Reviewer #1, Concern # 1:

According to the method description in Section 3, the ensemble framework and all algorithms are designed for multiclass (L classes) classification problems. So, in experiments, why you choose to convert multiclass data streams into binary ones, rather than deal with them directly? What is the difference between these two approaches?

Author response:

Thanks for pointing out this. In the previous manuscript, the multi-class real-world data streams were concerted into the binary class streams by selecting one category as the majority class and another category as the minority class, which is the same as the approach in ref. [6], [10]. This method is mainly based on the following considerations

1. In the class imbalance learning problem, the classification performance on the minority class usually attracts more attention. And most of the research works on the joint problem of concept drift and class imbalance are focusing on the binary problem. Moreover, we applied PAUC as the evaluation index, which is designed as an overall performance measure for online drifting imbalance scenarios. And it can only be used in binary imbalance condition. In addition, the comparative method OOB and UOB are designed for binary problem. Therefore, in the manuscript, the classification performance comparison of ROALE-DI with other algorithms are carried out on binary data streams.

2. In the manuscript, we select the common-used real-world data streams for experimental comparison, GMSC, PAKDD, Covtype and Poker. And Covtype and Poker are multi-class imbalance streams which are recommended by the MOA official. The original class distribution of Covtype and Poker are as follows:

Table 1

Class distribution of the real-world data stream Covtype

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Total**
num instances	211840	283301	35754	2747	9493	17367	20510	581012
Percentage	36.46%	48.76%	6.15%	0.47%	1.63%	2.99%	3.53%	100%

Table 2

Class distribution of the real-world data stream Poker

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**	**Class 7**	**Class 8**	**Class 9**	**Class 10**	**Total**
num instances	415526	350426	39432	17541	3225	4657	1180	195	11	2	829201
Percentage	50.11%	42.26%	4.76%	2.11%	0.39%	0.56%	0.14%	0.024%	0.0013%	0.0002%	100.%

It can be seen from the Table 1 and Table 2 that in these two real-world data streams, the classes with the smallest number of instances only have a little percentage. Especially in the Poker, the class 10 only has 2 instances and when it compared with class 1, the class imbalance ratio is 207763:1. It hard for an active learning model to correctly find and classify such minority class instances. By selecting specific categories, the class imbalance ratio of the experimental real-world data streams could be set, which can make the experimental results more reliable. (The class imbalance ratio of the experimental data streams in this manuscript is refer to imbalance ratio setting in Wang Shuo's review article [6]). According to the class imbalance ratio and the length of the manuscript, we only showed the classification performance on covtype36, covtype46, poker23 and poker35. Results on other data streams that are generated by converting Covtype and Poker are also provided in the supplement files [Link: https://github.com/saferhand/ROALE-DI].

3. In the previous manuscript, we did not fully explain why the experiment focused on the binary imbalance problem and why should the multi-class real-world data streams be converted in that way. So, we modified the description in the manuscript to solve these problems.

4. The proposed ROALE-DI can address the multi-class imbalance real-world data streams. Therefore, we make the comparative experiments on original real-world data streams Covtype and Poker. In this experiment, the recall is used as the indicator for evaluating the classification performance. Table 3 and Table 4 shows the classification results of recall. The comparative algorithms include all the methods in the manuscript except OOB and UOB.

   As shown in Table 3, ROALE-DI achieves the best recall value than other active learning methods on class 3,4,5, and 7. OALE performs best on class 1 and 2, which are two majority classes. Although OAL-DI has the best recall result on class 6, it consumes 7.5% more real labels than ROALE-DI and OALE. From the perspective of the average recall result, ROALE-DI is the best active learning methods and BOLE is the best supervised methods.

