How to make sense of multi-class imbalanced classification in Confusion Matrix

Timothy Jeoung

Faculty of Science

Queensland University of Technology

2, George Street, Brisbane, 4001, QLD, Australia

***Abstract*— Imbalanced data problems can lead to bias while making decisions or implementing policies [3]. Therefore, the Confusion matrix can be difficult to understand when there is an imbalance in the data. In this paper, other methods based on the confusion matrix will be introduced to solve the imbalance problem by improving the performance on evaluating. The methods that will help to improve the evaluation of the confusion matrix are SMOTE oversampling to balance the imbalanced datasets and confusR to gain insight into the contribution of the classifiers. The results for the proposed methods have improved the performance of evaluation by balancing the imbalanced dataset. The imbalance problems with datasets and classes can be solved by using the proposed methods along with the confusion matrix.**

Keywords—Confusion matrix, imbalance, SMOTE oversampling, confusR

# Introduction

The confusion matrix is one of the well-known and common methods being used in Machine Learning. It’s a significant method used to describe the performance of a classification model on training and test data. Also, the confusion matrix gives an insight into the accuracy of a predictive model and which classes are being predicted correctly and incorrectly [4].

 However, it can be difficult to understand and recognize the results of the confusion matrix when there are various possible classes that can lead to an imbalance classification.

One of the causes that introduce bias in the interpretation of results is data imbalance [3]. It’s an issue that mainly occurs in classification performance. Furthermore, it is noticed that often the data causes a problem of imbalance which means one class has a higher percentage compared to another class. To simply define the class imbalance, it's an imbalanced dataset with unequal class distributions [3]. Hence, imbalanced data impacts the classification accuracy which means the confusion matrix predictions will be led to a poor degree of accuracy.

Thus, many researchers studied this research problem to deal with imbalance classification. However, the gaps and limitations of current research are that most of them only mainly focused on binary classifications problems and there is only few research about multi-class classification problems.

It is significant for this problem to be solved as there are significant issues in modern days with imbalanced data which leads to poor classification accuracy.

One of the research questions that can be asked to solve this problem is the following *‘What tools and techniques exist for measuring the performance of multi-class classification models for imbalanced datasets?’*. Answering this question will contribute to the research problem which is dealing with imbalance classification on multinomial. Therefore, gaining knowledge on what tools & techniques are used for performance on multinomial classification will help to understand how to deal with imbalanced multi-class datasets and improve the accuracy of the confusion matrix.

To fill the gaps and limitations, in this research to deal with class imbalance a method called SMOTE Oversampling will be used to balance the imbalanced dataset. SMOTE oversampling is a Synthetic Minority Oversampling Technique; it’s useful to address imbalanced datasets by oversampling the minority class.

Furthermore, another method called confusR will be applied for confusion matrix visualization. The method confusR is developed by academics named David Lovell and Dimity Miller.  The confusR is a method that visualizes the confusion matrix and makes clear contributions of the classifier and the contribution of the prior abundance of different classes [1].

Therefore, applying SMOTE oversampling will help to address the problem of dealing with a class imbalance in multinomial classifications. Moreover, the confusR method from academics will help to gain an insight into the change of discriminative ability of each class when given different prior probabilities of data by applying SMOTE oversampling to the imbalanced datasets.

The purpose of this research is to understand how to deal with the class imbalance in multinomial classification and improve the accuracy of the confusion matrix by making sense of the evaluation. The next section of this research will introduce the related works of the class imbalance problems. Afterward, the overall, specific methods and the experiment will be implemented to show how to deal with the multinomial imbalanced classification problems. Lastly, the results of the experiment and limitations will be discussed and implemented.

.

# Review of the literature or Related Works

[1]David Lovell. (2021). confusR draft vignette.

In this literature, David Lovell introduces a new development method called ‘confusR’ to visualize the confusion matrix. The visualization of the confusion matrix shows the contributions of classifiers and different classes. Also, the literature helps to gain an understanding of multinomial classifier performance. This literature is significant for this research as it’s the explanation of the new method ‘confusR’ which is one of the main methods to be used for this research. Thus, it relates to the research problem to give an insight into the contributions of the classifier and the contribution of the prior abundance of different classes. However, this literature does not solve the problem of imbalance it only shows the contributions of classifiers. Thus, the research will fill the gaps by implementing the method that could balance the class imbalance and then visualize it with confusR to see the effectiveness.

[2] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix.

Pattern Recognition, 91, 216–231.

