SIMPOR: SUMMARY OF CHANGES

We are immensely pleased to be offered helpful suggestions and have thoroughly revised the manuscript according to the reviewer's comments. In this revision, we have taken the opportunity to address the reviewer's concerns that were kindly drawn to our attention. We thoroughly considered each comment and made changes to clarify the manuscript accordingly. The following items summarize the main changes in the latest revision conducted to the paper.

1) Section II (Related Work) was revised, and two more related works have been included in this revision.
2) Section III.C (Entropy-based Active Learning) was revised to improve the reading comprehension of the informative sampling mechanism.
3) Section V.B (Generate minority Synthetic Data) was revised to improve the manuscript's readability.
4) A source code repository link is included for references on GitHub. The footnote containing the link is mentioned in Section VII.A Implementation detail (Page 8).

I. RESPONSE TO REVIEWER 1

Comment 1.1 Eq. (13) seems to be incorrect. According to the manuscript, the length of vector v is r, which is supposed to be a fixed value. In this case, why sample r from a Gaussian distribution? Moreover, the length of vector v is r and the proposed methods generates a minority sample within the radius r.

If the length is large, will that generate samples located in the area of the majority samples? (In Section V.B, the d-sphere's radius is set by the length of vector v.)

Response 1.1: Thank you for your comment, and sorry for the confusion. The main idea is determining vector $\vec{v}$ length and direction. For the length of $\vec{v}$, we first sample r from a Gaussian distribution as depicted in Equation 13. This generates different distances between synthetic samples and the original minority samples (thus, enriching synthetic samples). The length of $\vec{v}$ then is directly set to r. In other words, we can consider r as an alias for $\vec{v}$

As mentioned in Section "V.B.Finding synthetic samples surrounding a minority sample," r is sampled from a Gaussian distribution with the standard deviation of αR (Equation 13), and R is computed by the average Euclidean distance of the minority sample k-nearest neighbors. Thus, r should not be too large. On the other hand, as the synthetic samples are generated in the direction (vector $\vec{v}$'s direction) toward maximizing the posterior ratio (Equation 7), they should be placed into the minority class and away from the majority class.

To avoid any confusions, we revised Section V.B accordingly as follows.

(The change can be viewed at page 5, line 41, and page 6, line 10)

"‘ ... A synthetic neighbor x' and its label y' can be created surrounding a minority sample x by adding a small random vector v to the sample, $x' = x + v$. Thus, x' can be selected on the d-sphere's surface centered at x with a radius of $|\vec{v}|$. For notation convenience, let $r = |\vec{v}|$ be the radius of the d-sphere. To enrich the synthetic samples, r is sampled from a defined Gaussian distribution to generate a new synthetic sample distance each time. This section describes how the direction and distance of a synthetic sample are determined, which can also be represented via the direction and length of vector $\vec{v}$.

It is critical to generate synthetic data in the informative region because synthetic samples can unexpectedly jump across the decision boundary. This can be harmful to models as this might create outliers and reduce the model's performance. Therefore, we safely find vector $\vec{v}$ towards the minority class, such as $\vec{v}_0$ and $\vec{v}_1$ depicted in Figure 1. ...

”””

“‘ ... The relationship between a synthetic sample x' and a minority sample can be described as follows,

$$\vec{x'} = \vec{x} + \vec{v}, \tag{12}$$

where the length of $\vec{v}$ is equal to r, and r is sampled from a Gaussian distribution,

$$r \sim \mathcal{N}(0, (\alpha R)^2), \tag{13}$$

where αR is the standard deviation of the Gaussian distribution and $0 < \alpha <= 1$. The range parameter R is relatively small and computed as the average distance of a minority sample x to its k-nearest neighbors. This will ensure that the generated sample will surround the minority sample. The Gaussian distribution with the mean of zero and the standard deviation αR controls the distance between the synthetic samples and the minority sample. The standard deviation is tuned from 0 to R by a coefficient $\alpha \in (0, 1]$. The larger the α is, the farther synthetic data is placed from its original sample. ... ”””

Comment 1.2 The paragraph on p.4 (Line 39) regarding the description of using entropy scores to select samples in a batch needs to be clarified. The k-highest entropy samples are determined. Is this process in a batch? Why does this batch is concatenated with the other data as stated in Line 41, “This batch is then concatenated with the initial training data for the training classifier parameter ($\theta^{(1)}$) in the next phase ($t^{(1)}$) and also accumulated to the informative set.”.

