DOI: 10.15439/2022R10
ISSN 2300-5963 ACSIS, Vol. 33

Dealing with Imbalanced Data for GPS Trajectory Outlier Detection

Nguyen Van Chien
Faculty of Information Technology
Ho Chi Minh City University of Transport
Ho Chi Minh City, Vietnam
chiennguyensrdn@gmail.com

Van-Hau Nguyen
Faculty of Information Technology
Hung Yen University of Technology and Education
Hung Yen, Vietnam
haunv@utehy.edu.vn

Le Van Quoc Anh[1]
Faculty of Information Technology
Ho Chi Minh City University of Transport
Ho Chi Minh City, Vietnam
anh@ut.edu.vn

***Abstract*—Detecting abnormal GPS trajectories derived by the mobility of people, cars, buses, and taxis plays a crucial role in developing applications for intelligent transportation systems. Outlier detection based on classification models is among promising approaches but it faces the imbalanced data problem, where instances labeled as abnormal have a very low number of observations. In this paper, we propose a framework that employs methods to deal with imbalanced data to the problem of GPS trajectory outlier detection. Our experiments show that dealing with imbalanced data beforehand can improve the performance of outlier detection models.**

***Index Terms*—component, formatting, style, styling, insert.**

I. Introduction

For the past few years, with the popularity of smart devices with positioning technology like GPS (Global Position System), more and more GPS trajectory data are collected and available for analysis. Basically, trajectory datasets contain sequences of time-stamped points, each of which consists of latitude, longitude and altitude information. Such data represent space-time information for tracking moving objects.

Mining from GPS trajectory data plays an important role to develop applications in intelligent transportation systems. Among data mining tasks from GPS trajectory data, outlier detection, which aims to identify abnormal moving behaviors, has received considerable attention [1]. Due to the characteristics of the GPS trajectory data and there is no clear definition of trajectory anomaly, the outlier detection task is facing many challenges. One direction to solve the unclear definition of trajectory anomaly is to allow users specify which trajectories are abnormal. In this way, trajectories or subsequences of trajectories are labeled as normal or abnormal, and then the labeled datasets are used to train a machine learning model. The biggest advantage of such an approach is that the machine learning models capture the same definition of trajectory anomaly as humans do. However, one drawback of this approach is that training datasets may contain less abnormal instances than normal ones. This leads to the problems of imbalanced data in machine learning.

The data set is Imbalanced when the majority label component is much superior to the remaining label component [2]. Using imbalanced data for training a machine learning model poses a challenge since the model can bias towards the majority class. This results in models that have inferior predictive performance, especially for the minority class [3]. In the problem of GPS trajectory outlier detection, there is more sensitive to classification errors for the abnormal instances than the normal ones.

To handle imbalanced data, several approaches have been proposed, such as changing the performance metrics, modified the algorithms or the data. In this paper, we study significant techniques for dealing with imbalanced data and apply them to GPS trajectory data. The results show that such techniques can be utilized and integrated to the mining process to improve the performance of the outlier detection from GPS trajectory data.

II. Background and Related Work

A. GPS trajectory data

This paper uses the following concepts to develop the issue:

- **GPS point**: we represent a GPS point by a tuple *<id, latitude, longitude, timestamp>*, where *id* is the identifier of the moving object; and the last three components describe the position and the timestamp of the moving object.
- **GPS trajectory:** a series of GPS points that have the same *id* component and the points are arranged in time order.
- **GPS dataset or GPS log:** a set of GPS trajectories.

B. Outlier detection from GPS trajectory data as classification problem

In this paper, we focus on approaches to the problem of outlier detection from GPS data that are based on classification methods. In this way, sub-trajectories are labeled as normal or abnormal in some way, for example a manual method or using histogram-based approaches. We describe the framework in detail in Section III.

[1] Corresponding author. Email: anh@ut.edu.vn

Since the above approach depends on manually labeling, the number of instances labeled as abnormal has a very low number of observations in comparison with the number of instances labeled as normal. Therefore, we need to handle imbalanced data in this case.