This literature introduces a study of the impact of class imbalance on classification performance matrices. Moreover, it also introduces the definition of class imbalance and the impacts of imbalance in certain metrics. The literature is mainly about how confusion matrices work in detailed explanation, along with how bias can be detected on imbalance classes in the binary confusion matrix. The literature will be helpful to understand what the confusion matrix is specifically through detailed explanation. However, it mainly focuses on binary classes thus there are limitations as it doesn’t focus on multi-class. Thus, this literature will be helpful on the research problem as it gives deep knowledge on the confusion matrix. Moreover, the gaps can be filled by this research by applying a SMOTE oversampling method that could deal with multi-class imbalanced classification.

[3] Kulkarni, A., Feras, A., Chong, D. (2020). Foundations of data imbalance and solutions

for a data democracy.

This literature introduces the complexity of the concept of a confusion matrix. Moreover, it also explains imbalance data basics & the degree of class imbalance. It mainly focuses on how to deal with imbalanced data and under-sampling. This literature will be useful on the research problem to understand how to deal with classifications issues. There are various methods introduced to deal with imbalanced data. Thus, these methods can be compared to the new method ‘confusR’. It will be helpful in answering the following research question ‘What tools and techniques exist for measuring the performance of multinomial classification models for imbalanced datasets?’ as it introduces various techniques and tools that can be used for measuring the performance of the classification model. This will help to learn various methods and know what kinds of methods are suitable in terms of class imbalance. However, as there are various methods introduced it only introduces the method briefly. Therefore, in the research, it will introduce specific method such as SMOTE oversampling in specifically to deal with class imbalance.

[4] Brownlee, J. (2020). Multi-Class Imbalanced Classification.

This literature introduces the method SMOTE oversampling and Cost-Sensitive Learning. It mainly focuses on explaining how SMOTE oversampling works with an example of an experiment applying SMOTE on imbalanced classification. This literature is related to the research problem of dealing with class imbalance by introducing the method SMOTE oversampling to deal with a class imbalance on multi-class classifications. The gap and limitation of this literature were that it did not apply the confusion matrix method to see the prediction accuracy. To fill the gaps and limitations the research will apply a confusion matrix to see the prediction accuracy compared with the multi-class datasets that have applied SMOTE oversample and not. Furthermore, it will use the academic method confusR to see the contributions of classes as well.

[5] Beauxis-Aussalet, E., Hardman, L. (2014). Visualization of Confusion Matrix for Non-Expert Users.

This literature introduces the use of various matrices in evaluating errors in classification problems. It mainly focuses on the confusion matrix and discusses the issues with the confusion matrix as well. The literature also focuses on non-expert users to avoid them from using inappropriate metrics for their use case or misinterpreting them. This literature has relevant to finding suitable metrics to deal with issues in the confusion matrix. However, the gaps and limitations are that there is no specific information on dealing with multi-class classification issues on the confusion matrix. Thus, this research will work intend to fill this gap by applying SMOTE oversampling and clear visualizations of dealing with a class imbalance on multi-class datasets.

[6] Branco, P., Ribeiro, Rita P. (2017). Relevance-Based Evaluation Metrics for Multi-class Imbalanced Domains.

This literature proposes a relevance-based evaluation that incorporates user preferences by allowing the assignment of differentiated importance values to each class [6]. The focus of this literature is to find the relevance of each class by detecting discrimination in two classes and multi-classes. The literature has the relevance of dealing with the same problem of class imbalance. It also introduces only a few solutions that exist for the multi classes’ imbalance problem. However, it does talk about how to deal with a class imbalance on the multi-class problems but briefly without any examples. Thus, in the research, it will specifically use one of the methods such as SMOTE oversampling that can deal with the class imbalance and show an example of an experiment to explain for non-experts to understand.

# MAterials and Methods

## Data-Oriented Research

The overall research method of this project is *data-oriented research* as it has used a design data collection for the experiment.

## Type of data

The type of data that was used is an imbalanced multi-class classifications dataset, as the research problem is to solve class imbalance on multi-class classifications problems. The data that was collected is a glass multi-class classification dataset referred to as ‘Glass identification’. Moreover, the data was collected via online sources from GitHub credited by Vina Spiehler in 1987.