Response 1.2: Thank you for your comment, and sorry for the confusion. In our study, it was processed in batches, gradually updating the model and selecting a new batch of informative samples. One can fully train the model with the entire data and estimate the high-entropy scores for the same data. But this might potentially cause bias in selecting the informative set as the model training with imbalanced data might be biased by itself. Inspiring by the idea of exploring critical data batch-by-batch from active learning, we proposed a mechanism that gradually explores informative samples. First, the model is trained with an initial batch of data. The model is then used to estimate entropy scores for the remaining unseen data to select the first set of high-entropy samples (this set is considered informative examples relative to the current model parameters). This high informative data is then accumulated to previous training data to continue fine-tuning the current model and select the next informative set from the rest of the data. The process is repeated until reaching the desired amount of informative samples.

To improve the readability of the manuscript, we revise the regarded section as follows.

(The change can be viewed at page 4, line 23)

“‘ ... To select informative samples, one can fully train the entire original dataset and estimate entropy scores on the same data to select high-entropy samples. However, the model fully training on the entire dataset might be biased due to the data imbalance. This could lead to a bias in selecting informative samples because the model might only recognize the densest minority area and ignore other areas containing fewer informative examples. To avoid this issue, we proposed to explore more informative samples batch by batch gradually; this mechanism was inspired by the idea of exploring critical data by batches from active learning. First, the model is trained with an initial batch of data. The model is then used to estimate entropy scores for the remaining unseen data to select the first set of high-entropy samples (this set is considered informative examples relative to the current model parameters). This high informative data is then accumulated to previous training data to continue fine-tuning the current model and select the next informative set from the rest of the data. The process is repeated until reaching the desired amount of informative samples. The remainder of this section describes more detail on the mechanism.

The proposed approach implementation requires repeated phases, and a batch of informative data is selected for each phase. At the first phase $t^{(0)}$, a classifier with parameter $\theta^{(0)}$ (Note that this classifier differs from the classifier for the final classification problem) is trained on an initial batch of data (at least one sample in each class is required) and use the model $\theta^{(0)}$ to estimate the entropy for the remaining data. The entropy scores are then estimated for the remaining samples based on Equation 4. The first batch of informative samples is determined by selecting k highest entropy samples. This batch is then concatenated with the initial training data for the training classifier parameter ($\theta^{(1)}$) in the next phase ($t^{(1)}$) and also accumulated to the informative set. In the next phase, similarly, the classifier is fine-tuned with new data and used to estimate the entropy of the remaining data. The next informative batch is selected and also added to the informative set. Phases are repeated until the number of accumulated informative samples reaches a pre-set informative portion (IP). For example, $IP = 0.3$ will select 30% training samples as informative samples. ...

”””

Comment 1.3 The authors can share their code on GitHub to help other researchers reproduce and perform comparisons.

Response 1.3:

Thank you for your suggestion. The implementation is now publicly available on Github (https://github.com/nsh135/_SIMPOR_).

The repository link is mentioned in Section VII.A ”Implementation Detail’ and attached to Page 8 footnote.

Comment 1.4 4. The proposed method uses entropy as a criterion to select informative samples and uses KDE to approximate the likelihoods, which are related to the following two papers. The authors should discuss these two papers in the manuscript.

Liu, C. L., & Chang, Y. H. (2022). Learning from imbalanced data with deep density hybrid sampling. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(11), 7065-7077. DOI: 10.1109/TSMC.2022.3151394

Liu, C. L., & Hsieh, P. Y. (2019). Model-based synthetic sampling for imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1543-1556. DOI: 10.1109/TKDE.2019.2905559

Response 1.4:

Thank you for your suggestion. In this revision, we have considered the two related works in our study and briefly discussed them in Section II.A Related Work. The change can be viewed at page 3, line 9, and the corresponding references are [9] and [10].