A NOVEL APPROACH USING INCREMENTAL MULTI MODAL OVERSAMPLING FOR DATA STREAM MINING

ANUPAMA N[1],RAVI SANKAR V[2], SUDARSON JENA[3]

Research Scholar,Gitam School of Technology, Department of CSE, Hyderabad, Telangana, India
Associate professor, Gitam School of Technology, Department of CSE,Hyderabad,Telangana,India
Associate professor,SUIIT, Department of Computer Science Engineering & Application,Sambalpur, Odissa, India
E-mail: [1]anupama.niranjan@gmail.com, [2]rvadali@gitam.edu, [3]sjena@suiit.ac.in

ABSTRACT

Data mining is the process of discovering hidden knowledge from the existing datasets. The process of knowledge discovery is a complex task when the data source is in the form of data streams and more tough when the data source is of class imbalance in nature. To find an optimal solution for these problems many research proposals are formulated by researchers. Some of the unsolved problems in the literature for the above said problem are for very large data sources of data streams with class imbalance nature. In this paper, a novel proposal for class imbalance large data streams is presented with novel techniques of oversampling and a unique multi modal filtering technique known as Multimodal Increment over Sampling for Data Streams (MIOSDS). The experimental simulations are conducted on three large datasets with different domains with high class imbalance ratio. The results generated are very impressive in terms of accuracy, AUC, precision, recall and F-measure validation metrics.

Keywords: *Knowledge Discovery, Data Streams, Imbalanced data, oversampling, Multimodal Increment Over Sampling for Data Streams (MIOSDS).*

1. INTRODUCTION

Data streams have become the part and parcel of day-to-day life with complex and huge internet systems used in different areas like security, finance, hospitals, e-commerce, networking, surveillance etc. The raw data collected from the real world has properties like noise, outliers, skewness, spread, redundancy and correlation etc, which are to be carefully handled by the machine learning algorithms. Class imbalance nature is such a property which can drastically degrade the performance of any classifier due to insufficient data for model building. This problem also arises in the real world scenario due to unavailability of instances for training and model building for prediction of unseen instances. Algorithms proposed for data streams have to work in dynamic environments where data is evolved in huge volumes and with a swift arrival rate, this means the data received is to be analyzed online as it is received. One of the most well-known and widely used online learning benchmarks for evolving data streams is Massive Online Analysis (MOA) [1]. The intrinsic strengths of MOA are in the implementation techniques incorporated for bi-directional data transfer from Waikato Environment for Knowledge Analysis (WEKA) [2].

A definite practical and thoughtful approach, with an insight of modern technology is required to solve the problem of data stream learning in real world applications. Furthermore, in this work, we extended the core of the recent approaches for solving the problem of skewed data stream learning. The proposed algorithm's improved empirical results of Multimodal Increment Over Sampling for Data Streams (MIOSDS) are presented with solid theoretical background.

The rest of this paper is organized as follows: Section 2 presents the recent works on data stream mining. Section 3 presents the main framework of the proposed MIOSDS algorithm. Section 4 provides a detailed explanation of the datasets and the evaluation criterias used in the experiments. Section 5 presents the algorithms used for comparison and experimental setup. Section 6 presents the detailed experimental results and discussion. Section 7 draws the conclusions and points out future research directions.

2. RELATED WORK

This section presents the most recent works in the research area of imbalance data learning and data stream learning in classification.

2.1. Literature Review On Imbalance Data Learning:

Alberto Fernández et al. [3] have reviewed the class imbalance big data problem and presented several challenges and solutions for different existing issues. Kapil K. Wankhade et al. [4] have presented different methods to handle concept drift in data stream learning. They have also presented different limitations for the existing methods and their proposed solutions. Z. Li et al. [5] have proposed a novel ensemble algorithm which dynamically updates the model for class imbalance dataset learning with concept drift. Xiangjun Li Li et al. [6] have proposed a new approach using cohesiveness and separation index for effective solution of concept drift in classification process. Eréndira Rendón et al. [7] have proposed a hybrid approach using heuristic sampling techniques for multi class big data imbalance datasets. They also considered the applicability of deep learning techniques on imbalanced big data sets.

Dariusz Brzezinski et al. [8] have investigated different class imbalance metrics for dynamic analysis of gradients, diverging and distribution of the datasets for varying class proportions. Gregory Ditzler et al. [9] have used replacing by smote technique and ensemble technique for minority class improvement. They also implemented weighted voting for gaining class imbalance adjusted accuracy. Rebeen Ali Hamad et al. [10] have implemented a solution for class imbalance for smart home data using deep neural network learning. The temporal video technique is used for learning minority data more sensitively.

T. Ryan Hoens et al. [11] have presented different challenges in class imbalance especially in data drift scenarios and presented some common solutions. Bartosz Krawczyk [12] have discussed wide areas of class imbalance learning such as classification, regression and clustering and also reviewed proposed data level and algorithmic level for effective knowledge discovery. Lei Zhu et al. [13] have proposed an incremental version of LPSVM for class imbalance learning using matrix coefficient updating technique. The learning process is also aimed for loss less learning for the updated information for the model.

Yang Lu et al. [14] have proposed an online data stream chunk based learning for minority weighted technique and specific data size selection. The proposed approach is well suitable for concept drift and least memory utilization with incremental building of the model. Rafiq Ahmed Mohammed et al. [15] have proposed an auto balancing framework for class imbalance data using racing algorithms and incremental learning. The approach is applicable for incremental stream frameworks using static batch learning technique.