Table 3

Classification results on original Covtype of recall (%)

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Ave.Recall**	**labels**
ROALE-DI	82.5	88.9	**89.7**	**84.0**	**74.3**	68.1	**89.9**	**82.5**	**15.6**
OALE	**83.4**	**90.9**	89.1	78.2	54.6	65.8	83.9	78.0	**15.6**
OAL-DI	78.9	86.1	86.7	76.3	59.0	**68.5**	80.6	76.6	23.1
LB	91.2	94.2	90.1	**83.0**	75.9	71.7	**91.4**	85.3	-
ARF	91.2	**95.3**	90.9	76.5	57.8	66.7	83.4	80.3	-
OAUE	89.2	92.5	88.0	84.3	**76.5**	70.5	**91.4**	84.6	-
BOLE	**93.4**	94.1	**91.4**	79.2	70.4	**83.7**	86.9	**85.6**	-

   As shown in Table 4, ROALE-DI achieves the best recall on class 1, 5, 7, 9, and 10. In addition, ROALE-DI has better average recall than OALE with less real labels. Although the average recall of ROALE-DI is 0.2% lower than that of OAL-DI, it uses 7.4% fewer real labels. In addition, BOLE is the best supervised algorithm. It is worth mentioning that the active learning algorithm obtained recall results on class 9 and class 10. And other supervised algorithms did not achieve results.

Table 4

Classification results on original Poker of recall (%)

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Class8**	**Class9**	**Class10**	**Ave.Recall**	**labels**
ROALE-DI	**90.9**	70.0	21.5	49.7	**67.8**	18.7	**16.1**	26.7	**4.5**	**20.0**	38.6	**17.7**
OALE	88.2	**72.5**	16.9	41.5	50.9	7.4	12.0	16.4	3.6	0.0	30.9	18.7
OAL-DI	86.8	68.4	**24.0**	**53.7**	62.7	**22.3**	15.9	**40.4**	3.6	10.0	**38.8**	25.1
LB	**95.5**	**85.7**	36.7	**67.3**	78.9	15.9	17.6	37.9	0.0	0.0	43.5	-
ARF	94.1	74.0	16.4	43.5	7.8	**40.0**	3.8	29.7	0.0	0.0	30.9	-
OAUE	91.9	76.6	31.3	59.3	**79.7**	12.4	**24.5**	**42.1**	0.0	0.0	41.8	-
BOLE	87.4	78.7	**51.0**	60.4	75.9	38.4	24.0	41.5	0.0	0.0	**45.7**	-

   From the experiments on the multi-class real-world data streams, ROALE-DI can also achieve good performance on the multi-class data streams. Considering that the main problem of our research is about the classification

performance of majority and minority classes in the binary imbalance condition, we intended to provide this part as the supplement materials for the manuscript.

5. Thanks again for pointing out this problem. In the process of thinking about the problem, we found many problems in the multi-class imbalanced data streams. For example, in multi-class imbalance streams, the majority class and the minority class are hard to clearly define. One class can be the minority class when compared with a lager class and be the majority when compared with a smaller class. It may influence the class imbalance handing mechanism. In addition, finding the classes that have only few instances is also a daunting task.

   We think that the multi-class imbalance drifting problem worth further study. And we have added some ideas inspired by your opinions to the conclusion and future work section.

Author action:

1. In page 2 5th paragraph of section 1, we modified the manuscript to explain why the paper is mainly focused on the binary imbalance problem.

   "As the class imbalance learning aims at the classification performance on the minority class, and most of the research works on the joint problem of concept drift and class imbalance are focusing on the binary problem. This paper also focuses on binary classification problems that have two classes, namely, the minority and majority."

2. In page 9 5th paragraph section 4.1, we modified the manuscript to show why and how we converted the multi-class real-world streams to binary streams. And we also provided the additional experiments results in footnote 1.

   "As most of existing research for drifting imbalance streams are only for binary condition, in this paper, the experiments are also focusing on the binary streams. Therefore, the multiclass real-world data streams are converted into a binary data stream using the same method in [1], which selects one class as the minority and another class as the majority. In this approach, the class imbalance ratio of the real-world streams can be set and selected more appropriately. Covtype have 7 classes and the covtype36 selects class 3 and class 7. And Poker has 10 classes. In general, covtype36, covtype46, poker23 and poker35 are selected in the experimental comparative according to the class imbalance ratio. And results on other streams generated from Covtype and Poker can be found at 1."

