Transactions on Machine Learning and Artificial Intelligence - Vol. 9, No. 2
Publication Date: April, 25, 2021
DOI:10.14738/tmlai.92.9964.
Alfhaid, M. A., & Abdullah, M. (2021). Classification of Imbalanced Data Stream: Techniques and Challenges. Transactions on Machine Learning and Artificial Intelligence, 9(2). 36-52.

SOCIETY FOR SCIENCE AND EDUCATION

Classification of Imbalanced Data Stream: Techniques and Challenges

Mashaal A. Alfhaid, Manal Abdullah
[1, 2] King Abdulaziz University, Faculty of Computing and Information Technology, Jeddah, Kingdom of Saudi Arabia

ABSTRACT

As the number of generated data increases every day, this has brought the importance of data mining and knowledge extraction. In traditional data mining, offline status can be used for knowledge extraction. Nevertheless, dealing with stream data mining is different due to continuously arriving data that can be processed at a single scan besides the appearance of concept drift. As the pre-processing stage is critical in knowledge extraction, imbalanced stream data gain significant popularity in the last few years among researchers. Many real-world applications suffer from class imbalance including medical, business, fraud detection and etc. Learning from the supervised model includes classes whether it is binary- or multi-classes. These classes are often imbalance where it is divided into the majority (negative) class and minority (positive) class, which can cause a bias toward the majority class that leads to skew in predictive performance models. Handles imbalance streaming data is mandatory for more accurate and reliable learning models. In this paper, we will present an overview of data stream mining and its tools. Besides, summarize the problem of class imbalance and its different approaches. In addition, researchers will present the popular evaluation metrics and challenges prone from imbalanced streaming data.

Keywords: Data Mining, Imbalanced data, Data Stream Classification, Concept Drift, Data preprocessing.

INTRODUCTION

Nowadays, millions of people around the world use Information and communication technology (ICT)a to share data anywhere at any time. With the increase in technical devices and their rapid development, including their applications, this expansion has led to the rises of the importance of data mining and knowledge extraction which become one of the treasures of our time in many fields, including tourism [1], medicine [2], business [3] and many more. Several models can be used to gain insights whether it is supervised, semi-supervised or unsupervised [4]. Therefore, as this data becomes bigger every day, it cannot be stored in memory for processing and extraction as traditional data. So, due to the change in the nature of the data over time, the concept of streaming data has emerged which forms a quantity of unstoppable data that increases in frequency every second and can be processed in a real-time environment [5]. Data streams can be extracted via Sensors, Social Media, the Internet of Things (IoT) and etc [6, 7]. Dealing with this type of data is a challenging task regarding limited memory, single scanning, concept drift occurrence with its different types and most

importantly dealing with pre-process especially the common phenomenon in every real data is class imbalanced [8],[9].

In knowledge discovery specially supervised learning models that are mainly used for prediction tasks, their performance highly depends on the dataset. Data pre-processes consider one of the major critical phases and must be taking into consideration when dealing with data mining [10]. Besides, it considers one of the stages that consumes most of the time and resources [11]. However, the importance of this phase is crucial since it reflects the quality of the data and the accuracy of the decisions making process [12].

In the real-world scenarios both traditional and non-stationary data environment suffers from the problem of skewed data that almost always exists, where the classes distribution is unequal. This kind of phenomenon is known as imbalanced data where the number of the majority class exceeds the number of the minority class [4]. The classes that include frequently occurring are known as majority classes or negative classes while the rare ones are known as minority classes or positive ones [4] which can contribute to bias toward the dominant class. This kind of learning includes labels of data categorized into classes whether it is binary classes that include only two types of classes or multi-class that include more than two classes.

Solving the problem of class imbalance is important for a more reliable and efficient machine learning accuracy model. For example, in medical diagnosis predicting the cancer disease are mainly concern with the rare classes to detect the probability of rare disease occurrence, the unreliable prediction model can cause danger to human life that can end with death. This paper aims to review the data stream mining concept, challenges, techniques, and tools. Also, we will address the problem of imbalanced data classification, its existing methods and evaluation metrics. Moreover, we will address the different proposed solutions to address this issue. The paper will be organized as follow, section 2 will be an overview of data stream mining and its challenges with its different tools. Section 3 will present the class imbalance problem, approaches and proposed techniques in different studies will be presented. Section 4 will include the conclusion.

OVERVIEW OF DATA STREAM MINING

Data mining is defined as the method for discovering hidden patterns to extract meaningful knowledge from numerous numbers of data using machine learning techniques [13]. In general, the process of data mining has five stages [14], it starts by **(1) Discovering**, where the problem needs to be defined in addition define the hypothesis and understand how to address these problems. **(2) Collect the data,** by define how the data will be generated and the collection method (**3) Prepare the data**, through cleaning the noise, missing value and transformation if needed. **(4) Modelling**, which selects and implements the most suited model for the problem defined (e.g., classification) and validates it. **(5) Interpret the results,** by drawing a conclusion and knowledge that will be helpful in the decision-making process. Data mining can be divided into three main techniques [15]. **(1) Supervised learning**, which is used for a predictable model and has training label data in order to identify the relationships such as classification and regression. **(2) Unsupervised learning**, where it does not have an outcome to predict or classify. However, its main goal to find the relationship between the data point itself such as clustering and association. **(3) Semi-supervised learning,** which combined both label and un label. Below are the most common methods used in knowledge extraction [16]:

- Cluster: is a type of unsupervised learning technique that used to group similar object into clusters. Since data stream has a continuous arrive of data, dealing with cluster must be updated incrementally with the data arrives [17]. Stream clustering can be online and offline. It basically has several algorithms divided into five categories [18] **(1)** Hierarchical algorithm. **(2)** Partitioning-based. **(3)** Grid-based. **(4)** Density-based. **(5)** Model-based. The most popular algorithms to deal with stream cluster are ClueStream [19] that consider model-based algorithm, E-Stream [18] which is hierarchical algorithm and DenStream [17] that is considered to be density-based.
- Classification: is a type of supervised learning algorithms where it basically concerns in prediction tasks and consists of many algorithms wither decision trees, Bayesian classifiers, etc. Classification in data stream must deal with streaming arrives instances to deal with. It basically divided data into two main steps **(1)** Training set that is used in order to build the model. **(2)** Testing set in order to predict the label of class. There are several algorithms to deal with the stream classification methods include CVFDT and VFDT [20].

