Handling Class Imbalance Problem in Time-series Data

Presented By
Vinay Gupta
Enrollment No. 21535036

Under the Supervision of
Prof. Pradumn K. Pandey
Department of CSE, IIT Roorkee

January 19, 2024

Table of Contents

Introduction

- For Example: If dataset contain 1000 sample in which 995 positive sample and 5 negative sample .Let, the model predict all samples as positive samples.
  - Then, $TP = 0$, $FP = 0$, $FN = 5$, $TN = 995$.
  - Accuracy= $\frac{(TP+TN)}{(TP+FP+FN+TN)} = \frac{(0+995)}{(0+0+5+995)}$ =99.5%
  - Precision = $\frac{(TP)}{(TP+FP)} = \frac{(0)}{(0+0)}$ =0%
  - Recall = $\frac{(TP)}{(TP+FN)} = \frac{(0)}{(0+5)}$ =0%

  where, TP is True Positive, FP is False Positive, TN is True Negative and FN is False Negative.

Application

- Rare medical diagnoses
- Fraud detection in banking
- Anomaly detection

Approach to Handle Imbalance Class Data

- Different Type of Approach to Handle Imbalance Class Data:
  - Algorithm level approach.
  - **Data level approach**[1].
  - Ensemble Approach.
  - Cost-Sensitive Approach.

[1] Li,W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, Inf. Sci. (Ny) 409–410 (2017).

Data level approach

- Data level algorithms convert the imbalanced dataset by balancing the distribution of the data using sampling techniques :
  - Oversampling.
  - Undersampling.
  - Hybrid-Sampling[2].

[2]S. Choirunnisa and J. Buliali, “Hybrid method of undersampling and oversampling for handling imbalanced data,” 11 2018.

Contd.

- **Oversampling**

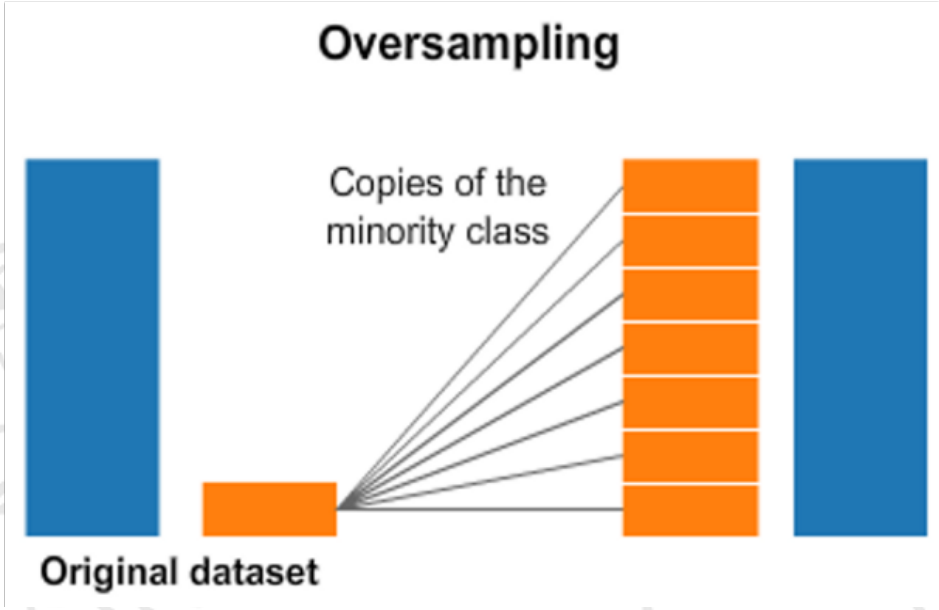

Figure: 2 Oversampling of data

- Limitation: It replicates the same minority samples multiple time to balance the dataset which leads to over-fitting.

Contd.

- **Undersampling**[3].