   "FootNote1: ROALE-DI can also address multi-class imbalance streams. Results on the original multi-class real-world data streams Covtype and Poker can be found in the website: https://github.com/saferhand/ROALE-DI. And results on other generated binary real-world data stream are also available."

3. In page 13 conclusion, we modified the manuscript to show that there are more content worthy of research in multi-class problems and expand future research.

   "In future work, we intend to further study the drifting multi-class imbalance problem where the size relationship between categories will be relative and variable."

Reviewer#1, Concern # 2:

Training instances are selected randomly from the data block. I am wondering whether the training set can be representative enough, as its class distribution and consequently calculated parameters (such as DCIR) will be different from those of original data.

Author response:

Thanks for your pointing out whether the class imbalance ratio estimate is accurate.

Estimating the class imbalance ratio in the active learning condition is a difficult task as the learning model cannot get the full view of the real labels. In addition, the random active learning labeling strategy unbiasedly selects instances from the data stream, which can provide a data stream subset which is closed to the real distribution.

In this manuscript, we combined the random active labeling strategy with the class imbalance ratio calculation process. Therefore, the Damped Class Imbalance Ratio (DCIR) is proposed to reflect the real-time class imbalance ratio. As the class imbalance ratio in the data streams could change over-time, the old class imbalance information could be outdated. So, we design a time-decayed factor to eliminate the influence of historical information.

One factor that affects statistical accuracy is the selection ratio of random strategies. Although increasing the random strategy ratio can increase the number of selected samples, thereby improving the statistical accuracy of class imbalance ratio. However, active learning is to obtain the highest performance at the smallest label cost. Therefore, we also control the label ratio of the random strategy as much as possible.

Table 5

The class imbalance varying condition of the experimental data streams

Dataset	No. Inst	Class ratio	Type	Drifts	Drift Occurs Postion
Agrawal$_{RS}$	100 k	3/7,2/8,1/9,3/7	sudden	3	1/4; 2/4; 3/4
Sine$_{RS}$	100 k	3/7,2/8,1/9,3/7	sudden	3	1/4; 2/4; 3/4
HYP$_{RG}$	100 k	1/1,1/9	gradual	1	The entire process

To verify the estimation accuracy of DCIR for class imbalance ratio, we carried out the following experiments. First, we modified the source code of ROALE-DI to output the DCIR value of each classes. Then, the experiments are carried out on three synthetic data streams with varying class imbalance ratio Agrawal$_{RS}$, Sine$_{RS}$ and HYP$_{RS}$. Table 5 shows the class imbalance varying condition of the data streams. The class imbalance ratio of Agrawal$_{RS}$ and Sine$_{RS}$ are varying from 3/7, 2/8, 1/9 to 3/7 at the position of the 1/4, 2/4, and 3/4 of the data stream, and the changes are completed within 25 instances. Then, the class imbalance ratio of HYP$_{RS}$ gradually changes from 1/1 to 1/9 in the entire streams. And the parameters setting of ROALE-DI is the same with the default setting in the comparative experiments (Section 4.3). ROALE-DI conducted five parallel experiments on each data streams and averaged the results. Figure 1 shows the Two-dimensional area chart of the DCIR results.