Data Stream Evolution

For many years data has been surrounding us, the nature of the data has changed over time due to the continuous development of numerous new technologies and real-time applications. Many applications appeared that converted static data into non-stoppable data or the so-called streaming data, which led to the need to change the way of extracting knowledge from it through different methods and techniques. Data can play a crucial role in decision-making for the community's development and prosperity. Studies have shown that data has divided into static data and stream data. As shown in table 1, static data is a type of data that can be fully accessed and available. It focuses on analysing static repositories for traditional data mining which can be accessed many times. Therefore, the increase of online applications and the continuous data generated every day brought the concept of the data stream.

Data Stream is defined as a type of continuous and non-stoppable data that produced incrementally [5]. The continuous flow of the data stream is one of the main sources of big data [21]. Big data refers to the enormous amount including high volume, velocity, and variety of data [22]. One of the most important sources for extracting data streaming is social networks such as Twitter, Instagram, Facebook and etc. These kinds of applications include real-time data about posts, likes, comments, tweets, retweets, etc. Social network has become an important part of human life with a rich source of streaming data. While on the other hand, the Internet of things (IoT), cloud computing and sensors provide numerous amounts of streaming data[6]. Finding and understanding the methods to use in this kind of data is crucial since different industry depends on its real-time data include the stock market and financial industries [23].

Table 1. Traditional data Vs Steam data [24].

Static Data	Stream Data
Offline	Arrive online
Can control the data order	Cannot control the order
Any time access	Single pass
Bounded size	Unbounded in size
Random access	Sequential access

Challenges

Data streams involve many challenges that must be aware of to overcome. Some of the most popular challenges are listed below include:

- Handling memory space is unlike static data, streaming data face a challenge that needs more efficient techniques to handle these extremely massive or infinite data. Besides, storing real-time data can be hard and costly [8].
- Single scan, unlike traditional data, streaming data is keeping growing over time it cannot be processed in a single scan, using a single scan in streaming data require to drop some of the data [24].
- Pre-processes include class imbalance which consider one of the important phases in knowledge discovery which highly support the model accuracy. Dealing with pre-processing in stream data mining requires specific techniques for more accurate machine learning model [12].
- Dealing with concept drift consider one of the challenges to handle and detect the changing properties of the data to build an efficient model, as it keeps increasing which causes changes in unforeseen ways [9]. Concept drift means the environment in which the statistical properties of data changes over time [25]. There are different types of drift to deal with include [26]: **(1) Sudden** the data changes unexpectedly in the distribution of class without alteration. **(2) Incremental** were the data changing over time. **(3) Gradual** where the state of data is not stable, but it does not occur suddenly. While the **(4) Recurring concept** where the state changes and repair after a time.

Techniques for Data Stream Mining

The process of selecting the data has been classified into two categories [24]. **(1) Data-based solutions**, which refer to summarizing the stream of data otherwise choose a subset to reduce the streaming data. It includes sampling, load shedding, sketching, synopsis data and aggregation. **(2) Task-based solutions** refer to the process of modifying an existing technique or invent a new one for the data stream. It includes approximation algorithms, sliding windows, algorithm output granularity.

Table 2: Techniques Data Selection [24] [11].

Technique	Type	Description	Obstacles
Sampling	Data-based	Refers to the process of probabilistic choice of a data item to be processed which consider very fast technique.	Sampling cannot determine the data size.
Load shedding		Refers to the process of dropping a sequence of data streams.	It cannot be used in mining algorithm
sketching		the process of randomly project a subset of the features	The main obstacles are the accuracy
synopsis data		the process of summarization techniques for streams (e.g., histograms, quantiles) for further analysis	It does not represent all data characteristics
aggregation		the process of computing statistical measures (e.g., means) that summarize the stream	it does not perform well with highly changing data distributions
approximation	Task-based	concerned with design algorithms for computationally hard problems	the problem of data rates with regard with the available resources could not be solved
sliding windows		The process of analysis most recent data streams.	----------
output granularity		It can face with turning high data rates according to the available memory and speed in time constraints.	--------

Data Stream Tools

There are many open-source tools and packages to deal with the dynamically changing environment of streaming data. These tools and packages used in the data stream environment are listed below:

- Spark tool [27], is one of large-scale data processing that has high performance to deal with streaming data and advance analytic which could be written using python and R languages. It can be standalone or runs into Hadoop or apache or cloud.
- Scikit-Multiflow [28], is open-source free library for machine learning algorithms package for streaming data basically in python. It supports supervised, unsupervised learning and concept drift detection. It also compatible with Jupyter notebooks.
- Apache Storm [29], tool is a free open-source system for process unbounded real-time streaming data, it considers to be easy, and It is scalable and is easy to set up and operate. Moreover, can be implemented in any programming language.
- MOA [30], is a shortcut for massive online analysis which is a software environment that deals with the massive data stream and mining techniques such as classification and clustering. It basically uses Java programming language.
- Weka [31], is open-source software that using Java programming language it can deal with machine learning algorithms such as regression and clustering and can run on any platform. Weka also can train classifiers without the need for coding commands.

