

A Gaussian Mixture Model–Driven Hybrid Resampling Technique for Imbalanced Data Classification

Tahmina Asghar

The University of Faisalabad

Ahmed Shahzad

The University of Faisalabad

Qurat Ul Aen Naqvi

Riphah International University Faisalabad

Ikraam Asghar

Government College University Faisalabad

Hamza Saleem

Government College University Faisalabad

Mudassir Riaz

The University of Faisalabad

Sajawal Khan

The University of Faisalabad

ABSTRACT

The issue of class imbalance is prevalent in various real-world applications, including spam filtering, fraud detection, and medical diagnosis. It has a significant impact on how well learning algorithms function. Under sampling at random is a well-known way to overcome the issue. However, because of their randomness, it is unable to effectively extract the samples close to the majority and minority class cross-edge, despite the fact that these samples are crucial to a classifier since they affect the classification performance. This class rarity leads to classification methods behavior biased towards majority classes that cause misclassification. There are different methods for handling imbalanced data like data-level, algorithm-level, and hybrid. For handling this issue we propose a GMM-based approach, which is performed in three ways as GMM- based under sampling, GMM-based oversampling, GMM-based combination of under sampling and oversampling. For evaluation of results we use four matrices such as precision, recall, f-measure, and accuracy after GMM-based oversampling precision, recall, F1 is 83%, 86%, 84% in fraud detection respectively, accuracy is 0.43 and in GMM-based under sampling precision, recall, F1 is 0%, 100% and 1% and the accuracy is 0.50, while combining the GMM-based under sampling and oversampling the precision, recall, F1 is 0%, 99%, 1% respectively and the

accuracy of fraud detection is 0.43.

Key words: Data Mining, Machine learning, Imbalanced Data, Data Classification

Introduction

Most of the real-world problems of classification having big data follow a lengthy distribution tail i.e. few of the object class data is rare on the other hand few of them are abundant. This behavior tends to “Imbalance class distribution”. A dataset is called to be imbalanced when few of its ‘minority’ classes are under representations in the comparison of most of its ‘majority’ classes. A data set that contains an inadequate distribution of data among its classes is called an imbalanced dataset. This class rarity leads to mislead the characteristics of the minority class. [15] In various problems, the minority classes are interesting or positive classes for example in credit card fraud detection where most of the transactions are not fraud but only a few of them and in medical diagnosis for disease identification where most of the patients data are healthy and only a few is un-healthy then the disease identification is more difficult having great interest. [12] Working on these types of datasets is a very difficult task especially in the scenario of big data whereas the desired outcomes often focus to achieve with non-standard machine learning approaches. This is indispensable to thoroughly study and understanding class imbalanced techniques of machine learning because most of the real-world problems contain such skewed data. [3, 4]

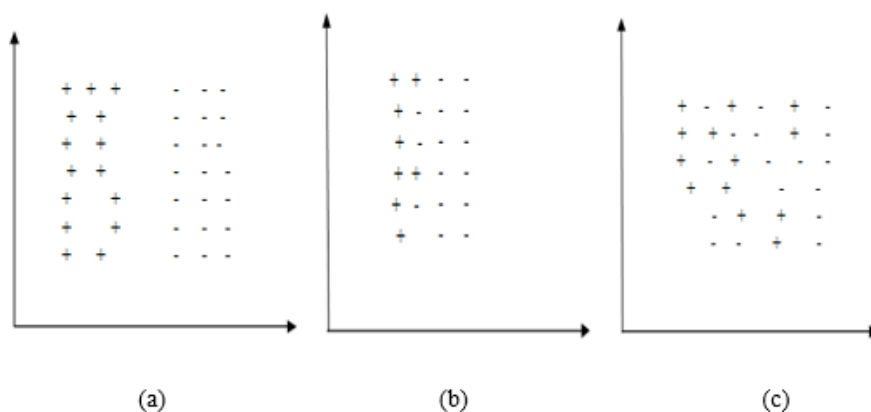


Figure 1: Classification imbalanced data



Figure 2: Characteristics of imbalanced data

Minor size of samples: Usually imbalanced datasets not contain sufficient samples of minority class the error rate that is effected due to the imbalance distribution of class's decreases when number of the samples from minority classes is increased in its illustrations (fixing its imbalance ratio). In this way, positive samples defined through these patterns can be learned better regardless irregular distribution of class. Nevertheless, this is an unreachable fact in different problems of real-world. [8]