C. Handling Imbalanced data

Resampling methods in disequilibrium learning applications alter the data set with mechanisms and techniques for a more balanced distribution [2]. Basic classification algorithms have better performance on balanced data sets, as mentioned in previous studies [3],[4].

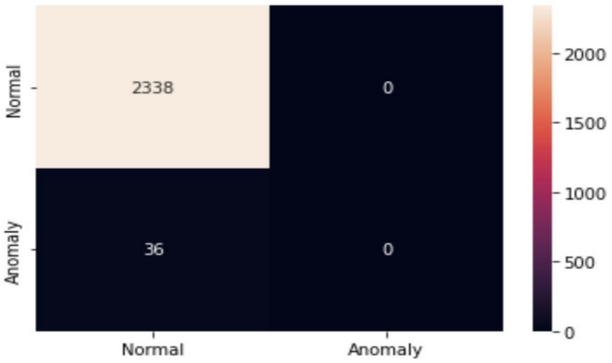


Figure 2. The original data classification model confusion matrix (SVM).

Figure 2 shows that the prediction rate of the majority class is very high, but the minority class is very low. Therefore, even distribution between classes is very important for classification algorithms.

Sampling Methods for Imbalanced Learning

- **Tomek Links(T-Link)**

Tomek links is a data cleaning technique to remove overlapping, noise applied from sampling methods.

```
$x \in S_{maj}$ , $y \in S_{min}$
d(x,y) distance between x and y
If there is no instance k, such that d(x,y)<d(x,k) or d(x,y)< d(y,k):
    (x, y) is a T-Link
Else:
    x or y is nosie or x and y is near border
```

So use Tmoke Link to "cleanup" unwanted noisy data, discarding until all the closest neighbor pairs that are farthest apart belong to a class. The removal of noisy data makes the classifier better and improves the performance of the system.

- **Synthetic Minority Oversampling Technique (SMOTE)**

SMOTE is an progressive technique of over-sampling developed by Chawala. Procedure consists of steps [2]:

```
● For each sampler $X_0$ in minority class:
– Pick one of its K nearest neighbors X ∈ minority class
– Create Z is a new sampler, as follows:
    $Z = X_0 + w(X - X_0)$    (1)
where w is Random Uniform range [0, 1].
(1) => Z is a random point on the line with equation:
    $X_0 + w(X - X_0) = 0$
```

The samples synthesized according to (1) is a point on the line connecting X_0under consideration and K-nearest is randomly selected X.

- **Adaptive Synthetic Sampling (ADASYN)**

ADASYN, on the different hand, uses a systematic procedure to adaptively create extra pieces of synthetic data according to their allotments. The algorithm is described in detail [5]. The main concept of the ADASYN algorithm is to use density distribution as a measure to automatically decide how many aggregate samples need to be developed for each minority example by adaptively modifying the weights of the examples different minority examples to compensate for the unequal distribution[6].

- **Random Oversampling and Undersampling**

One of the familiar approaches was to use resampling techniques to construct the dataset balanced. Oversampling or oversampling can be applied to the resampling of the data set. Reducing the number of elements in the data set is the idea of undersampling. The process of oversampling is the multiplication of minority cases by duplicating or repeating some cases. [7].

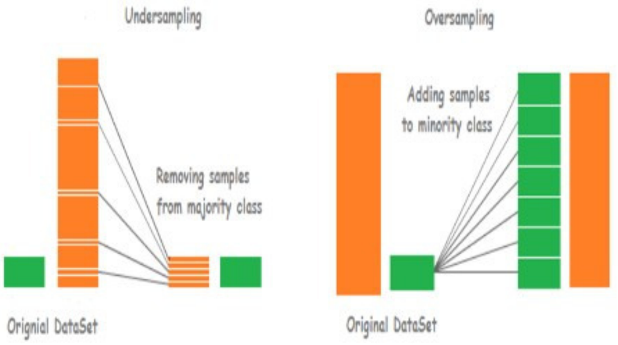

Undersampling
Removing samples from majority class
Orignial DataSet
Oversampling
Adding samples to minority class
Original DataSet

Figure 3. Differences between undersampling and oversampling

- **SVM-SMOTE (SVMs)**

A variant of the SMOTE algorithm which uses an SVM algorithm to detect samples to use for developing new synthetic samples as presented in [8]. SVM uses the concepts of support vector, maximize soft-margin, and hyperplane to classify between data samples, to improve performance, reduce classification errors during data resampling [9].