CLASS IMBALANCE PROBLEM

As the nature of data can produce challenges and difficulties in many practical data mining tasks, especially in a supervised approach. Learning from data can be affected by an important factor which called Imbalanced Classification that refers to a skewed data where the number of negative class are higher than the positive classes which usually a critical problem especially in predictive tasks context [32]. It can be identified by the imbalanced ratio (IR) as depicted by equation (1) [33], if the value of IR is equal to one this indicated that the data set is perfectly balanced otherwise it is needed for processing.

$$IR = \frac{Number\ of\ observation\ of\ Majority}{Number\ of\ observation\ of\ Minority} \qquad \textbf{(1)}$$

For example, if we want to obtain a prediction model for medical diagnoses and the dataset include 100 minority class and 4000 majority class, this requires a classifier that gives a balance predictive accuracy for both classes.

Nowadays, many application areas obtain a prediction model that requires a high accuracy to rely on. Many real-life applications suffer from an uneven representation of the data which is listed in table 3. As Real-world data is not static, the need of efficient algorithms to handle the issue of imbalanced stream data is mandatory especially when dealing with supervised learning such as prediction tasks through classification algorithms. The prediction result can have highly skew performance, insufficient and less accurate outcomes due to imbalanced data.

Table 3: Applications for Imbalanced Data

Application Area	***Reference***
Cancer Diagnosis	[34]
Face Recognition	[35]
Social network	[36]
Fraud Detection for E-commerce Transactions	[37]

Imbalanced data classes can have different two approaches as shown in figure 2, this includes [38]: **(1) Binary-Classes** where the dataset has only two classes, one positive and one negative [39]. **(2) Multi-Classes** that has more than two classes [39].

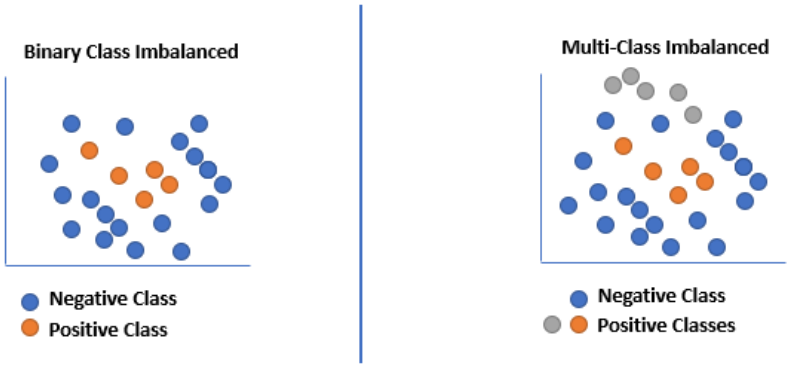


Figure 2: Imbalanced classes.

Many factors can influence the classifier to identify the rare events or minority samples this include sample size, where a small data size can cause difficulties in finding inherent to small classes [40]. In addition to class separability, an overlapping sample may cause difficulties in classifying them correctly [40] as shown in figure 3. And the last common phenomenon which refers to within-class imbalance which contain several sub-clusters with a different number of examples [41] as shown in figure 4.

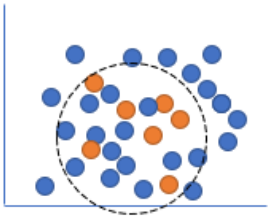


Figure 3: Class Separability factor.

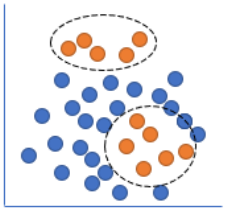


Figure 4: within Class Clusters.

Methods to Address Class Imbalance

The way of handling imbalanced data has been divided into four main groups as shown in figure 5 [42] include:

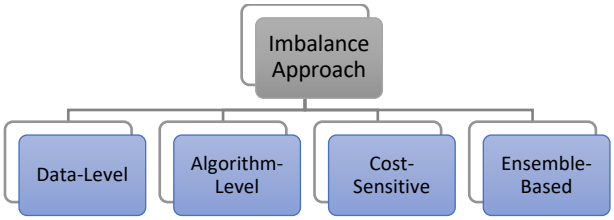


Figure 5: Approaches to Address Class Imbalance

(1) Data-Level, where the data is modified in the training set in order to produce more balanced data for better classifier performance. Data pre-process for data-level techniques in imbalanced data consider most popular which divided into three main algorithms [43] and it been listed in figure 6 with most popular algorithms.

- Over-sampling Algorithms ROS, which create a new instance from existing once the most popular one is Synthetic Minority Over-sampling Technique SMOTE [43]. One of the drawbacks in this technique is over generalization is largely attributed to the way in which it creates synthetic samples.
- Under-sampling Algorithms RUS, this algorithm is the opposite than the first one, which create a subset of the original data set and remove majority class to reduce the imbalance ratio. However, this method can discard number of useful information that could be critical in decision making process.
- Hybrid Algorithms, it basically combines the two algorithms above over and under sampling techniques. They either combine over- and under sampling at one time or perform under sampling after oversampling as a data cleaning step.