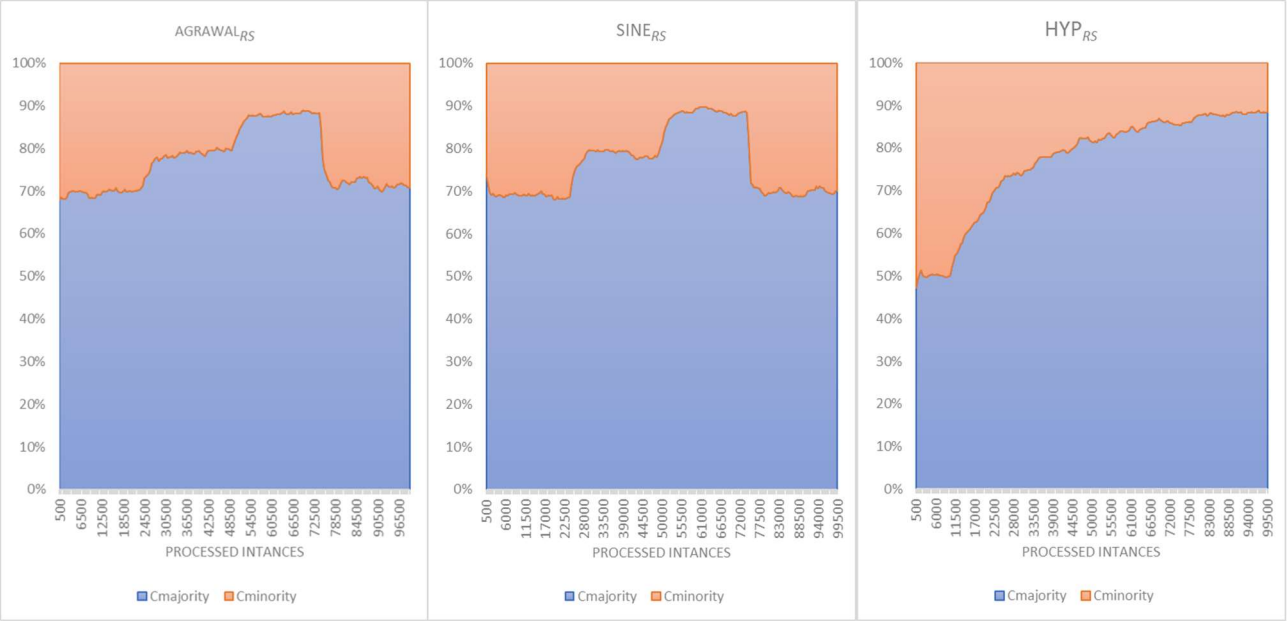

Figure 1 Two-dimensional area chart of the DCIR results

As shown in Figure 1, the DCIR two-dimensional area chart shows the class size relationship of the majority class and the minority class. As for Agrawal$_{RS}$ and Sine$_{RS}$, the DCIR area chart is divided into four stages, and the corresponding class imbalance ratios are 3/7, 2/8, and 1/9. Between each stage, the class imbalance ratio changes rapidly. As for the HYP$_{RS}$, the DCIR area chart gradually from 50% to 90%, which shows the class imbalance ratio of HYP$_{RS}$ is change from 1/1 to 1/9. Therefore, it can be conducted that the DCIR value can provide the class imbalance ratio information for the learning model.

Author action:

1. In page 9 section 4.1, we provided the experimental of DCIR accuracy as supplementary material. [Link: https://github.com/saferhand/ROALE-DI].

Reviewer#1, Concern # 3:

It seems that when a minority class appears rarely, it will have less effect on the ensemble model. On one hand, it is hard to be randomly selected to train the ensemble classifier. On the other hand, it will take quite a long time for the labeled instances buffer to collect enough instances of this class. Is there any mechanism to cope with this situation?

Author response:

Thanks for pointing out this.

As for the class imbalance data stream, the learning model could lack the training instances of the minority class. And the classification capacity on the minority class could be weak. Therefore, we designed a labeled instances buffer to store instances of all categories. Provide training instances of the minority categories when initializing a new classifier. Improve the classification ability on the minority class of the base classifiers.

However, it may take quite a long time to collect enough instances of minority categories by active learning labeling strategy. In this manuscript, we proposed two solutions to solve this problem. On the one hand, we designed an imbalance labeling strategy that tends to request instances of the minority class. On the other hand, for the first data block in the data stream, ROALE-DI will learn the first data block in a supervised manner. That is, ROALE-DI will select the real labels of all instances in the first data block to initialize the first classifier. Therefore, instances in the first data block can be used to fill up the labeled instances buffer. (The implementation of the initialization process is on line 178 to 183 in the ROALE-DI code we provided.)

As the active learning algorithms provided by MOA usually set a period of supervised learning time for initializing the classifiers, we also used this idea to initialize the classifier without taking it as an innovative point, and at the same time, we did not realize its role in filling the labeled instance buffer in class imbalance condition. We are sorry that the initialization process was not described in detail in the previous manuscript.

