Methods for Imbalanced Classification

Harsh Choudhary
Dept. of Electrical Engineering
IIT Bombay
Mumbai
Email: 200070023@iitb.ac.in

Hiranmay Mondal
Dept. of Electrical Engineering
IIT Bombay
Mumbai
Email: 213070019@iitb.ac.in

***Abstract*—Skewed class distribution is a common occurrence in data used in real-world applications, which presents a significant problem for machine learning. When there is an imbalance in the distribution of the data, conventional classification methods are ineffective. In this project, we will use various general classification methods and will compare them Deep RL using various performance metrics.**

1. Introduction

Several problems like Medical Anomaly Detection, Credit Card Fraud Detection, Cancer Detection, etc. have a highly imbalanced dataset with significantly less number of elements in positive classes and thus generally, there is the problem of False positives while doing such classifications. The performance metrics like Precision, Recall, F1-Score, AUC-ROC, etc are useful in such cases. The general approach involves data augmentation techniques like SMOTE (Synthetic Minority Over-sampling Technique) , Threshold Moving, etc.

2. Methodology

2.1. SMOTE (Synthetic Minority Over-sampling Technique)

It is one of the the most widely used approach to synthesizing new examples. It works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. The approach is effective because new synthetic examples from the minority class are created that are plausible, that is, are relatively close in feature space to existing examples from the minority class. A general downside of the approach is that synthetic examples are created without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes.

2.2. Deep Reinforcement Learning

[3] proposes a general imbalanced classification model based on deep reinforcement learning. solve it by deep Q-learning network. The agent performs a classification action on one sample at each time step, and the environment evaluates the classification action and returns a reward to the agent. Imbalanced data classification has been widely researched in the field of machine learning. Most machine learning algorithms are suitable for a balanced training data set. When facing imbalanced scenarios, these models often provide a good recognition rate to the majority instances, whereas the minority instances are distorted. The methods to tackle these issues are mainly divided into two groups: the data level and the algorithmic level.

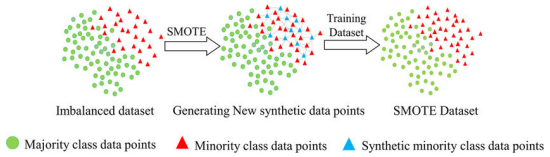

Figure 1. SMOTE

For classification problems, deep reinforcement learning has served in eliminating noisy data and learning better features, which made a great improvement in classification performance. A Deep Q-learning network (DQN) based model for im-balanced data classification is proposed in this paper. In our model, the imbalanced classification problem is regarded as a guessing game that can be decomposed into a sequential decision-making process. The reward from the minority class is higher than that of the majority class. Now we formalize the Imbalanced Classification Markov Decision Process (ICMDP) framework which decomposes an imbalanced data classification task into a sequential decision-making problem.

$$R(s_t, a_t, l_t) = \begin{cases} +1, & a_t = l_t \text{ and } s_t \in D_P \\ -1, & a_t \neq l_t \text{ and } s_t \in D_P \\ \lambda, & a_t = l_t \text{ and } s_t \in D_N \\ -\lambda, & a_t \neq l_t \text{ and } s_t \in D_N \end{cases}$$

In this project, we mainly study the binary imbalanced classification with deep reinforcement learning. We used a credit card data set. We have 364 fraud data points and

179986 regular data points. We use deep neural networks to learn the feature representation from the imbalanced and high-dimensional datasets Before the research of imbalanced data learning, we compare our DQNimb model to the DNN which is a supervised deep learning model in balanced data sets. The experiments were conducted on the credit card data set.

$$\begin{aligned}\frac{\nabla L(\theta_k)}{\nabla \theta_k} = & -2\sum_{m=1}^{P+N}((1-t_m)\gamma \max_{a'_m} Q(s'_m, a'_m; \theta_{k-1}) \\ & - Q(s_m, a_m; \theta_k))\frac{\nabla Q(s_m, a_m; \theta_k)}{\nabla \theta_k} \\ & -2\sum_{i=1}^{P}(-1)^{1-I(a_i=l_i)}\frac{\nabla Q(s_i, a_i; \theta_k)}{\nabla \theta_k} \\ & -2\lambda\sum_{j=1}^{N}(-1)^{1-I(a_j=l_j)}\frac{\nabla Q(s_j, a_j; \theta_k)}{\nabla \theta_k}\end{aligned}$$