(2) Algorithm-Level, where the training set is not modified as the previous approach. Therefore, in which a new algorithm is created or altered which consider less flexible, these algorithms can include support vector machine SVMS and its variants [44] [45]

and Bayesian classifiers [46]. **(3) Cost-sensitive level** is a way of minimizing the cost error by using data level and algorithm level one example include cost-sensitive support vector machines [47]. **(4) Ensemble-based methods**, which use a combination of ensemble learning that use a combination of classifier in order to enhance the accuracy in addition one of the three level data-level, algorithm-level and cost-sensitive to improve the performance. Two most popular ensemble methods to deal with imbalanced problem include:

- Bagging, where the weak leaner model learns in parallel and then combine their output such as UnderOverBagging which combine the data level with ensemble-based [42].
- Boosting, where the weak learner model works sequentially such as Cost-Sensitive Boosting [42].

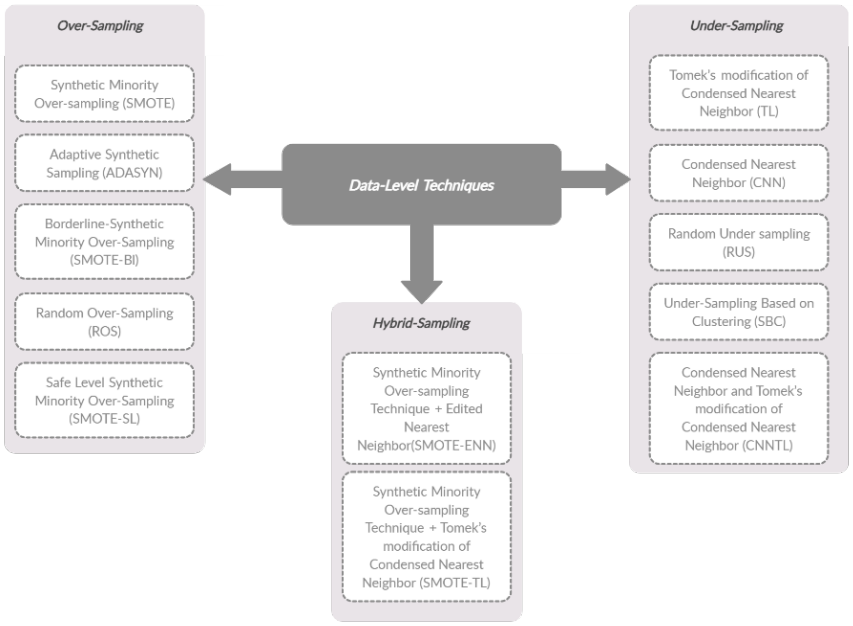


Figure 1: Popular Data-level Techniques [43, 48, 49].

Evaluation Criteria for Class Imbalance

In order to produce an optimal classification model, one of the important steps to handle the imbalanced classification problem is using evaluation metrics [50, 51]. The most famous method to measure the classification is accuracy however, dealing with imbalanced data cannot be measured through accuracy because it cannot determine the effect of small classes since small classes has very little impact on accuracy compared to majority classes. Nevertheless, several evaluation metrics can be used to determine the number of positive and negative classes one of them is the Confusion matrix [52] where the number of correctly classified classes represented using a true positive and true negative. In addition to wrongly classified classes that represented by the number of a false positive and false negative. Besides, F-measure [4] that combined both Precision and recall [4], when the number of precision and recall high then we can say the metric is having high value. G-Mean [4] is also another metric which refers to geometric mean that measure the accuracy of two classes and ROC [53] which refer to the Receiver Operating Characteristics.

Proposed Solutions to deal with Class Imbalance

In the last few years, many researches have focus on the class imbalance phenomenon to solve regarding traditional data. Nevertheless, with the numerous numbers of data that has produced every day and the massive growth in technology has risen the imbalanced issue in the streaming data which require a specialized way to handling it because of its different challenges including concept drift appearance. Several studies from (2014-2021) have applied various algorithms in order to balance the datasets and get more optimal results. Several libraries had been used in these studies to extract the data include Keel [54], UCI Machine Learning repository [55], PROMISE Software Engineering repository [56], Stream-Learn generator [57], MOA [58] and Kaggle [59].

A. Class Imbalance in Traditional Data:

Several studies have applied oversampling techniques to solve the problem of imbalanced data and integrate it with other methods. As authors in [32] has applied a combination of density-based spatial clustering of applications with noise (DBSCAN) [60] with the popular data-level method the over-sampling algorithm that refer to synthetic minority over-sampling technique to propose a density-based SMOTE (DSMOTE). According to the metrics the results show that DSMOTE has higher accuracy performance than the state-of-art algorithms SMOTE and Borderline-SMOTE. However, it needed for manually setting some parameters which consider not practical specifically when it deals with large size of data. While similar authors in [61] try to solve imbalanced problem in static data set. However, they consider in the presence of other impact such as noise or overlapping. They have proposed an algorithm called Multi-class Combined Cleaning and Resampling; their algorithm combine cleaning operation with SMOTE for more optimal oversampling that the traditional ones.

On the contrary, authors in [62] has proposed a hybrid sampling approach specially for binary classification problem. The framework included grouping by divide the training set into two clusters then using sampling approach to each cluster include under-sampling and over-sampling depending on the number of positive and negative class while the final step will be gathering the clusters together in one new training set. They also have done an experiment for both decision tree and naïve bays models and it performs highest accuracy in term of F-measure and predictive accuracy. Similar [63] has proposed a framework to solve binary problems for highly imbalanced data with imbalanced ration up to 40. The method combines both over-sampling and classifier ensemble. Authors in [64] has found an interesting point is that the under-bagging methods perform better than the over-bagging methods. Besides, they proposed an under-sampling-based bagging ensemble (EUS-Bag) that solve the problem of two class imbalanced in static datasets. Their method is depending on searching for the more optimal majority class and they have compared it with the state-of-art method and prove it have the highest performance. While [65], has also applied hybrid sampling approach to propose a framework CMTFSVM that combine fuzzy support vector machine (FSVM) to eliminate any noise in the datasets with over-sampling technique SMOTE for imbalanced static datasets. They have proved that combining both over- and under-sampling technique provide optimal results. While better performance is done by integrating FSVM classifier.