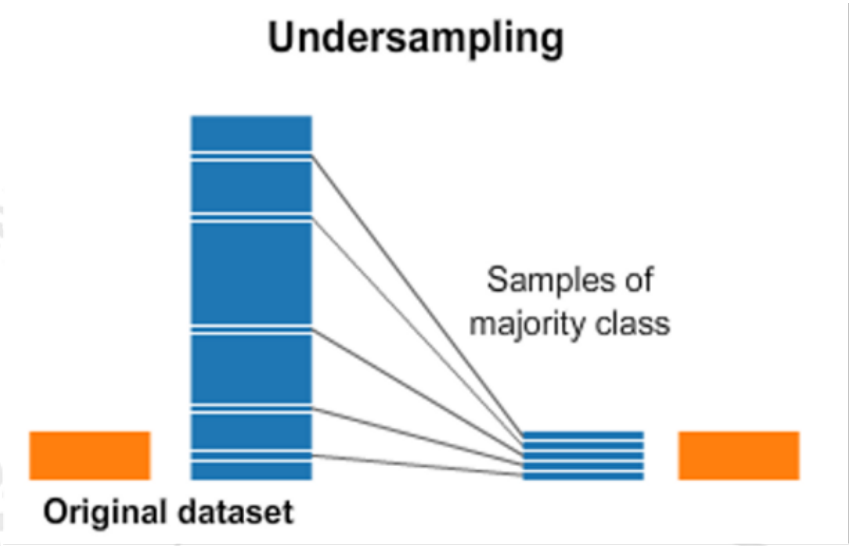

Figure: 3 Undersampling of Data

- Limitation.
  Due to deletion of samples there is a loss of information which is not good for training the model.

[3]X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539–550, 2008

Baseline Oversampler

- SMOTE (Synthetic Minority Oversampling Technique)
- Borderline-1
- Borderline-2
- Random SMOTE
- ADASYN SMOTE
- KMeans SMOTE
- ASN SMOTE

SMOTE(Synthetic Minority Oversampling Technique)[4].