Glass dataset is about identifying what types of glasses they are among seven types of glass which are [4]:

* Class 1: building windows (float processed) 70 examples
* Class 2: building windows (non-float-processed) 76 examples
* Class 3: vehicle windows (float processed) 17 examples
* Class 4: vehicle windows (non-float processed) 0 examples
* Class 5: containers 13 examples
* Class 6: tableware 9 examples
* Class 7: headlamps 29 examples

Referring to the above seven class list, it has 214 examples and the number of examples in each class is imbalanced. Although there are minority classes, all classes are equally significant in the prediction problem [4]. Class 4 will be avoided in the experiment as there are not any examples. Thus, Class 1 was labelled as 0 and Class 2 was labelled as 1 and the following will be numbered in order up to 5 which is total of six classes for the experiment.

## Tools & Environments

To solve the class imbalance on the glass identification dataset the experiment was run in two languages. The first language is *Python* to use SMOTE oversampling technique to balance the imbalance multi-class which will be introduced shortly after this section. The *Python* work was done in the integrated development environment (IDE) called Jupyter Notebook.

Another language that was used is *R language* in the *R studio* environment. *R language* was used to use the academic method confusR to see the contributions of the classes in visualization.

## Specific methods

In this research, there will be two specific methods to be used to solve the class imbalance in multi-class classification problems.

One of the main methods to deal with a class imbalance on multi-class classification is a method called *Synthetic Minority Oversampling Technique (SMOTE).* The problems with class imbalance are that there are only a few examples of the minority class for a model to efficiently learn the decision boundary [4]. One of the effective ways to solve this problem is to oversample the examples from the minority class in the training dataset prior to fitting a model [7]. This process improves on examples that have been duplicated from the minority class which makes to synthesize new examples from the minority class.

To simplify, it’s a type of data augmentation for tabular data which can be extremely effective. Therefore, the most used and well-known data augmentation technique is SMOTE.  The SMOTE was described by Nitesh Chawla in 2002 from their research paper titled ‘SMOTE: Synthetic Minority Over-Sampling Technique’ [7]. SMOTE oversampling works by choosing examples that are near in feature space, then it draws a line between the examples in the feature space, and along the line, the new example is drawn. In specifically, SMOTE chooses minority class examples first at a random, and finds its *k* nearest minority class neighbors. The synthetic example is then formed by selecting one of the *k* nearest neighbors *b* at a random. Then it connects *a* and *b* to create a line segment in the feature space. The synthetic examples are created as a convex combination of the two selected examples *a* and *b* [7]. By applying the procedure as above, it will allow to the formation of as many synthetic examples as possible for the minority class which is oversampling the minority class to make it balanced to solve the class imbalance problem. Therefore, the SMOTE oversampling technique was applied on Glass identification datasets to deal with class imbalance.

Another, main method that was used in the problem is the academic method called *confusR* by David Lovell. The motivation of confusR is to understand the multinomial classifier performance. Comparing the binary models, it’s not simple to have well-established measures of performance. Thus, confusR is still in development but it is effective to measure the multinomial classifier performances.

The method of confusR works by approaching the Bayes rule with odds to have a better understanding of the discriminative performance in terms of prior odds and Bayes factors [1]. This was done by the following equation.

*Fig. 1. Equation of Bayes rule with odds [1]*

Text

Description automatically generated with low confidence

In the case of glass identification datasets, the *posterior odds* are the odds of being the types of glasses after the prediction then this equals to a *prior odd* whichare the odds of being the glasses which is before the prediction. Afterward, it multiplies both odds for the test to predict which types of glass are used. Finally, to make the visualization the logarithmic scale is applied as *Figure 2* equation.

*Fig. 2. Logarithmic scale [1]*



Hence, the confusR method was applied to gain an insight of understanding into the change of discriminative performance of each class in glass identification datasets that applied SMOTE oversampling.

## Description of eact stage

Using the methods, tools, environment on glass identification imbalanced datasets, the experiment was tested as the following steps.

1.        Understanding the Glass identification dataset

2.        Apply SMOTE oversampling on the dataset using Python

3.        Visualize with Matplotlib to create a bar chart

4.        Export confusion matrix of the result

5.        Save the confusion matrix into csv file

6.        Load the confusion matrix that was saved to csv file into R studio

7.        Visualize the confusion matrix using confusR in R Studio

8.        Evaluated the effectiveness of the SMOTE

9.        Repeated the steps excluding step 2 & step 8

10.     Made comparison with Imbalanced (default) vs balanced (applied SMOTE)

# Results and Discussion

## Results of experiment

Applying the SMOTE oversampling technique to glass identification datasets that were imbalanced has solved the problem of class imbalance. The following works using Python shows a clear difference of default version which is imbalanced vs balanced which applied SMOTE oversampling:

Fig.3. Matplotlib bar chart Imbalanced

Chart, bar chart

Description automatically generated

Referring to Figure 3, it’s a default version of the glass identification dataset which clearly shows that 0 and 1 have the most examples compared to other classes. Classes 0 and 1 have more than 70 examples, and other classes such as 3 and 4 have less than 15 examples. Thus, it is clearly shown that there is a class imbalance that will impact classification accuracy.

Fig. 4. Imbalanced confusion matrix default version

Chart

Description automatically generated

The above default version of the confusion matrix (figure 4) has shown that the majority of predictions are mostly only occurred in classes 0 and 1. Therefore, the above confusion matrix has predicted the classes with a poor degree of accuracy due to class imbalance.

Fig.5. Matplotlib bar chart balanced with SMOTE

Chart, bar chart

Description automatically generated

By applying SMOTE oversampling, figure 5 shows the distributions of examples in each class now have the same examples which are 76 examples. The bar chart provides a strong visual indication that all classes now have an equal number of examples.

Fig. 6. Balanced confusion matrix applied SMOTE

Chart

Description automatically generated

Now by applying SMOTE oversampling to the glass identification dataset it shows a high diagonal value in figure 6. In a balanced confusion matrix (Figure 6) the 6 classes in the dataset have been predicted to be classed with a good degree of accuracy compared to the Figure 4 confusion matrix. Thus, balancing the glass identification dataset with SMOTE has improved the performance of classification accuracy.

After applying SMOTE technique to the glass identification dataset, the academic method confusR also has been applied for the visualization. In the next part, it will show the two visualizations that have been created by confusR method in R studio using R language. It has given an insight into the change of discriminative ability of each class when given different prior probabilities.

Fig.7. Imbalanced logscale visualization from confusR

Chart, box and whisker chart

Description automatically generated

To avoid confusion, the classes from 0 to 5 have changed class labels to alphabets characters A to F. Figure 7 is an imbalanced confusion matrix visualized in log scale created from confusR method. The left panel shows the prior odds of glasses and the effect that the classifier has on updating those odds [1]. The right panel shows the effect of the classifier alone [1].

According to figure 7, shows class A with prior probability at 0.3 and posterior at 0.68, class B prior at 0.4, and posterior at 0.65. Class C has no records as there wasn’t any example predicted in the confusion matrix due to class imbalance. Moreover, class D has a prior probability at 0.05 and a posterior at 0.55. Lastly, class F has a prior probability at 0.15 and a posterior at 0.9. As it's imbalanced they all have huge differences in prior probabilities and class A & B has the highest prior probabilities as they had most examples in the datasets.

Fig. 8. Balanced logscale visualization from confusR that has applied SMOTE

Chart, box and whisker chart

Description automatically generated

 Figure 8 is a log scale that has applied SMOTE oversampling which has balanced the glass identification dataset. Thus, the confusion matrix had a good degree of accuracy for each class. Compared to the figure 7 log scale, it can be clearly identified that all classes of prior and posterior probability on the left panel have increased. One of the big differences is that all classes now have a similar amount of prior probability around 0.2. Moreover, class C now has a record of being oversampled by SMOTE. This approach separates the characteristics of the classifier from the prior distribution of the classes it is applied to [1]. Therefore, confusR has given an insight on the performance of using SMOTE to address imbalance classification. It has given a clear view into the change of discriminative ability of each class when given different prior probabilities.

## Discussion

The most important findings of this research are that class imbalance on multinomial classification problems can be solved. It has been proven in the experiment by applying SMOTE method it oversamples the minority classes to balance each class in glass identification datasets which have multi-class and imbalanced examples for each class.

Moreover, the academic method confusR is also an important finding because it gives a clear insight of understanding the discriminative performance of a model in terms of prior odds and Bayes factors. As it shows in the experiment comparing Figures 7 and 8, it gives a clear view of the prior & posterior probabilities and the likelihoods (LR+).

Therefore, by using SMOTE oversampling to deal with class imbalance it supports to increase the performance of confusion matrix prediction accuracy. Moreover, the confusR method supports understanding and recognizing the results of the confusion matrix when there are various possible classes. Hence, the two methods have supported to answer the research problem to make sense of multi-class imbalanced classification in confusion matrix and answered the research question *‘What tools and techniques exist for measuring the performance of multi-class classification models for imbalanced datasets?’.*

Compared to other related literature reviews, this research has clearly demonstrated an experiment dealing with a class imbalance on multinomial classification problems. Also, it has provided a clear insight into understanding the complexity of the confusion matrix based on multi-class classification with confusR method. The other works of literature majority only show dealing with binary classification imbalance problem and have not gained any insight into the change of discriminative ability of each class when given different probabilities.