III. Proposed Framework

In this section, we describe our framework for detecting sub-trajectories with anomalous motion behavior. The framework consists of four steps.

The first step is a processing step that extracts trajectories by time frame. The second step is a feature extraction. We derive objects in the form of a tuple <*s,m,d,t,l,v*> when *s* is the start point, *m* is a midpoint, *d* is the destination point, *t* is the timestamp, *l* is the distance length, and *v* is the velocity. Each object is labeled as normal or abnormal by humans or by an automatic method like histogram-based method as described in [10].

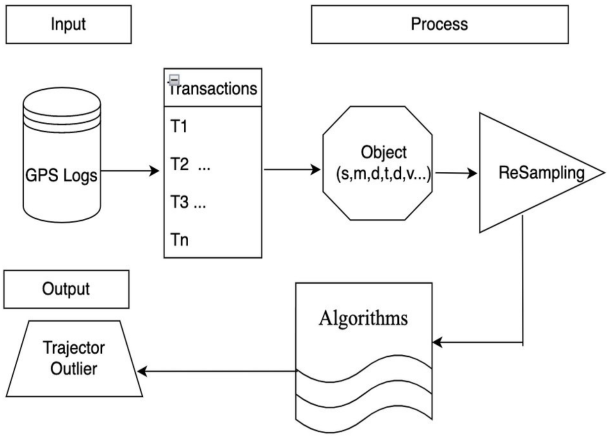


Figure 1. Framework to detect abnormal movement behavior

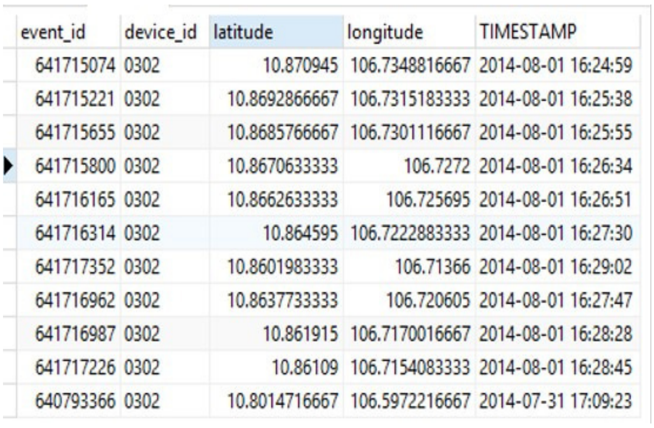
The third step applies techniques to deal with imbalanced data. Finally, in the last step, one can apply any classification algorithm to classify the objects as abnormal or not.

IV. Experimental Results

D. Dataset Description and Experimental Setup

We use two data sources to demonstrate our proposed approach. The first dataset is provided by the OTS transport service monitoring company. Itinerary data is exploited on Ho Chi Minh City routes. The dataset has 411 vehicles, mined from June 01, 2015, to June 07, 2015. This dataset is the same as the dataset using in [10],[11].

The second dataset is provided by Kaggle website. This data is exploited on Beijing routes. The dataset has 10,357 taxis, mined from June 01, 2015, to June 07, 2015. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches to 9 million kilometers. Figure 4 demonstrates the format of the GPS Log of the data in Ho Chi City, where each record represent a GPS point in a trajectory.