Table 5: Imbalanced static data studies

	Algorithm	Technique	Compared State-of-art algorithms	Dataset	Key points
1	DSMOTE [32]	Density-based Spatial Clustering with SMOTE	-SMOTE -Borderline-SMOTE	-2 class datasets -2 datasets with multi-class	It performs better than state-of-art. However, it needed for manual setting some parameters
2	MC-CCR [61]	Cleaning and Resampling algorithm	-S-SMOTE -Borderline-SMOTE -S-SMOTE+ENN	20 datasets with multi-class	Perform well for multi-class imbalanced in static data.
3	Hybrid sampling [62]	Over- and Under-sampling	-SMOTE -Tomek Link (TL)	Binary Classification for 12 datasets	Perform well for Binary-classification imbalanced in static data.
4	UMCE [63]	Under-sampling and class ensemble	none	40 datasets	Solve highly class imbalanced static data
5	EUS-Bag [64]	-Under-sampling based bagging ensemble	-Bagging -ROS-Bag -SMO-Bag -EB-Bag -RUS-Bag -Esy-Ens -Bal-Cad	22 datasets with two-class imbalanced	Highest performance than the stat-of-art bagging methods in static datasets
6	CMTFSVM [65]	-FSVM -SMOTE	-NN classifier -SVM classifier -FSVM classifier	5 datasets with binary classification	Better performance for noise and imbalanced binary classification static dataset problem

B. Class Imbalance in Streaming Data:

As authors in [66] used hybrid data pre-process that include both over- and under-sampling techniques to propose a Deterministic Sampling Classifier (DSC), they were especially interested in binary classification problem in 86 data streams of both real and synthetic data. On the other hand [67] has relied on ensemble approach to propose a classification method minority driven ensemble (MDE) using classifier selection and chunk-based to deal with non-stationary imbalanced data stream in a form of fixed chunks that consist of both gradual and sudden drift based on binary classification. Their statistical analysis result shows that the proposed solution is significant when the imbalance ratio is large. Like authors in [45] proposed a one-class support vector machine classifier ensemble (OCEIS) based on using one-class support vector machine (OCSVM) to solve binary imbalanced problem in data stream. It considers to be a chunk-based method. The algorithm has applied on generated and real streams data that do not exceed 60,000 samples.

While some has integrated both pre-processing techniques with ensemble based as authors in [68] has integrate six of the pre-processing methods and dynamic ensemble selection to apply it on different stream datasets that include both sudden and incremental drift with a fixed size chunks and label noise up to 20%. They also prove the

accuracy of their ensemble without the use on any pre-process technique. Also, authors in [69] has compared the ensemble approach streaming ensemble (SEA), modified weighted accuracy (WEA) and accuracy updated ensemble (AUE2) with over-sampling algorithms SMOTE and its variants. While the SEA has been tested with gradual, sudden, and recurring drift types. They also have applied the experiment using the test then train approach for nine datasets each with 100,000 samples where each chunk has 250 samples. Study [70] proposed an Ensemble Classifier of imbalanced streaming data (ECISD) which consider to be block-based that integrate ensemble learning dynamic weighting strategy that deal with the occurrence of concept drift with both over-sampling TL and under-sampling SMOTE technique in order to balance the classes. The datasets include sudden concept drift type. They also, has study the block size of the proposed algorithm from 500 to 2000 and done an experiment on it on each of the datasets where they observe that the proposed framework is not sensitive to the size of the block.

On the other hand, authors in [71] has proposed a framework that combine both data-level and algorithm-level to yield a hybrid framework that tackle the problem of imbalanced data. This framework consists of three components include resampling where the SMOTE is employed in this component, classifier ensemble for updating mechanism and cost-sensitive classifier. Their experiment was based on efficiency and effectiveness with window size 100 to 2000 of their proposed framework where they found that hybrid was the fastest than the other two resampling and cost sensitive. They also design a data generation to generate three kinds of concept drift include gradual, reoccurring, and abrupt. In terms of authors in [72], they has proposed a framework for solving the issue of imbalanced data stream in form of fixed size data chunk in binary classification with the occurrence of single concept drift with it gradual, sudden, or incremental occurrence. They have integrated dynamic ensemble selection stratified bagging with prepossessing techniques include over- and under-sampling. They have developed an experiment based on three part first to test the ensemble classifier as a stand-alone then pair it with pre-processing technique and the final one is to compare it with state-of-art methods.

In study [73], authors have focused on binary classification problem in non-stationary streaming data by proposing selection-based resampling ensemble classifier (SRE) [73]. It basically has four steps first they divide the training data into fixed size of blocks to avoid exceeding memory usage then a selectively resampling technique will be used in the second step will evaluate the weight. While third update ensemble members which designed to react to drifts type. The method proves it effectiveness again the different chunk size and drift type for binary classification problem. Authors in [74] has also focus on two class problem of imbalanced data that include sudden and gradual drifts by proposed an ensemble static classifier that handle gradual changes and multiple dynamic classifiers sliding window were the dynamic suits sudden drift. It proves its effectiveness in terms of accuracy an AUC compared to state-of-art methods. While for multi-class [75] has focus on three issues include drift, imbalanced for multi-class data and limited access ground truth for multi-class. Their experiments were against the state-of-art methods of both supervised and sparsely labelled data. Their method was sufficient for only single classifier.