Class Overlapping or separability: When class overlapping occurs, it very difficult to encourage discriminative directions. As a result, the general rules are convincing to misclassify the instances with low quantity (samples from minority class). If dataset does not contain any overlapping classes, one of the simple classifier can classify appropriately. [1]

Small Disjuncts: In a dataset small disjuncts are represented where minority class is designed through sub concepts. In many problems there are representation of some small hidden disjuncts. The representation of these small disjuncts as minority class increased the complexity of dataset because these are unbalanced classes.[20]

Noisy Data: Having imbalanced data problem, the minority class has great influence due to noise as compared to the casual cases, because the positive class consists of fewer examples so that it will yield some noisy samples for impacts on learning sub concepts. [19]

Dataset Shift: For following the different distributions on training and testing phase is known as dataset shifting. It's an accustomed problem that effects on all kinds of data classification problems, this oftenly held due to biased behavior in the selection of data samples. Dataset shift problem is commonly rise in the

Online ISSN

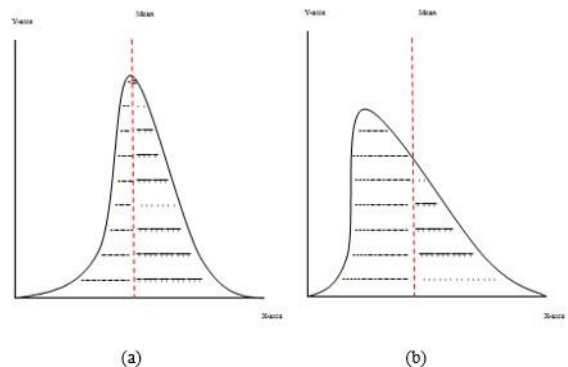
3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

scenario of big data classification handling due to high level imbalanced domains, the minority group is sensitive due to errors in classification because of the low numbers of samples typically. [19]



High Skewness: Mostly the minority and majority classes are not distributed normally that cause the data skewness. Data is skewed in positive and negative manner. Skewed data in positive manner consists of disadvantage to the right line of distribution, on the other hand skewed data in negative manner shows the disadvantage on the left of the normal line distribution. [9]**Figure 3: (a)** Normal skewness **(b)** Positiveskewness.

Redundant and Irrelevant features in dataset: The redundant or irrelevant features in a dataset also impacts on the imbalance level that can affects upon the classifier performance. Irrelevant features also affects upon the learning algorithms accuracy. Different approaches or methods are applied upon datasets for selecting appropriate features. [9]

Class imbalance methods:

Different methods for handling class imbalance are categorized into three groups that are:

Data level:

Data-level approaches to reduce the imbalance level consist of under sampling and oversampling. These types of techniques minimize noise like mislabeled
Expressing and seeing the thoughts of yourself and others is a major factor in human behavior. Whether we want to communicate with others or make a decision, we must first try to understand the concept that concerns us. In the world of computing, feedback is also known as emotion analysis. It involves exploring the language written by a man for the emotions, attitudes and predictable appearance of people. Not only can he view information on a sentence or the subject of a document, but he also views the views expressed [27]. Characteristics of Imbalanced Data: These are some characteristics of imbalanced dataset that creates difficulties during learning. [15] The latest spam discovery was not really about fake reviews, as initial reviews were not common online reviews. Numerous studies have been done to detect false news and spam, but the detection of false comments has never been the main benefit of spam detection. Yet, as online reviews have changed the perception of many products and

services, many people have purchased products and services based on reading online reading reviews [27]. Only 350 million people worldwide are depressed can lead to suicide. According to research, nearly one million people who suffer from depression die each year. it is not a conference to represent a private patient, but is a forum for raising the voice of authorities so that relevant authorities can come together to identify the candidate rather than their job. State psychological certificate. We respect the restrictions on the use of social media data, so do not use any user information [28].