YANGE SUN et al. [16] have presented a two stage classifier for online data stream classification using window adjustment technique for concept drift. The proposed framework uses both cost sensitive information at feature selection and classification stages. Shuo Wang et al. [17] have presented a systematic literature review for class imbalance data streams for mitigating and adjusting the concept drift phenomenon. Hang Zhang et al. [18] have proposed an online learning strategy for ensemble learning for class imbalance data with uncertainty and semi-supervised techniques.

3. PROPOSED MULTIMODAL INCREMENT OVER SAMPLING FOR DATA STREAMS (MIOSDS) ALGORITHM

The basic definition for multi modal is for using multiple methods for different contexts. The proposed Multimodal approach built on different models is used for classification of data streams of class imbalance nature. The proposed system makes use of a multi modal approach for extraction of the main features for improved classification of data streams. The extraction of the features subset is also influenced by the strategy of feature to class association.

The proposed MIOSD approach performs classification of the data streams by considering several aspects. One of the aspects is on account of formation and merging of novel class instances from the income data streams. Here the C4.5 algorithm approach is used for identification of instances for merging into the classes. Another aspect is for performing oversampling of the minority sub class for reducing class imbalance in the forming data stream. Figure 1, presents the framework of MIOSDS.

The following sections discuss the phases in MIOSDS algorithm.

3.1. Multi Modal Inspection Of The Data Stream For Concept Drift:

In the initial phase, the chunks of data stream are collected and provided for the algorithmic processing. These algorithms should be able to handle two scenarios of the data stream. First, as the chunks of data streams arriving are in class imbalance in nature and there is a need for continuous model improvement in the minority class, second the arriving data composition can also change depending upon the time and the minority class can be changed as majority and vice versa. The incoming chunks of data stream is passed through different evaluators for detecting the class imbalance nature using a simple technique of continuous summation and evaluation. The class imbalance nature or drift detection can be done by finding the percentage of classes and evaluating the changeover of class from majority to minority or vice-versa. The following issues of drift detection, noisy or outlier's detection and removal, detecting and replacing missing values, selecting the appropriate feature subset, generating required synthetic instances for reducing class imbalance nature are efficiently handled by the proposed framework.

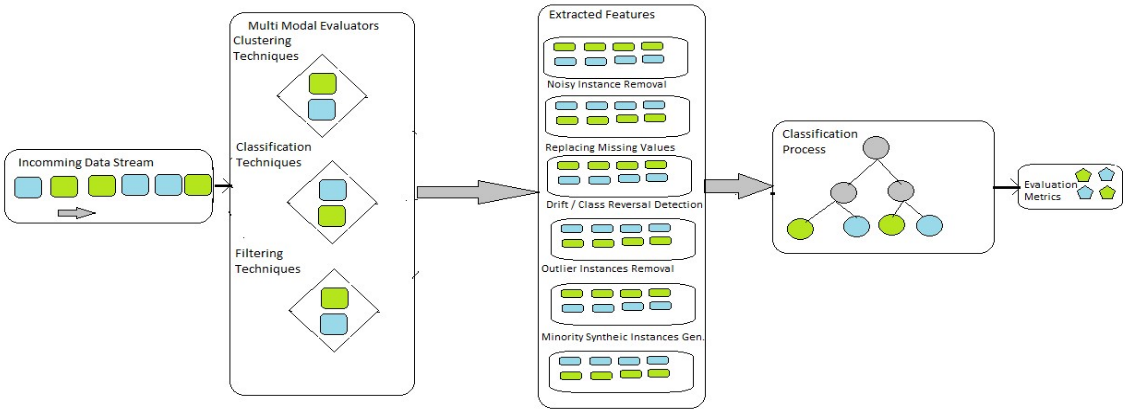


Figure 1 :Framework of Multimodal Increment Over Sampling for Data Streams (MIOSDS)

3.2. Multi Modal Approach For Preprocessing Stage For Removing Noisy And Missing Values:

The basic definition of the pre-processing is to analyze the dataset for the preliminary level, for finding the deviations from standards. The two main issues are regarding noisy and missing value instances. The multi modal approach is used for detection and elimination of noisy and missing values. The multi modal approach will be effective in terms of effective up gradation of the data stream. The technique used for noisy values detection is by investigating the deviation of instances from the class margins, either by projection technique or intra class distance measure. In our proposed research, we have adapted both the techniques of class margin deviation using support vector machine projection technique and intra class distance margin using remove misclassified filter technique for detection and removal of noisy or outlier instances.

3.3. Multi Modal Identifying Influential Feature Subset:

The process for identifying influential feature subset is one of the factors for better performance of the knowledge discovery process. The multi modal approach of filter and wrapper techniques can work effectively for selection of the best influential features subset. The features which are common in multi modal approach are selected and an improved data stream is prepared. In our work, the features are selected using correlation between features and class by passing the incoming data stream through correlation based feature subset selection. This technique is implemented by analyzing the individual predictive ability of features with the level of redundancy between the features. The searching of total features for discovering a novel subset is performed using greedy hill climbing technique with backtracking. The level of backtracking is limited to utmost 5 non performing attribute selection in the subset.

3.4. Multi modal Incremental oversampling on minority subset:

The oversampling of the minority samples in the data source will definitely reduce the problem of class imbalance nature in the data stream. The incremental and multi-modal approach is applied for synthetic instance generation, duplicating existing instances and creating hybrid instances from two or more instance features. A specific strategy can be implemented to find the level of oversampling and should not cross from the threshold level in every chunk of data streams. The synthetic minority over sampling instance generation technique is used for generating pure synthetic instances. The synthetic oversampling is restricted to a maximum of hundred percent for all the datasets. The level of oversampling required for specific datasets is correlated with the imbalance ratio of the dataset.

3.5. Preparing improved data stream for effective knowledge discovery:

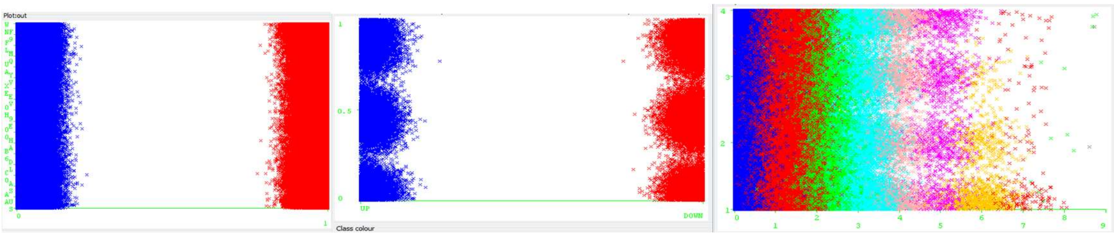
The minority chunk's of data stream which was prepared using the above phases is joined with majority chunks of data streams collected. The improved data source will be applied to the base algorithm and different evaluation metrics are generated. In our proposal, we have used C4.5 as our base learner due to the merits of easy generalization, better interpretability and compact model generation. The different evaluation criteria such as accuracy, AUC, precision, recall and f-measure are generated for model evaluation.

4. DATASETS AND EVALUATION CRITERIA'S

In the experiments, we run our proposed MIOSDS on three large data streams from MOA [19]*. Table 1 presents the details of the data stream used in the experimental study.

Table 1 : Details of the Data Stream from large datasets

S No	Data Set	Instances	Majority	Minority	Ir
1	Airline	539383	240264	219119	1.09
2	ElecNormNew	45312	26075	19237	1.35
3	Poker-lsn	829201	415526	17541	23.06



ElecNormNew Classes Airline Classes Poker-lsn Classes

Figure 2: The summary of classes of ElecNormNew, Airline and Poker-Isn datasets.

The different properties of datasets such as serial number, name of the dataset, total number of instances, number of majority instances, number of minority instances and imbalance ratio are given below in Table 1. Table 2 presents the detailed description of the data stream formed for 't' time. In this, we considered the maximum time 't' as '10' chunks of arrival for total instance. Figure 2 shows the summary of the classes of ElecNormNew, Airline and Poker-Isn datasets.

In the same way, the data stream chunks arrive and total instances to be processed and the imbalance ratio varies for each and every dataset depending upon the total instances. The complete details of all the datasets are given in Table 2.

Table 2: Number of samples for each class (indicated by indices 0–10) for the three inspection data streaming sets from the datasets.

Data	1	2	3	4	5	6	7	8	9	10	Sum
Airline	53939	53939	53939	53939	53939	53939	53939	53939	53939	53939	539383
ElecNorm New	4532	4532	4531	4531	4531	4531	4531	4531	4531	4531	45312
Poker-lsn	82921	82920	82920	82920	82920	82920	82920	82920	82920	82920	829201

Our goal is to explore stability of our algorithm on data stream learning datasets with machine learning techniques across different levels of imbalance. Moreover, we want to evaluate potential sources of bias in study design by constructing a number of experiments in which we diverse one parameter per experiment. To integrate conclusions obtained from each experiment a meta–analytic statistical analysis is proposed. These methods are suggested by a number of authors as tools for generalizing the results and integrating knowledge across many studies. The implementation of the proposed MIOSDS is done on WEKA [2] workbench with a system unit of i5-2410M CPU working on 2.30 GHz and 4.0 G RAM on Windows 7 operating system.

In this paper, we used AUC, Precision, Recall and F-measure for performance evaluation. We used the experimental dataset and stratified 10 fold cross validation evaluation methods which are used by the previous studies [20-24]. The 10 fold cross validation technique splits the data into 10 folds and in each run it uses 9 folds for training and 10th fold for testing. The process is repeated 10 times and in each run the testing data is replaced with untested fold.

The evaluation metrics used in the paper are detailed as follows, Let us revise a few well known and widely used measures: True Positives (TP) are actual positives, which are correctly predicted as positive by classification algorithm. True Negatives (TN) are the actual negatives which are correctly predicted as negative by classification algorithm. False Positives (FP) are actual negatives that are wrongly predicted as positives and False Negatives (FN) are actual positives wrongly predicted as negatives. TP rate can be defined as the ratio between true positive and total positive instances which is given in the equation (1). TN rate can be defined as the ratio between true negative and total negative instances which are given in the equation (2).

$$\text{TP}_{\text{RATE}} = \frac{TP}{POSITIVE} \tag{1}$$

$$\text{TN}_{\text{RATE}} = \frac{TN}{NEGATIVE} \tag{2}$$

Receiver Operating Characteristic (ROC) curve is the recent evaluation metric used for supervised learning dealing with imbalanced data study. This ROC curve can be used for projecting results depending upon the user perspective with different combinations of basic components such as true positives, false positives, true negatives and false negatives. The summary of the ROC curve can be given as the area under it, which is known as Area Under Curve (AUC). AUC can be computed simply as the micro average of TP rate and TN rate when only a single run is available from the classification algorithm. The Area Under Curve (AUC) measure is computed as given in the equation (3) and (4),

$$\text{AUC} = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \tag{3}$$

Or

$$\text{AUC} = \frac{TP_{RATE} + TN_{RATE}}{2} \tag{4}$$

The Precision measure is computed as given in the equation (5),

$$\text{precision} = \frac{TP}{(TP)+(FP)} \quad (5)$$

The Recall measure is computed as given in the equation (6),

$$\text{recall}=\frac{TP}{(TP)+(FN)} \quad (6)$$

The F-measure Value is computed as given in the equation (7),

$$\text{F-measure} = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

5. COMPARISON OF ALGORITHMS AND EXPERIMENTAL SETUP

We have conducted the comparative study using C4.5 [25] a classical benchmark decision tree algorithm and with a massive data stream learner Hoeffding tree [26]. The reason for choosing these two algorithms for comparison is to validate the MIOSDS approach in both terms of classical decision tree approach and data stream learning approaches. The details of the C4.5 and Hoeffding tree are given below.

C4.5 is a decision tree algorithm used to perform classification. C4.5 uses the training data to build the model for efficient classification. The model is built by splitting the data in a recursive manner till it reaches a leaf node. The criterion used for splitting the model is known as splitting criterion. C4.5 uses normalized information gain i.e. gain ratio as the splitting criterion. After the model is built, the unnecessary branches generated due to noisy, missing values etc are deleted. The technique used for removing unnecessary branches from the built decision tree is known as pruning technique. In C4.5 error based pruning technique is used. A Hoeffding tree is an efficient decision tree algorithm capable of classifying the instances from a huge data stream. The rationale behind the working of the Hoeffding tree is the mathematical principle of Hoeffding bound. A Hoeffding tree isn't able to address the data streams in which there is swift drift in the classes. In the experimental simulation, we considered the data stream which doesn't have drift in the classes so as to make a fair comparison with the proposed MIOSDS approach. One of the advantages of the Hoeffding tree is to learn from a small subspace of samples. The Hoeffding tree assures to generate an efficient output as equivalent to non-incremental learning algorithms.

6. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the proposed approach with the two benchmark comparative algorithms C4.5 and Hoeffding Tree. One of the two best known algorithms is chosen for comparison so as to expose the strengths and the weakness of the proposed approach on different evaluation criterias.

The experiments are evaluated by the measures: accuracy, AUC, precision, recall and f-measure with formulae stated in section 4. Each value in the table is split into two parts: mean and standard deviation values. For example, consider the first value in Table 3, 64.39±0.60 of accuracy for the Airline dataset using the C4.5 algorithm. The 64.39 is the mean value and 0.60 is the standard deviation value for 10 runs of 10 fold cross validation.

In the first row of Table 3, the accuracy values for each of the 10 chunks of Airline dataset are 64.39±0.60, 0 64.28±0.52, 64.36±0.52, 64.15±0.57, 64.38±0.52, 64.29±0.58, 64.29±0.58, 64.24±0.53, 64.31±0.48, 64.19±0.54 and average value is 64.28±0.54 which is also followed with ● indicating the value is less than MIOSDS algorithm**.** Here three average values of C4.5, Hoeffding Tree and MIOSDS are 64.28±0.54, 62.16±0.65, 98.38±0.16 and MIOSDS value is better than C4.5 and Hoeffding Tree which is bold faced indicating the best value in the row. In general, the algorithm that produces higher measure values is better than the other algorithm. The results show that our proposed MIOSDS approach performs better than both C4.5 and Hoeffding Tree algorithms on all the three data stream datasets from MOA. The graphical representation of results for the tables 3, 5, 6 and 7 are presented below the corresponding tables for better understanding of the readers as figures 3,4,5,6,7.

*(https://moa.cms.waikato.ac.nz/datasets/).

Table 3 :Normalized Accuracy (in %) of all classifier variants in case of three data streams from large datasets

data stream #	Airline#			elecNormNew#			Poker #		
	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS
1	64.39 ±0.60	61.98 ±0.69	90.17 ±0.40	79.35 ±1.72	74.98 ±1.99	93.77 ±1.02	99.11 ±0.14	62.14 ±1.99	**80.00 ±0.35**
2	64.28 ±0.52	62.13 ±0.63	87.79 ±0.40	79.02 ±1.94	74.42 ±2.31	87.91 ±1.41	99.11 ±0.12	62.03 ±2.16	**79.96 ±0.29**
3	64.36 ±0.52	62.11 ±0.72	89.20 ±0.42	79.37 ±1.81	74.72 ±3.91	87.12 ±1.39	99.10 ±0.14	62.83 ±2.09	**79.78 ±0.36**
4	64.15 ±0.57	62.05 ±0.62	87.01 ±0.47	80.00 ±1.86	73.66 ±3.32	88.82 ±1.14	99.16 ±0.12	62.78 ±2.61	**79.74 ±0.35**
5	64.38 ±0.52	62.22 ±0.67	88.56 ±0.49	78.59 ±2.37	73.92 ±4.79	84.46 ±1.44	99.15 ±0.11	62.76 ±2.00	**79.82 ±0.35**
6	64.29 ±0.58	62.36 ±0.64	**89.60 ±0.36**	79.20 ±2.12	73.66 ±2.67	84.05 ±1.67	99.14 ±0.11	62.81 ±1.68	79.82 ±0.45
7	64.29 ±0.58	62.51 ±0.67	**91.78 ±0.28**	79.17 ±1.77	74.63 ±2.05	84.97 ±1.98	99.12 ±0.11	63.17 ±1.17	79.86 ±0.37
8	64.24 ±0.53	62.21 ±0.71	**88.49 ±0.43**	79.86 ±1.60	74.81 ±2.07	92.20 ±1.24	99.10 ±0.17	63.00 ±1.48	79.88 ±0.37
9	64.31 ±0.48	61.99 ±0.67	**89.39 ±0.43**	78.68 ±1.76	74.46 ±3.94	90.43 ±1.24	99.13 ±0.15	63.01 ±1.58	79.88 ±0.31
10	64.19 ±0.54	62.11 ±0.62	**89.18 ±0.44**	78.68 ±1.79	74.16 ±1.93	86.22 ±1.50	99.16 ±0.12	62.56 ±1.95	79.88 ±0.35
Average	64.28 ±0.54●	62.16 ±0.65●	**89.12 ±0.16**	79.19 ±1.82●	74.34 ±2.33●	**88.65 ±0.83**	99.12 ±0.14●	62.07 ±0.86●	**99.56 ±0.17**

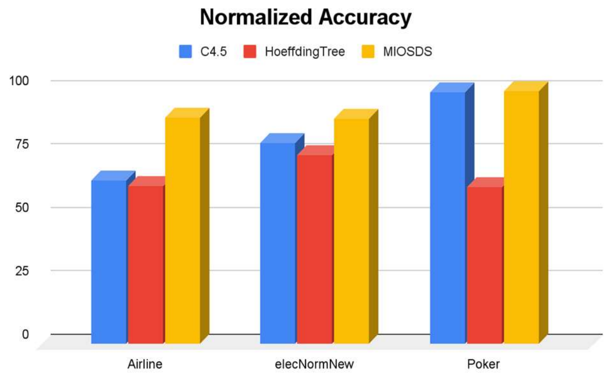


Figure 3: The summary of experimental analysis using Bar plots of average Normalized Accuracy.

Table 4 : Normalized AUC (in %) of all classifier variants in case of three data streams from large datasets;

data stream #	Airline#			elecNormNew#			Poker #		
	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS
1	0.653 ±0.008	0.636 ±0.008	0.967 ±0.002	0.824 ±0.021	0.731 ±0.036	0.962 ±0.010	0.998 ±0.000	0.758 ±0.047	**0.857 ±0.004**
2	0.653 ±0.008	0.636 ±0.008	0.955 ±0.004	0.814 ±0.024	0.735 ±0.040	0.941 ±0.011	0.998 ±0.000	0.756 ±0.051	**0.859 ±0.004**
3	0.659 ±0.008	0.637 ±0.008	0.963 ±0.002	0.821 ±0.019	0.746 ±0.046	0.946 ±0.010	0.998 ±0.000	0.765 ±0.046	**0.857 ±0.004**
4	0.651 ±0.009	0.636 ±0.007	0.943 ±0.004	0.829 ±0.021	0.737 ±0.055	0.928 ±0.012	0.998 ±0.000	0.770 ±0.050	**0.858 ±0.004**
5	0.657 ±0.008	0.639 ±0.008	0.961 ±0.003	0.802 ±0.026	0.741 ±0.060	0.902 ±0.013	0.998 ±0.000	0.775 ±0.035	**0.857 ±0.004**
6	0.650 ±0.009	0.640 ±0.007	**0.966 ±0.002**	0.811 ±0.025	0.739 ±0.052	0.822 ±0.020	0.998 ±0.000	0.769 ±0.043	0.858 ±0.005
7	0.653 ±0.007	0.641 ±0.008	**0.953 ±0.003**	0.809 ±0.021	0.743 ±0.045	0.879 ±0.031	0.998 ±0.000	0.779 ±0.027	0.859 ±0.004
8	0.652 ±0.007	0.640 ±0.008	**0.961 ±0.002**	0.827 ±0.022	0.750 ±0.036	0.961 ±0.011	0.998 ±0.000	0.774 ±0.037	0.857 ±0.004
9	0.653 ±0.006	0.636 ±0.009	**0.968 ±0.002**	0.812 ±0.021	0.746 ±0.046	0.960 ±0.009	0.998 ±0.000	0.779 ±0.034	0.858 ±0.003
10	0.652 ±0.008	0.637 ±0.008	**0.967 ±0.002**	0.814 ±0.023	0.724 ±0.035	0.846 ±0.022	0.998 ±0.000	0.763 ±0.043	0.859 ±0.004
Average	0.653 ±0.008●	0.637 ±0.007●	**0.995 ±0.001**	0.816 ±0.023●	0.739 ±0.048●	**0.964 ±0.011**	0.964 ±0.011●	0.768 ±0.043●	**1.000 ±0.000**

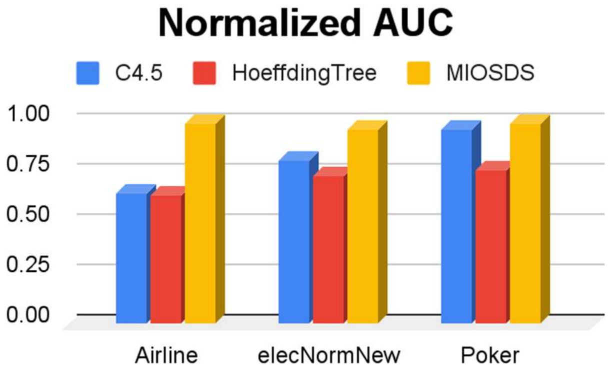


Figure 4: The summary of experimental analysis using Bar plots of average Normalized AUC.