In the personal interpretation of the findings in this research, the SMOTE oversampling is a simple method to implement and interpret. Thus, it is easy to use to deal with the class imbalance in classification performance. However, there are still limitations of SMOTE such as during the generating process, SMOTE does not take into consideration neighboring examples that can be from other classes [8]. This can lead to increasing the overlapping of classes and can introduce additional noise. In figure 6 confusion matrix that applied SMOTE shows slightly more false positives compared to the default confusion matrix. This could be because of the additional noises formed due to overlapping. Moreover, SMOTE can be not very practical for high dimensional data [8]. Thus, further work and research are required to experiment with SMOTE with high-dimensional data.

Furthermore, the academic method confusR was a useful method to gain an insight into the change of discriminative ability of each class in the glass identification dataset. However, the confusR is still in the development stage and it is not finalized. Thus, for non-experts and not having experience in using the R language in R studio could be one of the limitations. The confusR would be more useful to everyone including the non-experts when it’s finalized with a clear instruction to guide using the method. Despite this, confusR is a powerful method that should be used widely as there are not many methods that give a clear understanding of the complexity of the confusion matrix with visualization.

# Conclusion

In conclusion, the significant findings of this research were the two methods SMOTE oversampling and the confusR to deal with class imbalance in multinomial classification performance. The SMOTE method has advanced the field by oversampling the minority classes to balance the class imbalance in multinomial classifications. Moreover, the confusR method has advanced the field by giving an insight into the change of discriminative ability of each class when given different probabilities by applying SMOTE. Therefore, the two methods have supported to address the research problem which was knowing how to make sense of multi-class imbalanced classification in confusion matrix.

It was significant for this problem to be solved as there are significant issues in modern days with imbalanced data which leads to a poor classification accuracy. Also, the majority of the solutions were only focused on binary classification problems for class imbalance problems. Therefore, these research findings were extremely significant to know how to deal with a class imbalance on the multinomial classification to understand the confusion matrix.

Nevertheless, there are still limitations that exist in these two methods. The SMOTE method has a limitation of overlapping which can increase the noises and SMOTE can be not practical for high dimensional data. Therefore, for future work, there should be more tests and experiments applying SMOTE oversampling on high dimensional data to identify a clear reason for not being applicable to use SMOTE in high dimensional data. Furthermore, for confusR, as the method is still in the developing stage it is difficult to understand the method and apply it to suitable cases. To avoid this limitation the confusR needs to be finalized with a clear instruction to guide using the method for everyone including the non-experts.

However, despite the limitations, the two methods were essential in solving the problem of having a clear understanding of the confusion matrix with a class imbalance in multinomial classifications.

# References

[1]David Lovell. (2021). confusR draft vignette.

[2] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix.

[3] Kulkarni, A., Feras, A., Chong, D. (2020). Foundations of data imbalance and solutions for a data democracy.

[4] Brownlee, J. (2020). Multi-Class Imbalanced Classification.

[5] Beauxis-Aussalet, E., Hardman, L. (2014). Visualization of Confusion Matrix for Non-Expert Users.

[6] Branco, P., Ribeiro, Rita P. (2017). Relevance-Based Evaluation Metrics for Multi-class Imbalanced Domains.

[7] Brownlee, J. (2021). SMOTE for imbalanced classification with Python

[8] Sayak, P. (2018). Diving Deep with Imbalanced Data

[9] Fernandez, A., del Rio, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced Big Data classification: outcomes and challenges. Complex & Intelligent Systems, 3(2), 105-120.

[10] Baranco, P., Torgo, L., Ribeiro, R. (2017). Relevance-Based Evaluation Metrics for Multi-class imbalanced Domains. Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 698-710).

[11] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35.

[12] Radecic, D. (2020). How to effortlessly handle class imbalance with python and SMOTE

[13] Nabi, Javaid. (2018). Machine Learning – Multiclass classification with imbalanced dataset

[14] Brownlee, J. (2019). A Gentle introduction to imbalanced classification

[15] Salrite, R. (2018). Demystifying ‘Confusion Matrix’ Confusion