event_id	device_id	latitude	longitude	TIMESTAMP
641715074	0302	10.870945	106.7348816667	2014-08-01 16:24:59
641715221	0302	10.8692866667	106.7315183333	2014-08-01 16:25:38
641715655	0302	10.8685766667	106.7301116667	2014-08-01 16:25:55
641715800	0302	10.8670633333	106.7272	2014-08-01 16:26:34
641716165	0302	10.8662633333	106.725695	2014-08-01 16:26:51
641716314	0302	10.864595	106.7222883333	2014-08-01 16:27:30
641717352	0302	10.8601983333	106.71366	2014-08-01 16:29:02
641716962	0302	10.8637733333	106.720605	2014-08-01 16:27:47
641716987	0302	10.861915	106.7170016667	2014-08-01 16:28:28
641717226	0302	10.86109	106.7154083333	2014-08-01 16:28:45
640793366	0302	10.8014716667	106.5972216667	2014-07-31 17:09:23

Figure 4. GPS Log collected from a vehicle tracking device

E. Results

We give an example of an oversampling technique. Its main goal is class balance, by random repetition of minority samples.

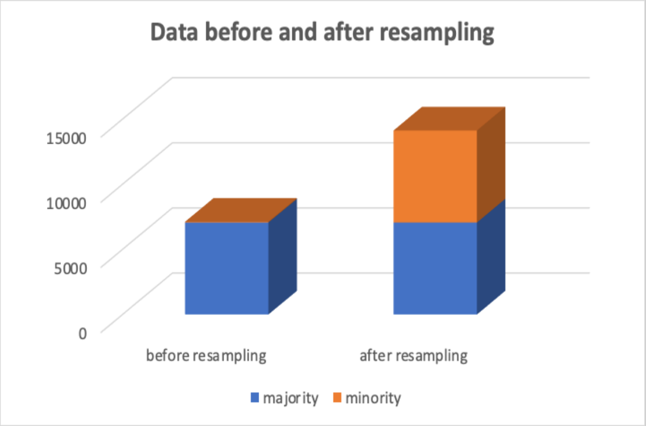


Figure 5. Data before and after resampling (oversampling)

Figure 5 shows how the class target is distributed after using this method on our dataset and it equals to 7,005.

However, this technique has two limitations. First, it will rise the probability of over-fitting, as it creates the same reproductions of the minority class instances [2]. Second, it makes the learning process take longer if the initial data set is very large, but balanced.

In this paper, we only give the imbalance handling solutions in order to show that it will achieve better minority class recognition performance. The need for preprocessing, resampling data is very important for the problem of data imbalance in machine learning.

The performance of the model is evaluated using measures such as Weighted accuracy, F-scoere, G-means. The following is a summary of each measure:

$$Sensitivity: The\ True\ Positive\ rate\,(TP) = \frac{TP}{TP+FN}$$

$$Specificity: The\ True\ Negative\ rate\,(TN) = \frac{TN}{FP+TN}$$

$$G-mean=\sqrt{Sensitivity*Specificity}$$

$$Weighted\ accuracy=0.5*(Sensitivity+Specificity)$$

$$Precíion=\frac{TP}{TP+FP};\ Recall=\frac{TP}{TP+FN}$$

$$F-Score=2\frac{Precíion*Recall}{Precíion+Recall}$$

F. *GPS route data Ho Chi Minh City-Viet Nam*

For real datasets, we obtain GPS Log from a company providing vehicle tracking services, called OTS. Itinerary data is exploited on Ho Chi Minh City routes. The dataset has 411 vehicles, mined from June 01, 2015, to June 07, 2015 [10].

Table I. Performance measures G-mean, F-score, Weighted accuracy (algorithm SVM).

SVM	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.6325	0.0507	0.1620
SMOTE/T-Link	0.6351	0.0515	0.1632
Over-Sampling	0.6390	0.0545	0.1683
Over/T-Link	0.6646	0.0587	0.1748
ADASYN	0.6328	0.0508	0.1621
SVMSMOTE	0.5865	0.0687	0.1968

Table II. Performance measures G-mean, F-score, Weighted accuracy (algorithm Logistic Regression).