3. EDA

We have demonstrated the imbalanced classification for Credit-Card Data [1] which has minority classes as Fraud and majority classes as Legitmate Transaction

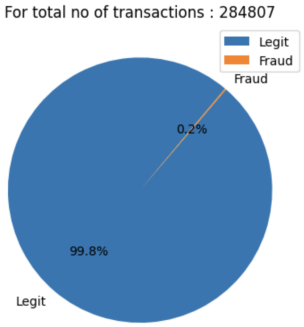

Figure 2. Pie chart for Imbalanced Distribution

The data has records of 284807 transactions and 30 differents features including time instance, amount, etc and one output column for class.

We will check the correlation of different features to remove redundant features

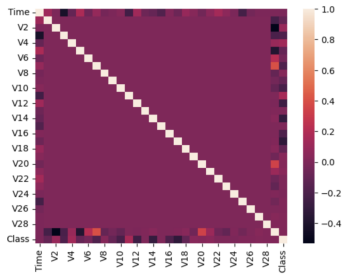

Figure 3. Heatmap for correlation between features

It implies the fraud and legitmate transactions both occur at similar times

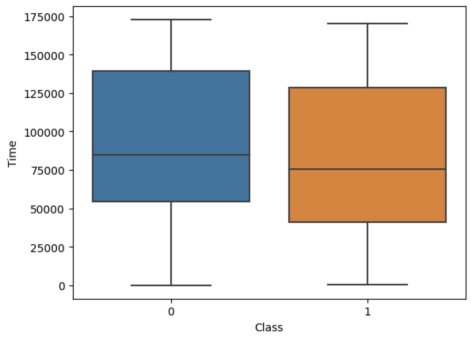

Figure 4. Boxplot for Timestamps of the Transactions

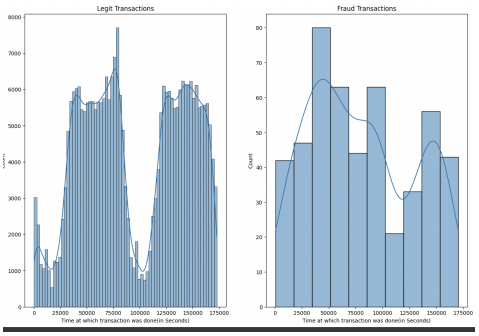

Figure 5. Uniform Distribution of Transactions with time

Fraud Transactions are generally of smaller amount typically of less than 100 Euros

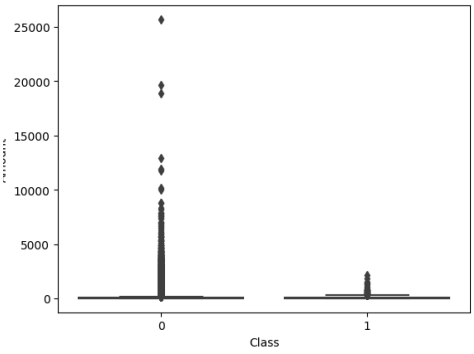

Figure 6. Boxplot for transaction Amount

4. Results

There is a significant increase in precision and recall in LR, Random Forest, and Neural Networks when we use SMOTE and there is a further increase in recall when we use Deep RL approach

Model	Recall	Precision	F1-score
Neural Network	0	0	0
Neural Network (with SMOTE)	0.31	0.91	0.28
NN (With Deep RL)	0.79	0.61	0.53
Logistic Regression	0.56	0.72	0.41
Logistic Regression (with SMOTE)	0.91	0.05	0.04
Random Forest	0.68	0.91	0.62
Random Forest (with SMOTE)	0.78	0.81	0.63