The Experiment contains a number of techniques for analyzing both classification and classification of SVM and nonlinear Bayesian classification. In addition, there are three columns, the first being binomial notions (not depressed, not depressed), and the second paragraph has some form of depression (one in nine during depression), and therefore the third column contains the education system[28].

LITERATURE REVIEW

Referenced Paper	Method	Level	Network Type	Description	Strengthens	Weaknesses
[5, 10]	ROS	Data	CNN	Randomly Repeating the minority class samples until the balanced level was achieved	Useful for small data sets	In the case of big data, scenario increasing the training time
[5]	RUS	Data	CNN	Undersampled all the majority classes until the balanced level was achieved	Reduce training time in big data problems	Redundancy occurs
[13]	Two-phase learning	Data	CNN	Fine-tuning with the whole dataset on pre-trained model with ROS or RUS	Show effectiveness in minority class performance as well as maintaining the majority class performance	Higher variance or higher bias impact on final model class wise performance

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

[18]	Dynam ic sampli ng	Data	CNN	Adjust the sampling ratio class wise F1 score based onprevious repetitions	The most nominated feature is the capability to self- adjust the sampling rates	Problematic in class rarity where few +ve samples occur
[26]	GM M- GMU S	Data	CNN	Combine the Gaussian Mixture Model (GMM) with undersamplin g the majority class and then applied Probability Density Function(PD F) as the cross edge of two classes	Shows better results than other models in credit card fraud detection Improve the recognition rate of abnormal samples improve the accuracy rate	Neglect the minority class samples Further improvement needed
[24]	MFE and MSFE loss	Algorith m	MLP	A loss function that permits+ve and – veclassescontr ib ute to loss equality	Minimized the difference between FNE and FPE and also minimize the training time	Problem to problem performanc e gain will vary that shows further tryouts to be needed
[14]	Focal loss	Algorith m	CNN	A loss function that minimizing the influence on overall loss by down weighting easy-to- classify samples	Minority classes to play a part by updating the weight without majority classes loss and also reduced	Addition in imbalance classes slower the model speed due to the adjustment with per class

					training time	learned speed
[23]	CSDNN	Algorithm	MLP	DNN with 2 hidden layers consisting of 128 and 64 neurons each layer composed with CE loss function integrate a predefined cost matrix	Suitable for small datasets	Recognizing a perfect cost matrix in DNNs and BD is very costly and also time-consuming
[16]	CoSe n CN N	Algorithm	CNN	From end to end backpropagation and integration	Ability to specify an appropriate	Increased the training time
				with the output layer, the cost matrix was attained	cost matrix throughout the training	due to iterations
[8]	CSDB N- DE	Algorithm	DBN	Use the DE algorithm to improve the cost matrix in each iteration and incorporate these	Ability to specify an appropriate cost matrix throughout training	Time-consuming and difficult to implement

				costs into DBN		
[5]	Threshold moving	Algorithm	CNN	The threshold was applied through the division of network outputs for the estimation of preceding probability in each class	Overall accuracy improved by moving the threshold	Accuracy score having no ability to discriminate class wise performance
[26]	Category centers	Algorithm	CNN	CC combine deep CNN extractor, transfer learning, and K-NN discriminator for removing imbalanced distributions	Minimize avoid the boundary error in class imbalance	Highly dependent on DNN ability to discriminative feature generation
[6]	Very-deep NNs	Algorithm	CNN	Observe the convergence tolls by using 50 layers of CNN network	Solutions are valid and acceptable	Very costly Required more memory Time taken process

[11]	LMLE	Hybrid	CNN	Produce more discriminate features with quintuplet sample and three-layered header pivot	Shows the higher performance outputs	Complexity occurs High cost Expensive Computations
[2]	DOS	Hybrid	CNN	Using KNN and micro cluster loss oversampled minority classes in deep feature space	May useful with further extending 's	More complex Very costly Discourage to use
[7]	CRL loss	Hybrid	CNN	Build more dominant features through rigid sample mining and class rectification loss	Suitable for imbalanced data on large scale throughout a comprehensive testing	Not good for low-level imbalance data Complexity exists Costly

All the hybrid methods developed for class imbalance will be more difficult and complicated to implement. Likely to be expected that both the combination of data and algorithm levels improve the performance of classification Due to the ease of use and flexibility decrement, this is challenging to implement them in new problems. [12]

MATERIALS AND METHODS

Dataset:

For the companies of credit card it is very important to be able in the detection of fraud transactions because peoples are not ready to be pay against the transactions that are based on fraud. For this purpose a dataset has been analyzed and collected during the collaboration in the research of world line and a group of machine learning of UniversiteLibre de Bruxelles (ULB) on fraud detection and big data mining.

The dataset consists of credit cards transactions via European credit card holders in the month of September 2013. Dataset represent the transactions based on 2 days only that contains on 284807 total transactions out of them only 492 transactions are

fraudulent. This dataset is imbalanced highly as our positive class (fraud class) with only 492 transactions with just 0.172% of total transactions. The imbalanced ratio of the dataset is 578.87%.

Dataset contains only numerical variables as input that are exist in the basis of PCA transformations result. Due to the trustworthy problems of customers dataset cannot provide the real information and the complete background of data. Features from V1, V2, V3,.....V28 are in the form of principal components achieved from PCA the only attributes like “Time” and “Amount” have not been transformed through PCA. The attribute such as “Time” consists of elapsed in seconds among every transaction existing in dataset. The attribute “Amount” contain the amount of transaction. The attribute “Class” is a variable of response that takes the value as 1 if the transaction is fraudulent and takes the value as 0 in the case of non-fraudulent.

Environment:

The imbalanced data classification is done with python programming language. There are also some other languages are available for imbalance data classification, but python Python is considered better as compared to some other languages due to its suitability for data handling and for implementing different machine learning methods. Machine learning techniques such as Gaussian Mixture Undersampling and Gaussian Mixture Oversampling are used for imbalanced data classification on credit card imbalanced dataset. The coding environment that is used for imbalanced data classification is Pycharm that gives the results immediately on screen after the implementation of model.

Gaussian Mixture Model (GMM):

A density function of parametric probability that is the weighted sum of Gaussian component densities is known as Gaussian mixture model (GMM). In biometric systems, GMMs are used commonly as parametric framework with the distribution of probabilities of continuous quantities or biometric features.

A GMM represent the weighted sum of its M components densities as shows in equation

$$P(x | \lambda) = (x) = \sum_{i=1}^M (w_i g(x|\mu_i, \Sigma_i)) \dots\dots\dots (\text{Eq 3.1}).$$

Where x data vector of continuous values of D- dimension (i.e. features or measurements) w_i , $i=1, \dots, M$ are the weights of mixture model and $g(x|\mu_i, \Sigma_i)$, $i = 1, \dots, M$ are Gaussian densities components Every density component Gaussian

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i)\right\} \dots\dots\dots (\text{Eq. 3.2})$$

function D-variate form of

With the mean vector μ_i and the matrix of covariance Σ_i the mixture weights satisfy the constraint that $\sum_{i=1}^M w_i = 1$.

Gaussian Mixture under sampling (GMUS) GMUS

Is a proposed approach for under sampling that consists of two steps as follows: a For well fitted of majority samples we use Gaussian Mixture Model and the Bayesian Information Criteria for model optimization. GMM containing both of the high training speed and the well-fitted ability

The well-fitted GMM for the maximization of probability density function for minority samples is selected from two classes as a cross-edge. For the distributions of Gaussian mixture, most of the data is dispersed nearby the maximum probability so that it can represent better the extremely focused maximum samples from minority class.