Thank for your opinion about how long it takes to gather enough minority class instances to fill up the labeled instances buffer. It expands our understanding of the issue. In the new manuscript, we modified the method section of ROALE-DI to show the initialization procedure using the first formed data block, and we also added the effect of the initialization procedure for collecting the minority class instances.

Author action:

1. In page 5 6th paragraph section 3.3, we modified the description of the algorithm to show the problem that the labeled instances buffer will consume lot of time to collect enough instances of the minority class. And the solution to this problem.

   "In the class imbalance condition, fill the labeled instances buffer by selecting real labels through labeling strategy may take a long time. Therefore, the real labels of the instances in the first block A_1 will be all requested. So, the algorithm will create the stable classifier C_s in the supervised manner and copy it as the first dynamic classifier C_d."

2. In page 5 algorithm 3, we modified the line 8 to show the supervised initialization process for the first data block A_1 to create the first stable classifier.

   "Cs=CreateNewBaseClassifier(1,U,θm,I,L,D,wd,DCIR[I]) // label all instances in the first block to create stable classifier"

Reviewer#1, Concern # 4:

Why do the performances of OAL-DI and ROALE-DI differ apparently in Fig.5 (Section 4.4)? I see in your previous paper [7] that the performance of OAL-DI is also "closed to the supervised methods". So what model change makes ROALE-DI better than OAL-DI apparently?

Author response:

Thanks for pointing out this.

In the work OAL-DI, we proposed to apply the hybrid labeling strategy that favors the minority class and a paired classifier to overcome concept drift and class imbalance. Experiments on both synthetic and real-world streams can show that the OAL-DI performs better than the traditional single or hybrid strategy. Compared with the OAL-DI algorithm, ROALE-DI algorithm has following improvements:

First, ROALE-DI have more dynamic base classifiers than OAL-DI, which can improve the overall classification performance.

Second, the weights of the component classifiers in ROALE-DI are adjusted by reinforcement mechanism and time decayed mechanism, which help the algorithms focus on the latest concept and minority class. And the weight of classifiers in OAL-DI are fixed and will not dynamically adapt to instances of new concepts and of minority class.

Third, new base classifiers in ROALE-DI are initialized by a sample-based procedure, which could get more information on the minority class. Therefore, the new base classifier in ROALE-DI will perform better than that in OAL-DI on the minority class.

The experimental streams are also improved in the manuscript. In the experiments of OAL-DI, we used some class balanced real world data streams (Air, Elec) and synthetic data streams ($Agrawal_1$, $AssetNegotiation_1$, HYP_1, SEA_1, $Sine_1$, $STAGGER_1$, and RBF). Although these data streams can show the performance of the algorithms on the class balanced data streams, it cannot fully meet the research topics of class imbalance in this manuscript of ROALE-DI. Moreover, we also improve the class imbalance ratio selection of the experimental data streams. Wang Shuo's work [10] is most recent survey article on the joint problem of concept drift and class imbalance. Therefore, we also referred to the class imbalance ratio design of the survey article.

In ROALE-DI's experimental design process, we want to make up for the previous regrets as much as possible and improve the fairness of the new experimental comparison. Compared with the experimental data streams of the OAL-DI algorithm, we remove the class balanced data and the class imbalance ratio of the data streams is more reasonable. Therefore, we believe that the improvement of the experimental comparison dataset can also make the results more reference significance.

In general, ROALE-DI has both improvements on the algorithm design and experimental data streams compared to OAL-DI. Therefore, it has a more obvious performance improvement.

Author action:

1. In page 3 4th paragraph of section 2.2, we modified the manuscript to tell about ROALE-DI solves the problems that OAL-DI cannot solve.

"In our previous work, we proposed an online active learning paired ensemble framework for concept drift and class imbalance [7]*, which applies a hybrid labeling strategy for selecting the representative instances of the minority class. However, as the active labeling strategy is unable to provide minority class instances stably, the OAL-DI still exhibits the problem that classifiers may lack the training samples of the minority class."*

2. In page 10 2th paragraph of section 4.1, we update the reason of class imbalance ratio setting on synthetic data streams.

   "To ensure that the comparison experiment of the synthetic data streams is more meaningful, the class imbalance ratio setting is refer to Wang Shuo's review article [6]."

3. In page 4 1st paragraph of section 4, we added content to describe the reasons for the choice of active learning comparative algorithms.

   "The performance of both semi-supervised and supervised methods is compared. For the semi-supervised method, the proposed ROALE-DI, OALE and OAL-DI are compared. The classic active learning strategies [25] are not included in the experimental comparison because OALE and OAL-DI have already outperformed them."

4. In page 13 2nd paragraph of section 4.4, we modified the manuscript to analyze the reason for the performances of the algorithms.

   "The statistical test on the AUC and accuracy value indicates that ROALE-DI has a nearly performance ability with the supervised algorithms, such as PAUC-LB, PAUC-BOLE and PAUC-ARF. And ROALE-DI is better than other semi-supervised algorithms apparently. On both AUC and accuracy indicator, PAUC-LB achieves the first place. OAL-DI and PAUC-UOB ranked lower in the performance comparison. But OAL-DI is an active learning method that only has 2 base classifiers. PAUC-UOB applies undersampling and it trains the instances with less times compared with other bagging-based algorithm OOB, BOLE and LB. And OAL-DI is also close to the supervised method UOB on both PAUC and accuracy."

Reviewer#1, Concern # 5:

Some typos:

-footnote 1 looks strange;

-Some sentences do not read smoothly, such as “However, if the OAL-DI still exhibits the problem that classifiers may lack the training samples of the minority class.” (Page 3 Section 2.2);

-please do not split table or algorithm apart (such as Table 3 and Algorithm 6).

Author response:

Thanks for pointing out these problems. We have checked the full text of the manuscript, fixed the footnote format according to the template, checked the writing and grammar, and modified the layout of the tables and algorithms. Thanks again for your rigorous work.

Author action:

1. The footnote looks strange according to the typesetting restrictions caused by software version and we have fixed this error.

2. We have eliminated the errors in this sentence.

   “However, as the active labeling strategy is unable to provide minority class instances stably, the OAL-DI still exhibits the problem that classifiers may lack the training samples of the minority class.”

3. We have re-adjusted the layout to ensure that tables and algorithms do not spread across pages

Reviewer: 2

Author response:

Thank you for your recognition and encouragement of our work. In the new manuscript, to respond to the comments of other reviewers, we made the following improvements to the article.

1. For the problem of converting multi-class data streams into binary streams, we modified the introduction and experiment to show the reason why the experiments are mainly on binary streams. Then, we made experiments on multi-class data streams to compare the performance ROALE-DI with other algorithms, which applies Recall as evaluator. Results shows that ROALE-DI can also achieve good performance on multi-class data streams. The experimental code, streams, and results are provided as supplement materials.
2. For the problem about class imbalance ratio estimating accuracy of ROALE-DI, we modified the source code to output the Damped Class Imbalance Ratio (DCIR) and made two-dimensional area chart to show the class imbalance ratio estimating results. The modified code, original results are also provided.
3. To solve the problem of collecting the minority class instances efficiently, as a stream active learning algorithm, ROALE-DI use the first data block as the initialization stage and request real labels of the instances in the first block. As the active learning algorithms provided by MOA usually set a period of supervised learning time for initializing the classifiers, in the previous manuscript, we also used this idea to initialize the classifier without taking it as an innovative point. Therefore, we modified the method section to show the initialization stage of ROALE-DI and the role of the initialization stage.
4. As for the improvements between ROALE-DI with OAL-DI, in related-work section, we modified the manuscript to show the innovations of ROALE-DI and corresponding problems to be solved. In experiments section, we add more content to compare the performance of ROAL-DI, OAL-DI with other supervised algorithms.
5. We modified the grammatical and writing mistakes through professional polishing institution. And the typesetting issues are also fixed.

We are uploading (a) our point-by-point response to the comments (below) (response to reviewers), (b) a clean updated manuscript without highlights (PDF main document).

In order to effectively disclose and prove the authenticity and repeatability of our work, we published the data, source codes and supplements results of our algorithm for review on ResearchGate.

Link: https://github.com/saferhand/ROALE-DI