LR	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.5526	0.0355	0.1343
SMOTE/T-Link	0.5523	0.0355	0.1343
Over-Sampling	0.5261	0.0324	0.1284
Over/T-Link	0.5399	0.0340	0.1315
ADASYN	0.5370	0.0337	0.1308
SVMSMOTE	0.5192	0.0351	0.1390

Table III. Performance measures G-mean, F-score, Weighted accuracy (algorithm Random Forest)

RF	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.6670	0.1032	0.2404
SMOTE/T-Link	0.6523	0.0958	0.2315
Over-Sampling	0.5274	0.1000	0.7020
Over/T-Link	0.5132	0.0500	0.4963
ADASYN	0.6504	0.0932	0.2279
SVMSMOTE	0.5558	0.0952	0.2674

Summary of results when applying data resampling techniques when applied to different data classification algorithms:

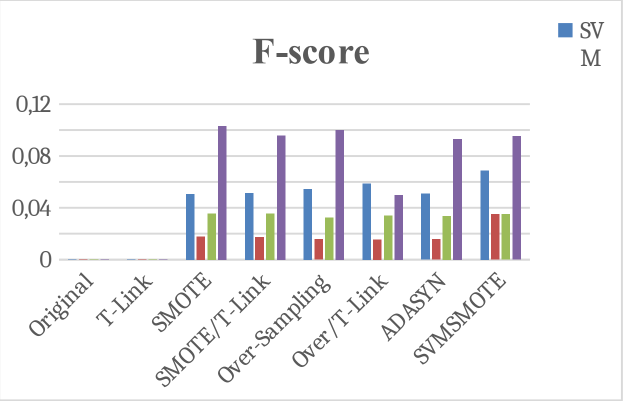


Figure 6. Weighted accuracy of various Machine learning algorithms using various sampling techniques

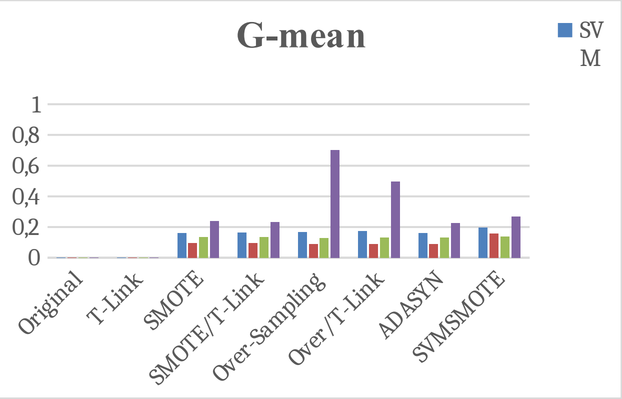


Figure 7. F-score of various Machine learning algorithms using various sampling techniques

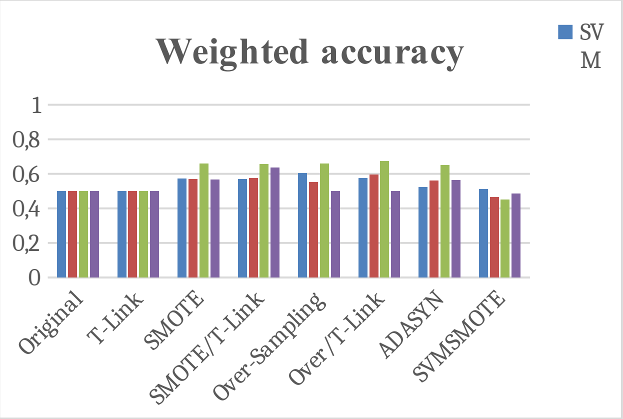


Figure 8. G-mean of various Machine learning algorithms using various sampling techniques

The results show that when applying data resampling techniques, such as minority reinforcement sampling, the majority group reduction, the model predicts better than the minority group on all the classification algorithms being studied. However, even though the prediction accuracy of the majority class was slightly reduced. But the indicators of sensitivity and accuracy of the model such as Weighted Accuracy, G-mean, F-score have changed markedly Table I Table II Table III. Achieving the original purpose of the article detected unusual journeys.

SVM: G-means F-score and Weighted accuracy the best results with SVM-SMOTE resampling technique of 19.68%, 6.87%, and 58.65%, respectively.