Gaussian Mixture Oversampling (GMOS) GMOS

Is a proposed approach for oversampling that consists of two steps as follows?

For well fitted of minority samples we use Gaussian Mixture Model and the Bayesian Information Criteria for model optimization. GMM containing both of the high training speed and the well-fitted ability

The well-fitted GMM for the maximization of probability density function for minority samples that is selected from two classes as a

Cross-edge. For the distributions of Gaussian mixture, most of the data is dispersed nearby the maximum probability so that it can represent better the extremely focused maximum samples from minority class.

Planned Methodology

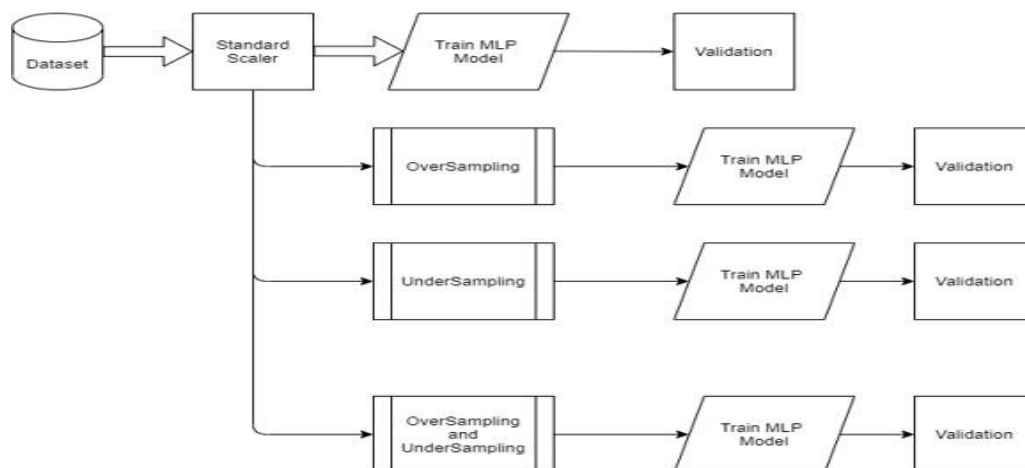


Figure 3 Flow diagram

In first experiment, I train my MLP on credit card fraud dataset. I conducted four different experiments with different techniques. In 1st experiment we used our original dataset, I normalized my dataset using standard scaler, which was then fed into MLP for training and after that we tested our model. In 2nd experiment, I used over sampling on our minority class of fraud dataset to make its sample equal to

majority class, which was normalized using standard scaler. After which we fed it into our MLP model for training, after which for testing purpose we used original dataset to evaluate the accuracy. In 3rd experiment, I used under-sampling on our majority class of fraud dataset to make its sample equal to minority class, which was normalized using standard scaler. After which we fed it into our MLP model for training, after which for testing purpose we used original dataset to evaluate the accuracy. In 4rd experiment I used Gaussian Mixture Model based under sampling and oversampling as well, first I oversampled minority class to increase its samples up to 100,000. After then I under sampled majority class samples to make it close to minority samples. Next these samples were used to train our MLP model, for evaluation purpose I used original dataset.

Performance Metrics

Table 1 represents a confusion matrix that summarizes binary results for classification. FP and FN errors are correspondence to Type I and Type II errors correspondingly. Performance matrices recorded in this section are all derived from the confusion matrix.