Table 5 :Normalized Precision (in %) of all classifier variants in case of three data streams from large datasets

data stream #	Airline#			elecNormNew#			Poker #		
	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS
1	0.639 ±0.005	0.632 ±0.006	0.981 ±0.006	0.767 ±0.027	0.809 ±0.061	0.944 ±0.013	0.996 ±0.000	0.644 ±0.027	**0.759 ±0.004**
2	0.637 ±0.004	0.633 ±0.004	0.962 ±0.011	0.771 ±0.029	0.776 ±0.068	0.900 ±0.020	0.996 ±0.000	0.644 ±0.028	**0.759 ±0.003**
3	0.639 ±0.005	0.634 ±0.006	0.964 ±0.009	0.765 ±0.028	0.774 ±0.073	0.885 ±0.021	0.996 ±0.000	0.647 ±0.026	**0.758 ±0.004**
4	0.637 ±0.006	0.632 ±0.005	0.943 ±0.008	0.772 ±0.026	0.770 ±0.092	0.908 ±0.013	0.996 ±0.000	0.651 ±0.026	**0.758 ±0.004**
5	0.640 ±0.004	0.635 ±0.005	0.958 ±0.006	0.752 ±0.029	0.772 ±0.091	0.879 ±0.017	0.996 ±0.000	0.652 ±0.019	**0.758 ±0.004**
6	0.637 ±0.005	0.635 ±0.005	0.973 ±0.005	0.759 ±0.028	0.754 ±0.063	0.946 ±0.024	0.996 ±0.000	0.648 ±0.024	0.758 ±0.005
7	0.639 ±0.004	0.636 ±0.005	0.921 ±0.003	0.767 ±0.025	0.770 ±0.053	0.872 ±0.034	0.996 ±0.000	0.654 ±0.017	0.759 ±0.004
8	0.636 ±0.004	0.634 ±0.005	0.962 ±0.008	0.762 ±0.023	0.759 ±0.045	0.941 ±0.016	0.996 ±0.000	0.652 ±0.021	0.760 ±0.004
9	0.637 ±0.004	0.633 ±0.006	0.962 ±0.006	0.745 ±0.025	0.771 ±0.072	0.917 ±0.017	0.996 ±0.000	0.652 ±0.019	0.759 ±0.003
10	0.639 ±0.004	0.633 ±0.005	0.972 ±0.009	0.761 ±0.027	0.793 ±0.062	0.942 ±0.021	0.996 ±0.000	0.647 ±0.025	0.760 ±0.004
Average	0.638 ±0.005●	0.632 ±0.006●	0.960 ±0.003	0.765 ±0.027●	0.771 ±0.078●	0.913 ±0.018	0.996 ±0.000●	0.649 ±0.000●	**1.000 ±0.000**

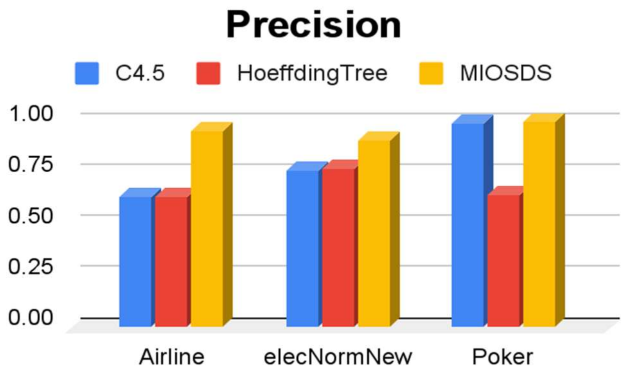


Figure 5: The summary of experimental analysis using Bar plots of average Normalized Precision.

Table 6 :Normalized Recall (in %) of all classifier variants in case of three data streams from large datasets

data stream #	Airline#			elecNormNew#			Poker #		
	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS
1	0.822 ±0.010	0.752 ±0.015	0.835 ±0.008	0.739 ±0.034	0.550 ±0.087	0.948 ±0.013	1.000 ±0.000	0.834± 0.058	**1.000 ±0.000**
2	0.827 ±0.010	0.756 ±0.015	0.799 ±0.011	0.722 ±0.036	0.576 ±0.078	0.885 ±0.021	1.000 ±0.000	0.833 ±0.068	**1.000 ±0.000**
3	0.820 ±0.016	0.750 ±0.014	0.833 ±0.011	0.744 ±0.034	0.598 ±0.090	0.884 ±0.022	1.000 ±0.000	0.847 ±0.055	**0.999 ±0.001**
4	0.820 ±0.017	0.754 ±0.015	0.794 ±0.009	0.752 ±0.036	0.579 ±0.110	0.896 ±0.019	1.000 ±0.000	0.826 ±0.086	**1.000 ±0.000**
5	0.816 ±0.012	0.750 ±0.013	0.818 ±0.010	0.741 ±0.042	0.592 ±0.125	0.850 ±0.025	1.000 ±0.000	0.829 ±0.049	**1.000 ±0.000**
6	0.826 ±0.010	0.754 ±0.014	0.830 ±0.007	0.749 ±0.041	0.581 ±0.089	0.651 ±0.035	1.000 ±0.000	0.840 ±0.050	**1.000 ±0.000**
7	0.824 ±0.011	0.757 ±0.016	0.984 ±0.003	0.734 ±0.034	0.584 ±0.074	0.727 ±0.039	1.000 ±0.000	0.833 ±0.035	1.000 ±0.000
8	0.831 ±0.008	0.752 ±0.013	0.821 ±0.009	0.766 ±0.035	0.604 ±0.063	0.927 ±0.017	1.000 ±0.000	0.837 ±0.044	0.999 ±0.000
9	0.828 ±0.010	0.750 ±0.012	0.839 ±0.008	0.758 ±0.036	0.591 ±0.081	0.922 ±0.020	1.000 ±0.000	0.836 ±0.046	1.000 ±0.000
10	0.816 ±0.010	0.754 ±0.011	0.826 ±0.010	0.727 ±0.040	0.544 ±0.084	0.690 ±0.037	1.000 ±0.000	0.838 ±0.060	1.000 ±0.000
Average	0.823 ±0.012●	0.752 ±0.014●	0.838 ±0.002	0.743 ±0.035●	0.579 ±0.095●	**0.838 ±0.017**	1.000 ±0.000	0.835 ±0.073●	1.000 ±0.000