In study [76], authors proposed a framework Reduced Label Samples (RLS) that detect concept drift in imbalance data. The framework includes two stages at first the model is bult when the first chunk of streaming data arrives by using support vector machine SVM, and since it sensitive to imbalance data used SMOTE oversampling method to solve the imbalanced problem. Second stage is when second chunk of stream data arrive apply the same unless concept drift detected it then needed to update the model. Detecting concept drift by comparing the F-measure to the first chunk if it higher than 13% it will be updated. The update model performs by critical samples. However, the experiment was only train for conditional drift. Authors in [77] proposed Concept drift detection toward imbalanced stream data (CD2A) that is used to detect various drifts type for binary classification by using Cuckoo search-based incremental binary classifier. They have detected the changes by divide the streams of data into chunks similar to classifier and apply Jaccard similarity [77] if the distance of chunks and window records larger than threshold then remove the old records and update the model.

Table 6: Imbalanced Streaming Data Studies

	Algorithm	**Technique**	**Compared State-of-art algorithms**	**Dataset**	**Key points**
1	DCS-R DSC-S [66]	Over- and Under-sampling SMOTE and NearMiss	-Learn++CDS -Learn++NIE -REA	-Multi-dimensional binary classification -26 real stream -60 synthetic	Perform well for multi-dimensional classification problem and it utilize the fixed memory buffer
2	MDE [67]	-Chunk based -ensemble learning -classifier selection	-Rank -LCA -KNORAE -KNORAU	-Five concept drifts	Highly class imbalanced for stream data. However, statistically it was not significantly better than KNORA in small IR
3	OCEIS [45]	-OCSVM -Chunk-based	-L++CDS -L++NIE -KMC -REA -OUSE -MLPC	-24 generated -30 real streams	Used to solve binary imbalanced problem in the exist of sudden, incremental drift and label noise
4	Data Pre-processing and DESL [68]	-pre-processing -dynamic ensemble selection [78]	-SMOTE -SVM-SMOTE -Borderline-SMOTE -SL-SMOTE -ADASYN	72 artificially generated	It performs well for self-updating ensemble model
5	Compared ensemble with SMOTE variants [69]	------	-SEA -AUE2 -WAE -SMOTE -Borderline-SMOTE -SL-SMOTE	9 datasets	Proved SMOTE Effectiveness against another ensemble approach

6	ECISD [70]	-Over- and Under-sampling - Block-based ensemble	-Learn++.NIE -AUE2 -VFDT	-3 synthetic -2 real-world	Perform well in dynamic environment and recurring drift
7	Hybrid framework [71]	-Resampling -Classifier updating -Cost sensitive	none	-2 real-world -Synthetic -concept drift types	The framework considers to be effective predictor in case of stream data
8	Pre-processed dynamic classifier ensemble selection [72]	-Dynamic ensemble selection -Stratified bagging - Over- and Under-Sampling	- Online Bagging - Oversampling-Based Online Bagging -Learn++.NIE -Learn++CDS -REA Over\Under sampling Ensemble -KMC	-90 Generated Data -45 Generated Data	It performs well for single concept drift binary classification problem
9	SRE [73]	- Selective resample mechanism - Periodical update mechanism	-AWE -SERA -MuSeRA -SMOTE -UB -OOB -SRE	-7 synthetic and 1 real-world datasets -(1,3,4,9) drifts	SRE in learning nonstationary imbalanced data streams for different type of concept drifts in binary classification problem and for only continuous attribute.
10	RE-DI [74]	-Static Classifier -Reinforcement mechanism	-OOB _UOB -LB -ARF	-4 real and 8 synthetic data streams -(1 Gradual,3 Sudden) drifts -2 class	focus on two class problem of imbalanced data
11	multi-class data streams [75]	-Active learning -Oversampling	-LB -OZABAG -OB-ADW -DWM	-1 real data and 12 semi-synthetic data. -(1,2) concurrently Drift	focus on multi class problem of imbalanced data with one type of drift
12	RLS-A [76]	-SVM -critical samples	none	synthetic and real-world	Decrease the label samples effectively from binary classification problem
13	CD2A [77]	-Binary classifier cuckoo search - under-sampling	-AWE -SMOTE -OOB	2 synthetic and 1 real	Detect drift types for binary problems and make the classifier up to date

CONCLUSION

Data stream imbalance are a common concern lately, especially when dealing with knowledge discovery projects. In order to produce a more optimal machine learning model, we have to reprocess this data to get more balance datasets since the interested point goes to the abnormal class behavior. In this paper, several approaches to solve the problem of class imbalance had been summarized whither its existing approach such as SMOTE or proposed once. Data stream imbalance classification consider a challenge task because of the nature of streaming data as it

keeps grow all the time specially with the existing of concept drift as not many studies had been focusing on.

References

[1] R. Zhou, S. Lin, and M. Lin, "Design of evaluation system of residents' tourism quality of tourism area based on fuzzy evaluation," in *Ecosystem Assessment and Fuzzy Systems Management*: Springer, 2014, pp. 189-197.

[2] T. Mahmud, J. Sikder, and S. Tripura, "Knowledge-based Decision Support System to Select Hospital Location."

[3] C. Caserio, "Knowledge generating decision support systems: managing the trade-off between generating knowledge and supporting decisions," in *Information systems: crossroads for organization, management, accounting and engineering*: Springer, 2012, pp. 303-311.

[4] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[5] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion,* vol. 37, pp. 132-156, 2017.

[6] C. Widanage *et al.*, "Anomaly detection over streaming data: Indy500 case study," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, 2019, pp. 9-16: IEEE.