Table 3.1 Confusion Matrix

	Actual positive	Actual Negative
Predictive Positive	True Positive	False Positive
Predictive Negative	False Negative	True Negative

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (Eq. 3.12)$$

$$Precision = \frac{TP}{TP+FN} \dots\dots\dots (Eq. 3.13)$$

$$Recall = TPR = \frac{TP}{TP+FN} \dots\dots\dots (Eq. 3.14)$$

$$F - measure = \frac{2 \times recall \times precision}{Recall + Precision} \dots\dots\dots (Eq. 3.15)$$

RESULTS AND DISCUSSION

Dataset

Our Original dataset represented in figure 4.1, with unbalanced samples.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

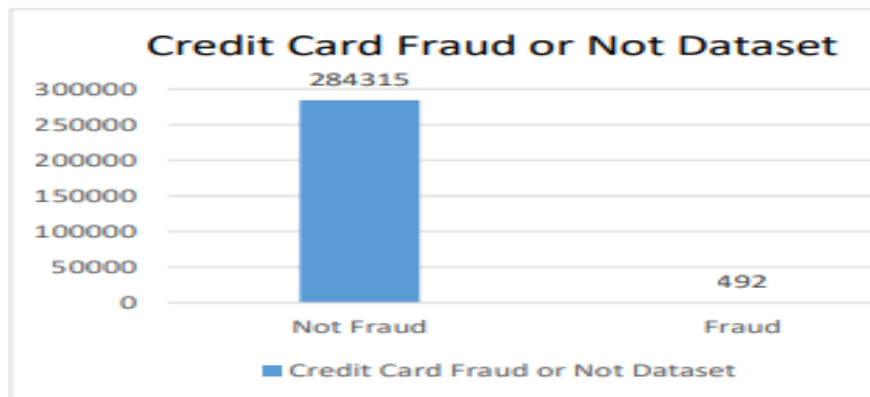


Figure 4 Original Dataset with unbalanced samples

After GMM-based Oversampling

Samples after GMM-based oversampling are represented with figure 4.2.

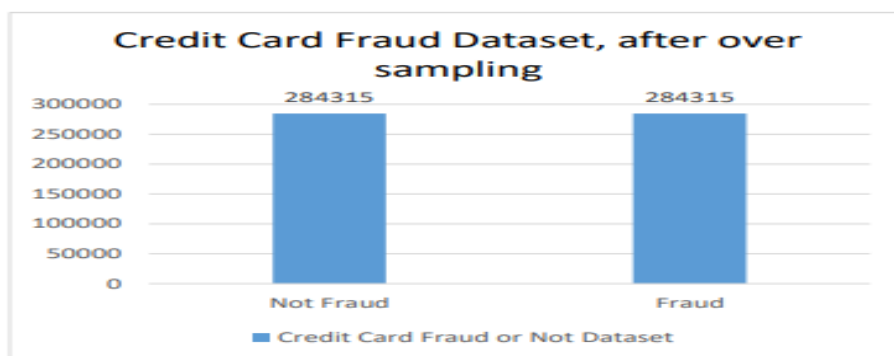


Figure 5 after oversampling

After GMM-based Under sampling

Samples after GMM-based oversampling are represented with figure 4.3.

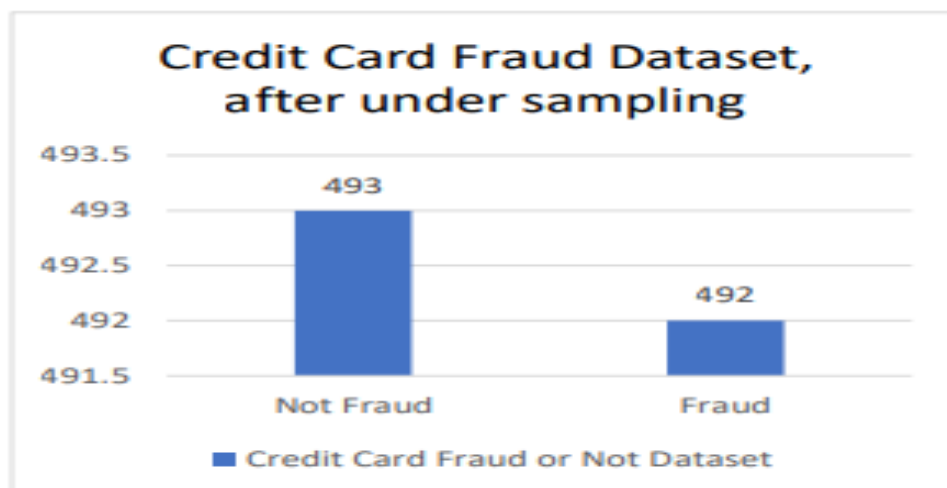


Figure 6 after under sampling

After Combination of GMM-based under sampling and Oversampling

Samples after GMM-based combination of under sampling and oversampling are represented with figure 4.4.