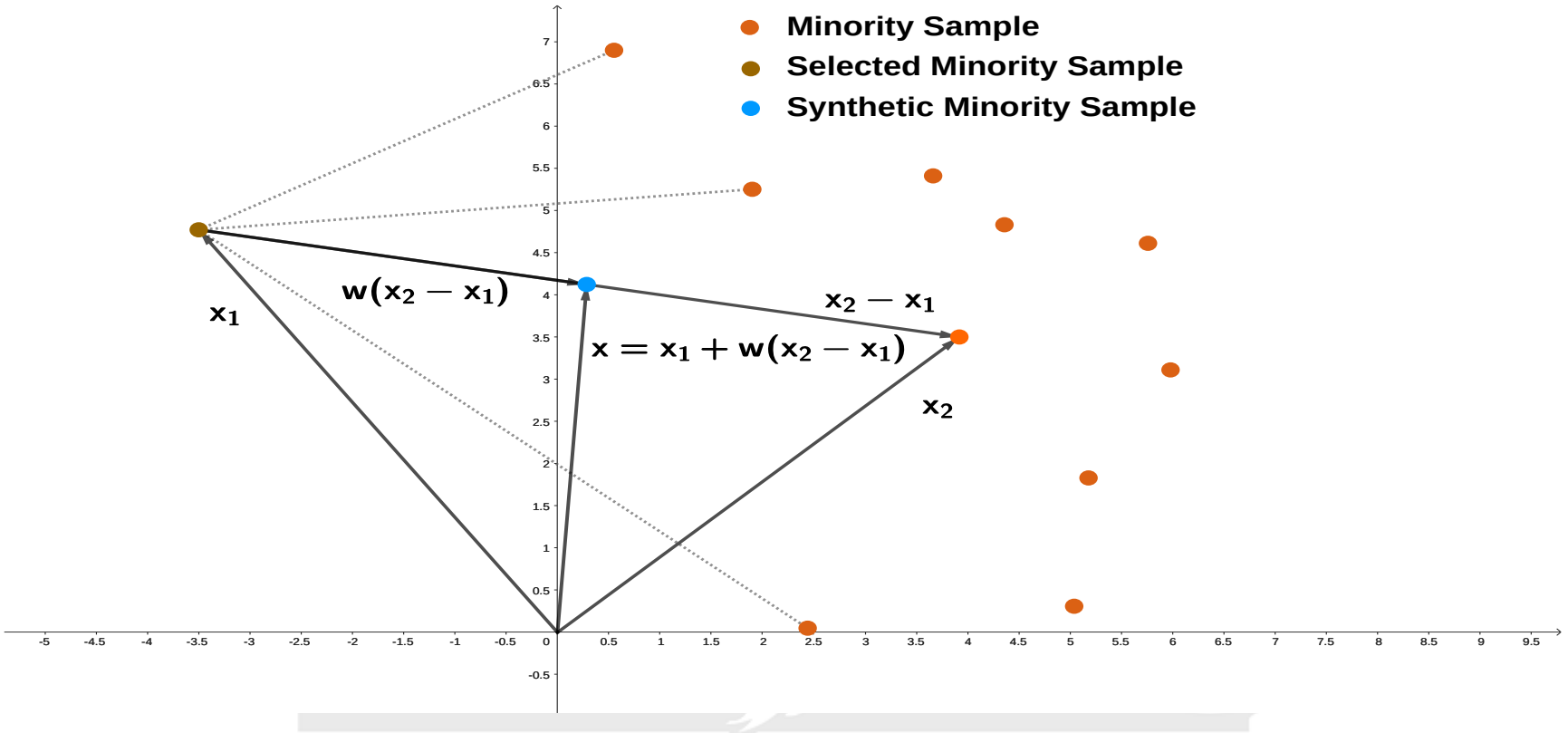

Figure: Underlying concept of SMOTE

[4]N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Arti. Intell. Res. 16 (2002)

Limitation

- SMOTE may generate minority samples in majority regions in the presence of noise. Most non-noise samples are generated in already dense minority areas, contributing to within-class imbalance.

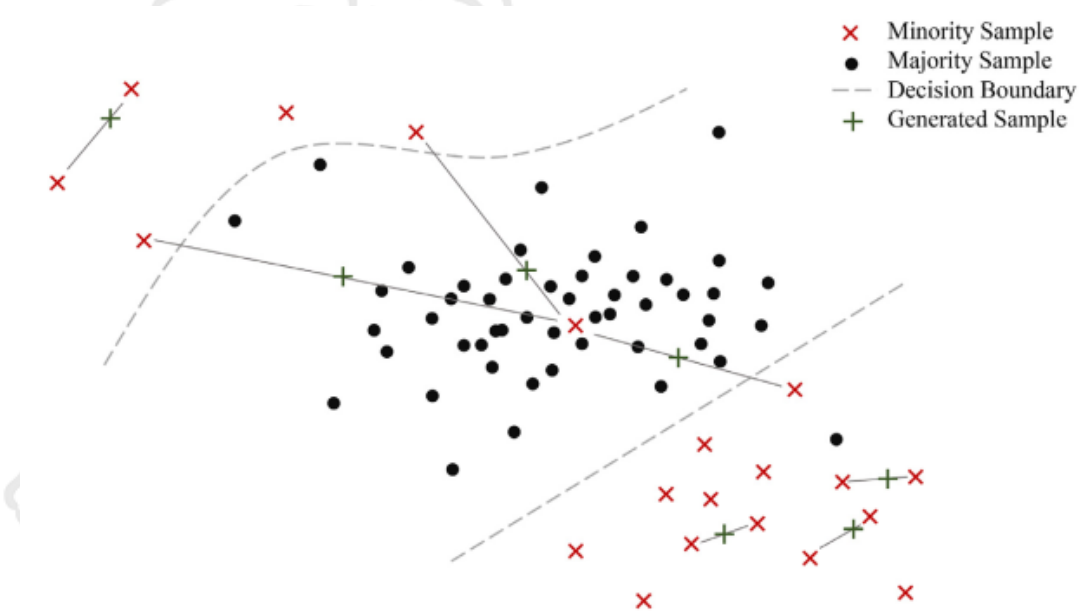

Figure: 4 Noise Minority Sample

Heuristic Oversampling method based on k-means and SMOTE

- K-means SMOTE consists three steps[5]:
  - Clustering.
  - Filtering.
  - Oversampling.