[7] D. Le-Phuoc, J. X. Parreira, and M. Hauswirth, "Linked stream data processing," in *Reasoning Web International Summer School*, 2012, pp. 245-289: Springer.

[8] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: a review," *ACM Sigmod Record,* vol. 34, no. 2, pp. 18-26, 2005.

[9] R. Jani, N. Bhatt, and C. Shah, "A survey on issues of data stream mining in classification," in *International Conference on Information and Communication Technology for Intelligent Systems*, 2017, pp. 137-143: Springer.

[10] E. Ikonomovska, S. Loskovska, and D. Gjorgjevik, "A survey of stream data mining," in *Proceedings of 8th National Conference with International participation, ETAI*, 2007, pp. 19-21.

[11] M. Kholghi and M. Keyvanpour, "An analytical framework for data stream mining techniques based on challenges and requirements," *arXiv preprint arXiv:1105.1950,* 2011.

[12] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing,* vol. 239, pp. 39-57, 2017.

[13] J. J. Cochran, *Informs analytics body of knowledge*. John Wiley & Sons, 2018.

[14] G. Mariscal, O. Marban, and C. Fernandez, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review,* vol. 25, no. 2, p. 137, 2010.

[15] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *Ieee Access,* vol. 5, pp. 20590-20616, 2017.

[16] H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng, "A survey on data stream clustering and classification," *Knowledge and information systems,* vol. 45, no. 3, pp. 535-569, 2015.

[17] S. Laohakiat, S. Phimoltares, and C. Lursinsap, "A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction," *Information Sciences,* vol. 381, pp. 104-123, 2017.

[18] A. Zubaroglu and V. Atalay, "Data stream clustering: a review," *arXiv preprint arXiv:2007.10781,* 2020.

[19] M. Carnein and H. Trautmann, "Optimizing data stream representation: An extensive survey on stream clustering algorithms," *Business & Information Systems Engineering,* vol. 61, no. 3, pp. 277-297, 2019.

[20] M. M. YACOUB, A. REZK, and M. SENOUSY, "ADAPTIVE CLASSIFICATION IN DATA STREAM MINING," *Journal of Theoretical and Applied Information Technology,* vol. 98, no. 13, 2020.

[21] G. Casalino, G. Castellano, and C. Mencar, "Data Stream Classification by Dynamic Incremental Semi-Supervised Fuzzy Clustering," *International Journal on Artificial Intelligence Tools,* vol. 28, no. 08, p. 1960009, 2019.

[22] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data,* vol. 6, no. 1, p. 44, 2019.

[23] R. S. Barros, D. R. Cabral, P. M. Gonçalves Jr, and S. G. Santos, "RDDM: Reactive drift detection method," *Expert Systems with Applications,* vol. 90, pp. 344-355, 2017.

[24] M. Lange, A. Koschel, and I. Astrova, "Dealing with Data Streams: Complex Event Processing vs. Data Stream Mining," in *International Conference on Computational Science and Its Applications*, 2020, pp. 3-14: Springer.

[25] N. R. Pal, E. S. Corchado, L. T. Kóczy, and V. Kreinovich, "Advances in intelligent systems and computing," 2014.

[26] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering,* vol. 31, no. 12, pp. 2346-2363, 2018.

[27] (Nov 20). *Apache spark Lightning-fast unified analytics engine*. Available: https://spark.apache.org/

[28] scikit-multiflow. (2019-2020). *scikit-multiflow A machine learning package for streaming data in Python*. Available: https://scikit-multiflow.github.io/

[29] A. S. Foundation. *Apache Storm*. Available: http://storm.apache.org/

[30] MOA. (Nov 20). *Machine Learning for Data Streams*. Available: https://moa.cms.waikato.ac.nz/

[31] (Nov 20). *Data Mining: Practical Machine Learning Tools and Techniques*. Available: https://waikato.github.io/weka-wiki/citing_weka/

[32] X. Xiaolong, C. Wen, and S. Yanfei, "Over-sampling algorithm for imbalanced data classification," *Journal of Systems Engineering and Electronics,* vol. 30, no. 6, pp. 1182-1191, 2019.

[33] R. Zhu, Z. Wang, Z. Ma, G. Wang, and J.-H. Xue, "LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test," *Pattern Recognition Letters,* vol. 116, pp. 36-42, 2018.

[34] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *Journal of biomedical informatics,* vol. 90, p. 103089, 2019.

[35] C. Huang, Y. Li, C. L. Chen, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE transactions on pattern analysis and machine intelligence,* 2019.

[36] J. Zhao and N. Liu, "Semi-supervised Classification Based Mixed Sampling for Imbalanced Data," *Open Physics,* vol. 17, no. 1, pp. 975-983, 2019.

[37] S. Carta, G. Fenu, D. R. Recupero, and R. Saia, "Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model," *Journal of Information Security and Applications,* vol. 46, pp. 13-22, 2019.

[38] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM),* vol. 3, no. 3, pp. 1-13, 2007.

[39] M. J. Er, R. Venkatesan, and N. Wang, "An online universal classifier for binary, multi-class and multi-label classification," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 003701-003706: IEEE.

[40] A. Fernández, S. García, and F. Herrera, "Addressing the classification with imbalanced data: open problems and new challenges on class distribution," in *International conference on hybrid artificial intelligence systems*, 2011, pp. 1-10: Springer.

[41] N. Japkowicz, "Concept-learning in the presence of between-class and within-class imbalances," in *Conference of the Canadian society for computational studies of intelligence*, 2001, pp. 67-77: Springer.

[42] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018.

