have been positive.

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives}+\text{FalseNegatives}}$$

F1-score shows us how accurate a model is by showing how many correct classifications are made. F1-score has a range between 0 and 1. The greater the score, the better the performance of the model.

$$\text{F1-score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

Conclusion

Literally advisable for the researchers to collect huge data with low distribution ratio to deal an imbalanced dataset on any. However, data collection is often an expensive, tedious, and time-consuming process. Imbalanced datasets can deceive both human beings and the model itself into believing that it generalizes well. To avoid such a scenario, it is important to understand how to correct the datasets imbalance. Paper has explored a few of the possible techniques to carry out this correction. However, the choice of technique is dependent on the nature of the problem which should be taken into consideration.

REFERENCES

[1] Jason Brownlee "SMOTE for Imbalanced Classification with Python", machine learning mastery, January 17, 2020

[2] Rushi Longadge et.al "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) Volume 2, Issue 1, February 2013

[3] Graczyk M, Lasota T, Trawinski B, Trawinski K. "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal", Asian conference on intelligent information and database systems. 2010. pp. 340–50.

[4] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F " A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches", IEEE Trans Syst Man Cybern C Appl Rev. 2012;42(4):463–84

[5] Learning from imbalanced data. Retrieved from https://www.jeremyjordan.me/imbalanced-data/

[6] Benoît Frénay et al , "Classification in the Presence of Label Noise: A Survey", IEEE Transactions on Neural Networks and Learning Systems 25(5):845-869, 2014.