[5]Georgios Douzas and Fernando Bacao and Felix Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE" , Information Sciences, (2018).

Contd.

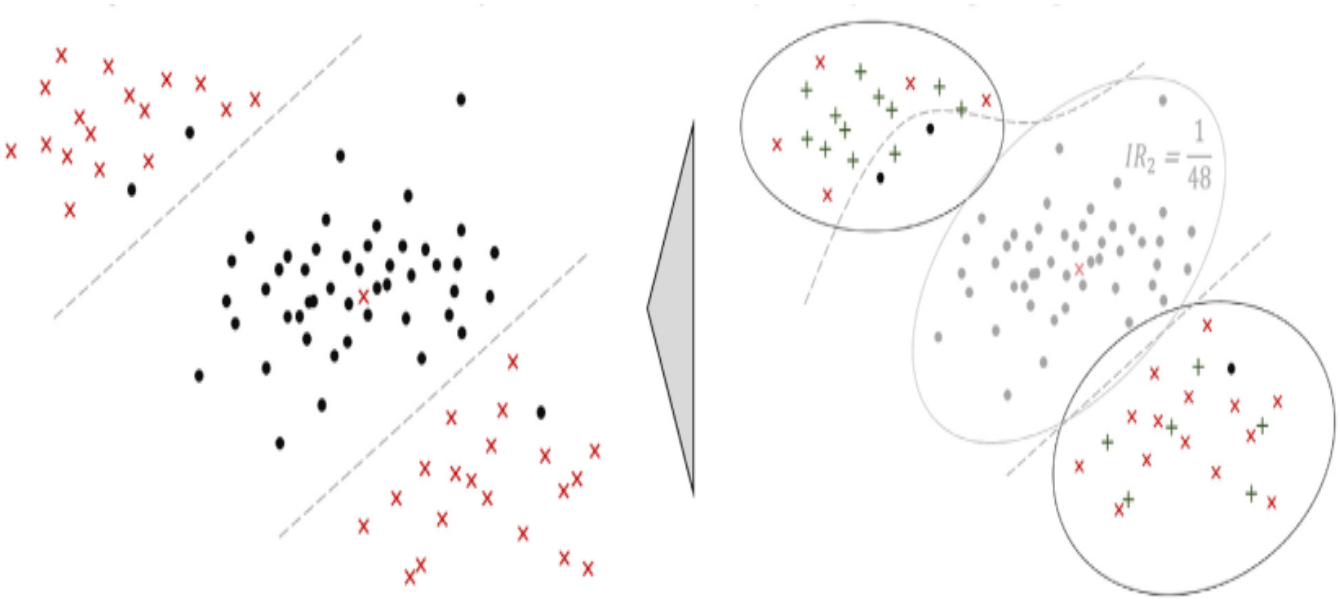

Figure: 6 K-means SMOTE

Research Gap

- Time attribute plays a crucial role in building models that can make decisions based on the latest data.
- Over the period of time nature of the data is changed so our model should also evolve according to the data.
- We proposed a solution for these kind of application in which models are trained according to the recent data sample.