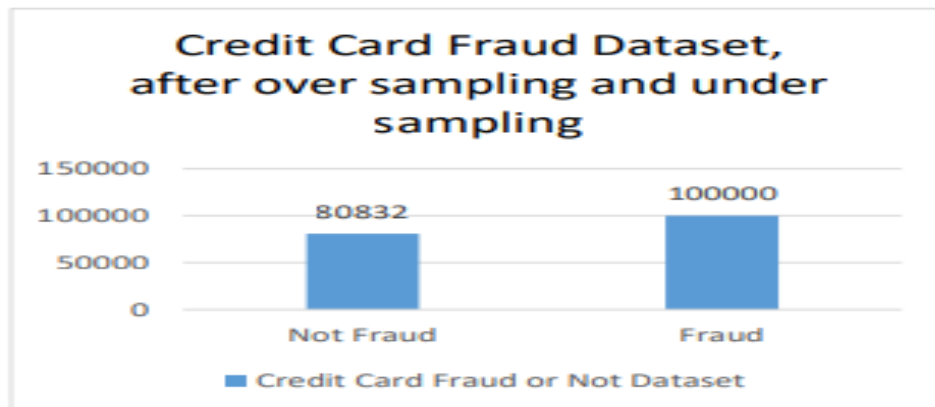


Figure 7 after combination of oversampling and under sampling

Results

After training MLP on different samples of our dataset we test on original unbalanced dataset and compared our results.

Results after Oversampling

Metrics	Not Fraud	Fraud
Precision	100%	83%
Recall	100%	86%
F1-Score	100%	84%

Metrics	After oversampling
Accuracy	0.43

Results after under sampling Table 3 Under sampling Results

Metrics	Not Fraud	Fraud
Precision	100%	0%
Recall	0.5%	100%
F1-Score	66%	1%

Metrics	After Undersampling
Accuracy	0.50

Results After the Combination of Under sampling and Oversampling

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Table 4 Results with combining under sampling and Oversampling

Metrics	Not Fraud	Fraud
Precision	100%	0%
Recall	43%	99%
F1-Score	60%	1%

Metrics	After Oversampling and Undersampling
Accuracy	0.43

CONCLUSION

A dataset is called to be imbalanced when few of its ‘minority’ classes are under representations in the comparison of most of its ‘majority’ classes. A data set that contains an inadequate distribution of data among its classes is called an imbalanced dataset. In several problems, the minority classes are interesting or positive classes for example in credit card fraud detection where most of the transactions are not fraud but only a few of them in this problem fraud detection is more difficult having great interest. This class rarity leads to classification methods behavior biased towards majority classes that cause misclassification. For handling this issue we propose a GMM-based approach, which is performed in three ways as GMM-based under sampling, GMM-based oversampling, GMM-based combination of under sampling and oversampling. For evaluation of results we use four matrices such as precision, recall, f-measure, and accuracy after GMM-based oversampling precision, recall, F1 is 83%, 86%, 84% in fraud detection respectively, accuracy is 0.43 and in GMM-based under sampling precision, recall, F1 is 0%, 100% and 1% and the accuracy is 0.50, while combining the GMM-based under sampling and oversampling the precision, recall, F1 is 0%, 99%, 1% respectively and the accuracy of fraud detection is 0.43. In future we can improve the results and enhance the accuracy measurements and also check the behavior of model after applying it on other imbalanced datasets.

REFERENCES

- W. Chen, Y. Liu, and X. Zhang, “A survey on imbalanced learning: Latest research and future directions,” *Artificial Intelligence Review*, 2024.
- Y. Zhang, H. Li, and J. Sun, “An ensemble oversampling method based on GAN for imbalanced classification,” *Information Fusion*, 2023.
- Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, “Credit card fraud detection using improved VAE–GAN models,” *IEEE Access*, 2023.
- S. F. Taskiran et al., “A comprehensive evaluation of oversampling techniques for imbalanced classification,” *PeerJ Computer Science*, 2025.
- M. Tayebi et al., “Generative modeling for imbalanced credit card fraud detection,” *Security Analytics*, 2025.
- H. A. Ahmed and S. Khan, “Network intrusion detection using oversampling techniques and machine learning algorithms,” *PeerJ Computer Science*, 2022.

- L. Scrucca, "On the influence of data imbalance on supervised Gaussian mixture models," Algorithms, 2023.
- G. Chen, B. Hou, and T. Lei, "Monte Carlo sampling using Gaussian Mixture Models for imbalanced data classification," Mathematical Biosciences and Engineering, 2023.
- S. Shi et al., "A balanced variational autoencoder with attention for fraud detection," Applied Intelligence, 2025.
- O. Volk et al., "Adaptive cost-sensitive learning for imbalanced data classification," Journal of Intelligent Information Systems, 2024.
- M. Ahmadlou et al., "Cost-sensitive machine learning models for class-imbalance problems," International Journal of Computational Intelligence Systems, 2024.
- F. Brishti et al., "Imbalanced classification with label noise: A systematic review," Information Sciences, 2025.
- M. Imani et al., "Imbalanced data classification using graph-based feature extraction," Scientific Reports, 2025.
- F. Zhang et al., "GMM-based undersampling and its application in credit card fraud detection: Updated evaluation," Neural Computing Proceedings, 2022.
- C. Wang et al., "Adaptive ensemble of classifiers with regularization for imbalanced learning: Enhanced 2023 version," Information Fusion, 2023.
- J. Tanha et al., "Boosting methods for multi-class imbalanced data: Updated experimental review," Journal of Big Data, 2022.
- S. Pouyanfar et al., "Dynamic sampling in deep neural networks for imbalanced learning: Recent updates," IEEE Multimedia Processing and Retrieval, 2023.
- R. A. Bauder et al., "Extended analysis of class rarity in big data," International Journal of Data Science, 2024.
- W. C. Sleeman and B. Krawczyk, "Scalable methods for multi-class imbalanced big data classification on Spark," Knowledge-Based Systems, 2024.
- E. L. A. Guven, "A comparative analysis of autoencoder-based models for credit card fraud detection," Journal of Financial Data Science, 2023.
- H. Zhang et al., "CNN-GMM-SMOTE based intrusion detection: Extended experiments," Computer Networks, 2022.
- H. Wang et al., "Predicting hospital readmission using updated cost-sensitive deep learning frameworks," IEEE/ACM Trans. Comput. Biol. Bioinf., 2023.
- S. F. Ahmad et al., "Ensemble oversampling and feature selection for imbalanced medical datasets," Health Information Science, 2023.
- H. Du et al., "Autoencoder-LightGBM pipeline for handling imbalanced credit card fraud datasets," Symmetry, 2023.
- X. Liu et al., "BM-WGAN: Bootstrap Wasserstein GAN for minority oversampling," Mathematical Biosciences and Engineering, 2024.
- A. Patel et al., "A comprehensive survey on imbalanced data learning: Gaps and future trends," arXiv preprint, 2024.
- C. Wu et al., "Logistic regression vs GMM with resampling: A modern comparative analysis," Journal of Supply Chain and Operations Management, 2023.
- S. Khan et al., "Synergizing digital twin technology for depression categorization via social-media data mining," International Journal of Innovations in Science and Technology, 2024.