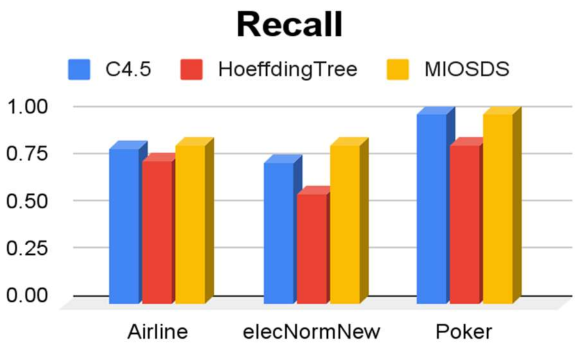


Figure 6: The summary of experimental analysis using Bar plots of average Normalized Recall

Table 7 :Normalized F-measure (in %) of all classifier variants in case of three data streams from large datasets

data stream #	Airline#			elecNormNew#			Poker #		
	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS	C4.5	HoeffdingTree	MIOSDS
1	0.719 ±0.005	0.687 ±0.007	0.902 ±0.004	0.752 ±0.021	0.647 ±0.053	0.945 ±0.009	0.998 ±0.000	0.725 ±0.018	0.863 ±0.002
2	0.720 ±0.005	0.689 ±0.007	0.873 ±0.005	0.745±0.025	0.654 ±0.041	0.892 ±0.013	0.998 ±0.000	0.724 ±0.021	0.863 ±0.002
3	0.718 ±0.005	0.687 ±0.007	0.894 ±0.005	0.754 ±0.022	0.666 ±0.043	0.884 ±0.012	0.998 ±0.000	0.732 ±0.019	0.862 ±0.003
4	0.717 ±0.006	0.688 ±0.007	0.862 ±0.005	0.761 ±0.024	0.647 ±0.052	0.902 ±0.010	0.998 ±0.000	0.725 ±0.053	0.862 ±0.003
5	0.718 ±0.005	0.688 ±0.007	0.882 ±0.006	0.746 ±0.030	0.654 ±0.068	0.864 ±0.013	0.998 ±0.000	0.729 ±0.017	0.862 ±0.002
6	0.719 ±0.004	0.690 ±0.007	0.896 ±0.004	0.753 ±0.027	0.649 ±0.053	0.771 ±0.027	0.998 ±0.000	0.730 ±0.015	0.862 ±0.003
7	0.720 ±0.005	0.691 ±0.008	0.951 ±0.002	0.749 ±0.023	0.659 ±0.043	0.792 ±0.029	0.998 ±0.000	0.732 ±0.009	0.863 ±0.003
8	0.720 ±0.004	0.688 ±0.007	0.885 ±0.005	0.763 ±0.020	0.669 ±0.036	0.934 ±0.011	0.998 ±0.000	0.732 ±0.012	0.863 ±0.003
9	0.720 ±0.004	0.686 ±0.006	0.896 ±0.004	0.751 ±0.022	0.661 ±0.039	0.919 ±0.011	0.998 ±0.000	0.731 ±0.017	0.863 ±0.002
10	0.717 ±0.005	0.688 ±0.006	0.893 ±0.005	0.743 ±0.024	0.637 ±0.050	0.796 ±0.026	0.998 ±0.000	0.729 ±0.018	0.863 ±0.003
Average	0.719 ±0.005●	0.687 ±0.006●	**0.894 ±0.002**	0.751 ±0.022●	0.654 ±0.057●	0.870 ±0.014	0.998 ±0.000●	0.731 ±0.012●	**1.000 ±0.000**

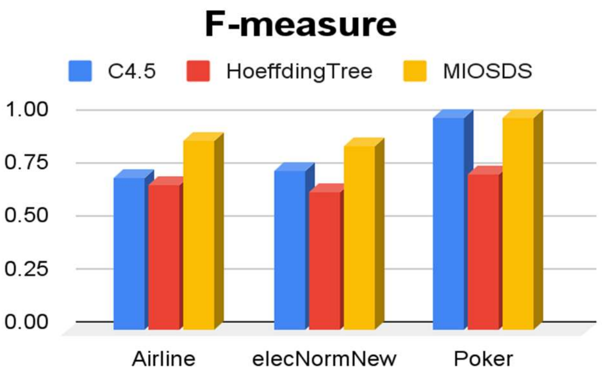


Figure 7: The summary of experimental analysis using Bar plots of average Normalized F-measure

Finally, we can conclude that MIOSDS approach is efficient to solve the real time issues of knowledge discovery for imbalance and noisy data streams. The experimental results clearly indicate that the MIOSDS approach performs better or similar on the data streams than the classical approaches.