[43] S. Vluymans, *Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods*. Springer, 2019.

[44] L. G.-A. a. H. N. a. C. A. a. F. Velasco, "GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems," *Applied Soft Computing,* vol. 17, 2014.

[45] J. Klikowski and M. Woźniak, "Employing One-Class SVM Classifier Ensemble for Imbalanced Data Stream Classification," Cham, 2020, pp. 117-127: Springer International Publishing.

[46] P. R. a. S. Jaiyen, "Hybrid ensembles of decision trees and Bayesian network for class imbalance problem," in *2016 8th International Conference on Knowledge and Smart Technology (KST)*, 2016, pp. 39-42.

[47] C. Lemnaru and R. Potolea, "Imbalanced classification problems: systematic study, issues and best practices," in *International Conference on Enterprise Information Systems*, 2011, pp. 35-50: Springer.

[48] O. Komori and S. Eguchi, *Statistical methods for imbalanced data in ecological and biological studies*. Springer, 2019.

[49] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *Int J Recent Trends in Eng & Res,* vol. 3, pp. 444-449, 2017.

[50] P. Branco, L. Torgo, and R. P. Ribeiro, "Relevance-based evaluation metrics for multi-class imbalanced domains," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2017, pp. 698-710: Springer.

[51] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences,* vol. 513, pp. 429-441, 2020.

[52] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process,* vol. 5, no. 2, p. 1, 2015.

[53] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J Inf Eng Appl,* vol. 3, no. 10, 2013.

[54] Keel. *Knowledge Extraction based on Evolitionary Learning*. Available: http://www.keel.es/

[55] UCI. *Welcome to the UC Irvine Machine Learning Repository!* Available: https://archive.ics.uci.edu/ml/index.php

[56] *PROMISE Software Engineering repository* Available: http://promise.site.uottawa.ca/SERepository/

[57] *Stream-Learn*. Available: https://pypi.org/project/stream-learn/

[58] *Machine Learning for Data Streams*. Available: https://moa.cms.waikato.ac.nz/

[59] Kaggle. Available: https://www.kaggle.com/

[60] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, vol. 96, no. 34, pp. 226-231.

[61] M. Koziarski, M. Woźniak, and B. Krawczyk, "Combined Cleaning and Resampling Algorithm for Multi-Class Imbalanced Data with Label Noise," *arXiv preprint arXiv:2004.03406,* 2020.

[62] A. Hanskunatai, "A new hybrid sampling approach for classification of imbalanced datasets," in *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, 2018, pp. 67-71: IEEE.

[63] P. Ksieniewicz, "Undersampled majority class ensemble for highly imbalanced binary classification," in *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2018, pp. 82-94.

[64] B. Sun, H. Chen, J. Wang, and H. Xie, "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification," *Frontiers of Computer Science,* vol. 12, no. 2, pp. 331-350, 2018.

[65] R. Pruengkarn, K. W. Wong, and C. C. Fung, "Imbalanced data classification using complementary fuzzy support vector machine techniques and smote," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 978-983: IEEE.

[66] B. Bobowska, J. Klikowski, and M. Woźniak, "Imbalanced Data Stream Classification Using Hybrid Data Preprocessing," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019, pp. 402-413: Springer.

[67] P. Zyblewski, P. Ksieniewicz, and M. Woźniak, "Classifier selection for highly imbalanced data streams with minority driven ensemble," in *International Conference on Artificial Intelligence and Soft Computing*, 2019, pp. 626-635: Springer.

[68] P. Zyblewski, R. Sabourin, and M. Woźniak, "Data preprocessing and dynamic ensemble selection for imbalanced data stream classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019, pp. 367-379: Springer.

[69] B. Gulowaty and P. Ksieniewicz, "SMOTE Algorithm Variations in Balancing Data Streams," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2019, pp. 305-312: Springer.

[70] Y. Sun, Z. Wang, H. Li, and Y. Li, "A Novel Ensemble Classification for Data Streams with Class Imbalance and Concept Drift," *International Journal of Performability Engineering,* vol. 13, no. 6, 2017.

[71] W. Zhang and J. Wang, "A hybrid learning framework for imbalanced stream classification," in *2017 IEEE International Congress on Big Data (BigData Congress)*, 2017, pp. 480-487: IEEE.

[72] P. Zyblewski, R. Sabourin, and M. Woźniak, "Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams," *Information Fusion,* vol. 66, pp. 138-154.

[73] S. Ren *et al.*, "Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning," *Knowledge-Based Systems,* vol. 163, pp. 705-722, 2019.

[74] H. Zhang, W. Liu, S. Wang, J. Shan, and Q. Liu, "Resample-based ensemble framework for drifting imbalanced data streams," *IEEE Access,* vol. 7, pp. 65103-65115, 2019.

[75] Ł. Korycki and B. Krawczyk, "Online Oversampling for Sparsely Labeled Imbalanced and Non-Stationary Data Streams," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8: IEEE.

[76] E. Arabmakki, M. Kantardzic, and T. S. Sethi, "RLS-A reduced labeled samples approach for streaming imbalanced data with concept drift," in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 2014, pp. 779-786: IEEE.

[77] M. A. A. Abdualrhman and M. Padma, "CD2A: Concept Drift Detection Approach Toward Imbalanced Data Stream," in *Emerging Research in Electronics, Computer Science and Technology*: Springer, 2019, pp. 597-612.

[78] J. Montiel, J. Read, A. Bifet, and T. Abdessalem, "Scikit-multiflow: A multi-output streaming framework," *The Journal of Machine Learning Research,* vol. 19, no. 1, pp. 2915-2914, 2018.