LR: G-means the best results with SVM-SMOTE resampling technique of 13.9%. F-score and Weighted-accuracy the best results with SMOTE resampling technique of 3.55%, 55.26% respectively.

RF: G-means the best results with Over-Sampling resampling technique of 70.2%. F-score and Weighted-accuracy the best results with SMOTE/T-Link resampling technique of 9.58%, 65.23% respectively.

In each different classification algorithm, there will be data resampling techniques suitable for each algorithm.

G. *GPS route data Shanghai-China*

We have used the dataset provided by the competition from Kaggle website. This data is exploited on Beijing routes. The dataset has 10,357 taxis, mined from June 01, 2015, to June 07, 2015.

Table IV. Performance measures G-mean, F-score, Weighted accuracy (algorithm SVM)

SVM	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.5709	0.0320	0.1276
SMOTE/T-Link	0.5683	0.0316	0.1269
Over-Sampling	0.6042	0.0336	0.1307
Over/T-Link	0.5746	0.0313	0.1261
ADASYN	0.5219	0.0266	0.1162
SVMSMOTE	0.107	0.0259	0.1161

Table V. Performance measures G-mean, F-score, Weighted accuracy (algorithms Logistic Regression)

LR	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.6574	0.0389	0.1408
SMOTE/T-Link	0.6553	0.0386	0.1403
Over-Sampling	0.6596	0.0373	0.1378
Over/T-Link	0.6733	0.0413	0.1345
ADASYN	0.6490	0.0377	0.1386
SVMSMOTE	0.4505	0.0	0.0

Table VI. Performance measures G-mean, F-score, Weighted accuracy (algorithms Random Forest).

RF	Weighted accuracy	F-score	G-mean
Original	0.5	0.0	0.0
T-Link	0.5	0.0	0.0
SMOTE	0.5655	0.0469	0.1600
SMOTE/T-Link	0.6351	0.0515	0.1632
Over-Sampling	0.4989	0.0	0.0
Over/T-Link	0.500	0.0	0.0
ADASYN	0.5634	0.0455	0.1573
SVMSMOTE	0.4847	0.0	0.0

Summary of results when applying data resampling techniques when when applied to different data classification algorithms:

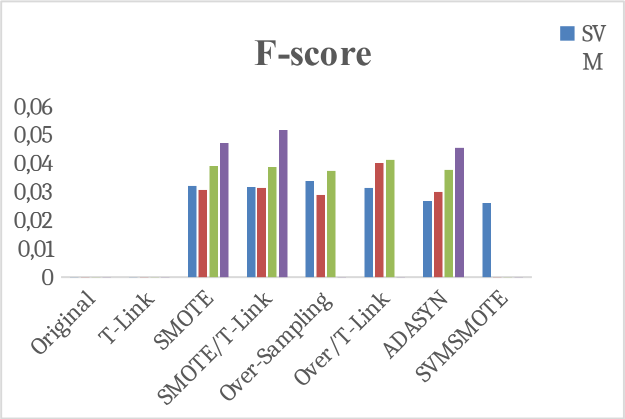


Figure 9. Weighted accuracy of various Machine learning algorithms using various sampling techniques

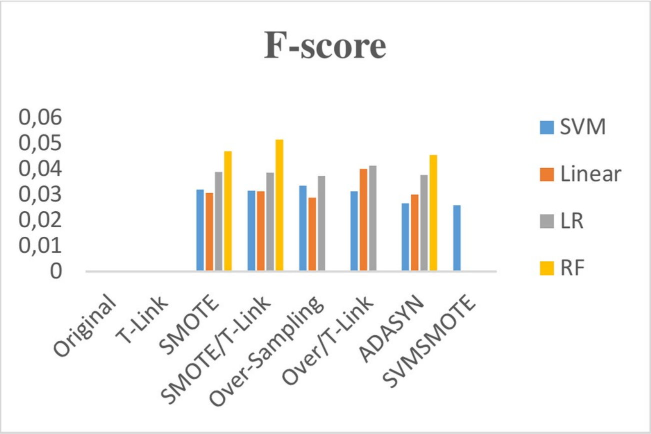


Figure 10. F-score of various Machine learning algorithms using various sampling techniques.