Experiment

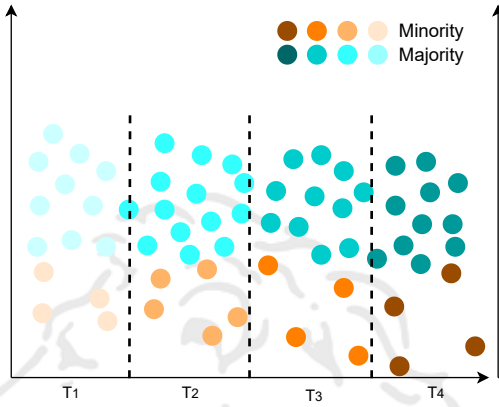

Figure: Base Dataset

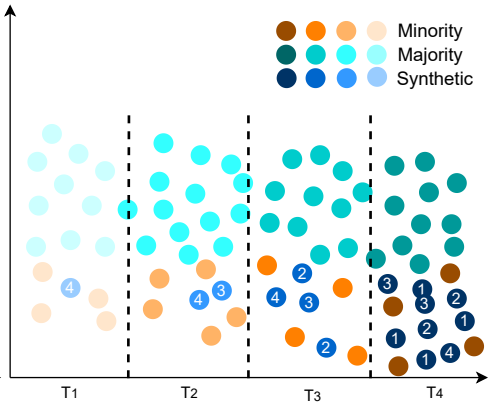

Figure: TCluster_4

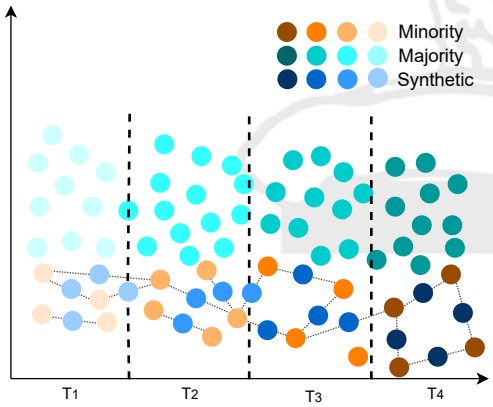

Figure: 1.SMOTE

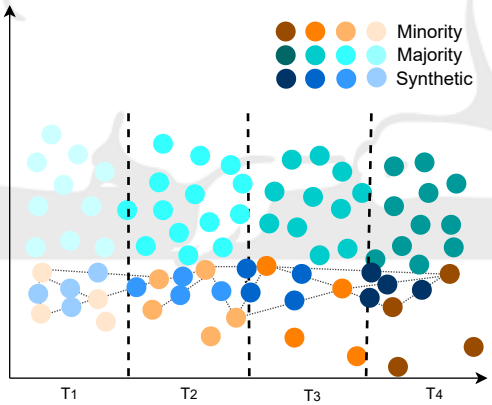

Figure: 2. Borderline[3]

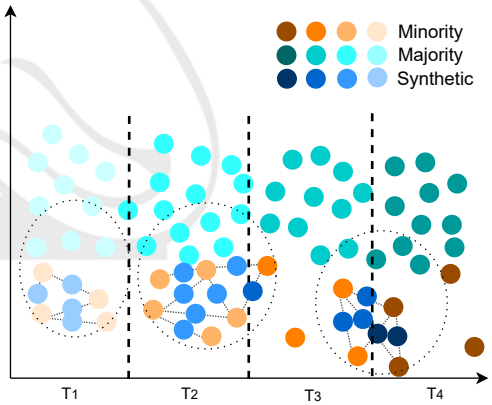

Figure: 2. Kmeans[3]

Temporal Clustering

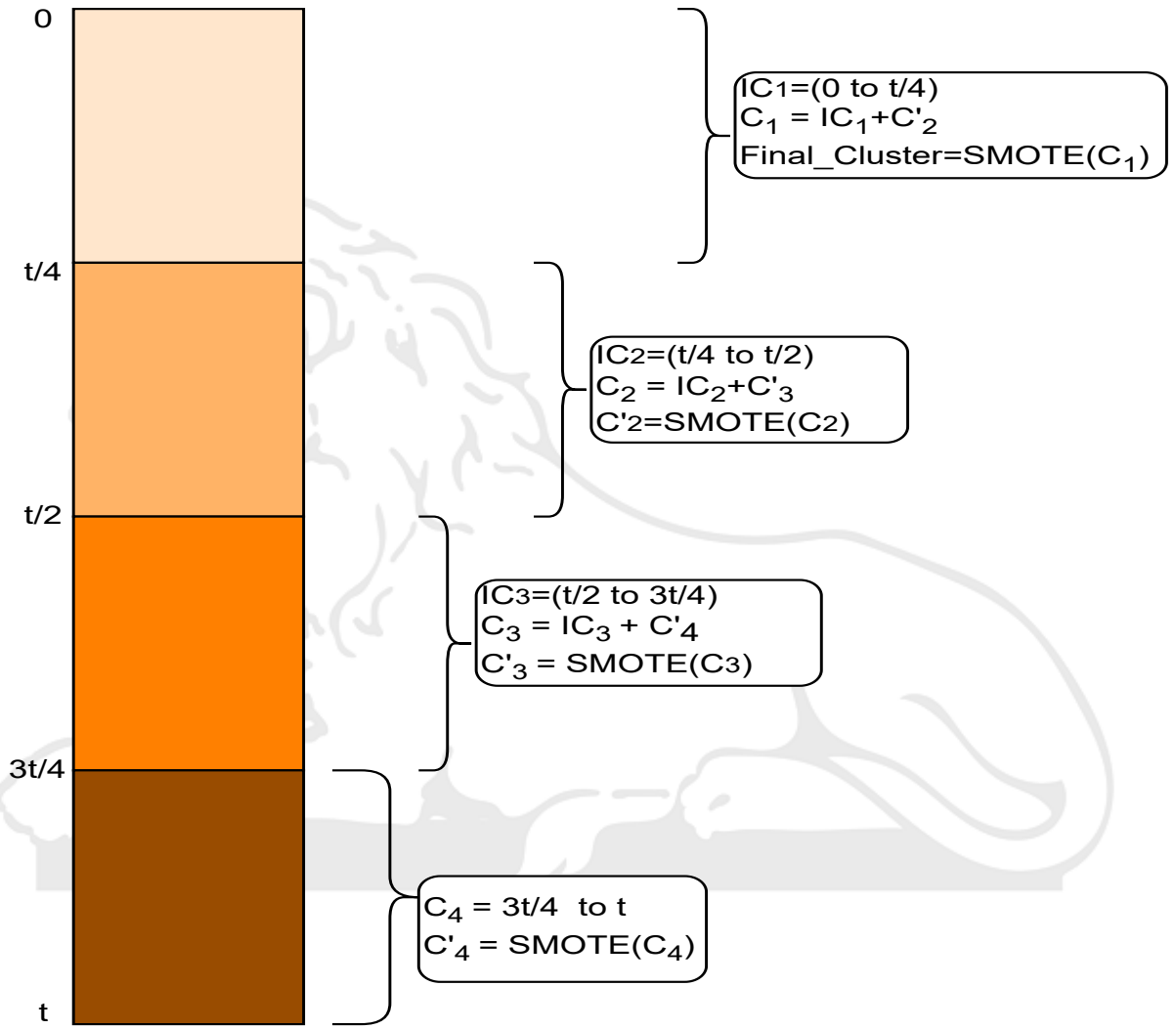

Figure: 7 Cluster on the basis of time

Proposed Method

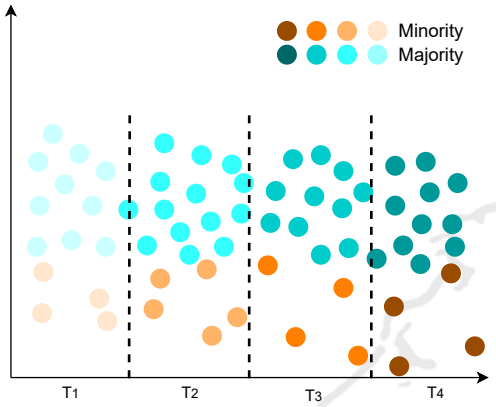

Figure: 1 Base Dataset

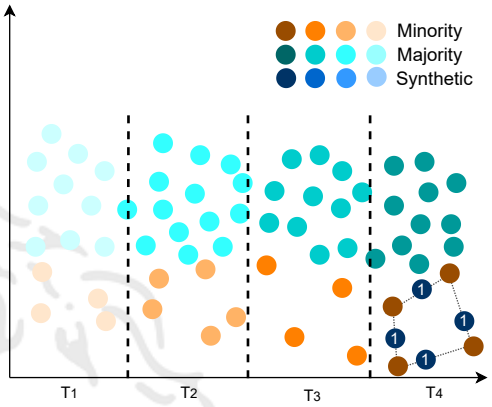

Figure: 2 Stage-1

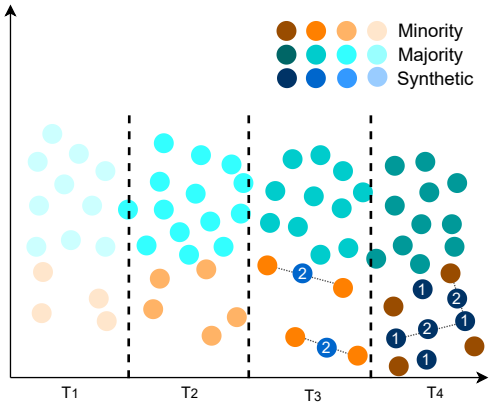

Figure: 3 Stage-2

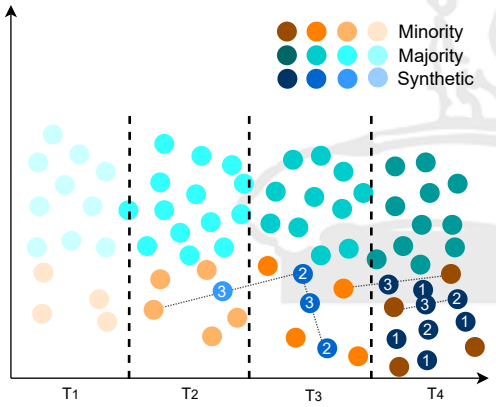

Figure: 4 Stage-3

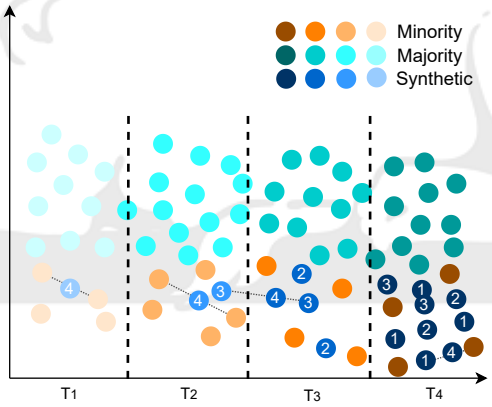

Figure: 5 Stage-4

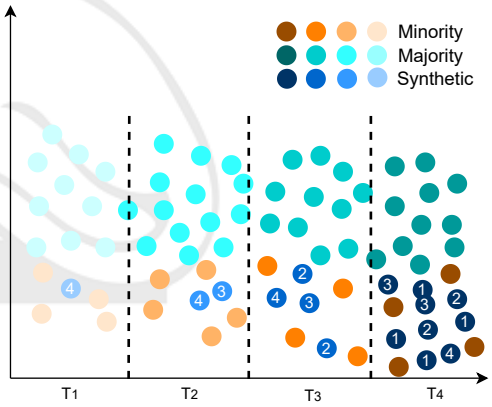

Figure: 6 Final

Experimental Setup

Table: Datasets used to evaluate the proposed model

Dataset	# Features	# Samples	# Minority	# Majority	IR
haberman	3	306	81	225	2.77
pima	8	768	268	500	1.86
creditcard	31	284807	492	284315	577.87
tackling	27	590540	20663	569877	27.57

- Classifier
  - MLP (Multi layer Perceptron)
  - LR (Logistic Regresssion)
  - KNN (K Nearest Neighbours)
  - SVM (Support Vector Machine)
  - CNN (Convolution Neural Network)
- 7 Baseline oversampler are used to compare to our results.

Result and Comparison

Table: Result on Pima Dataset of Kmeans-SMOTE using MLP classifier vs TCluster_4 using MLP classifier

Parameter	Baseline	TCluster_4
Precision	0.863636364	0.826923077
Recall	0.745098039	0.86
Accuracy	0.811881188	0.84
F1-score	0.8	0.843137255
AUC	0.81254902	0.84
G-mean	0.809744574	0.839761871

Table: Result on Haberman Dataset of Kmean-SMOTE using MLP classifier vs TCluster4 using MLP classifier

Parameter	Baseline	TCluster_4
Precision	0.842105263	0.875
Recall	0.695652174	0.913043478
Accuracy	0.777777778	0.891304348
F1-score	0.761904762	0.893617021
AUC	0.779644269	0.891304348
G-mean	0.775106776	0.891039197

Table: Result on Creditcard Dataset of Random-SMOTE using KNN classifier vs TCluster$_4$ using KNN classifier

Parameter	Baseline	TCluster$_4$
Precision	0.999824167	0.999683444
Recall	1	0.999683444
Accuracy	0.999912069	0.99968345
F1-score	0.999912076	0.999683444
AUC	0.999912071	0.99968345
G-mean	0.563612232	0.99968345

- Disadvantage of Random SMOTE is Overfitting[6].

[6] Dong, Yanjie, and Xuehua Wang. "A new over-sampling approach: random-SMOTE for learning from imbalanced data sets." Knowledge Science, Engineering and Management: 5th International Conference, KSEM 2011, Irvine, CA, USA, December 12-14, 2011. Proceedings 5. Springer Berlin Heidelberg, 2011.

Table: Result on tackling Dataset of Random-SMOTE using KNN classifier vs TCluster_{16} using KNN classifier

Parameter	Baseline	TCluster_{16}
Precision	0.937596	0.976655132
Recall	1	0.975643995
Accuracy	0.966721	0.976161648
F1-score	0.967793	0.976149302
AUC	0.966721	0.976161648
G-mean	0.966148	0.976161511

Limitation

- Determining the optimal number of clusters and balancing image dataset is a topic for future research.
- Proposed method works on time series data.

References

[1] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Arti. Intell. Res. 16 (2002).
[2] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, Inf. Sci. (Ny) 409–410 (2017).
[3] Georgios Douzas and Fernando Bacao and Felix Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE" , Information Sciences, (2018).
[4] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009).

Contd.

[5] S.J. Yen, Y.S. Lee, Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset, in: D.-S. Huang, G. William Irwin (Eds.), Intelligent Control and Automation, Springer, Berlin, Heidelberg,(2006).
[6] N.V. Chawla, N. Japkowicz, P. Drive, Editorial: special issue on learning from imbalanced data sets, ACM SIGKDD Explor. Newslett. 6 (1) (2004).
[7] F. Provost, “Machine learning from imbalanced data sets 101,” in Proceedings of the AAAI’2000 workshop on imbalanced data sets, vol. 68, no. 2000. AAAI Press, 2000, pp. 1–3
[8] S. Choirunnisa and J. Buliali, “Hybrid method of undersampling and oversampling for handling imbalanced data,” 11 2018.
[9] A. Asuncion and D. Newman, “Uci machine learning repository,” 2007.

Contd.

[10] B. Santoso, H. Wijayanto, K. Notodiputro, and B. Sartono, “Synthetic over sampling methods for handling class imbalanced problems: A review,” in IOP conference series: earth and environmental science, vol. 58, no. 1. IOP Publishing, 2017, p. 012031.

[11] Dong, Yanjie, and Xuehua Wang. "A new over-sampling approach: random-SMOTE for learning from imbalanced data sets." Knowledge Science, Engineering and Management: 5th International Conference, KSEM 2011, Irvine, CA, USA, December 12-14, 2011. Proceedings 5. Springer Berlin Heidelberg, 2011.

Thanks.