7. CONCLUSION

In this paper, a novel Incremental Over Sampling algorithm for Data Stream using an efficient and unique oversampling strategy is proposed. MIOSDS approach oversamples the prominent and class oriented instances from the minority subset for better knowledge discovery from the data streams. The experimental validation is conducted in two different scenarios for UCI and MOA datasets of small, moderate, huge and very huge data streams. The experimental results suggest that the MIOSDS approach efficiently discovers knowledge from the imbalanced data streams. Some important future directions of this work are: i) to mitigate the solution for the problem of concept drift for imbalance data streams; ii) the present proposal's limitation is that it is only applicable for bi-class problems and are not applicable to multiclass cases; iii) modification of proposal for handling uncertain and noisy environments is one of the direction for the reconstruction of the future classifier.

REFERENCES:

[1]. A. Bifet and G. Holmes and R. Kirkby and B. Pfahringer, MOA: Massive Online Analysis, Journal of Machine Learning Research, vol. 11, pp. 1601--1604, 2010.

[2]. Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[3]. Alberto Fernández, Sara del Río, Nitesh V. Chawla, Francisco Herrera," An insight into imbalanced Big Data classification: outcomes and challenges", Complex Intell. Syst. DOI 10.1007/s40747-017-0037-9.

[4]. Kapil K. Wankhade, Snehlata S. Dongre, Kalpana C. Jondhale," Data stream classification: a review", Iran Journal of Computer Science, https://doi.org/10.1007/s42044-020-00061-3.

[5]. Z. Li, W. Huang, Y. Xiong et al., Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm, Knowledge-Based Systems (2020), doi: https://doi.org/10.1016/j.knosys.2020.105694.

[6]. Xiangjun Li ,Yong Zhou , Ziyan Jin , Peng Yu, and Shun Zhou," A Classification and Novel Class Detection Algorithm for Concept Drift Data Stream Based on the Cohesiveness and Separation Index of Mahalanobis Distance", Journal of Electrical and Computer Engineering Volume 2020, Article ID 4027423, 8 pages, https://doi.org/10.1155/2020/4027423

[7]. Eréndira Rendón, Roberto Alejo, Carlos Castorena, Frank J. Isidro-Ortega and Everardo E. Granda-Gutiérrez," Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem", Appl. Sci. 2020, 10, 1276; doi:10.3390/app10041276.

[8]. Dariusz Brzezinski , Jerzy Stefanowski , Robert Susmaga, and Izabela Szcz¸ech "On the Dynamics of Classification Measures for Imbalanced and Streaming Data", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,2019.

[9]. Gregory Ditzler, and Robi Polikar " Incremental Learning of Concept Drift from Streaming Imbalanced Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2012.

[10]. Rebeen Ali Hamad · Masashi Kimura2 · Jens Lundström3," Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments", SN Computer Science (2020) 1:204 https://doi.org/10.1007/s42979-020-00211-1.

[11]. T. Ryan Hoens · Robi Polikar · Nitesh V. Chawla," Learning from streaming data with concept drift and imbalance: an overview", Prog Artif Intell (2012) 1:89–101, DOI 10.1007/s13748-011-0008-0.

[12]. Bartosz Krawczyk," Learning from imbalanced data: open challenges and future directions", Prog Artif Intell (2016) 5:221–232, DOI 10.1007/s13748-016-0094-0.

[13]. Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh," Class Imbalance Robust Incremental LPSVM for Data Streams Learning", WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15,2012 - Brisbane, Australia.

[14]. Yang Lu , Yiu-Ming Cheung , and Yuan Yan Tang," Adaptive Chunk-Based Dynamic Weighted Majority for Imbalanced Data Streams With Concept Drift", IEEE

TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,2019.

[15]. Rafiq Ahmed Mohammed, Kok-Wai Wong, Mohd Fairuz Shiratuddin, Xuequn Wang," PWIDB: A framework for learning to classify imbalanced data streams with incremental data re-balancing technique ", Procedia Computer Science 176 (2020) 818–827.

[16]. YANGE SUN, YI SUN, AND HONGHUA DAI," Two-Stage Cost-Sensitive Learning for Data Streams With Concept Drift and Class Imbalance", *10.1109/ACCESS.2020.3031603.*

[17]. Shuo Wang , Leandro L. Minku, and Xin Yao*,"* A Systematic Study of Online Class Imbalance Learning With Concept Drift*",* IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

[18]. Hang Zhang, Weike Liu, and Qingbao Liu," Reinforcement Online Active Learning Ensemble for Drifting Imbalanced Data Streams", DOI 10.1109/TKDE.2020.3026196, IEEE Transactions on Knowledge and Data Engineering.

[19]. HamiltonA. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* ±School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[20]. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.

[21]. Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Jesse Read, Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl," MOA: A Real-Time Analytics Open Source Framework ",Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2011: Machine Learning and Knowledge Discovery in Databases pp 617-620.

[22]. A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, MOA: massive online analysis, J.Mach. Learn. Res. 11 (2010) 1601–1604.

[23]. Edwin Lughofer, Eva Weigl, Wolfgang Heidl, Christian Eitzinger, Thomas Radauer ," Integrating new classes on the fly in evolving fuzzy classifier designs and their application in visual inspection ",Applied Soft Computing 35 (2015) 558–582

[24]. Edwin Lughofer, Eva Weigl, Wolfgang Heidl, Christian Eitzinger, Thomas Radauer,"Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelle d instances", Information Sciences 355–356 (2016) 127–151.

[25]. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[26]. Geoff Hulten, Laurie Spencer, Pedro Domingos: Mining time-changing data streams. In: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 97-106, 2001.