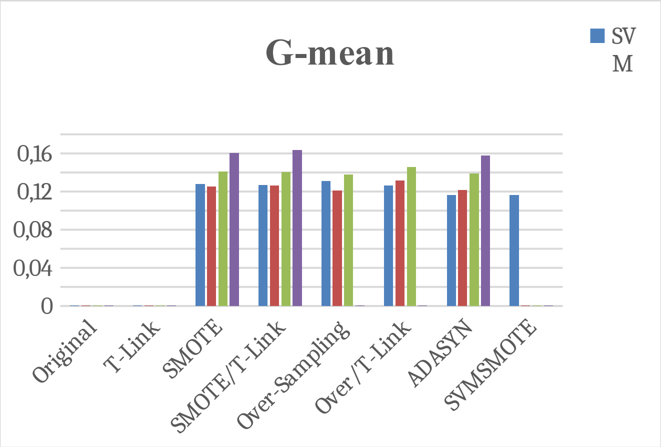


Figure 11 G-mean of various Machine learning algorithms using various sampling techniques

SVM: G-means F-score and Weighted accuracy the best results with SMOTE resampling technique of 12.76%, 3.2%, and 57.09%, respectively.

LR: G-means, F-score and Weighted accuracy achieved the best results with Over/T-Link resampling technique of 13.45%, 4.13%, and 67.33%, respectively.

RF: G-means the best results with SMOTE/T-Link resampling technique of 16.32%. F-score and Weighted-accuracy the best results with SMOTE resampling technique of 5.15%, 65.51% respectively.

Summary: Figure (6-11) We grouped by model evaluation techniques, for each resampling technique for each machine learning algorithm, to show that applying resampling techniques achieves results high when predicting the minority class.

Table (I-IV): G-mean, F-score all algorithms that do not apply resampling have the value 0, because the prediction of the minority cases is incorrect.

Weighted accuracy also shows that performance is improved when resampling techniques are applied, the results are show in Table (I-IV) and Figure (6-11).

Looking at (Table I-IV), our data showed that using SMOTE and T-Link as a combined sampling method has a better performance than T-Link sampling and the original data.

V. Conclutions and Future Work

In this paper, we solve imbalanced data problems in GPS outlier detection and prove that it will achieve better performance for overall process. The need for preprocessing, and resampling data is very important for the problem of data imbalance in machine learning. We demonstrate the effectiveness with real datasets, i.e., GPS logs from vehicle tracking services. We apply resampling techniques to the classification models and compare them with the original (unbalanced) data set. We find that the application of resampling techniques achieves better results for minority class prediction. For future work, we are planning to apply deep learning techniques, learn the optimal threshold of data resampling to deal with class imbalance to improve the predictive model.

References

[1] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 316-324.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[3] V. S. Spelmen and R. Porkodi, "A review on handling imbalanced data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2018: IEEE, pp. 1-11.

[4] P. Nair and I. Kashyap, "Hybrid pre-processing technique for handling imbalanced data and detecting outliers for KNN classifier," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019: IEEE, pp. 460-464.

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008: IEEE, pp. 1322-1328.

[6] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban, "Metric learning from imbalanced data," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019: IEEE, pp. 923-930.

[7] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th international conference on information and communication systems (ICICS)*, 2020: IEEE, pp. 243-248.

[8] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," in *Proceedings: Fifth International Workshop on Computational Intelligence & Applications*, 2009, vol. 2009, no. 1: IEEE SMC Hiroshima Chapter, pp. 24-29.

[9] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[10] C. Nguyen, T. Dinh, V.-H. Nguyen, N. P. Tran, and A. Le, "Histogram-based Feature Extraction for GPS Trajectory Clustering," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems,* vol. 7, no. 22, 2020.

[11] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Transactions on Knowledge and Data Engineering,* vol. 25, no. 1, pp. 220-232, 2011.