Figure 7. Summary of results of different Classifiers

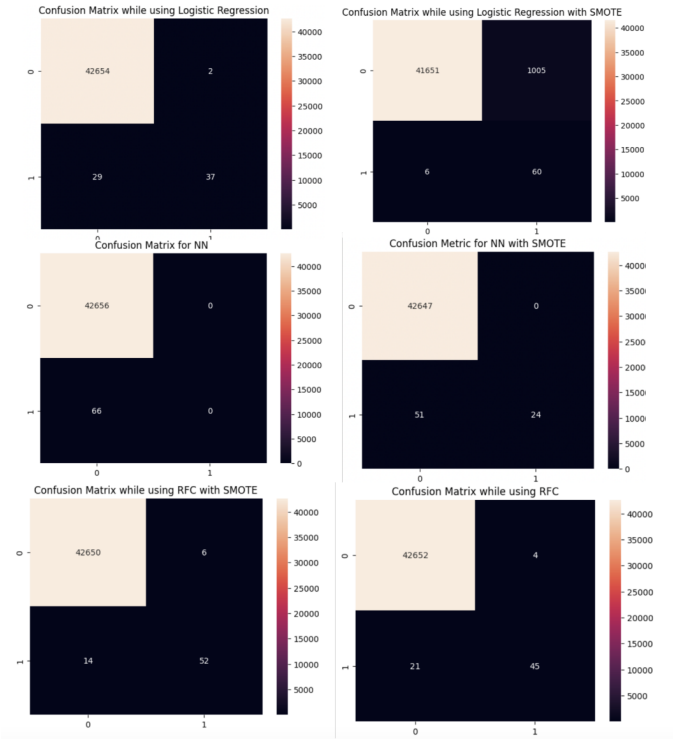

Figure 8. Confusion Matrices for 6 different Classifiers

5. Statement of Contribution

Harsh Choudhary have done EDA, Data Cleaning, Ran basic models, SMOTE, and studied the literature for Deep RL Hiranmay have studied and used Deep RL from [4]

6. Conclusions and Future Works

Commonly occurring imbalanced classification issues result in higher False Negatives and a lower recall value because general methods cannot classify them correctly. Confusion matrices for the classification of credit-card data from [1] were used to illustrate it. We showed that there is significant increase in recall value by using two approaches SMOTE and Deep RL.

Acknowledgments

We would like to Thank Professor Amit Sethi for providing us with a great experience throughout the course. The

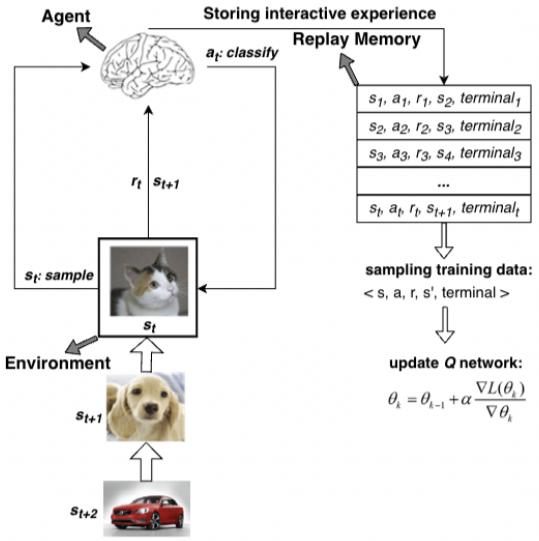

Figure 9. Block Diagram for Deep Reinforcement Learning[3]

course was highly planned and had a proper direction along with a good mathematical and application-based approach

References

[1] Dataset Link : $https : //datahub.io/machine - learning/creditcardresource - creditcard_zip$

[2] SMOTE: Synthetic Minority Over-sampling Technique https://arxiv.org/abs/1106.1813

[3] Deep Reinforcement Learning for Imbalanced Classification(2019) https://doi.org/10.48550/arXiv.1901.01379

[4] The official implementation https://github.com/linenus/DRL-For-imbalanced-Classification
Metrics Different Metrics for Imbalanced classification https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa